Gene expression

PathCluster: a framework for gene set-based hierarchical clustering
Tae-Min Kim¹, Seon-Hee Yim², Yong-Bok Jeong², Yu-Chae Jung¹ and Yeun-Jun Chung¹,², ∗

¹Department of Microbiology and ²Integrated Research Center for Genome Polymorphism, College of Medicine, The Catholic University of Korea, Seoul 137-701, Korea

ABSTRACT
Motivation: Gene clustering and gene set-based functional analysis are widely used for the analysis of expression profiles. The development of a comprehensive method jointly combining the two methods would allow for greater biological insights.
Results: We developed a software package, PathCluster for gene set-based clustering via an agglomerative hierarchical clustering algorithm. The distances between predefined gene sets are illustrated in a dendrogram in which the relationships between gene sets can be visually assessed. Valuable biological insights can be obtained according to the type of gene sets, e.g. coordinated action of molecular functions (functional gene sets) and putative motif synergy (promoter gene set) in a biological process. The combined use of gene sets further enables the interrogation of different biological themes and their putative relationships, such as function-versus-regulatory motif or drug-versus-function. PathCluster can also be used for knowledge-based sample partitioning or class categorization for clinical purposes. With extended applicability, PathCluster will facilitate the gleaning of meaningful biological insights and testable hypotheses in the contexts of given expression profiles.
Availability: PathCluster executable files can be freely downloaded at http://www.systemsbio.kr/PathCluster/.
Contact: yejun@catholic.ac.kr

1 BACKGROUND
The objective of gene clustering is to group genes with similar expression patterns or that are expressed in a coordinated manner (Eisen et al., 1998). Subsequent functional enrichment analysis can provide clues as to which molecular functions or annotation categories are associated with individual gene clusters using biological knowledge. Despite its potential utility, the treatment of gene clusters as exclusive units may raise a number of practical concerns in subsequent functional analysis. For example, a large list of candidate functionalities is obtained as the number of clusters increases, thus making it difficult to compare the results between clusters or to establish appropriate significance thresholds considering multiple testing adjustments. Also, the performance of enrichment analysis is profoundly dependent on prior clustering

∗To whom correspondence should be addressed.

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
genes between gene sets have peculiar interests (especially the case of promoter gene sets), the mean correlation can also be calculated only for the gene pairs within overlapping genes between gene sets. Detailed descriptions of the metrics utilized and examples are available in the online manual at the PathCluster homepage (http://www.systemsbiology.co.kr/PathCluster/Manual.pdf).

PathCluster provides default gene sets covering four kinds of gene annotation categories; molecular functions, the association with regulatory motifs corresponding to transcription factors or miRNA, as well as drug treatment-related expression changes. In addition, gene sets from public databases such as MSigDB or user-defined custom query sets can be readily included in the gene set reference, in order to ensure the versatility of the method.

3 BIOLOGICAL APPLICATION

3.1 Associated molecular functions or regulatory motif sequences in a biological process

Using functional gene sets, PathCluster can identify the putative associations between molecular functions, thereby providing clues on coordinated action of specific functions in a given expression profile. Similarly, in the case of promoter gene sets, PathCluster can identify the putative motif synergy between cis-regulatory motifs or corresponding transcription factors delineating the regulatory crosstalks in a transcriptional regulatory network. Moreover, using combined gene sets with different annotation categories, previously unknown, novel links can be revealed. In erythropoiesis-related expression profiles, a number of functionalities related with immunity and the major histocompatibility complex are observed in a cluster (Fig. 1A). Within the cluster, signal-related functionalities as well as drug treatment-related expression changes. In addition, genes between gene sets have peculiar interests (especially the case of promoter gene sets), the mean correlation can also be calculated only for the gene pairs within overlapping genes between gene sets. Detailed descriptions of the metrics utilized and examples are available in the online manual at the PathCluster homepage (http://www.systemsbiology.co.kr/PathCluster/Manual.pdf).

PathCluster provides default gene sets covering four kinds of gene annotation categories; molecular functions, the association with regulatory motifs corresponding to transcription factors or miRNA, as well as drug treatment-related expression changes. In addition, gene sets from public databases such as MSigDB or user-defined custom query sets can be readily included in the gene set reference, in order to ensure the versatility of the method.

3.2 Function-based sample classification

Knowledge-driven or function-based class categorization has recently emerged as a highly challenging subject. This strategy has already been employed to identify the functional relationships in a large cancer-derived expression compendium or to elucidate drug-signature relationships for clinical benefits (Wong et al., 2008). Adopting a user-friendly platform and extended reference of gene sets, PathCluster provides a platform for the classification or molecular diagnosis of clinical samples, also allowing for the interrogation of diverse biological knowledge in terms of gene sets. Figure 1B shows that function-based classification can successfully distinguish between the three lung cancer subtypes, including normal tissues. In this cluster, eight cancer-related functions are specifically up-regulated in small cell lung cancer and squamous cell carcinoma of the lung.

ACKNOWLEDGEMENTS

This work is supported by the grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0405-BC02-0604-0004) and (01-PJ3-PG6-01GN07-0004).

Conflict of Interest: none declared.

REFERENCES

Bhattacharjee,A. et al. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl Acad. Sci. USA, 98, 13790-13795.

Cheadle,C. et al. (2007) GSMA: gene set matrix analysis, an automated method for rapid hypothesis testing of gene expression data. Bioinform. Biol. Insights, 1, 49-62.

Eisen,M.B. et al. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95, 14863-14868.

Keller,M.A. et al. (2006) Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators. Physiol. Genomics, 28, 114-128.

Rapaport,F. et al. (2007) Classification of microarray data using gene networks. BMC Bioinformatics, 8, 35.

Wong,D.J. et al. (2008) Revealing targeted therapy for human cancer by gene module maps. Cancer Res., 68, 369-378.