Discovering novelty in sequential patterns: application for analysis of microarray data on Alzheimer disease
Sandra Bringay, Mathieu Roche, Maguelonne Teisseire, Pascal Poncelet, Ronza Abdel Rassoul, Jean-Michel Verdier, Gina Devau

To cite this version:
Sandra Bringay, Mathieu Roche, Maguelonne Teisseire, Pascal Poncelet, Ronza Abdel Rassoul, et al.. Discovering novelty in sequential patterns: application for analysis of microarray data on Alzheimer disease. MedInfo: Congress on Medical Informatics, Sep 2010, Cape Town, South Africa. pp.1314-1318. hal-00558156

HAL Id: hal-00558156
https://hal.archives-ouvertes.fr/hal-00558156
Submitted on 21 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Discovering Novelty in Sequential Patterns: application for analysis of microarray data on Alzheimer disease

Bringay Sandra, Roche Mathieu, Teisseire Maguelonne, Poncelet Pascal, Abdel Rassoul Ronza, Verdier Jean-Michel, Devau Gina

*LIRMM, Laboratory of Informatics, Robotics, and Microelectronics, University of Montpellier 2, Montpellier, France
Mathematic and Informatics Department, University of Montpellier 3, Montpellier, France
CEMAGREF - Earth Observation and Geoinformation for Environment and Land Management research Unit, Montpellier, France
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 identifying 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 	riangledown<(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.
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 (e.g., Alzheimer), the tool generates a set of queries (for each sequence S) in order to extract documents associated with the topic denoted \( Q_{set_S,T} \).

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 visualization.

We now describe the process in more detail.

---

1. PubMed is an electronic library of publications in the field of biomedical research that is available to the public:
   

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 $\text{pop.min.gene}$. For each gene $G_i$ in $\text{DicG}_{all\_id}$, its popularity, $Pop_{G_i,T}$ according to a topic $T$ is obtained as follows:

$$\forall G_i \in \text{DicG}_{all\_id},
Pop_{G_i,T} = \begin{cases} 1 & \text{if } |Q(G_i,T)| > \text{pop.min.gene} \\ 0 & \text{otherwise} \end{cases}$$

For example, if $\text{pop.min.gene} = 100$ and $|Q(\text{ApoE, Alzheimer})| = 3805$ 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 $\text{pop.min.seq}$. For each sequence $S_i$, we compute its popularity, $Pop_{S_i,T}$ according to a topic $T$. Let $PS_i$ be the set of popular genes in a sequence $S_i$:

$$\forall S_i \in SP,
Pop_{S_i,T} = \begin{cases} 1 & \text{if } |PS_i| > \text{pop.min.seq} \\ 0 & \text{otherwise} \end{cases}$$

For example, $S = \langle (\text{ApoE}) (\text{NCSTN}) (\text{LOC153222}) \rangle$ is a popular sequence for $\text{pop.min.seq} = 0.5$ 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 $\text{new.min.gene}$. For each gene $G_i$ of $\text{DicG}_{all\_id}$, we compute its novelty, $\text{New}_{G_i}$, according to a topic $T$. Let $PS_{G_i}$ be the set of popular sequences containing $G_i$:

$$\forall G_i \in \text{DicG}_{all\_id},
\text{New}_{G_i} = \begin{cases} 1 & \text{if } |P_{seq} G_i > \text{new.min.gene} \\ 0 & \text{otherwise} \end{cases}$$

For example, $\text{PUM1}$ is an innovative gene because it is present in more than $\text{new.min.gene}$ 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 $\text{new.min.seq}$. For each $S_i$, we compute its innovative score, $\text{New}_{S_i,T}$ according to a topic $T$. Let $NS_i$ be the set of innovative genes in a sequence $S_i$:

$$\forall S_i \in SP,
\text{New}_{S_i,T} = \begin{cases} 1 & \text{if } |NS_i| > \text{new.min.seq} \\ 0 & \text{otherwise} \end{cases}$$

For example, if $\text{new.min.seq} = 0.5$ then $S = \langle (\text{ApoE}) (\text{LOC9-0624}) (\text{PUM1}) \rangle$ is an innovative sequence as $\text{PUM1}$ and $\text{LOC90624}$ 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 $S_{d_i}$ for a document $d_i$ published in the year $d$ and dealing with the $g$ genes. Let $\text{MinA}$ (resp. $\text{MaxA}$) 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 $D$. Let $\alpha$ be a coefficient. We then rank the documents according to the equation 5. $S_{d_i} \in [0,1]$ $\alpha \in [0,1]$. The value $\alpha = 1/2$ gives the same weight to both components of the formula.

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

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.5

**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 marinus*, 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.

---

5 Information on the visualization tool is available at: [http://www.lirmm.fr/~bringay/Bringay/MedInfo/MedInfo.html](http://www.lirmm.fr/~bringay/Bringay/MedInfo/MedInfo.html)
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 the parameters\(^4\). 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\(_p\) score ((=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-AP0E-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

<table>
<thead>
<tr>
<th>No. of doc.</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>%PubMed vs. S(_p)</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.2</td>
<td>3.7</td>
<td>3.4</td>
</tr>
<tr>
<td>%PubMed vs. Pareto</td>
<td>2</td>
<td>2</td>
<td>2.6</td>
<td>2</td>
<td>1.6</td>
<td>1.3</td>
<td>1.1</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>%Pareto vs. S(_p)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2.8</td>
<td>7</td>
<td>6.5</td>
<td>5.7</td>
<td>5.1</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table 1: Search for shared documents with the three methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>PubMed</th>
<th>S(_p)</th>
<th>Pareto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of doc.</td>
<td>404</td>
<td>537</td>
<td>225</td>
</tr>
</tbody>
</table>

Table 2: Number of documents retrieved with the 3 methods

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>PubMed</th>
<th>S(_p)</th>
<th>Pareto</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of the documents by the expert

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\(_p\). 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\(_p\) 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\(_p\) 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]).

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.

Queries based only on PubMed returned the best rate of relevant documents but the results of the two other methods (S\(_p\) 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 PTTP4, Tyrosine Phosphatase Receptor, which is tested with Affymetrix microarrays. These two different 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

---

\(^4\) Due to lack of space, all results are not reported here, but are available at: http://www.lirmm.fr/~bringay/Bringay/MedInfo/MedInfoResults.pdf
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. BioMinIT [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

Address for correspondence
S. Bringay, LIRMM, 161 rue Ada, 34392 Montpellier France