Review Article

Text Mining Perspectives in Microarray Data Mining

Jeyakumar Natarajan

Data and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore 641 046, India

Correspondence should be addressed to Jeyakumar Natarajan; jeyakumar.natarajan@gmail.com

Received 19 July 2013; Accepted 4 September 2013

Academic Editors: Z. Su and Z. Yu

Copyright © 2013 Jeyakumar Natarajan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Current microarray data mining methods such as clustering, classification, and association analysis heavily rely on statistical and machine learning algorithms for analysis of large sets of gene expression data. In recent years, there has been a growing interest in methods that attempt to discover patterns based on multiple but related data sources. Gene expression data and the corresponding literature data are one such example. This paper suggests a new approach to microarray data mining as a combination of text mining (TM) and information extraction (IE). TM is concerned with identifying patterns in natural language text and IE is concerned with locating specific entities, relations, and facts in text. The present paper surveys the state of the art of data mining methods for microarray data analysis. We show the limitations of current microarray data mining methods and outline how text mining could address these limitations.

1. Introduction

DNA microarrays facilitate the simultaneous measurement of the expression levels of thousands of genes [1, 2]. As a result, this high-throughput technology has led to increased amount of gene expression data. Microarrays have been used for a variety of studies, including gene coregulation studies, gene function identification studies, identification of pathway and gene regulatory networks, predictive toxicology, clinical diagnosis, and sequence variance studies. For a complete description about microarrays and its analytical tasks, refer to the books [3–5]. Current microarray data mining methods such as clustering, classification, and association analysis are based on statistical and machine learning algorithms. Most of these techniques are purely data driven and do not incorporate significant amounts of biological knowledge. Considering the statistically ill-defined nature of microarray data (many more variables than observations) and the massive body of existing biological knowledge, it is imperative that we exploit that knowledge for analysis and interpretation of microarray data. Text mining techniques constitute a promising technology for automating the incorporation of scientific knowledge in the microarray data mining process.

Applying domain knowledge is fundamental in any scientific discovery process. In biology, domain knowledge

is available in vast collections of the literature in natural language form such as abstracts [6] and full-text journal articles [7, 8] and also as textual annotations in databases such as SwissProt [9] and GenBank [10]. For example, the biological abstract database PubMed comprises more than 19 million citations for biomedical articles from MEDLINE and life science journals. The rapid growth of literature databases and structured repositories renders it increasingly difficult for humans to access the required information within reasonable time constraints. Text mining and information extraction are computerized techniques facilitating the automated filtering and analysis of large amounts of electronic texts. Text mining can be defined as the process of identifying nontrivial, implicit, previously unknown, and potential useful patterns in natural language text [11]. Information extraction, on the other hand, focuses on the identification of specific predefined classes of entities, relations, or facts in natural language text, records, and presents this information in a structured format [12]. In a simple definition, the goal of information extraction is to extract nuggets of information from text and text mining is to find new knowledge. However, both text mining and information extraction are complex processes composed from multiple tasks approached by disciplines such as statistics, information retrieval, natural language processing, machine learning, and artificial intelligence. For

a complete description and review about TM and IE methods and natural language processing, refer to the books [13–16].

In recent years, there has been a growing interest in methods that attempt to simultaneously discover patterns occurring in multiple data sources. Gene expression data analysis is an example for such a scenario, as it is concerned with the analysis of the actual expression data in conjunction with existing relevant textual information of genes, proteins, diseases, and so on. The goal here is to come up with solutions for statistical issues and limitations in current microarray data mining with integrated text mining.

The following sections introduce in detail the current statistical limitations in all the three stages of microarray data mining pipeline, that is, (i) data preprocessing, (ii) data modeling or model construction, and (iii) postprocessing, and outline how text mining could address these limitations with related literature references.

2. Text Mining Perspectives in Microarray Data Mining

Can text mining be incorporated into the microarray analysis process: where and how? This review tries to provide an answer to this question. Common strategies of data mining in microarray are

- (i) data preprocessing;
- (ii) data modeling or model construction;
- (iii) data postprocessing.

Text mining techniques constitute a promising technology for automating the incorporation of scientific knowledge in all the above data mining process. For example, text mining outputs (i.e., new knowledge) could be combined and correlated with actual gene expression data in model construction (i.e., clustering, classification, and association analysis). Further, text mining can also be applied earlier in data preprocessing for feature selection, data transformation, and data enrichment and in the postprocessing phase for interpretation and knowledge-based validation microarray analysis results (see discussions below). The three major steps of microarray data mining and how text mining approaches could help are shown in Figure 1.

2.1. Data Preprocessing. The data preprocessing task data mining focuses on the preparation and transformation of the “raw” data for further analysis and processing. The most important aspects in microarray data preprocessing include data transformation (normalization, centralization, and standardization), the handling of missing value imputation and other forms of error correction, data discretization, feature selection and dimension reduction, and data enrichment.

Of major interest is the task of feature selection. Feature selection and dimension reduction techniques are important for tackling the dimensionality problem inherent in microarray data (a.k.a *curse of dimensionality* or *large- p -small- n problem*). Feature selection can also directly lead to the discovery of marker genes, that is, genes associated with

FIGURE 1: Typical processing flow and applications of text mining in microarray analysis. Text mining methods could be applied in all the three stages of data mining pipeline.

the phenotype under investigation (e.g., cancer class). To our knowledge, all publications addressing the issue of feature selection in the context of microarrays involve statistical or machine learning methods only, such as the p -metric [17], information gain ranking [18], and Chi square statistic [19]. These methods take into account only the numerical values of the microarray matrix. However, these purely statistical approaches have been criticized by King and Sinha and O’Neill et al. [20, 21]. West et al. [22] point out that not only highly expressed genes are of main interest, but also are those genes whose expression highly correlates with the phenotype, regardless of the level of expression. Clearly, statistical significance does not necessarily imply biological relevance.

The problem with the methods relying on the numerical representation of gene expression can be illustrated in a simple example. Suppose that we are interested in identifying those genes whose expression profile highly correlates with a cancer, and let us assume that we investigate two classes, A (cancer) and B (normal). If a gene X is highly overexpressed in samples of type A , but underexpressed in samples of type B , then the vast majority of currently applied feature selection methods would identify this gene as a marker gene. However, a large number (if not the majority) of genes found to be overexpressed in tumor cells simply reflect the fact that aggressive cells tend to be in active cell cycle and thus express genes that are known to play a pivotal role in the cell cycle. Thereby, genes associated with general homeostasis functions such as metabolism, protein synthesis, and cytoskeletal structures are more highly expressed and are likely to dominate the expression profile. Genes might play a pivotal role in cancerogenesis, yet they are not frequently missed from being identified [21]. Furthermore, it is known that cancer is a multistep process [23]. However, microarray experiments in cancer research deal with “mature” cancers, that is, tumors that have already reached a critical mass that can be diagnosed. Therefore, it is problematic to identify the “trigger genes,” that is, those genes that are responsible for cancerogenesis in the first place. How can we prevent the genes associated with general homeostasis functions from obscuring the global picture? Current number-based approaches are unable to readjust the focus on those genes.

Text mining could be the method of choice to address this problem. Text classification and clustering methods have the potential to identify those genes that are involved in general homeostasis functions using the literature data and thus allowing the exclusion of those genes from further analysis. Further, text clustering might be used to group genes according to their functions using related literature of genes, which allows to select functionally important genes from different clusters.

In contrast to number crunching statistical and machine learning approaches, text mining can also select genes on the basis of *other semantic* criteria. This facilitates the filtering of genes that are known to be involved in specific pathways, sharing similar functions, or have same cellular localization. For example, suppose that we are interested in comparing the expression of some genes of known or putative cancer inhibitory function in two different types of cancer. For each gene, good sources of scientific texts are available. By mining these resources, we can identify the genes that we are interested in and compare, for instance, their expression in different cancer types.

2.2. Data Modeling. The data modeling task focuses on construction of data models using expression data as clusters, association rules, and classification systems.

Current clustering of microarray data has been performed on the basis of gene expression measurements represented in numerical format, that is, mostly as real numbers [24–26]. However, the results of clustering and other data mining methods depend on a number of factors, not only on the data to be analyzed. Different clustering algorithms, for example, usually detect different, yet meaningful, groups in the same data set. Clustering algorithms can produce different result for different parameter initializations, for example, *k*-means and fuzzy *c*-means [26]. Most algorithms compute some kind of similarity or distance function as criterion for assigning objects to clusters. However, the notion of similarity is extremely difficult to define in high-dimensional data sets such as microarray data [27]. Further, biostatisticians do not agree over which number-based clustering technique should be applied under which circumstances [28]. Furthermore, biological validation and interpretation of cluster analysis results is a difficult task [29].

Text mining can pave the way to a fundamentally different concept of clustering instead of focusing on the numeric expression values, one could group genes according to a semantic concept membership, for example, with respect to their function, cell localization, or role in a disease, and the expression profiles are inspected in the next step. For example, Raychaudhuri and colleagues [30] investigated whether the genes within a gene expression cluster share a common biological function based on the associated scientific literature using text-clustering methods. Another system, PubGene [31], is used for the functional interpretation of genes expressed in gene expression clusters using the literature information. PubGene initially compiles a network of human genes by finding the cooccurrences genes names within the scope of relevant Medline abstracts. Next, it uses

genes names in this network to search for other literature references involving these genes and annotates the networks with that literature information. These compiled annotations were then compared to the gene expression cluster results for the correlation.

Various machine learning and data mining methods are currently used for classifying gene expression data. However, these methods have not been developed to address the specific requirements of gene microarray analysis. As pointed out by Sabatti [32], the main issues for the microarray data analysis by machine learning and data mining methods include experimental design, noise level, measurement errors, and quality. Furthermore, such statistical techniques do not address the biologist's requirement for sound mathematical confidence measures and also misclassification costs; that is, they are indifferent to the costs associated with false positive and false negative classifications [33]. Therefore, purely statistical models are prone to a variety of problems. We believe that text mining methods have a great potential to complement current statistical and machine learning techniques, thus representing a more adequate framework for analysis and interpretation of microarray data. Following are few examples on how text mining could help to answer these questions.

Classification of microarray data requires methods for dimension reduction to overcome the *curse of dimensionality*. Using text mining, it may become possible to build more sophisticated classifiers that take into account existing background knowledge. For example, assume a classification problem in cancer subtype or outcome prediction [34]. In such classification problem, the classifier might weigh the genes differently, depending on their function, instead of focusing on the numeric expression values alone. This will be the potential area of research, and as per our knowledge no work has been done along these lines. Instead, many groups explored the possibilities of complementing these kinds of studies with text mining [35–37]. However, integrated analysis is a challenging and open problem.

For example, consider the KDD cup 2002 task 2 [35]. It explores the possibility of predicting a gene expression classification system using the literature and other data resources. The task was to analyze the knockout behavior of yeast genes in expression data using text literature of about 15,000 scientific abstracts and other data resources such as gene interaction information, subcellular localization hierarchies, and gene functions. The winning solution combines information extraction and text mining methods. IE methods were used to extract key features/terms from the given literature, which were used by the text classification algorithm to classify the genes.

Mining microarray data for association between genes in the form of pathways can reveal new insights into underlying biological or pathological systems, but it is a computationally complex task because thousands of potentially interesting associations can be derived from a data set that comprises a large number of genes. However, text mining methods could come up with predefined sets of relations and interactions between genes and proteins. Resultant interactions and associations may be combined and correlated with gene

expression analysis. This problem has been conferred by Natarajan et al. in [36]. They have shown how text mining could be a complement to a purely data-driven approach in finding the impact of SIP on invasivity of a glioblastoma cell lines. Another text mining system, Literature Lab [37], was used to find the association between the pathway related to FOSB and genes with increased expression in metastatic prostate cancer.

2.3. Data Postprocessing. This task focuses on biological validation, and interpretation of microarray analysis results applying statistical and/or machine learning methods for the analysis of microarray data requires a sound validation of the results. This validation could in principle involve three basic steps:

- (1) validation using statistics (P values, confidence intervals, measures like specificity, sensitivity, etc.);
- (2) validation based on further biological experiments;
- (3) validation based on available domain knowledge (provided by human expert, obtained in semistructured format from scientific literature repositories, or as highly structured records in information bases such as SwissProt, GenBank, and others).

The validation of the obtained results using statistics is the most straightforward type of validation. For example, many studies focus on the classification of cancer types based on gene expression profiles [38, 39]. To assess the statistical significance of their results, many researchers use statistical tests, including random permutation tests. The validation of the expression study may also trigger additional biological experiments. Yeoh et al. [19], for example, validated their results by FISH experiments. Combining multiple cytogenetic experiments may also validate the results of a microarray experiment; see, for example, Berrar et al.'s work [40], where inductive classification methods often employ simple accuracy measures (e.g., precision, lift) and cross-validation to validate the induced classifiers. Statistical validation is relatively straightforward and inexpensive (both in terms of money and time), while to validate and interpret microarray analysis results in terms of their biological relevance, on the other hand, is not. It requires the combined expertise of the involved scientists, such as pathologists, cytogenetists, chemists, and biologists. For example, a potential outcome of a microarray study focusing on cancer-related genes might reveal that a specific gene is statistically strongly correlated to the cancer under investigation. Here, text mining comes into play—by validating the result using existing knowledge bases such as scientific text. The analysis of the text repositories may reveal that the identified gene does play a pivotal role in a specific pathway that is related to the cancer under investigation. This approach is far from being new; in fact, it is common practice to validate the results by referring to already published material. However, this type of validation is still being performed in a highly “manually” fashion by the researchers. By sifting through the already existing vast amount of information, text mining could be the tool for

finding new nuggets of knowledge and become a valuable complement to the biological validation.

3. Conclusion

The analysis of microarrays has become one of the main research areas in computational biology, and new methods and applications are continuously being developed. Here, we have reviewed current limitations of microarray data mining using statistical and machine learning methods and how text mining could come up with solutions to overcome these limitations. The fact that free text information is far from being unambiguous remains an open problem. As Chang and Altman pointed out [41], maybe the most difficult problem is that our knowledge about living systems is fluid as definitions of words and conceptual paradigms change.

Microarray technology has become a mature method for monitoring the expression of thousands of genes, yet the incorporation of existing domain knowledge remains an open problem. We believe that currently applied statistical and machine learning methods are too limited to exploit the full potential of microarray data. A plethora of information is already out there in publicly accessible databases, and TM can help to incorporate this knowledge in the analysis process. The high-throughput technologies of the next generation, such as protein chips, are most likely to entail similar problems than those related to gene microarray data. We believe that text mining will become the tool to achieve the next quantum leap in analyzing these high-throughput data.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author wishes to thank Werner Dubitzky and Daniel Berrar at the University of Ulster, UK, for their valuable suggestions and fruitful discussions.

References

- [1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray,” *Science*, vol. 270, no. 5235, pp. 467–470, 1995.
- [2] J. L. DeRisi, V. R. Iyer, and P. O. Brown, “Exploring the metabolic and genetic control of gene expression on a genomic scale,” *Science*, vol. 278, no. 5338, pp. 680–686, 1997.
- [3] “The chipping forecast I,” *Supplement to Nature Genetics*, vol. 21, no. 1, 1999.
- [4] “The chipping forecast II,” *Supplement to Nature Genetics*, vol. 32, 2002.
- [5] *A Practical Approach to Microarray Data Analysis*, Kluwer Academic, Boston, Mass, USA, 2002, edited by D. Berrar, W. Dubitzky and M. Granzow.
- [6] National Library of Medicine, “PubMed literature abstract database,” <http://www.ncbi.nlm.nih.gov/pubmed>.

- [7] National Library of Medicine, "PubMed Central full text repository," <http://www.pubmedcentral.gov>.
- [8] BioMed Central, "The open access publisher," <http://www.biomedcentral.com>.
- [9] Swiss institute of Bioinformatics, "SwissProt protein sequence database," <http://us.expasy.org/sprot>.
- [10] National Library of Medicine, "GenBank sequence database," 2012, <http://www.ncbi.nlm.nih.gov/genbank>.
- [11] M. A. Hearst, "Untangling text data mining," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL '99)*, University of Maryland, 1999.
- [12] J. Cowie and W. Lehnert, "Information extraction," *Communications of the ACM*, vol. 39, no. 1, pp. 80–91, 1996.
- [13] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*, John Benjamins, Amsterdam, The Netherlands, 2002.
- [14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Mass, USA, 1999.
- [15] J. Allen, *Natural Language Understanding*, The Benjamin/Cummings, Menlo Park, Calif, USA, 1995.
- [16] J. Natarajan, D. Berrar, C. J. Hack, and W. Dubitzky, "Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications," *Critical Reviews in Biotechnology*, vol. 25, no. 1-2, pp. 31–52, 2005.
- [17] D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander, "Class prediction and discovery using gene expression data," in *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB '00)*, pp. 263–272, Universal Academy Press, Tokyo, Japan, April 2000.
- [18] E. P. Xing, M. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 601–608, 2001.
- [19] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.
- [20] H. C. King and A. A. Sinha, "Gene expression profile analysis by DNA microarrays: promise and pitfalls," *Journal of the American Medical Association*, vol. 286, no. 18, pp. 2280–2288, 2001.
- [21] G. M. O'Neill, D. R. Catchpole, and E. A. Golemis, "From correlation to causality: microarrays, cancer, and cancer treatment," *BioTechniques*, vol. 34, no. 3, pp. S64–S71, 2003.
- [22] M. West, C. Blanchette, H. Dressman et al., "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11462–11467, 2001.
- [23] B. Vogelstein and K. W. Kinzler, "The multistep nature of cancer," *Trends in Genetics*, vol. 9, no. 4, pp. 138–141, 1993.
- [24] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.
- [25] P. Tamayo, D. Slonim, J. Mesirov et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999.
- [26] M. Granzow, D. Berrar, W. Dubitzky, A. Schuster, F. J. Azuaje, and R. Eils, "Tumor classification by gene expression profiling: comparison and validation of five clustering methods," *ACM SIGBIO Newsletter*, vol. 21, no. 1, pp. 16–22, 2001.
- [27] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proceedings of the 8th International Conference on Database Theory (ICDT '01)*, pp. 420–434, 2001.
- [28] C. Tiltstone, "Vital statistics," *Nature*, vol. 424, no. 6949, pp. 610–612, 2003.
- [29] F. Azuaje, "Clustering-based approaches to discovering and visualising microarray data patterns," *Briefings in Bioinformatics*, vol. 4, no. 1, pp. 31–42, 2003.
- [30] S. Raychaudhuri, J. T. Chang, P. D. Sutphin, and R. B. Altman, "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature," *Genome Research*, vol. 12, no. 1, pp. 203–214, 2002.
- [31] T.-K. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," *Nature Genetics*, vol. 28, no. 1, pp. 21–28, 2001.
- [32] C. Sabatti, "Statistical issues in microarray analysis," *Current Genomics*, vol. 3, no. 1, pp. 7–12, 2002.
- [33] D. Berrar, C. S. Downes, and W. Dubitzky, "Multiclass cancer classification using gene expression profiling and probabilistic neural networks," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 8, pp. 5–16, 2003.
- [34] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.
- [35] "KDD Cup 2002 task 2 of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining," <http://www.biostat.wisc.edu/~craven/kddcup/index.html>.
- [36] J. Natarajan, D. Berrar, W. Dubitzky et al., "Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line," *BMC Bioinformatics*, vol. 7, no. 1, p. 373, 2006.
- [37] P. G. Febbo, M. G. Mulligan, D. A. Slonina et al., "Literature Lab: a method of automated literature interrogation to infer biology from microarray analysis," *BMC Genomics*, vol. 461, pp. 8–18, 2007.
- [38] J. Khan, J. S. Wei, M. Ringnér et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.
- [39] S. Ramaswamy, P. Tamayo, R. Rifkin et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.
- [40] D. Berrar, M. Granzow, W. Dubitzky et al., "New insights in clinical impact of molecular genetic data by knowledge-driven data mining," in *Proceedings of the 2nd International Conference on Systems Biology*, pp. 275–281, Omni press, 2001.
- [41] J. T. Chang and R. B. Altman, "Extracting and characterizing gene-drug relationships from the literature," *Pharmacogenetics*, vol. 14, no. 9, pp. 577–586, 2004.