MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1314

Discovering Novelty in Sequential Patterns: application for analysis of microarray data on Alzheimer disease

Bringay Sandra[a,b], Roche Mathieu[a], Teisseire Maguelonne[c], Poncelet Pascal[a], Abdel Rassoul Ronza[d], Verdier Jean-Michel[d], Devau Gina[d]

[a]*LIRMM, Laboratory of Informatics, Robotics, and Microelectronics, University of Montpellier 2, Montpellier, France*
[b]*Mathematic and Informatics Department, University of Montpellier 3, Montpellier, France*
[c]*CEMAGREF - Earth Observation and Geoinformation for Environment and Land Management research Unit, Montpellier, France*
[d]*Molecular Mechanisms in Neurodegenerative Dementia, Inserm U710, Montpellier 2 University, EPHE*

Abstract

Analyzing microarrays data is still a great challenge since existing methods produce huge amounts of useless results. We propose a new method called NoDisco for discovering novelties in gene sequences obtained by applying data-mining techniques to microarray data. Method: We identify popular genes, which are often cited in the literature, and innovative genes, which are linked to the popular genes in the sequences but are not mentioned in the literature. We also identify popular and innovative sequences containing these genes. Biologists can thus select interesting sequences from the two sets and obtain the k-best documents. Results: We show the efficiency of this method by applying it on real data used to decipher the mechanisms underlying Alzheimer disease. Conclusion: The first selection of sequences based on popularity and innovation help experts focus on relevant sequences while the top-k documents help them understand the sequences.

Keywords:

Information retrieval, DNA microarrays, Alzheimer disease

Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia. In 2006, more than 26.6 million cases of Alzheimer were declared. Due to the increasing number of cases (expected to be multiplied by 4 in 2050), discovering genes involved in AD is becoming a priority for the biomedical community [1,2].

In recent years, DNA microarrays have been successfully used for numerous applications. They allow researchers to compare gene expression in different tissues, cells or conditions [3,4] and provide some information on the relative levels of expression of thousands of genes among samples (usually less than a hundred). Nevertheless, due to the amount of data available, processing them in a way that makes biomedical sense is still a major issue. Data mining techniques, such as [5,6,7], play a key role in discovering previously unknown knowledge and, in this context, it has been shown that they could be of great help to biologists in identify subsets of microarray data that could be useful for further analysis [8]. However, the amount of results obtained with these techniques is still huge and cannot be easily analysed by the experts.

In [8], we proposed a general process, called *GeneMining* based on the mining of sequential patterns. The process starts with a table produced thanks to static experiments we conducted to check the levels of expression of the genes. Each column corresponds to a microarray and each line to a gene. Each microarray measures the intensity of the gene that corresponds to the numerical value in a given cell. We describe in [9] an efficient algorithm to extract frequent patterns of correlated genes ordered according to their level of expression. We extract only patterns that distinguish classes of individuals (e.g. AD vs. healthy). An example of such a pattern is *<(MRVI1)(PGAP1,GSK3B)>,80% AD,10% H* meaning that *"For 80% of AD individuals and 10% of healthy individuals, the level of expression of gene MRVI1 is lower than those of PGAP1 and GSK3B, whose levels of expression are very close".* Although this method was useful, the way to select relevant patterns was not efficient. Depending on the values of parameters, we obtained from 1,000 to 100,000 patterns that were not easy to interpret.

In addition to the problem of the number of patterns, biologists have to face other difficulties. First, they have to link the spot on the microarray to a gene. As no standard exists for specifying names of genes, this is a difficult task. Second, they have to look for relevant publications concerning the genes that interest them. Although some tools are now available to automatically extract information from microarray data [11,12], there is no user-friendly tool to search the literature for sequential patterns.

In this paper, we focus on sequential patterns and our aim is to discover novelties to help biologists analyze how genes interact. Our contribution is three-fold: (i) We first help biologists select relevant sequences according to a specific topic and then to identify both popular genes (often available in the literature) and innovative genes (associated with popular genes in the patterns), (ii) for each sequence, we propose the top-k

relevant documents in the literature for their interpretation, (iii) we propose a visualization tool to underline the relationship between a pattern and its associated documents.

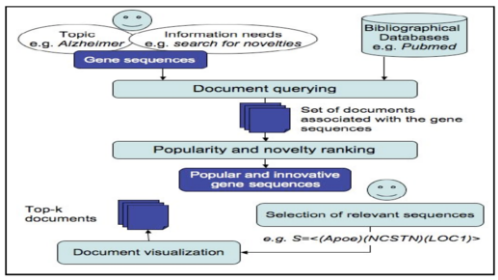

Figure 1- The NoDisco general process

General Process

Figure 1 illustrates the general process of NoDisco, an aid-tool for discovering innovative genes. Entries of the system are gene patterns obtained with an algorithm such as [9], some user information (e.g. study of Alzheimer disease) and a web-service to a bibliographical database such as PubMed. The workflow is organized in the following steps:

1. Document querying: Depending on the topic of interest `T` (e.g., Alzheimer), the tool generates a set of queries (for each sequence `S`) in order to extract documents associated with the topic denoted $Qset_{S,T}$.
2. Popularity and Novelty Ranking: For genes and gene sequences, a popularity rank (taking into account the number of references to the gene in the literature) and a novelty rank (non popular genes linked to popular ones) are computed.
3. Selection of Relevant Sequences: Popular and innovative sequences are proposed to the expert so he can select some of them depending on the information he is looking for.
4. Document Visualization: Top-k documents associated with selected sequences are organized in a sophisticated way for visualisation.

We now describe the process in more detail.

Document querying

For each sequence, a query is submitted to the PubMed Web Service to compute a popularity and novelty score for the sequence. Queries are defined as follows:

Query syntax: A query based on *n* terms with *n-1* operators returns *m* documents:

$$Q(term_1\ op_1\ term_2\ op_2\ \ldots\ op_3\ term_n) \rightarrow \{d1,\ldots,dm\}$$

The operators can be: 'AND, 'NOR' and 'OR'. The number of documents retrieved by a query is denoted $|Q(Terms)|$.

Gene designation: As detailed in [13], recognizing biological objects in natural language is a very difficult task for many reasons: The general lack of annotator agreement and naming conventions, excessive use of abbreviations, frequent use of synonyms and homonyms, biological objects often have names consisting of many individual words, such as '*human T-cell leukaemia lymphotropic virus type 1 protein*', etc. For all these reasons, in the query, it is not possible to directly use the names of a gene embedded in such a sequence. So, with the Entrez gene[1] Web Service, we first look for all aliases of the genes and store them in a contextual Gene Dictionary called $DicG_{all\_id}$ (a dictionary by type of microarray). For example, in $DicG_{all\_id}$ one alias for the gene *Apoe* is *Apolipoprotein E*.

Query about sequences: To build a query associated with a gene (e.g. *Apoe*), we group all aliases found in $DicG_{all\_id}$ with the operator ' *OR*' (e.g. *Q(Apoe OR Apolipoprotein E)*) . To build a query associated with a sequence (e.g. *<(Apoe)(VAMP2)>*) , we compose the previous aliases of the two genes with the operator '*AND*' (e.g. *Q((Apoe OR Apolipoprotein E) AND (VAMP2)))*. In the rest of this article, we use the term *Q(gene1)* for $Q(gene_1\ OR\ al_1\ OR\ al_2\ OR\ \ldots)$ where $al_1, al_2, \ldots$ are aliases of the gene $gene_1$.

Topics of interest: Not all the documents retrieved using the name of a given gene will be relevant for the biologist. Their number can be reduced by using the parameters available in the PubMed search engine such as:

- Standard parameters: Author, date, journal publication, language, accessibility (full or free text, abstract).
- Parameters about the topic: Type of article (clinical trial, editorial, etc.), species (human, animal), sex (male, female), journal topic, etc.

For example, to build a query associated with the sequence *<(Apoe)(VAMP2)>* and the topic *T= ' Alzheimer'* , we compose the preceding query and the topic with the operator *' AND'* : *Q((Apoe OR Apolipoprotein E) AND (VAMP2) AND Alzheimer).* Operators can also be used to specify the terms of the topics (e.g. *T= " human AND female").*

Popularity and Novelty Ranking

We use the number of documents retrieved for each gene to rank them according to their popularity and novelty.

Popularity of a gene: A gene G_i is *popular* if the number of documents dealing with this gene in the literature is greater than the defined threshold *pop_min_gene*. For each gene G_i in $DicG_{all\_id}$, its popularity, $Pop_{Gi,T}$, according to a topic T is obtained as follows:

$$\forall G_i \in DicG_{all\_id},$$

$$Pop_{G_i,T} = \begin{cases} 1 & \text{if } |Q(G_i,T)| > pop\_min\_gene \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

[1] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

For example, if *pop_min_gene=100* and $|Q(ApoE,Alzheimer)| = 3805$ then *ApoE* is popular.

Popularity of a sequence: A sequence S_i is *popular* if the proportion of popular genes in S_i is greater than the defined threshold *pop_min_seq*. For each sequence S_i, we compute its popularity, $Pop_{Si,T}$, according to a topic T. Let PS_i be the set of popular genes in a sequence S_i:

$$\forall S_i \in SP,$$
$$Pop_{S_i,T} = \begin{cases} 1 & \text{if } \frac{|PS_i|}{|S|} > pop\_min\_seq \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For example, *S=<(Apoe)(NCSTN)(LOC153222)>* is a popular sequence for *pop_min_seq=0.5* because *Apoe* and *NCSTN* are popular ($|PS_i|/|S| = 2/3$).

Innovative genes: A gene G_i is in *an innovative relation* with popular genes if the number of sequences associating G_i with popular genes is greater than the defined threshold *new_min_gene*. For each gene G_i of $DicG_{all\_id}$, we compute its novelty, $New_{Gi,T}$, according to a topic T. Let $Pseq_{Gi}$ be the set of popular sequences containing G_i:

$$\forall G_i \in DicG_{all\_id},$$
$$New_{G_i,T} = \begin{cases} 1 & \text{if } |PseqG_i| > new\_min\_gene \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

For example, *PUM1* is an innovative gene because it is present in more than *new_min_gene* popular sequences.

Innovative sequences: A sequence S_i is *innovative* if the proportion of innovative genes in S_i is greater than the defined threshold *new_min_seq*. For each S_i, we compute its innovative score, $New_{Si,T}$, according to a topic T. Let NS_i be the set of innovative genes in a sequence S_i:

$$\forall S_i \in SP,$$
$$New_{S_i,T} = \begin{cases} 1 & \text{if } \frac{|NS_i|}{|S|} > new\_min\_seq \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

For example, if *new_min_seq*=0.5 then s*=<(Apoe) (LOC90624)(PUM1)>* is an innovative sequence as *PUM1* and *LOC90624* are two innovative genes ($|NS_i|/|S| = 2/3$).

At the end of this step, we have reduced the initial set of sequences to two sets, popular and innovative sequences, which can be proposed as relevant sequences for the experts.

Top-k documents: To help the expert analyze a sequence S_i, we look for the Top-*k* documents. To this end, we ask PubMed to retrieve all documents D associated with genes in S_i and we rank them using the two following methods.

First, we compute the score S_{di} for a document d_i published in *the year d* and dealing with the *g* genes. Let *MinA* (resp. *MaxA*) be the year of publication of the oldest document (resp. the year of publication of the most recent document). Let *MinG* (resp. *MaxG*) be the minimum (resp. maximum) number of genes cited in the documents of D. α is a coefficient. We then rank the documents according to the equation 5. $S_{di} \in [0,1]$. $\alpha \in [0,1]$. The value α*=1/2* gives the same weight to both components of the formula.

$$S_{d_i} = (1-\alpha) * \frac{(MaxA-d)}{(MaxA-MinA)} + \alpha * \frac{(MaxG-g)}{(MaxG-minG)} \quad (5)$$

Second, in the documents described by the two criteria (year of publication and number of genes), we look for the Pareto points [14]. These points correspond to documents that are not dominated by others considering both criteria (i.e. they are the best ones considering one criterion alone and the best compromises based on both criteria). We then select k documents in these points. Finally, for the top-k documents, we obtain a rank that can be used by the expert to analyze a sequence[2].

Experiments

Case study

In the framework of the PEPS-ST2I Gene Mining project, we mined real data produced by analysis of DNA microarrays (Affymetrix DNA U133 plus 2.0) [10]. The aim was to decipher brain aging mechanisms. Aging is the primary risk factor in neurodegenerative disorders such as Alzheimer's disease. We analyzed the transcriptome from the temporal cortex of *Microcebus murinus*, a relevant primate model because as they age, some of them present the same lesions observed in human brains affected by Alzheimer's disease. Primates were divided into 3 groups: 6 young adults, 10 healthy aged and 2 aged with Alzheimer's disease lesions. We used DBSAP [10] to discover sequential patterns with several parameters. In the worst case, we obtained approximately 50,000 gene sequences. The longest sequence was composed of eight genes. These sequences can be used to distinguish between AD animals and healthy animals. However, as this number of sequences is too huge, the process of interpretation described in [9] cannot be directly applied on these sets of sequences.

Evaluation of popular and innovative sets

To identify relevant sequences, we analyzed popular and innovative genes and gene sequences. The topic we sued was "Alzheimer" and we varied the four other parameters: *pop_min_gene* (10, 50, 100), *pop_min_seq* (0.25, 0.5, 0.75), *new_min_gene* (5, 10, 30), and *new_min_seq* (0.25, 0.5, 0.75). We obtained quantitative results that varied with the values of

[2] Information on the visualization tool is available at: http://www.lirmm.fr/~bringay/Bringay/MedInfo/MedInfo.html

the parameters[3]. For example, from a set of 50,000 sequences, with the parameters *pop_min_gene=100, pop_min_seq=0.5, new_min_gene=10,* and *new_min_seq=0.5*, we obtained *336* popular and *208* innovative sequences. The important issue is that we defined two sets of sequences in a quantity which allows the use of the process described in [9]. The choice of the parameters depends on the number of sequences we define at the beginning of the process.

Evaluation of the ranking documents

In the first part of these experiments, we showed that we were able to help experts to select relevant patterns. The next step was to evaluate the quality of the NoDisco documents associated with these patterns. To this end, we arbitrarily selected five popular sequences (see figure 2) and studied them in collaboration with experts. We built three sets of ranked documents: (i) We ranked them according to their S_{df} score ($\langle$=1/2), (ii) We chose the first Pareto points, (iii) We extracted the first documents returned from PubMed using the names of the genes.

seq 1: ADAMTS9-APOE-KCNC1-PTPRA-LOC284214
seq 2: GSTO1-VAMP2-SMARCA2-PTPRA-UBE1DC1-CART
seq 3: ADAMTS9-PML-UBN1-FAT-SRRM2
seq 4: ADAMTS9-PML-PRLH-FAT-NBS1-RBX1-LOC284214
seq 5: ADAMTS9-PLXNA2-GSK3B-FAT-FLJ11029-DNAJB6

Figure 2- Five gene sequences

Are the documents returned by the different methods the same? For each pair of methods, we computed the number of shared documents considering the 10, 20 … 100 first documents in each ranking list (see Table 1). Results corresponded to the average number of documents obtained by the five sequences. For instance, we compared the 30 first documents sorted by PubMed and S_{doc}. In this case, we obtained an average of 4% of shared documents with both methods. Table 1 shows that the three approaches returned different documents (i.e. we extracted new knowledge that was not discovered by querying PubMed alone). These experiments were based on a large number of documents (1,083 different documents returned using our approaches).

Finally, the number of documents returned by the different approaches with the five sequences was very different. Table 2 shows that method S_{doc} returned a larger number of documents. This specific retrieval task (i.e., by generating a specific query) may be very useful for experts. This result can be explained by the fact that our method S_{doc} takes into account synonyms to extract relevant documents. The number of documents returned by the Pareto method was low because this method rejects all documents that are not in the Pareto front (i.e. dominated documents [14]).

[3] Due to lack of space, all results are not reported here, but are available at: http://www.lirmm.fr/~bringay/Bringay/MedInfo/MedInfoResults.pdf

Table 1- Search for shared documents with the three methods

No. of doc.	10	20	30	40	50	60	70	80	90	100
%PubMed vs. S_{doc}	0	2	4	4	4.8	4.3	4.2	4.2	3.7	3.4
%PubMed vs. Pareto	2	2	2.6	2	1.6	1.3	1.1	1	0.9	0.8
%Pareto vs. S_{doc}	0	0	2	2	2.8	7	6.5	5.7	5.1	4.6

Table 2- Number of documents retrieved with the 3 methods

Methods	PubMed	S_{doc}	Pareto
Number of doc.	404	537	225

Are the documents returned by our methods relevant? To go deeper into the analysis of the documents, we asked an expert to analyze the abstracts of the first documents retrieved. He manually analyzed the 10 first abstracts retrieved by *seq1* and *seq2* using our three ranking methods (60 documents were manually analyzed). He classified them in five groups: (1) *Relevant*; (2) *Too old* (e.g. documents published before 2000 were not relevant because they were published before the creation of the Affymetrix DNA microarray) (3) *Semantically not relevant* (e.g., documents with the term CART in their summary meaning Classification And Regression Tree instead of the gene CART) (4) *Off the topic* (e.g. documents retrieved in a journal of acupuncture are not relevant for biologists) (5) *Not related to the sequence.* When no term corresponding to one of the genes or to one of the aliases occurred in the abstract, the expert was unable to evaluate the relevance of a document. The classification is summarized in Table 3.

Table 3- Evaluation of the documents by the expert

Evaluation	PubMed	S_{doc}	Pareto
1	13	6	9
2	0	4	0
3	2	6	0
4	1	1	2
5	4	3	9

Queries based only on PubMed returned the best rate of relevant documents but the results of the two other methods (S_{doc} and Pareto) can be easily improved: The noise corresponding to irrelevant documents can be easily reduced by adding domain knowledge to our method.

First, we can consider documents published before 2000 to be less important. For example, we retrieved several documents dealing with PTP (Pancreatic Thread Protein), published before 1999. These documents were not relevant for the biologists concerned, who were looking for information about PTPR4, Tyrosine Phosphatase Receptor, tested with microarrays. These two proteins are linked by the same alias, but we can distinguish between them by the publication date.

Second, we can extend the topic to similar topics. For example, we did not retrieve any documents with the association VAMP2 and AD but had better result with “aging”. When a query does not produce the expected result, the topic can be extended by consulting a list of related topics. The concept of

family can be used in the same way. The genes are organized according to their properties or functionalities. For example, KCNC1 did not produce result with AD, but KCNC (subunit of the potassium channel family) did produce results. Thus, when there is no result, the query can be extended by using the family as the term of the query instead of its alias.

Discussion

Some tools are now able to mine the biological literature. BioMinT [15] is an easy-to-use information retrieval and extraction tool targeted at online biomedical literature. This tool retrieves relevant documents and proposes a range of relevant outputs. However, the tool is not dedicated to the analysis of genes. From a set of genes defined by the user, MedMiner [11] filters and organizes large amounts of textual and structured information retrieved by public search engines (GeneCards and PubMed). GoMiner [12] goes a step further and uses the Gene Ontology (GO) to identify biological processes, functions and components in a list of genes, and generates hypotheses to guide further searches.

Although existing tools are very powerful, they are not dedicated to the analysis of gene sequences produced by analysis of DNA microarrays. With NoDisco, popular sequences can be identified that will be useful to biologists to validate the gene sequences they identify. Indeed, these sequences are composed of genes that have already been linked in the literature and are well known. It is also possible to identify innovative sequences, revealing surprising associations of genes, which can draw the attention of the biologist to unknown gene interactions. For example, the expert who collaborated with us identified an innovative gene ADAMTS9. This gene is not yet known to be involved in Alzheimer disease (i.e. there is no publication dealing with ADAMTS9 and "Alzheimer"). However, in the sequences, this gene is linked to popular genes that are well known for their implication in Alzheimer disease. Moreover, the expert underlined the fact that two other genes in the same family, ADAM9, ADAM10 and ADAM17, have already been linked to Alzheimer disease, so the link between ADAMTS9 and Alzheimer needs to be studied.

Conclusion

The development of DNA microarray technologies and the explosion of online scientific biological literature overwhelm the ability of researchers to take full advantage of available knowledge. In this paper, we presented the NoDisco process which enables biologists to select relevant sequences obtained from DNA microarray analysis. According to a topic, biologists can identify popular and innovative genes along with the sequences in which these genes appear. We also linked gene sequences to the top-k documents in order to facilitate their interpretation. As discussed in the Experiments section, the relevance of the ranking in NoDisco can be easily extended to include domain knowledge. Moreover, NoDisco can be extended to other areas involving medical or pharmacological information. More generally, NoDisco can be used to organize the information retrieved from any arbitrary PubMed search.

References

[1] Eisen M, Spellman P, Brown P, and Botstein D. Cluster analysis and display of genome-wide expression patterns. National Academy of Science 1998: 85(25): 14863-14868

[1] Madeira S, and Oliveira A. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 2004: 1(1), 24-45.

[2] Blalock EM, et al. Harnessing the power of gene microarrays for the study of brain aging and alzheimer's disease: statistical reliability and functional correlation. Ageing Res. Rev., 2005: 4(4):482-512.

[3] Hoerndli F, David DC, and Götz J. Functional genomics meets neurodegenerative disorders. part ii: Application and data integration. Progress Neurobiol., 2005:76:169–188.

[4] Cong GA, Tung X, Pan F, Yang J. Farmer : Finding interesting rule groups in microarray datasets. in ACM (ed.), SIGMOD Conference, 2004:143-154.

[5] Khan J, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, 2001:673-679.

[6] Pensa RG, Besson J, Boulicaut J.-F. A methodology for biologically relevant pattern discovery from gene expression data, in E. Suzuki, S. A. (Eds.), Discovery Science, 2004: 3245:230-241.

[7] Korotkiy M, Middelburg R, Dekker H, Van Harmelen F, Lankelma J. A tool for gene expression based PubMed search through combining data sources, Bioinformatics, 2004:20(12):1980-1982

[8] Salle P et al. GeneMining: Identification, Visualization, and Interpretation of Brain Ageing Signatures. Stud Health Technol Inform. 2009; 150:767-71.

[9] Salle P, Bringay S, Teisseire M. Mining Discriminant Sequential Patterns for Aging Brain. In Proceedings of the 12th Artif Intell in Medicine proceedings 2009: 365-369.

[10] Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, andWeinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 1999:27, 1210-4, 1216-7.

[11] Zeeberg BR et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4, 2003, R28.

[12] Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform. 2005;6(4):357-69.

[13] Collet P, Rennard JP. Introduction to Stochastic Optimization Algorithms, in Handbook of Research on Nature-Inspired Computing for Economics and Management, JP. Rennard, IDEA Group Inc, 2006.

[14] Pillet V, Zehnder M, Seewald AK, Veuthey AL, Petrak J. GPSDB: a new database for synonyms expansion of gene and protein names Bioinformatics.2005; 21: 1743-1744.

Address for correspondence

S. Bringay, LIRMM, 161 rue Ada, 34392 Montpellier France