BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 22 2007, pages 3100–3102
doi:10.1093/bioinformatics/btm460

Gene expression

TAPPA: topological analysis of pathway phenotype association

Shouguo Gao* and Xujing Wang

The Max McGee National Research Center for Juvenile Diabetes & The Human and Molecular Genetics Center, The Medical College of Wisconsin and Children's Hospital of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, USA

Received on January 25, 2007; revised on September 1, 2007; accepted on September 5, 2007

Advance Access publication September 21, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: Extracting biological insight from microarray data is important but challenging. Here we describe TAPPA, a java-based tool, for identification of phenotype-associated genetic pathways utilizing the pathway topological measures. This is achieved by first calculating a Pathway Connectivity Index (PCI) for each pathway, followed by evaluating its correlation to the phenotypic variation. Our PCI definition not only efficiently captures the contributions from genes that show subtle but consistent changes in expression, but also naturally overweighs the hub genes that interact with a large number of other genes in the pathway. TAPPA also allows evaluation of sub-modules within a pathway and their association to phenotypes.

Availability: TAPPA and data for Figure 1 are freely available from http://watson.mcgee.mcw.edu:8080/~sgao

Contact: sgao@mcw.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Since its introduction more than a decade ago, the challenge of gene expression profiling using microarrays has changed from obtaining quality data to elucidating biological meanings of the expression data. Pathways are sets of genes that act together to achieve certain cellular or physiologic functions. Prioritizing pathways relevant to a particular phenotype can help researchers focus on the subset of most relevant genes, and generate further biological hypotheses. Functional pathway enrichment analysis has become popular in advanced microarray data analyses. Usually significant gene lists are created with *P*-value or fold change, followed by identification of pathways that have enhanced representation in the gene lists. Genes belonging to the same pathway often exhibit subtle, coordinated changes in their expressions. Such approaches are not sensitive to these genes. To address this issue, Gene Set Enrichment Analysis (GSEA) was developed by examining the overall differences in expression patterns between predefined gene sets and the whole gene list on the array (Subramanian *et al*., 2003).

*To whom correspondence should be addressed.

All these methods ignore the topological property of gene interaction networks within the pathway, thus treating all genes in the pathway equally. However, it is believed that 'hub' genes with a high degree of connections with other genes (i.e. topologically important) usually are most critical to network function (Carter *et al*., 2004). What is worse that 'hub genes' often show low level of changes, their changes are easily overshadowed by the non-hub genes (Lu *et al*., 2007). Recently, topological properties of networks have been utilized to characterize disease states, based on number of regulatory links between transcription factors and target genes (Tuck *et al*., 2006); and in cancer classification and identification of class-specific pathways, based on the fraction of network edges that are specific to the subclasses (Liu *et al*., 2006). Here, we define a novel pathway connectivity index (PCI) to characterize the topology of pathway at expression level, and then utilize it to identify pathways that are significantly associated to a certain phenotype.

2 METHOD

2.1 Pathway connectivity index

Our approach adopts the molecular connectivity concept in chemoinformatics. The molecular connectivity index is a widely used topological descriptor of chemical compounds and has been successfully used in many other fields, including protein structure and drug discovery. Typically, for a chemical molecule the zero and first-order indices are defined by:

$$^0\chi = \sum_{i=1}^{n(\text{atoms})} g(i)^{0.5}; \quad ^1\chi = \sum_{(i,j)\ \text{in bonds}} g(i)^{0.5} * g(j)^{0.5}, \qquad (1)$$

where $g(i)$ and $g(j)$ are the contributions attributed to atom i and j, first defined as the vertex degrees (or their reciprocals) of atoms, and later extended to other chemical properties(Hu *et al*., 2003). n(atoms) is the number of atoms in the molecule, quantity $g(i)^{0.5}$*$g(j)^{0.5}$can be considered as the description of the $i-j$ bond. $^0\chi$ and $^1\chi$ can be regarded as the contributions of all atoms and all chemical bonds, respectively, and $^1\chi$ reflects the inner atomic connectivity in the molecule.

Following this idea, we define PCI for a specific pathway:

(1) A pathway, such as hsa00010 in KEGG, can be represented by a graph $G(V, E)$, where $V = \{g_1, g_2, \ldots, g_n\}$

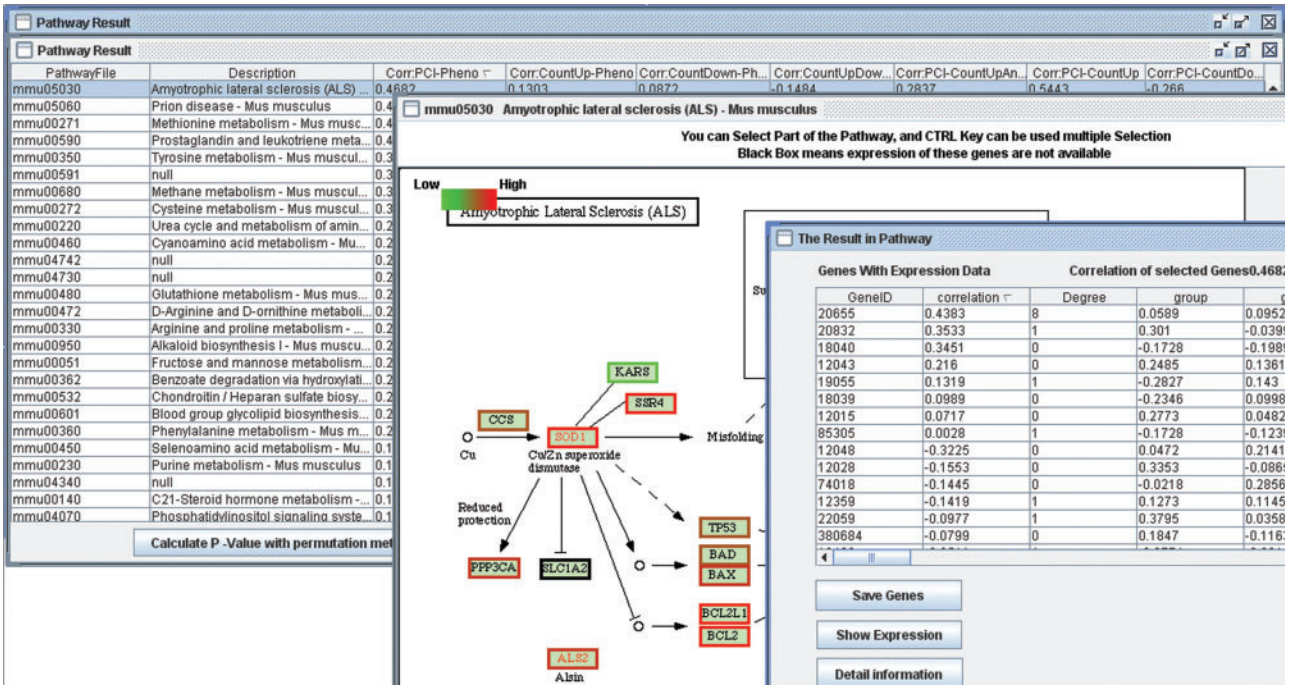

Fig. 1. An exemplary result of TAPPA to prioritize and visualize pathways. The data are from GEO (http://www.ncbi.nlm.nih.gov; accession number: GSE3330), gluc8 is selected as phenotype. The results accords with clinical view that type 2 diabetes may lead to ALS, both the PCI and hub gene SOD1 (geneid: 20655) highly correlate with the phenotype values (04682 and 0.4383). In mmu05030, the correlation coefficient between the average of normalized gene expression of all genes and phenotype is only 0.199.

represents the vertex set, and $E = \{(g_i, g_j) |$ genes g_i and g_j interact$\}$ represents the edge set. The adjacency matrix is defined as $A = (a_{ij})$, where $a_{ij} = 1$ if $i = j$ or $(g_i, g_j) \in E$ and $a_{ij} = 0$ if $(g_i, g_j) \notin E$.

(2) Assuming that x_{is} is the normalized log expression measurement [e.g. each column expression values are normalized to zero mean and same scope with $x_{is} = (x_{is}^{\text{orig}} - \bar{x}_s^{\text{orig}})/\sigma_s$, and further normalized to $(-0.5, 0.5)$ with Sigmoid function (Sigmoid $(x_{is}) - 0.5$) to lower the effects of extremely large/small values] for gene i in sample s. For each pathway we define:

$$\text{PCI} = \sum_{i=1}^{N} \sum_{j=1}^{N} \text{sgn}(x_{is} + x_{js}) \ * \ |x_{is}|^{0.5} * a_{ij} \ * \ |x_{js}|^{0.5}, \qquad (2)$$

where N is the number of genes in it. $\text{Sgn}(x_{is} + x_{js})$ represents the overall expression status (up- or down-regulation) of the gene pair, and helps to reduce the information loss resulted from using absolute values.

It is important to note that x_{is} and x_{js} tend to have same signs (positive or negative), e.g. in KEGG pathways, most of interacting gene pairs have been found to be evolutionally conserved, and thus often show high correlation in their expressions (Bhardwaj and Lu, 2005). Evidently, PCI incorporates information of all available genes in the pathway, subtle yet consistent gene expression modification would lead to a significant change of PCI. Further, PCI captures well the topological property of the pathway, as hub genes contribute more to PCI. Normalized PCIs (divided by the gene number in pathway) follow roughly a normal distribution (Fig. S6).

2.2 Association between pathway and phenotype

TAPPA is designed to handle both binary and quantitative traits. For binary traits, Mann–Whitney test is used to evaluate the significance of association between pathway PCI and phenotype. For continuous traits, Spearman correlation is used. The false discovery rate (FDR) is further determined through a permutation test. In a pathway, higher correlation can be obtained between PCI and phenotype if more genes correlate with the phenotype. PCI defined in this article would degenerate into the average of expression values if there is no (or if we do not consider) connections among genes. This corresponds to the approach described by Li *et al.*, where correlation between average expression values of transcription modules and classical traits is used to evaluate the strength of association between them (Li *et al.*, 2006). Note that this is conceptually equivalent to the definition of ${}^0\chi$. PCI is expected to be advantageous as it incorporates consideration of the network topological structure.

3 IMPLEMENTATION

TAPPA was written in JAVA, a platform-independent language. The main functions of TAPPA include: (1) prioritize pathways associated with discrete or quantitative phenotypic

traits, and calculate the corresponding FDR; (2) visualize the pathway with different zoom ratios and highlight the genes closely related to the phenotype and (3) examine phenotype association of the sub-modules within a pathway. This step is particularly helpful to determine the biological relevance of the associated genes, as correlation with genes in a sub-module is more meaningful than with the genes are dispersed in the pathway.

4 CONCLUSION AND FUTURE WORK

Pathway analysis is a useful approach to uncover the biological meaning from expression data. However, the inner connectivity should not be ignored, and hub genes need to be emphasized. PCI utilizes expression information of all genes in a pathway, and can well capture its topological properties. We have evaluated the ability of TAPPA to identify genetic pathways associated with clinical traits (both binary and continuous) with three published gene expression datasets, and compared it with existing pathway mining programs, the results can be found in the Supplementary Material. So far, we have collected KEGG pathways with gene number higher than 8, and are in the process of parsering the BioCarta pathways (http://cmap.nci.nih.gov/Pathways/BioCarta). We will also extend our tool, such that it is suitable for protein interaction networks.

ACKNOWLEDGEMENT

The project described was supported by the National Institute of Diabetes and Digestive and Kidney Diseases under grant No. R01DK080100.

Conflict of Interest: none declared.

REFERENCES

Bhardwaj,N. and Lu,H. (2005) Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, **21**, 2730–2738.

Carter,S.L. *et al*. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.

Hu,Q.N. *et al*. (2003) The matrix expression, topological index and atomic attribute of molecular topological structure. *J. Data Sci*., **1**, 361–389.

Li,H. *et al*. (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. *Hum. Mol. Genet*., **15**, 481–492.

Liu,C-.C. *et al*. (2006) Topology-based cancer classification and related pathway mining using microarray data. *Nucleic Acids Res*., **34**, 4069–4080.

Lu,X. *et al*. (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol*., **3**, 98.

Subramanian,A. *et al*. (2003) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tuck,D.P. *et al*. (2006) Characterizing disease states from topological properties of transcriptional regulatory networks. *BMC Bioinformatics*, **7**, 236.