National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

*An Approach to Automated Knowledge Discovery in Bioinformatics **

Ouyang, J., Famili, F., and Xu, W.
September 2005

* published in the Proceedings of the Conference on Artificial Intelligence and Innovations (AIAI2005). Beijing, China. September 7-9, 2005. NRC 48248.

Canada

AN APPROACH TO AUTOMATED KNOWLEDGE DISCOVERY IN BIOINFORMATICS

Junjun Ouyang[1], A. Fazel Famili[1], Weiling Xu[2]
[1]Institute for Information Technology, National Research Council Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6 Canada; [2]118 Shady Grove Street, Nepean, ON K2G 6Z5 Canada (a co-op student with National Research Council Canada)

Abstract: Extensive data mining applications to bioinformatics research have shown that knowledge discovery requires repeated manual interventions, and that conglomerating and summarizing the results would be time consuming and sometimes error prone. To assist in efficiently applying data mining technologies in bioinformatics, we have developed Automation facilities in our data mining software suite. Experiences gained from case studies are extracted and presented as scenarios, which are sets of data processing and analysis operations for specific data mining objectives. Built as sequences of these predefined scenarios, procedures apply previously established data mining strategies to new data sets in an automated way. Automation also highlights the results particularly related to researchers' own areas of interest. We present insights into our automated knowledge discovery and two example scenarios extracted from one case study to demonstrate the usefulness of our approach.

Key words: data mining; bioinformatics; automation

1. INTRODUCTION

Advances in molecular biomedical technologies, e.g. large scale gene expression profiling[1,2] and high throughput sequencing[3,4], are producing enormous amount of data. Various knowledge discovery tools, techniques and algorithms have been applied to expression profile clustering[5], disease pattern classification[6], and DNA/protein sequence analysis[7].

How to efficiently apply various knowledge discovery technologies and strategies, however, is a challenge to bioinformatics researchers. As a solution, some tools or services that are used for sequence analysis and annotation[8,9], or for microarray data management and analysis[10-12] have developed automated pipelines, or provided the flexibility to build workflows based on specific analysis protocols. These developments focused mainly on building a workflow/pipeline as a connection of basic function modules. The knowledge discovery process, however, often involves interactive and iterative applications of complex computational operations, e.g. clustering analysis and various data transformation. Moreover, many biomedical researchers do not have enough knowledge and experience applying data mining techniques, while computer scientists usually lack the required expertise in biomedical research.

In our data mining software suite, BioMiner (http://iit-iti.nrc-cnrc.gc.ca/projects-projets/biomine_e.html), we are developing Automation facilities to encapsulate data analysis strategies, automate some operations, and highlight the most interesting discoveries. This approach involves: capturing real research experiences from case studies; structuring and presenting them as building blocks of knowledge discovery processes for new data sets.

2. DEVELOPMENT OF AUTOMATION

2.1 Motivation

To support knowledge discovery in our bioinformatics research, we have in-house developed a data mining software suite, BioMiner. Integrated into one environment, there are thirteen functional modules in data processing and analysis layers. These modules, e.g. statistics, visualization, clustering, and pattern recognition, are collections of data mining algorithms and tools. More specifically, clustering module has various clustering algorithms, e.g. hierarchical, K-Means and SOM; and pattern recognition module contains algorithms of decision trees, association rules, and so on. These modules are not only the units of software development, but also those of data analysis.

Working with biomedical researchers, we have applied BioMiner to microarray and sequence data analyses[13-18]. These case studies required well-defined data mining strategies and various functional modules. Results were reported in tables and displayed in, e.g. histograms and cluster centroid plots. One functional module may have been applied to a data set many times but with different settings. We then compared all the results to find common

patterns and important discoveries. In these studies, the repeated manual operations and result summarization were usually time-consuming and error-prone. We, nevertheless, treasure the analysis strategies established in these studies and believe that preserving and reusing them in the future is a way to cope with the complexity of biomedical knowledge discovery.

2.2 Design

Designing any automated facility requires an in-depth understanding of the task(s) for which the system is built. Based on completed case studies, sets of operations of usually one functional module are identified and generalized as Scenarios for specific data mining objectives or tasks. A scenario relies on running one module and summarizing the outputs. This helps researchers use the module appropriately and focus on important results. Different applications of the same module may be designed as different scenarios (Figure 1a). These scenarios, sets of data mining operations, are building blocks for a Procedure, or components of a knowledge discovery process (KDP). Figure 1b shows a procedure as a sequence of one or more scenarios. For instance, a procedure of a microarray data analysis could consist of the following scenarios in sequence: characteristic checking, clustering analysis, and pattern recognition.

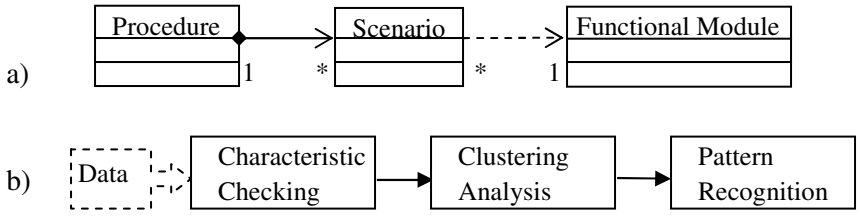

Figure 1. a). Object-oriented design: relationships between Procedure and Scenario, and between Scenario and Functional Module. b). A knowledge discovery process.

Other Automation features are also designed for creation, execution, and management of KDP's. Researchers may build a procedure consisting of several scenarios, set up appropriate parameters, and invoke the procedure. Automation executes scenarios in sequence, monitors the progress, and summarizes the outputs. Researchers then collect and study the results with highlighted discoveries.

Here, we use one of our genomics case studies to demonstrate what experiences are captured and how they are abstracted into Automation scenarios.

2.3 Case-Study-Based Scenario Design

In one case study on breast cancer[19], biologists isolated and characterized a murine mammary epithelial tumor cell line, which undergoes an epithelial-to-mesenchymal transition (EMT) as a result of a transforming growth factor (TGF-β1) exposure. We analyzed gene expression experiments to investigate TGF-β1 induced EMT and the role of p38 mitogen-activated protein kinase (p38MAPK). The expression data consisted of 331 genes (selected from a list of 15264 genes) obtained from cells treated with TGF-β1, or the p38MAPK inhibitor SB203580 (SB), or both of them (TGF-β1+SB). There were 6 experimental repeats for each treatment. In the next two subsections, we describe our data analyses and the designs of two scenarios accordingly.

2.3.1 A Pattern Recognition Scenario

We were interested in genes differentially expressed between two classes (treatments): (a) between TGF-β1+SB and SB; or (b) between TGF-β1+SB and TGF-β1. Two data sets were constructed that contained data for (a) or (b). Both expression data sets contained 331 genes (attributes) in 2 treatments (12 cases) respectively.

We used the pattern recognition module to generate decision tree models. A tree model consisted of one or more genes (nodes) that contain a particular threshold to discriminate between the classes in certain accuracy. These genes are considered informative as to the classification task. We removed (masked) these genes from the data and reran the algorithm to generate a new model involving some other genes. This operation of "discover-and-mask" was repeated until no more tree models could be built from the remainder of the data. With this approach, we highlighted informative genes in data set (a) and (b), respectively.

We captured operations of "discover-and-mask" into an Automation scenario, called EPR (Exhaustive Pattern Recognition). EPR automatically builds all possible models and identifies informative genes for the classifications. EPR also summarizes in a table the genes involved, discriminating threshold, class assignments and accuracy for each model. To assess the importance and informativeness, the models and involved genes can be ranked by their classification accuracies.

2.3.2 A Clustering Scenario

We also applied K-Means algorithm in the clustering module to select genes based on their expression profiles under the three treatments: TGF-β1,

TGF-β1+SB and SB. Here genes are objects and their expression levels in 18 experiments (3 treatments with 6 repeats for each) are attributes.

Our analysis started with running a series of K-Means clustering with different K (the number of clusters). Quality Measure (QM) facility[13] then helped us to determine the appropriate number of clusters for the 331 genes. With the aid of visual plots, we identified clusters of genes with interesting expression patterns. The expression pattern of one cluster is down regulated from TGF-β1 to TGF-β1+SB and to SB, while those of other clusters appeared relatively flat. We highlighted this cluster of genes for further biological validations.

We captured operations of this approach into an Automation scenario, called CQC (Clustering Quality Comparison). CQC automatically runs a series of clustering with different numbers of clusters. To help determine the most appropriate data division and choose interesting clusters, QM reports qualities for each clustering operation and for the generated clusters.

Additional EMT biological analysis brought the attention to some of the EPR and CQC highlighted genes for their modulation by TGF-β1 and the role of p38MAPK activities[19].

2.4 Implementation

Automation facilities are implemented within the BioMiner suite. When invoked from menu, Automation graphic user interface (GUI) is displayed next to the BioMiner main interface. Users therefore may easily switch between regular and automated data mining research activities.

Automation GUI has three components (Figure 2): Scenario List, Procedure Editor and Procedure List. Users drag an available scenario from Scenario List, and drop it to the panel of Procedure Editor. A procedure is built as a connected sequence of the selected scenarios. Users may set up or modify the parameters, if any, in a dialog box specific for each scenario. Users may also edit a procedure by adding and deleting scenarios, or cancel the editing and restore a procedure to its previous setting.

A procedure can be saved to a file and later reloaded into Automation for editing. Automation only saves the sequence of scenarios of a procedure and associated parameter settings, which do not require much storage space. Java serialization is the technology for saving and loading procedures. With graphic Procedure Editor in Automation, it is unnecessary to manually edit a text version of a procedure and its scenarios.

Once a procedure is started, Automation will execute embedded scenarios in sequence. In the case of extremely computation-intensive applications, we rely on a parallel (a computing cluster) version for the scheduling and allocation of the computing resources. Automation reports

the outcome and summarizes the results of each scenario and the procedure. A procedure can be applied to different data sets, and the same procedure on the same data may have different outputs that for instance may result from random initial seeds in K-Means clustering. Automation, therefore, saves the output files in a new folder for each run of a procedure.

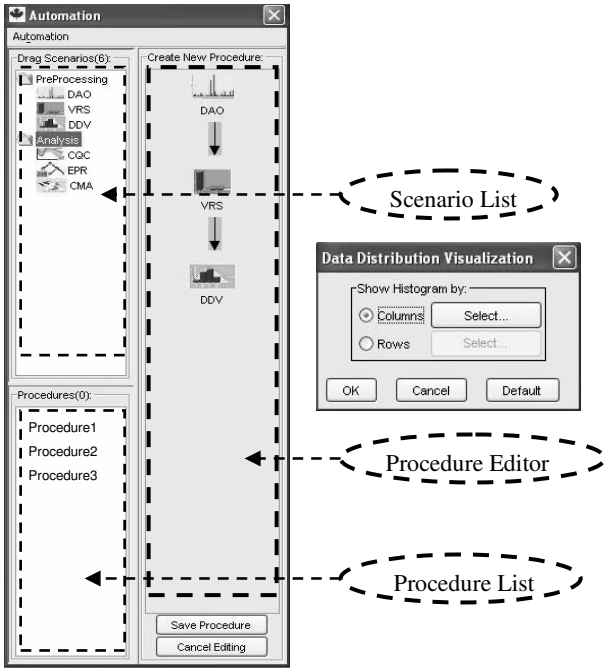

Figure 2. Automation interface: Scenario List; Procedure Editor; and Procedure List. A scenario's settings may be edited the left dialog box.

In case studies[13,17], the CQC based strategy helped to identify clusters of genes with targeted patterns. The EPR based technique generated lists of genes highly related to classification tasks in case studies[15,17,18]. Other scenarios were also developed for data quality examination and DNA sequence analysis. The more studies are carried out, the more experiences are accumulated as scenarios in Automation. In this way, Automation not only reduces manual intervention in the knowledge discovery processes, but also lowers the learning curve for biomedical researchers to plan advanced data analysis.

3. DISCUSSIONS

Production of high throughput genomics data and rapid advancements of biomedical research call for various data analysis technologies and application strategies. We introduce a novel approach to the research and development of automated data mining in bioinformatics. Processing operations and analysis strategies in case studies are abstracted and designed into Automation to reduce user's intervention and more importantly to lead the knowledge discovery activities in bioinformatics.

The development of Automation is an evolving process, as new scenarios are identified as a result of more case studies. A scenario editor would offer users the tool to create scenarios based on their own research experiences. More features for workflow management and results summarization will also be designed and implemented as our research and development advance.

Scenarios in Automation play a key role in transforming research experiences into powerful methods for later applications. The advantage of our approach is to guide bioinformatics researchers on how to plan and apply knowledge discovery processes. Other researchers and developers may wrap various data analysis experiences in the forms of scenarios or agents in order to automate their own technology workbenches.

ACKNOWLEDGEMENT

For biological experiments and analysis in the EMT case study, we would like to thank Dr. Maureen O'Connor-McCourt and Dr. Anne Lenferink (Biotechnology Research Institute, National Research Council Canada) for their special contributions. We would like to thank Dr. Youlian Pan for the valuable comments on an earlier version of this paper. We also acknowledge the contributions of all members of the BioMine project from the Institute for Information Technology and the Institute for Biological Sciences, National Research Council Canada.

REFERENCES

[1] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, 1996, 14: 1675-1680.

[2] Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K.W. Serial analysis of gene expression. *Science*, 1995, 270: 484-487.

[3] Lander, E.S., et al. Initial sequencing and analysis of the human genome. *Nature,* 2001, 409: 860-921.

[4] Venter, J. C., et al. The sequence of the human genome. *Science,* 2001, 291: 1304-1351.

[5] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA,* 1998, 95: 14863-14868.

[6] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* 1999, 286: 531-537.

[7] Zhu, Z., Pilpel, Y., and Church, C. M. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol*., 2002, 318: 71-81.

[8] Rowe, A., Kalaitzopoulos, D., Osmond, M., Ghanem, M., Guo, Y. The discovery net system for high throughput bioinformatics. *Bioinformatics,* 2003, 19 (Suppl. 1):i225-231.

[9] Shah, S. P., He, D. Y., Sawkins, J. N., Druce, J. C., Quon, G., Lett, D., Zheng, G. X., Xu, T., Ouellette, B. F. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics,* 2004, 5: 40.

[10] Grant, J. D., Somers, L. A., Zhang, Y., Manion, F. J., Bidaut, G., Ochs, M. F. FGDP: functional genomics data pipeline for automated, multiple microarray data analyses. *Bioinformatics,* 2004, 20: 282-283.

[11] Hokamp, K., Roche, F. M., Acab, M., Rousseau, M. E., Kuo, B., Goode, D., Aeschliman, D., Bryan, J., Babiuk, L. A., Hancock, R. E., Brinkman, F. S. ArrayPipe: a flexible processing pipeline for microarray data. *Nucleic Acids Res.,* 2004, 32: W457-9.

[12] Knudsen, S., Workman, C., Sicheritz-Ponten, T., Friis, C. GenePublisher: Automated analysis of DNA microarray data. *Nucleic Acids Res.,* 2003, 31: 3471-3476.

[13] Famili, A. F., Liu, G., Liu, Z. Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics*, 2004, 20: 1535-1545.

[14] Famili, A. F., Liu, Z., Ouyang, J., Walker, R. R., Smith, B., O'Connor, M., and Lenferink, A. A novel data mining technique for gene identification in time-series gene expression data. presented at the 16th European Conference on Artificial Intelligence, 2004, Valencia, Spain.

[15] Famili, A. F., Ouyang, J. Data mining: understanding data and disease modeling. *Applied Informatics,* 2003: 32-37.

[16] Famili, A. F., Ouyang, J., Kryworuchko, M., Alvarez-Maya, I., Smith, B., and Diaz-Mitoma, F. Knowledge Discovery in Hepatitis C Virus Transgenic Mice. *IEA/AIE,* 2004: 29-39.

[17] Pan, Y., Pylatuik, J. D., Ouyang, J., Famili, A. F., and Fobert, P. R. Discovery of functional genes for systemic acquired resistance in *Arabidopsis Thaliana* through integrated data mining. *J. Bioinform. Comput. Biol.,* 2004, 2: 639-655.

[18] Walker, P. R., Smith, B., Liu, Q. Y., Famili, A. F., Valdes, J. J., and Liu, Z. Data mining of gene expression changes in Alzheimer brain. *Artif. Intell. Med.,* 2004, 31: 137-154.

[19] O'Connor-McCourt, M., Lenferink, A., Nantel, A., Cantin, C., Magoon, J., Ouyang, J., Lui, Z., and Famili, A. F. Analysis of Transforming Growth Factor (TGF)-ß Modulated Genes Involved in the Epithelial to Mesenchymal Transdifferentiation of Murine Mammary Epithelial Cells. Poster presentation at the 94th Annual Meeting of the American Association for Cancer Research, 2003, Washington, USA.