Mining literature to improve biological knowledge extraction by microarray transcriptional profiling

R. A. Calogero¹, G. Iazzetti², S. Motta³, G. Pedrazzi⁴, S. Rago⁴, E. Rossi⁴, R. Turra⁴
¹Department of Clinical and Biological Sciences, Turin University, Italy.
²Department of Genetics, General and Molecular Biology, Naples, Italy.
³Department of Mathematics and Information Science, Catania, Italy.
⁴CINECA, Italy.

Abstract

DNA microarray technology is a high throughput method for gaining information on gene function. It allows the simultaneous collection of quantitative data about the differential expression of thousands of genes at a time. This large amount of data is analysed to identify clusters of genes that share common expression characteristics, but the obtained results provide little information regarding the biological similarities of genes within clusters. The published literature, on the other hand, provides a potential source of information to assist in interpretation of clustering results.

We present a tool that enables the validation and improves the comprehension of the experimental results by identifying the main functions among a group of genes, as are reported in MEDLINE documents. It also makes easier the identification of function/disease-specific genes, thus simplifying the design of specially-devoted microarray.

The tool relies on two components: a gene name extractor and a mining algorithm. The name extractor is based on existing dictionaries of gene names and aliases. The mining algorithm analyses the co-occurrences of words in the selected documents in order to automatically interpret the context where the gene names appear and to map documents (and genes) into functional classes.

We test our tool on a set of 400,000 MEDLINE abstracts related to cell cycle. Here we show present results, generated applying the text mining tool to gene-oriented MEDLINE searches, and future perspectives.
1 Introduction

1.1 The problem in biomedical research

In recent years, the "microarray" technology (also known as DNA arrays or DNA chips) for gene expression identification has been developed. A microarray is typically a few-inches glass slide on to which DNA sequences are attached at fixed locations (spots). In one single array, there could be tens of thousand of such spots, each containing a huge number of DNA molecules, whose length ranges between few dozens to many hundreds nucleotides. In principle, each DNA molecule should identify a gene of a genome.

Microarray transcriptional profiling is a powerful tool in the study of transcriptional control mechanisms. The typical application of microarrays consists of comparing gene expression in two samples of cells (e.g. cells of the same kind grown under different conditions, or cells belonging to an healthy tissue and cells belonging to a diseased tissue). An important point in the analysis of microarray data is the identification of hidden correlations between the differentially expressed genes generated upon some kind of cell stimulus. However, in complex organisms (e.g. homo, mouse) microarrays analysis gives less informative results with respect to those obtainable analysing genetically well characterized lower eukaryotes (e.g. yeast). One of the weak point in human and mouse microarray data mining is linked to gene functional annotation because, of the predicted 30-40 K human genes, only 4-5 K have a known function. However, functional data are rapidly accumulating in the scientific literature and biologist needs to retrieve this information which has been collected by MEDLINE (www.ncbi.nlm.nih.gov), a database that contains over 11,000,000 biomedical journal citations.

A microarray analysis usually generates few hundreds of differentially expressed genes and after statistical validation of the data and transcription profiles clustering, biologists enter in the “genes-correlations” nightmare trying to identify genes functionally correlated by scientific literature analysis. This part of the work is typically very time consuming, even if in recent years some tools have been developed to perform automated information extraction on MEDLINE database (MedMiner [1], PubCrawler [2]).

Another task that requires the analysis of vast portions of literature is the design of disease-specific microarrays. In order to identify the genes set to be spotted on the array, a powerful approach could be the identification of gene names in a specific subset (e.g. lung cancer related papers) of MEDLINE database. However, reading every article even in a specific field requires too much time and labour. Therefore it is necessary to have some kind of intelligent information extracting system that recognises the gene names inside the texts.

This paper describes a tool that makes simpler the design of microarray to be used in translational oncogenomics and the extraction of functional knowledge by literature abstracts directly/indirectly related to differentially expressed genes identified by commercial microarrays.
1.2 The approaches to the analysis of unstructured texts

The analysis of textual documents can be approached by two different points of view: text mining and information extraction.

The former aims at the automatic identification of groups of documents that share the same patterns of words, and thus refer to the same topic or theme. Text mining is performed by applying data mining algorithms to texts that have previously been prepared (tagged) by some sort of linguistic analysis [3].

The latter aims at providing a structured representation of the textual information and requires a pre-definition of entities and relationships to be looked for inside texts [4][5]. For example entities might be proteins and drugs and the relationship might be “activate” or “inhibit” or “suppress”, ... In this case, the information extraction approach enables the generation of a database that contains all names of proteins and drugs and all kind of interrelations that have been found inside the examined documents.

Thus while the text mining algorithms are general purpose, the information extraction algorithms are specific to the application. Furthermore, the text mining approach is explorative and enables the discovery of new concepts [6] and relations while the information extraction only extracts those elements that have already been defined.

Anyway, the information extraction is not an easy task and usually requires various steps of syntactic analysis as well as semantic analysis to be performed on texts.

These two approaches can be integrated: information extraction tools generate databases that can be analysed using data mining techniques, and, on the other side, text mining tools might take advantage of specific domain information extracted using I.E. techniques, as will be shown.

2 Method

In our specific application it was necessary to let documents group on the basis of their content, expressed in terms of nouns, verbs and adjectives, as the functional meaning of genes should emerge from that context. For this reason we decided to apply text mining tools. It was necessary indeed to recognise the names of genes inside the texts for both querying purposes and for interpreting the results, i.e. being able to relate each group of documents (or biological function) to the involved genes.

The gene name recognition is a particularly hard task, even for the information extraction tools, as these names don’t follow any predefined rule and have many aliases. We then built a “gene name extractor” based on a dictionary.

In order to let scientists mine only the subset of MEDLINE abstracts that are of their own interest, we built a prototype, MedMOLE, that makes available online a big portion of MEDLINE database already processed by our algorithms and ready for the mining phase.

The following paragraphs describe how we built this prototype, following the phases of a knowledge discovery process (KDD) [7].
2.1 Data selection

We selected from NCBI Pubmed all MEDLINE documents dealing with the cell cycle, by means of the query “cell cycle OR cell proliferation OR cell death OR oncogenes OR tumor suppressor OR apoptosis OR PARP OR caspase OR CDK OR PCNA OR Fas ligand OR cytochrome C”.

To do this, we used PubCrawler, a crawler that has been developed specifically for the extraction of documents from Pubmed and Genbank (http://www.pubcrawler.ie/).

We obtained 400,000 documents (about 4% of the whole database). The choice of the cell cycle topic is due to the fact that the main cellular functions are now quite well known and yet still of great interest among researchers and thus constitute a suitable topic for validating text mining results.

2.2 Data preparation

The data preparation phase has been performed in three steps by applying a filtering process, a linguistic tool and a “gene name extractor”. After extracting all the relevant information from texts, we made a transformation into the appropriate format for the data mining algorithm to process it, thus concluding the preparation phase.

The filtering process aims at cleaning the documents from the meta-information: author names, affiliation, publication type, name of journal, date of publication, country, ... This kind of information is not relevant for our application. After this step, we obtained the textual parts (i.e. title and abstract) for subsequent processing.

The linguistic analysis has been performed by means of a part-of-speech tagger and a stemming tool. Our objective was to eliminate the less significant words. For each document, words that are rich of semantic content (like nouns, verbs and adjectives) and that have been identified by the POS tagger, are then stemmed and stored as list of keywords. These lists characterize and describe the original documents in a synthetic form and are the basis for the mining process.

To identify the names of genes, that are so important in the biological research, another step of analysis is necessary. We call this step “information extraction” as it has the same goal of the I.E. techniques, although we are not, at present, applying any semantic analysis. This step is described in detail in the next sub paragraph and aims at producing, for each document, the list of genes that have been recognised inside the text, whether present in their official name or as alias name.

2.2.1 The gene name extractor

Information extraction tools may be grouped in two main typologies: those relying on hand encoded domain knowledge and those relying on training methods. The former is based on the human coding of semantic, syntactic and morphological rules that apply to the specific application being developed. It requires a big effort in terms of human resources and, even if it already provided
good results in the recognition of protein names [8], it is particularly unsuitable for our application as there are no rules or standards in the assignment of names to genes, nor in the construction of sentences containing them.

On the other hand, tools relying on training methods require a training corpus of documents already annotated by domain experts. The harder the recognition task, the larger is the corpus needed for the algorithms to learn the underlying rules. Techniques that have already succeeded with the protein recognition task have had poor results, in terms of precision and recall, with the gene recognition task [9][10]. In our case, we didn’t have an annotated corpus to try the automatic identification of rules.

Furthermore, beside the term recognition problem, we had the problem of linking together terms that refer to the same gene and it is very rare that this link (synonymy) is made explicit by the text. The most suitable solution was then to rely on a dictionary of gene names and aliases, as much complete as possible and updated.

We created such a dictionary on the basis of the LocusLink database, the most stable and complete source of information on genes that have been discovered up to now (whose sequences are stored in the linked database RefSeq), derived from UniGene database.

Our gene name extractor is a three step process that applies a filter, an indexer and a meta-information generator.

The filter is a perl script that extracts and selects from the LocusLink entries the “Official Symbols” together with the “Alias symbols” (until Official Symbols defined by the HUGO Nomenclature Committee have been recently introduced, many authors had not referred to genes by their official gene symbol). The filter runs monthly and generates the dictionary, which at present contains 14296 H. sapiens LocusLink entries and 97627 official symbols and alias symbols satisfying the following conditions:

- Symbols should be at least 3 letters long
- Alias symbols equal to the official symbol of a different gene were discarded as misleading.

The filter output is a two-column file where the first column, “gene”, contains the official name and the second, “alias”, contains an alias name of the same gene. The gene name on the left column is repeated as many times as there are aliases.

The indexer searches each name on the right hand column (alias) into the MEDLINE abstracts and stores the corresponding left hand column name (the official one) together with the resulting document identifiers.

After cleaning this list from duplicates and sorting by document identifier, we obtain the list of gene names that belong to each document. This list is integrated with the keyword list of the previous step (linguistic analysis).

The major problem we encountered in this step was due to the ambiguity of certain terms that are indicating genes but also have a meaning in common English language (e.g. FAR, GAS, RED, ...). The indexer, in fact, searches for an exact match of the terms inside the documents and would, in these cases, retrieve too many false positives. We restricted, in these cases, the results to
documents that contain a word that matches the gene name, but also contain the term “gene” or “protein” or “product” within a distance of no more than six words.

2.3 Data mining

After the pre-processing phase, each document has been described in a synthetic format by a list of keywords (automatically extracted) and a list of genes. It is now possible to group documents on the basis of these descriptors by means of a clustering algorithm.

The main difference between the available algorithms is on the applied metric, which is affected by the choice of the data representation: binary (presence or absence of a specific keyword in a specific document) or quantitative (frequency of the keyword in the document).

As our documents where actually quite short in length (abstracts and not full text papers), we noticed that the frequency of keywords tended to be 1. We then opted for a binary representation. To measure the distance of documents we used a similarity index (the Dice index) and we applied a partition algorithm based on the relational analysis.

With this method, document similarity is based on the number of co-occurrences of keywords: the more keywords are shared between two documents the more similar they are. The gene names do not take part into the clustering process, but they provide information on the obtained clusters.

The mining process is regulated by a few parameters: maximum number of clusters, similarity threshold, weighting system.

MedMOLE makes this process available on-line, so that the user can select his own set of documents to mine for similarities. The query language for selecting documents (among the 400,000 that have been pre-processed) and the possibility of tuning the parameters make it a very flexible tool.

3 Test and results

Although MedMOLE was designed as a tool to facilitate microarray data mining on MEDLINE it has a broader range of applications, e.g. correlating, by scientific literature analysis, genes characterized by common transcriptional regulative elements or sharing a common functional motif (e.g. SAM, leucine zipper, etc) or having peculiar enzymatic activities (e.g. kinase). Here we summarise some examples of the use of MedMOLE.

One of the simplest way to use MedMOLE is the identification, by literature scanning, of genes somehow related to a specific gene within a specific biological problem. Using the query “BRCA1 and (“breast cancer” or "breast tumour") on a cell cycle-oriented MEDLINE abstracts data set, after optimisation of the number and homogeneity of the clusters, we obtained 2 groups of correlated clusters and 2 isolated clusters (fig. 1). Observing the keywords, used by the clustering algorithm to generate the abstract groups, it is clear that clusters
Figure 1: cluster map of "BRCA1 and ("breast cancer" or "breast tumour")"

2, 5, 6, 7, 9 are linked to each other by keywords related to the association of breast cancer to women, age and genetic risk. Instead clusters 3, 8, 10 are related to each other due to molecular and cellular characteristics of breast cancer. Cluster 4 relates to the chromosomal locus of the disease and cluster 1 is bound gene mutations affecting the pathology. Observing the genes distributions in the various clusters (fig. 2) the association breast cancer and BRCA1 pops-up an other gene involved in breast cancer predisposition: BRCA2. Furthermore, p53, which is frequently mutated in breast cancer, and ERBB2, which can be amplified by p53 mutational pathway during tumour progression in breast
cancer, are strongly present in the grouped clusters. Even this simple example shows two important features of MedMOLE, the possibility to extract genes functionally related to the genes used in the query and the ability of the text mining approach to correlate the selected abstracts on the basis of their informational content.

Another use of MedMOLE is related to the identification of subset of genes which can be used to make specifically-devoted cDNA arrays. The design of microarrays containing all genes associated to specific signal transduction pathways is a way to refocus a microarray data analysis on specific targets, e.g. having found as differentially expressed genes JAG1, a gene associated to the Notch pathways, a transcriptional profiling of all genes of this pathway will offer a better picture of the effects of the cell stimulus under study on that specific pathway. In addition, making arrays to profile genes related to drug resistance in patients undergoing to pharmacological treatment might allow the identification of gene expression patterns that, integrated to clinical data, might allow the design of patient-oriented treatment in the oncological field. Using the query ("lung cancer" or "lung tumour") and "drug resistance", after optimisation of the number and homogeneity of the clusters, we obtained 1 group of clusters and 4 isolated clusters (fig. 3). Interestingly clusters 1, 5, 7, 8, 9, 10, which are grouped together are characterized by keywords related to “cell cycle”, “resistance accumulation”, “inhibition” and “dose effect”. On the other hand the isolated clusters 2, 3 and 6 seems interesting for the identification of genes related to “patient survival”, “mutated genes” and “drug sensitivity and transport”. Evaluating the genes distribution in the clusters we have identified some genes which are part of the multi-drug resistance sub-family (e.g. ABCCl, ABCB1), or are related to DNA repair (e.g. CDKN1A) or cell cycle (MYC, RAS, ), which are other important cellular functions to be monitored to evaluate the effect of DNA binding drugs in cancer. Starting from this analysis we enlarged our MedMOLE analysis searching genes related to the query: (MRP or “multi-drug resistance”) and (“lung cancer” or “lung tumour”), data not shown.

Figure 3: cluster map of ("lung cancer" or "lung tumour") and "drug resistance"
In one of the microarray analysis carried out in our lab we have obtained this group of co-regulated genes: ATM, CIP1, CDKN1, D2S448, DKK1. Applying a query ATM or CIP1 or CDKN1 or D2S448 or DKK1 to MedMOLE, after optimisation of the number and homogeneity of the clusters, we have observed that clustering produces a group of correlated clusters (3, 4, 5, 7, 8, 9) linked to each other by cell cycle control related keywords (fig. 4). A smaller group of clusters (1, 6) are related to carcinoma keyword and the other two isolated clusters seems less informative. Observing the histogram representation a common element in the clusters pops-up: the oncosuppressor p53 (data not shown). Getting more in deep in the analysis and reading some of the abstracts was possible to define that p53 was a common element in the functional pathways of all genes used in the query.

4 Conclusion

With the present implementation of MedMOLE the optimisation of the clustering parameters is essential to obtain a clear picture of the abstracts correlations, especially if a larger number of genes is used in the query. We observed in our tests that results generated by MedMOLE are getting more difficult to be interpreted as the number of genes used in the query increases. To make easier the data evaluation we are developing a client interface which would allow the user to get more in deep into the raw data set generated by MedMOLE analysis. We are implementing a client-oriented visualization tool in which abstract clusters will be represented as ellipsoids containing subcluster of abstracts related to the gene names used in the query. Furthermore these subclusters will contain the genes names extracted from the abstracts and all gene names will be linked to LocusLink database (www.ncbi.nlm.nih.gov). This visualization approach should make easier the identification of correlation
existing between gene names used in the query and those associated to the extracted abstracts.

In conclusion, we think that MedMOLE analysis, combined with transcription profiles clustering and gene ontology categorization, can help researchers to unravel the role of genes found differentially expressed in microarray experiments.

References