Journal of Biological Systems, Vol. 19, No. 4 (2011) 607–616
© World Scientific Publishing Company
DOI: 10.1142/S0218339011004068

A NEW METHOD OF DETERMINING THRESHOLD OF GENE NETWORK BASED ON PHENOTYPES

YUANYUAN ZHANG*, SHUDONG WANG*,§, MEIXI YANG*,
DASHUN XU† and DAZHI MENG*,‡

*College of Information Science and Engineering
Shandong University of Science and Technology
579 Qianwangang Road, Economic and Technical Development Zone
Qingdao, Shandong 266510, China

†Department of Mathematics, Southern Illinois University Carbondale
Makanida, IL 62958, USA

‡College of Applied Sciences, Beijing University of Technology
Beijing 100124, P. R. China
§wangshd2008@yahoo.com.cn

Received 26 March 2011
Accepted 7 May 2011

With the rapid growth of microarray data, it has become a hot topic to reveal complex behaviors and functions of life system by studying the relationships among genes. In the process of reverse network modeling, the relationships with less relevance are generally not considered by determining a threshold when the relationships among genes are mined. However, there are no effective methods to determine the threshold up to now. It is worthwhile to note that the phenotypes of genetic diseases are generally regarded as external representation of the functions of genes. Therefore, two types of phenotype networks are constructed from gene and disease views, respectively, and through comparing these two types of phenotype networks, the threshold of gene network corresponding to a certain disease can be determined when their similarity reaches to maximum. Because the gene network is determined based on the relationships among phenotypes and phenotypes are external representation of the functions of genes, it is considered that relationships in the gene network may show functional relationships among genes in biological system. In this work, the thresholds 0.47 and 0.48 of gene network are determined based on Parkinson disease phenotypes. Furthermore, the validity of these thresholds is verified by the specificity and susceptibility of phenotype networks. Also, through comparing the structural parameters of gene networks for normal and disease stage at different thresholds, significant difference between the two gene networks at threshold 0.47 or 0.48 is found. The significant difference of structural parameters further verifies the efficiency of this method.

Keywords: Systems Biology; Gene Network; Phenotype Network; Threshold.

§Corresponding author.

1. Introduction

In both prokaryotic and eukaryotic cells, gene expression is a highly regulated process. This regulation of gene expression has a meaningful effect in maintaining cell activity and division and responding to the change of environment or external stimuli. So it is meaningful to study gene expression. Now there are two methods in studying gene expression in theory, forward and reverse modeling. Forward modeling describes the expression of one or several genes starting with transcription and translation using detailed mathematical model, which includes the binding of transcription factors and RNA polymerase with DNA, the effect of specific inhibitory or active factors, the forming of mRNA and proteins in different matured stages, and the regulatory effect of internal feedback loops or external regulators and so on. Reverse modeling can be used to construct network model using gene expression profiles and gene expression patterns. And then as some gene clusters and motifs which are sets of functional related genes are discovered, finally some mechanisms related to their functions are predicted. Compared with the former which is only used to deal with a few genes, the latter can process large number of genes simultaneously, even the whole genome in a cell. Now the models of gene network include Boolean network, Bayesian network, linear model, differential equation model and mutual information correlation model[1–5] and so on. In mutual information correlation model, if the mutual information values between two genes are greater than a given threshold, the relationship between them is considered existing. But it is difficult to determine the threshold. Zhang *et al.*[6] described a general framework that assigns a connection weight to each gene pair and provided empirical evidence that the "weighted" topological overlap measure leads to more cohesive modules than its "unweighted" counterpart. But data simplification is often essential to reduce the complexity. Butte *et al.*[7] extracted gene networks by discarding gene pairs with correlation below the threshold. Voy *et al.*[8] used distribution of correlations of genes with buffer spots on the arrays to select a threshold correlation value, and found cliques of gene network. Sanoudou *et al.*[9] used a correlation threshold 0.80 to obtain relevance network. Lee *et al.*[10] considered the top 0.5% of correlations to build a co-expression network. Langaton *et al.*[11] recommended use of ontological distance, statistical significance, and various graph structural attributes to arrive at a correlation threshold. Palla *et al.*[12] found that a threshold based on clique size was effective separating networks. Bhavesh *et al.*[13] compared six conceptually diverse methods and found that the relationships obtained by threshold selection approaches based on network structure of gene relationships have greater relevance to real biological relationships than those of approaches based on statistical pairwise relationships. It is believed that laws of nature always clearly emerge at a proper coarse-grained level, namely an appropriate threshold. Therefore, it is important to study the selection of threshold in modeling of gene regulatory networks.

Now, the studies of human diseases have accumulated abundant data of disease related phenotypes and plenty of relationships between phenotypes and genes.[14–18]

The changes of phenotypes are considered as external representation of those of genes' expressions. Therefore, the threshold of gene network can be determined based on the relationships between genes and phenotypes and the gene network after determining the threshold may manifest functional correlations in biological system. In this work, mutual information gene networks for Parkinson disease at different thresholds are constructed, and the corresponding Parkinson phenotype networks are obtained based on the relationships between genes and phenotypes. On the other hand, from the view of disease, another Parkinson phenotype network can be built, in which the relationships among phenotypes better reflect the biological system. Through comparing the two types of phenotype networks constructed in different ways, the maximum similarity between them is obtained. The threshold at which the similarity reaches to maximum is the right one of gene network. Finally, the validity of this method is further confirmed through calculating the sensitivity and specificity of two types of phenotype networks and comparing the structural parameters of gene networks with and without disease.

2. Modeling Method

2.1. *Mutual information*

In this work, a gene expression profile is a vector whose components are its expression in different sample cells. For convenience, we denote gene expression profiles by their corresponding genes. For example, the mutual information of genes A and B means the mutual information of their expression profiles. The mutual information of genes A and B is defined as follows:

$$I(A;B) = H(A) + H(B) - H(A,B), \tag{2.1}$$

where $H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$ is the entropy of $X, H(A,B)$ is the joint entropy of A and B. Larger values of $I(A;B)$ imply closer interrelation between genes' expressions. In the case of $I(A;B) = 0$, genes' expressions are irrelevant.

2.2. *Construction of phenotype networks*

Phenotype networks will be constructed in two different ways. On one hand, phenotype network can be constructed from the view of disease, which is called target phenotype network. Actually, every disease has a large number of phenotypes. We believe that two phenotypes are related to each other when they are connected with the same disease. Using disease-related phenotypes derived from HPO (Human Phenotype Ontology), phenotype network can be constructed. On the other hand, phenotype network also can be built from the view of gene, which is called inference phenotype network. Because diseases with similar phenotypes are caused by genes with related features, we can assume that the similarity of phenotypes has a positive correlation with the correlation of genes. Therefore, we can calculate the score of the similarity of two phenotypes based on the closeness among genes and then construct the phenotype network. The similarity of two phenotypes is defined as[19]

$$S_{pp^*} = C_p(p^*) + \sum_{g \in G(p)} \sum_{g^* \in G(p^*)} \beta_{pg}(p^* g^*) e^{-l_{gg^*}^2} \tag{2.2}$$

where $G(X)$ represents a set of genes related to phenotype X, $C_p(p^*)$ is a constant which could be explained as the basal similarity between phenotypes p and $p^*, \beta_{pg}(p^* g^*)$ represents the level of gene g contributing to the similarity between phenotype p and any other phenotype p^*, and l_{gg^*} is the topological distance between g and g^*. The detailed description can be seen in Ref. 19. From the comparison of the above two types of phenotype networks, the gene network corresponding to these phenotype networks with maximum similarity is considered to better reflect real relationships among genes in disease.

3. Materials and Numerical Experiment

3.1. *Data source and processing*

3.1.1. *Data source*

In this work, the data we work on contain phenotype and gene data related to Parkinson disease. Phenotype data are derived from HPO database, which embodies 48 phenotypes related to Parkinson disease and a gene list corresponding to each phenotype; gene data are from NCBI (National Center for Biotechnology Information) database. The sample data sets for Parkinson disease and normal tissues include 50 samples from GSE6613 and 45 samples from GSE20295, respectively, both belong to GPL 96.

3.1.2. *Data processing*

In HPO database, some phenotypes describe the genetic model of Parkinson disease. For example, phenotype HP: 0000006 is autosomal dominant inheritance. Because we focus on the relationships among disease feature–related phenotypes, and the above phenotypes without Parkinson disease features have nothing to do with our purpose, we exclude those phenotypes from our phenotype database. Besides, some phenotypes have no corresponding specific genes. These phenotypes are always irrelevant to the functional changes of genome, so we also exclude them. Finally, there are 36 phenotypes related to Parkinson disease remaining in our phenotype database, 495 genes related to these 36 phenotypes in our gene database and 1211 relationships between these phenotypes and genes.[a]

We choose genes in gene database which are related to phenotypes of Parkinson disease for our purpose. In the case where several probes correspond to one gene,

[a] http://cise.sdust.edu.cn/institute/isbbc/DTGP.htm/Re_GP.xls

the highest expression value is chosen to form the gene's expression profile. Among the above-mentioned genes, there are some genes' expression almost completely 0 or 1 in all samples. We focus on the structure of gene networks and the difference of the structures. These genes contribute little to the difference of the structures. In our study, if less than 15% or more than 90% of the total components of a gene's profile are equal to 1, we exclude the genes from our gene database. Thus, there are 116 and 139 genes left in normal and Parkinson disease databases respectively. Our work is based on the databases[b] and each of these data sets includes p-values.

To calculate mutual information between genes, we discretize p-values in each database as follows. (i) Select the range [Min, Max] for p-values and divide it into 20 portions such that each portion contains almost the same number of p-values. Order the portions in the number order and denote them by 1st, 2nd, ..., 20th interval, respectively. (ii) Replace the p-values in an interval by its labeling value. Obviously, the granularity of our discretization is finer than that of $0-1$ discretization, and hence our discretization loses less information than that contained in the 0-1 discretization. Therefore, it is reasonable to believe that the mutual information networks based on our discretization can better reflect the nature of the gene regulatory system.

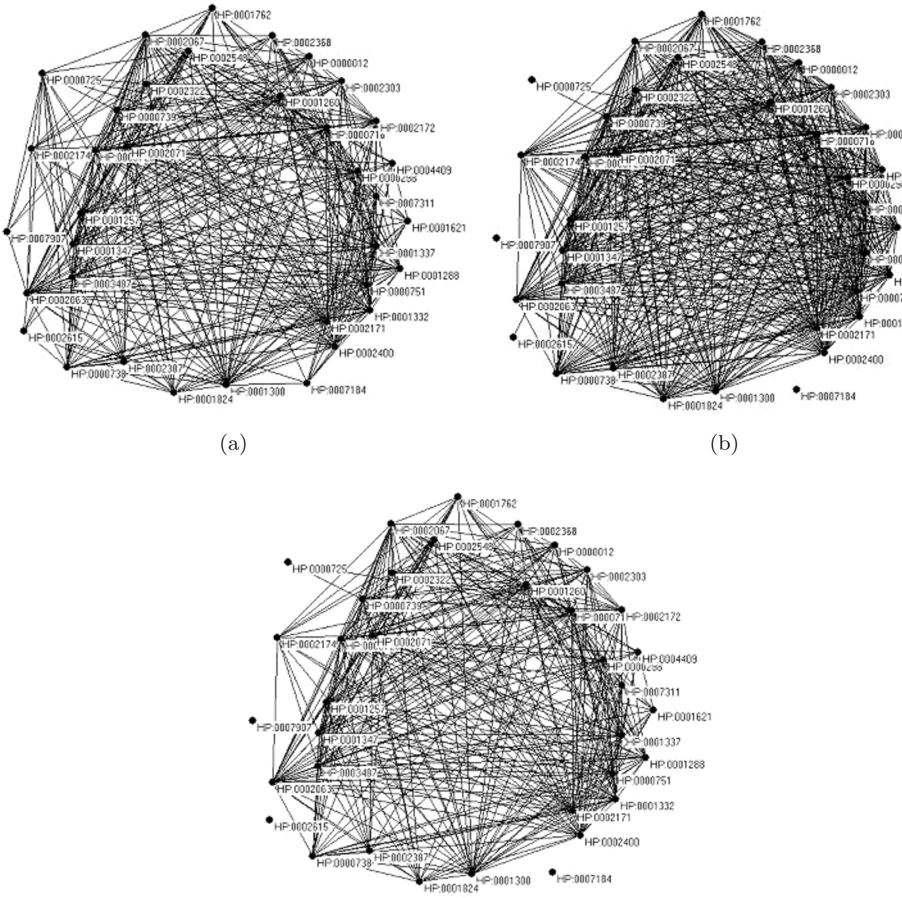
3.2. *Numerical experiment*

For each of the discretizated databases, we can calculate mutual information values and hence obtain a complete network of all genes in the database with mutual information values as edge weights. Note that the ranges for mutual information values for our two databases are different. For the purpose of comparison of networks, we normalize the mutual information values for each database as $x^* = (x - \min)/(\max - \min)$, where x and x^* represent the original and normalized mutual information values, and max and min are the maximum and minimum of the original mutual information values, respectively. Then we choose the threshold in the range $[0.1,\ 0.9]$ by step-length 0.01, and obtain 81 different mutual information networks in all. For each gene network for Parkinson disease, we obtain the corresponding phenotype network using the approach described in Sec. 2.2. Assume $C_p(p^*) = 0$ and $\beta_{pg}(p^*g^*) = 1/n$, where n represents the number of genes related to phenotype p. That is to say, each gene has the same contribution to the similarity between phenotypes p and p^*. In this research, we define the topological distance l_{gg^*} as the shortest path between g and g^*. Because the relationships among genes in gene network are mutual information correlations, the greater the value is, the closer the relation between two genes is and the less the distance is. Hence, we define the distance l_{gg^*} between g and g^* as $1 - I(g; g^*)$.

From the comparison of the phenotype network constructed from the view of disease phenotypes with each of 81 phenotype networks corresponding to gene

[b] http://cise.sdust.edu.cn/institute/isbbc/DTGP.htm/database.rar.

network, we can obtain that the maximum similarity between two types of phenotype networks is 66.6%. In this case, the threshold of gene network is 0.47 or 0.48. Here the similarity is measured by the ratio of the numbers of their common edges and the union-set of their edges. Because threshold 0.47 or 0.48 is obtained in the maximum similarity between phenotype networks, it is reasonable to believe that the gene network can reveal the close correlation between genome and disease. Taking 0.47 as the threshold value, we obtain 62 non-isolate nodes and 497 edges in the gene network. The corresponding phenotype network contains 499 edges, while the phenotype network from the view of disease includes 389 edges and two phenotype networks contains 355 common edges (see Fig. 1). Phenotypes related to the same disease should generally have closer relationships. The average degrees



(a)

(b)

(c)

Fig. 1. The phenotype networks related to Parkinson disease. (a) Target phenotype network. (b) Inference phenotype network when the threshold of gene network is 0.47. (c) The common relationships of two types of phenotype networks. Phenotypes have closer relationships.

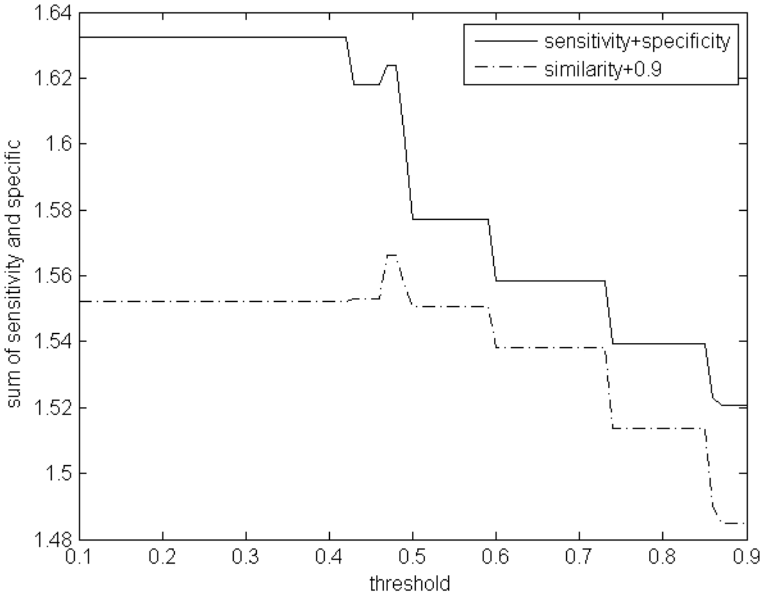


Fig. 2. The comparison of the sum of sensitivity and specificity with the similarity of phenotype networks at different thresholds. For convenience, we increase the similarity by 0.9 at each threshold.

of the two types of phenotype networks with 36 phenotypes are relatively great, 27.72 and 21.61, respectively. It is accordance with our above consideration.

The sensitivity is defined as the percentage of edges in target network which are also in inferred network, while the specificity is the percentage of edges in inferred network which are also in target network.[20] We believe that the greater the sum of the sensitivity and specificity are, the closer to target network the inference network is. Through computing the sensitivity and specificity of phenotype networks at different thresholds and comparing the similarity with the sum of sensitivity and specificity (Fig. 2), we observe that the sum of sensitivity and specificity reaches to the maximum value at thresholds 0.47 and 0.48, which is the same value obtained from the point of similarity. Note that the phenotype network is nearly a complete network when the threshold is small, so the maximum of the sum of sensitivity and specificity is a trivial fact. The inference networks at thresholds 0.47 and 0.48 are close to target network. This conclusion is identical to the above one, so this shows the effectiveness of the proposed method to determine the threshold of gene network based on phenotypes.

4. Comparison of Structural Parameters of Gene Network

In order to show the difference between gene networks corresponding to normal stage and Parkinson disease, we compute five structural parameters: average degree

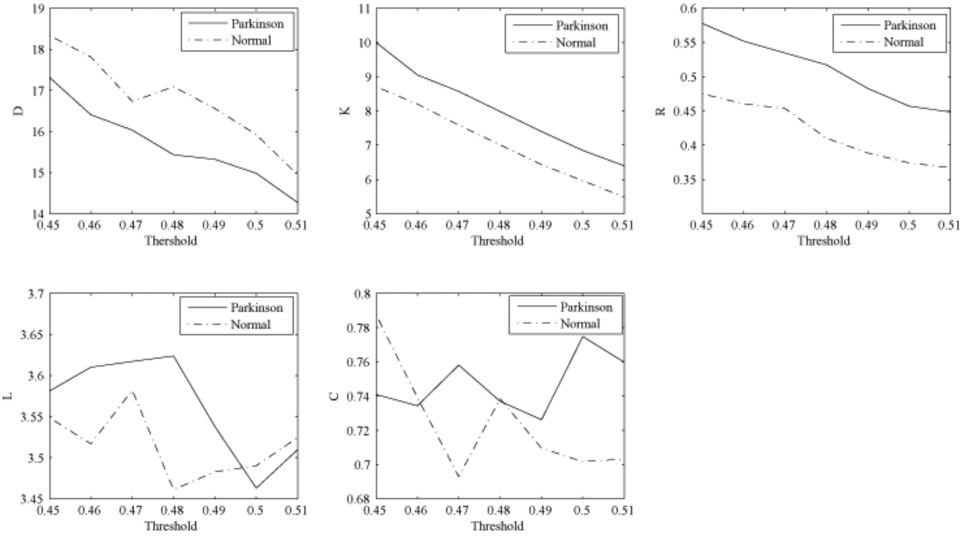


Fig. 3. Plots of five structural parameters versus the threshold values. The difference between the gene networks corresponding to normal stage and disease is most significant at the thresholds 0.47 and 0.48. The obvious difference reveals the functional one between the gene regulatory relationships of normal stage and Parkinson disease.

(K), average degree of non-isolated nodes (D), proportion of non-isolated nodes (R), average path length (L) and average clustering coefficient (C). These statistics are plotted versus the threshold values in Fig. 3. Comparing these statistics of the two networks, one can see that the difference between the gene networks corresponding to normal stage and disease is most significant at the thresholds 0.47 and 0.48. It is believed that the structure of a network dictates its functions. The obvious difference reveals the functional one between the gene regulatory relationships of normal stage and Parkinson disease.

5. Conclusions and Discussions

Disease phenotypes are external representation of genes' functions, and the change of phenotypes are derived from the variation of genes' expressions. In this work, we propose a new approach to determine the threshold of gene network. Through comparing two types of phenotype network constructed from the view of disease and gene respectively, we determine the threshold of gene network. When the similarity of two phenotype networks reaches to the maximum, the thresholds are 0.47 and 0.48. We verify the availability of the thresholds in two ways. On one hand, when the sensitivity and specificity are both higher, the two phenotype networks are closest with each other. The sum of sensitivity and specificity can reach the maximum value at thresholds 0.47 and 0.48, which are identical to the ones obtained from the point of similarity. On the other hand, we compare five statistics of gene networks for normal and disease stages, and find that the most obvious difference is also at thresholds 0.47 and 0.48. The significant difference of these statistics reveals the structural difference between gene networks corresponding to normal and disease stages. This further verifies the reliability and effectiveness of this approach.

Besides, we analyze the relationships of the gene network after determining the threshold through searching their shared Gene Ontology terms. For example, genes APTX, TRIM37, and SPG7 share common molecular function — zinc ion binding, and zinc plays a role in the central nerve system as a neurosecretory product, cofactor, or modulator. It has been reported that zinc levels are increased in substantia nigra, caudate nucleus, and lateral putamen in patients of Parkinson disease.[21] Genes ATXN3, CYP7B1, DCTN1, GARS, ITPR1, SETX, SPG7, PPP2R2B, TTBK2, SPAST, and APTX share common biological process — cell death, and in our gene network, there are many relationships among them.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China (Nos: 60874036, 60970065), SDUST Research Foundation of China and Zhejiang Provincial Natural Science Foundation of China (No: R1110261) for the support to this work.

References

1. Akutsu T, Miyano S, Kuhara S, Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, *Pac Symp Biocomput* pp. 17–28, 1999.
2. Husmeier D, Reverse engineering of genetic networks with Bayesian networks, *Biochem Soc Trans* **31**:1516–1518, 2003.
3. Van Someren EP, Wessels LF, Reinders MJ, Linear modeling of genetic networks from exprerimental data, *Proc Int Conf Intell Syst Mol Biol* **8**:355–366, 2000.
4. Chen ting, He Hongyu, Church GM, Modeling gene expression with differential equations, *Pac Symp Biocomput* pp. 29–40, 1999.
5. Butte AJ, Kohane IS, Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements, *Pac Symp Biocomput* **5**:415–426, 2000.
6. Zhang B, Horvath S, A general frame work for weighted gene co-expression network analysis, *Stat Appl Genet Mol Biol* **4**:17, 2005.
7. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS, Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proc Natl Acad Sci USA* **97**:12182–12186, 2000.
8. Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B, Chesler EJ, Branstetter LK, Langston MA, Extracting gene networks for low-dose radiation using graph theoretical algorithms, *PLoS Comput Biol* **2**:0757–0768, 2006.
9. Sanoudou D, Haslett JN, Kho AT, Guo S, Gazda HT, Greenberg SA, Lidov HG, Kohane IS, Kunkel LM, Beggs AH, Expression profiling reveals altered satellite cell numbers and glycolytic enzyme transcription in nemaline myopathy muscles, *Proc Natl Acad Sci USA* **100**:4666–4671, 2003.

10. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P, Coexpression analysis of human genes across many microarray data sets, *Genome Res* **14**:1085–1094, 2004.
11. Langston MA, Perkins AD, Saxton AM, Scharff JA, Voy BH, Innovative computational methods for transcriptomic data analysis: A case study in the use of FPT for practical algorithm design and implementation, *Computer J* **51**:26–38, 2008.
12. Palla G, Derenyi I, Farkas I, Vicsek T, Uncovering the overlapping community structure of complex networks in nature and society, *Nature* **435**:814–818, 2005.
13. Borate BR, Chesler EJ, Langston MA, Saxton AM, Voy BH, Comparison of threshold selection methods for microarray gene co-expression matrices, *BMC Research Notes* **2**:240, 2009.
14. Becker KG, Barnes KC, Bright TJ, Wang SA, The genetic association database, *Nat Genet* **11**:753–757, 2004.
15. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA, Online Mendelian Inheritance in Man (OMIM): A knowledgebase of human genes and genetic disorders, *Nucleic Acids Res* **33**:D514–D517, 2005.
16. Kahraman A, Avramov A, Nashev LG, Popov D, Ternes R, Pohlenz HD, Weiss B, PhenomicDB: A multi-species genotype/phenotype database for comparative phenomics, *Bioinformatics* **21**:418–420, 2005.
17. Robison PN, Kohler S, Bauer S, Seelow D, Horm D, Mundols S, The human phenotype ontology: A tool for annotating and analyzing human hereditary disease, *Am J Hum Genet* **83**:610–615, 2008.
18. Lussier YA, Borlawsky T, Rappaport D, Friedman C, PhenoGO: A multi-strategy language processing system assigning phenotypic context to gene ontology annotations, *Pac Symp Biocomput* pp. 64–75, 2006.
19. Wu X, Jiang R, Zhang MQ, Shao Li, Network-based global inference of human disease genes, *Mol System Biol* **4**:189, 2008.
20. Kyoda KM, Morohashi M, Onami S, Kitano H, A gene network inference method from continuous-value gene expression data of wild-type and mutants, *Genome Inform Ser Workshop Genome Inform* **11**:196–204, 2000.
21. Dexter DT, Carayon A, Javoy-Agid F, Agid Y, Wells FR, Daniel SE, Lees AJ, Jenner P, Marsden CD, Alterations in the levels of iron, ferritin and other trace metals in Parkinson's disease and other neurodegenerative diseases affecting the basal ganglia, *Brain* **114**:1953–1975, 1991.