



KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data

Tetsuro Toyoda* and Akihiko Konagaya

Bioinformatics Group, Genomic Sciences Center, RIKEN, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan

Received on August 21, 2002; revised on September 12, 2002; accepted on September 24, 2002

ABSTRACT

Summary: KnowledgeEditor is a graphical workbench for biological experts to model biomolecular network graphs. The modeled network data are represented by SRML, and can be published via the internet with the help of plug-in module 'GSCOPE'. KnowledgeEditor helps us to model and analyze biological pathways based on microarray data. It is possible to analyze the drawn networks by simulating up–down regulatory cascade in molecular interactions.

Availability: KnowledgeEditor is available at <http://gscope.gsc.riken.go.jp/>

Contact: bioinfo@gsc.riken.go.jp

DNA microarrays are used as powerful instruments to measure the expression levels of thousands of genes simultaneously. Clustering expression profiles provides valuable insights into function analyses. Hierarchical clustering (Eisen *et al.*, 1998) and self-organizing mapping (Tamayo *et al.*, 1999) have often been applied for data analyses. Moreover, various attempts have been made to elucidate genetic regulatory networks underlining the data: modeling methods with Boolean network (Somogyi and Shiegoski, 1996), differential equations (Chen *et al.*, 1999; D'haeseleer *et al.*, 1999) are proposed as estimators of regulatory networks. Although those methods predict possible gene network models, it is often the case that some parts of them are falsely modeled and should be corrected by biological experts. Thus, a new tool is needed to edit the networks by interactive manner. In the course of network editing, relationship between the networks and expression data must be easily confirmed and analyzed by users. Several public and commercial pathway resources can relate expression data to biological networks. EcoCyc (Karp *et al.*, 2002a), MetaCyc (Karp *et al.*, 2002b) and the Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al.*, 2002) contain a large amount of curated information which can be used to view expression data on the context of preexisting pathways. GenMAPP is a stand-

alone software which helps us to view expression data on pathways. Unfortunately, most of the pathway data are in a binary file which is difficult to process by script languages such as Perl: a text-based pathway data is more desirable. Here, we present a new tool called 'KnowledgeEditor', which is an XML (eXtended Markup Language)-based authoring tool for biomolecular network modeling.

KnowledgeEditor imports a file (cdt: cluster data table) created by the program Cluster (Eisen *et al.*, 1998). KnowledgeEditor can deal with more expression data points than 50 000 genes (rows) \times 100 experiments (columns), and help us to pick up any genes by comparing similarity of expression patterns. By manipulating a mouse, a user can lay out the picked genes on the draw window on which each gene is represented as a square item (Figure 1). The drawn arrow between items means the existence of a certain relationship between the genes. By selecting an experiment name and a coloring condition, a user can see the levels of gene expressions displayed as item colors on the drawn networks. The drawn networks on KnowledgeEditor are easily converted to an XML-based representation schema (SRML: Simple Result Markup Language) which we designed to model a gene–gene interaction relationship. In Figure 1 each arrow has various sorts of information including Boolean (up or down)-regulation among genes. Based on the SRML, KnowledgeEditor immediately simulates the up–down cascade among thousands of genes described in the SRML file. SRML basically defines only a binomial graph (an arrow) between two genes. Further complex relationship among multiple genes can be described by the combination of the arrows. Since it tends to be incomprehensible to see multiple arrows intercrossing each other, KnowledgeEditor can group them up into a polynomial graph by jointing the centers of the arrows and give us a coherent view of multiple gene interactions. It is possible to publish SRML data files on the internet by using the web-browser plug-in 'GSCOPE', (<http://gscope.gsc.riken.go.jp>). GSCOPE helps a web-client user to browse pathway data on their personal

*To whom correspondence should be addressed.

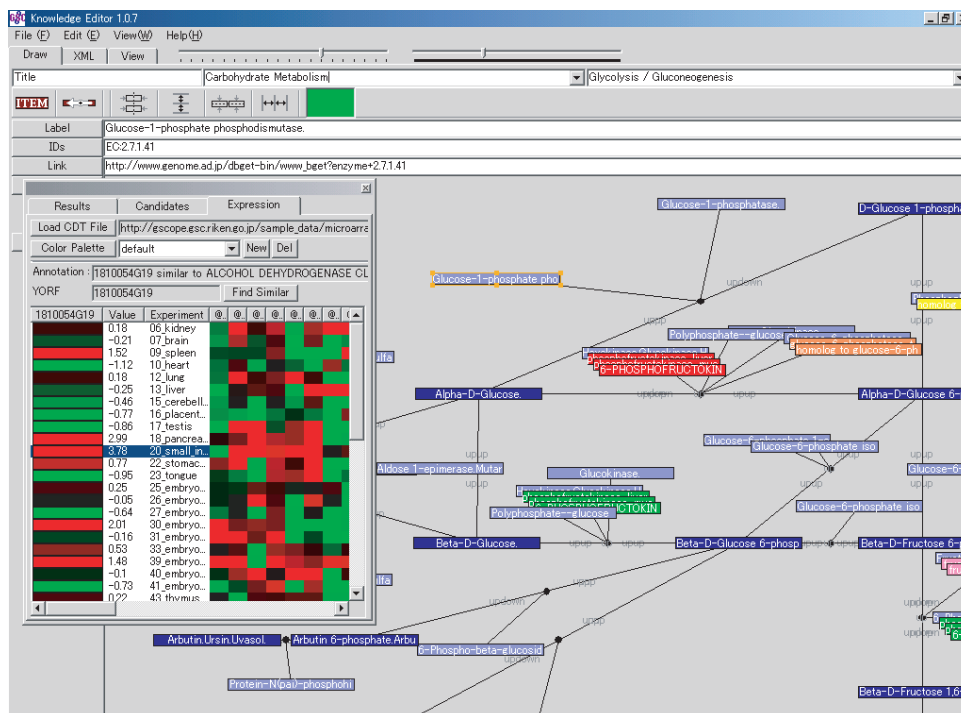


Fig. 1. 'Draw window' and 'gene-expression window' of the KnowledgeEditor are shown. 'Gene-expression window' is floating before the 'draw window.' Glycolysis metabolic pathway is sketched on the 'draw window,' herein squared items represent genes and compounds in the pathway. Lines drawn between the items are called 'arrows' in the text, because each line contains upward or downward regulatory information and displayed as an arrow by GScope. In the 'gene-expression window,' the listed genes have gene-expression patterns similar to the selected gene which is highlighted by six orange dots in the 'draw window.' Up-regulated genes are colored in red and down-regulated genes are in green in the current coloring condition. In the pathway modeling, a user can pick a gene from the 'gene-expression window' and easily put it on the 'draw window' by mouse operations.

computers, to simulate regulatory interactions of genes on the pathways, and analyze microarray data in the context of the pathways. For users' convenience, it is very easy to transfer SRML and microarray data between KnowledgeEditor and GScope by the mechanism which automatically saves snapshots of the analytical status on the software when they are terminated by a user, and can resume the snapshots on the request of a user. KnowledgeEditor works on Microsoft Windows2000 or XP. For more details about KnowledgeEditor and SRML, see <http://gscope.gsc.riken.go.jp>.

ACKNOWLEDGEMENTS

We thank Dr Hidemasa Bono and Dr Yasushi Okazaki for giving useful comments and discussion on the design of the software.

REFERENCES

Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput.*, 17–28.

D'haeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Proc. Pac. Symp. Biocomput.*, 41–52.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet.

Eisen,M.B., Spellman,P.T., Prown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002a) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.

Karp,P.D., Riley,M., Paley,S.M. and Pellegrini-Toole,A. (2002b) The MetaCyc database. *Nucleic Acids Res.*, **30**, 59–61.

Somogyi,R. and Shiegoski,C.A. (1996) Modeling the complexity of genetic networks: understanding multigene and pleiotropic regulation. *Complexity*, **1**, 45–63.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression pattern with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.