Integrated Data Mining and Text Mining In Support of Bioinformatics

M. Ghanem, Y. Guo and A.S. Rowe

Department of Computing, Imperial College London,
180 Queens Gate, London, SW7 2AZ
{mmg, yg, asr99}@doc.ic.ac.uk

Abstract

In this paper we present case studies in conducting integrated data and text mining activities within the Discovery Net project. We discuss how our open infrastructure provides a powerful workbench for the dynamic analysis and interpretation of bioinformatics data. Our examples include using text mining to aid the interpretation of gene expression data and also in relating metabonomic experimental results to those of gene expression data analysis.

1. Introduction

A fundamental problem facing biological researchers today is how to make effective use of the available wealth of online background domain knowledge to improve their understanding of complex biological systems. Effective use of such background domain knowledge plays a crucial role in all stages of integrative biological research studies.

Firstly, such background knowledge is an essential component of the prospective research phases where the aim is to survey, collect and understand the available published information on the state of the art in the research area with the aims of planning, designing and conducting the desired experiments.

Secondly, and more importantly, it also plays a crucial role during the research study itself where the aim is to collect and use such background information to validate, annotate and interpret the discovery results generated from analysing the experimentally generated data.

A plethora of online database sources provides curated background information in the form of structured (data tables) and semi-structured (such as XML) content about genes, their products and their involvement in identified biological systems. However, the main source of most background knowledge still remains to be scientific publication databases (e.g. Medline) that store the available information in an unstructured form; the required information is embedded within the free text found in each publication. The ability to automatically and effectively extract, integrate, understand and make use of this information embedded in such publications remains a challenging task.

In the remainder of this paper we describe two examples of conducting integrated data and text mining in the context of the bioinformatics research activities within the Discovery Net project [1]. Our aim is to investigate and develop methods whereby information integration methods and text mining methods can be used together to validate and interpret the results of a data mining procedure. The examples make use of Discovery Net's workflow model and execution engines as well as of the Discovery Net InfoGrid framework [2] that allows us to dynamically access and integrate data from a variety of online bioinformatics data sources. .

2. Text Mining for Interpreting Gene Expression Analysis Results

As a first example, consider a scientist engaged in the analysis of microarray gene expression data using traditional data clustering techniques. The result of this clustering analysis could be a group of co-regulated genes (i.e. genes that exhibit similar experimental behaviour) or could be groups of differentially expressed genes. Once theses groupings are isolated, the scientist may wish to investigate and validate the significance of his findings by:

a. Seeking background information on why such genes are co-regulated or differentially expressed, and
b. Identifying the diseases that are associated with the different isolated gene groupings.

Much of the required information is available on online genomic databases, and also

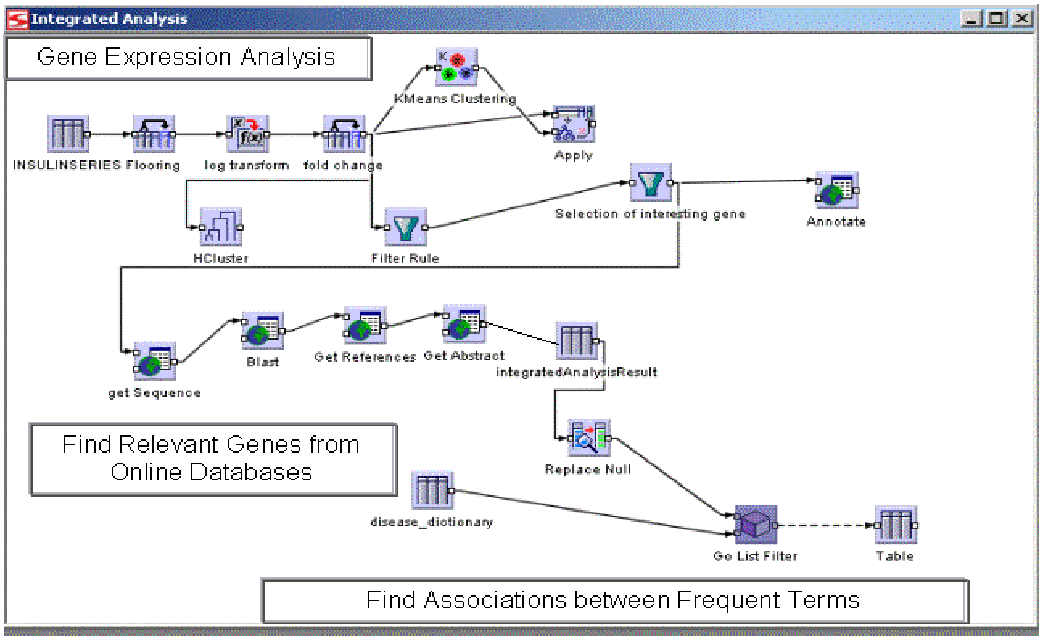

Figure 1 Integrated Data Mining / Text Mining Workflow for retrieving and summarising information on gene expression data

in scientific publications. What the user requires are interactive methods that enable him to access such information dynamically, summarise it and re-integrate it in his analysis.

This mode of scientific discovery is supported by the Discovery Net workflow shown in Figure 1. This workflow is divided into three logical phases:

1. The first phase ("Gene Expression Analysis"), corresponds to the traditional data mining phase, where the biologist conducts analysis over gene expression data using a data clustering analysis component to find co-regulated/differentially expressed genes. The output of this stage is a set of "interesting genes" or "gene groupings" that the data clustering methods isolate as being candidates for further analysis.
2. In the second phase of the workflow ("Find Relevant Genes from Online Databases") the user uses the InfoGrid integration framework to obtain further information about the isolated genes from online databases. In this phase, the workflow starts by obtaining the nucleotide sequence for each gene by issuing a query to the NCBI database based on the gene accession number. The retrieved sequence is then used to execute a BLAST query to retrieve a set of homologous sequences; these sequences in turn are used to issue a query to the SwissProt database to retrieve the PubMed Ids identifying articles relating to the homologous sequences. Finally the PubMed Ids are used to issue a query against PubMed to retrieve the abstracts associated with these articles, and the abstracts are passed through a frequent phrase identification algorithm to extract summaries for the retrieved documents for the gene and its homologues.
3. Finally in the third phase of the workflow ("Find Association between Frequent Terms") the user uses a dictionary of disease terms obtained from the MESH (Medical Subject Headings) dictionary to isolate the key disease terms appearing in the retrieved articles. The identified disease words are then analysed using a standard association analysis a priori style algorithm to find frequently co-occurring disease terms in the retrieved article sets that are associated with both the identified genes as well as their homologues.

3. Mapping Micro-array Probes to Metabolites

The second example shows how the Discovery Net infrastructure can support finding correlations between data sets obtained from different experiments. In this case, these are two data sets, one obtained from microarray experiments and the other from NMR-based metabonomic experiments. Both data sets are obtained from a project relating to studying insulin resistance in mice [3]. The microarray gene expression data measures the amount of RNA expressed at the time a sample is taken, and the NMR spectra are for metabolites found in urine samples of the same subjects. In this example the user is interested to find known associations between the genes that isolated as "interesting" from the first data set and the metabolites identified as "interesting" from the second.

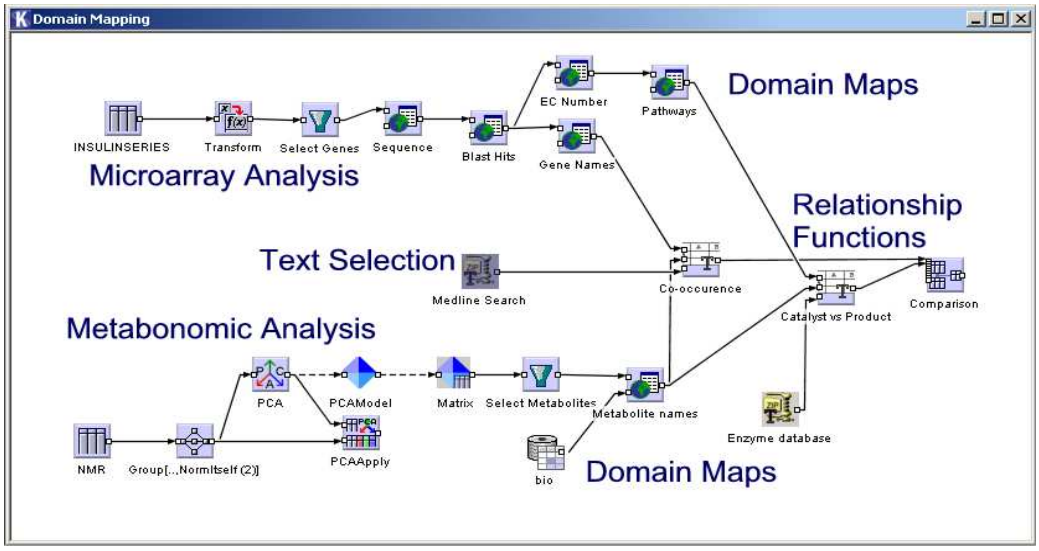

Figure 2 Workflow for the integrated analysis of gene expression and metabonomics data

This mode of analysis is supported by the workflow shown in Figure 2, and which proceeds into three logical phases:

1. The first phase ("Microarray analysis) uses standard gene expression analysis technique to filter interesting genes within the gene expression domain. The gene expression process that is used is starts by mapping the gene expression probe id to the sequence that would bind to that area. Using the sequence, we use BlastX to search the Swiss-Prot database. This provides a method of finding known genes. After the blast process, we use the hits from this database to download features from the actual records from the Swiss-Prot database to annotate the probe ID with possible gene names for the sequence and any Enzyme commission number when it exists.
2. In parallel, the second phase ("Metabonomic Analysis") proceeds by analysis the NMR data using multivariate analysis to study the NMR shifts, and mapping them to candidate metabolites using both manual processes and NMR shift databases. The output of this phase is a set of candidate metabolite names.
3. The third phase ("Text Selections and Relationship Functions") then proceeds based on the "joining" the outputs of the phases 1 and 2 to find known associations between the genes and the metabolites. This phase proceeds by a) Searching pathway databases for known relationships between the metabolites and the genes, and b) Searching scientific publications using a co-occurrence analysis approach to find the most general relationships possible between the metabolites and the genes. The outputs of both types of analysis is then merged and presented to the user.

4. Summary and Discussion

In this paper we have presented an approach for enabling the interpretation of bioinformatics data by using dynamic information integration techniques and workflows that mix both data and text mining methods. Our approach is generic and can be applied to various other case studies.

In addition to those examples described in this paper, we have recently been using the system to develop a large number of workflows that concentrate on the text mining aspects of result interpretation, examples include scientific document categorization [4], named entity extraction from text documents and studying the co-citation of genes and diseases within large document collections.

References

[1] V. Curcin, M. Ghanem, Y. Guo, M. Kohler, A. Rowe, J Syed, P. Wendel. Discovery Net: Towards a Grid of Knowledge Discovery. Proceedings of KDD-2002. The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23-26, 2002 Edmonton, Canada.

[2] Giannadakis N, Rowe A, Ghanem M and Guo Y. InfoGrid: Providing Information Integration for Knowledge Discovery. *Information Science*, 2003: 3:199-226.

[3] Rowe A, Ghanem M, Guo Y. Using Domain Mapping to Integrate Biological and Chemical Databases. *International Chemical Information Conference*, Nimes, 2003.

[4] Ghanem M. M, Guo Y, Lodhi H, Zhang Y, Automatic Scientific Text Classification Using Local Patterns: KDD CUP 2002 (Task 1), *SIGKDD Explorations*, 2002. Volume 4, Issue 2