*Medical Informatics in a United and Healthy Europe*
*K.-P. Adlassnig et al. (Eds.)*
*IOS Press, 2009*


*doi:10.3233/978-1-60750-044-5-767*

# GeneMining: Identification, Visualization, and Interpretation of Brain Ageing Signatures

Paola SALLE [a,1], Sandra BRINGAY [a,b], Maguelonne TEISSEIRE [a], Feirouz CHAKKOUR [a], Mathieu ROCHE [a], Ronza Abdel RASSOUL [c], Jean-Michel VERDIER [c], Gina DEVAU [c]

[a] *Montpellier Laboratory of Informatics, Robotics, and Microelectronics, Montpellier 2 University, National Center for Scientific Research, France*
[b] *Mathematic and Informatics Department, Montpellier 3 University, France*
[c] *Molecular Mechanisms in Neurodegenerative Disorders, Inserm U710, Montpellier 2 University, EPHE, France*

*Abstract.* Transcriptomic technologies are promising tools for identifying new genes involved in cerebral ageing or in neurodegenerative diseases such as Alzheimer's disease. These technologies produce massive biological data, which so far are extremely difficult to exploit. In this context, we propose GeneMining, a multidisciplinary methodology, which aims at developing new strategies to analyse such data, and to design interactive tools to help biologists to identify, visualize and interpret brain ageing signatures. In order to address the specific problem of brain ageing signatures discovery, we combine and apply existing tools with emphasis to a new efficient data mining method based on sequential patterns.

**Keywords.** bioinformatics, transcriptomic data, sequential pattern mining, data mining

## 1. Introduction

The DNA microarray technologies [1] allow to compare the expression of thousands of genes in different tissues, cells or physiological conditions. It can be used for diagnosis, therapy, follow-up of a treatment or even for characterizing physiological states. Indeed, the major interest of these technologies is to identify, among multiple candidate genes, which ones are the most likely to be involved in a considered trait. Likewise, online biological knowledge databases (KEGG, GO, RIKEN), biological repositories for gene expression array-based data (GEO) and bibliographical database (PubMed [2]) have recently been developed. However, the size and heterogeneity of such data sets remain problematic [1]. Therefore, many works have developed analysis software for huge amount of data [2–5]. Nevertheless, processing those data remains very challenging in terms of biological significance. Translating genes of potential interest to medicine

[1] Corresponding Author: Paola Salle, LIRMM, 161 rue Ada, 34392 Montpellier, France; E-mail: paola.salle@lirmm.fr.
[2] www.genome.jp/kegg; www.geneontology.org/; www.ebi.ac.uk/; www.riken.jp/engn/index.html; www.ncbi.nlm.nih.gov/geo/; www.ncbi.nlm.nih.gov/pubmed/.

discoveries is still an open issue [6]. We are convinced that the management of such volumes requires new methodologies.

Since 2008, in the framework of the GeneMining project, which gathers researchers from the LIRMM Laboratory (Computer Science) and the MMDN laboratory (Biology), we have developed a new method for extracting knowledge from massive data associated to microarray transcriptomic studies. Microarray datasets are very dense because they contain measurements for a large number of genes (e.g., 54,675 probesets for Affymetrix U-133 plus 2.0 Array), for each subject studied. Therefore, traditional methods become irrelevant for this type of data. Our methodology is based on data mining, visualization and interpretation techniques. Our aim is not only to offer efficient algorithms to discover characteristic signatures, but also to provide a process which enables experts to interpret them to produce relevant knowledge, (i.e., that shows biomedical significance). Our approaches have been applied to decipher mechanisms of brain ageing and associated pathologies (such as Alzheimer's diseases).

In this paper, after briefly presenting the complete methodology in the material and methods (Section 2), we present the original process developed to mine the transcriptomic data and two techniques for visualizating and interpretating. We experiment this methodology for the brain ageing study (Section 3). Section 4 discusses this methodology and in particular its generalization.

## 2. Material and Methods: A New Methodology to Analyse Transcriptomic Data

### *2.1. General Process*

In order to extract useful knowledge from massive biological data, we propose a three-step process, detailed in the next subsections: ***(1) Data mining:*** Although knowledge extraction methods have been successfully applied in different areas (marketing, web...), these methods cannot be extended as such to transcriptomic studies due to the huge volumes and density of digital data. Therefore, we propose an efficient data mining method based on sequential patterns. ***(2) Clustering and Visualization:*** Given the amount of results returned by data mining methods, we also add an interface that eases the discovery process by allowing experts to identify smaller sets of meaningful patterns from more general sets of patterns. ***(3) Interpretation:*** We integrate in our tool existing knowledge bases (GO, KEGG) as well as bibliographic databases (PubMed), to assist the biologists in interpreting the selected patterns.

### *2.2. Sequential Pattern Mining*

In the literature, three ways for analysing transcriptomic data are proposed: ***(1) Case-control methods:*** Such studies compare two groups: diseased patients (cases) and healthy controls. The aim is to identify which factors could be associated to the disease, or, more specifically, for transcriptomic studies, the identification of candidate genes with respect to a specific subset of state. SAM [5] is an appropriate method used to identify candidates. However, this method does not allow the interpretation to change according to gene relationships. ***(2) Clustering methods:*** A common method of clustering has been proposed by [2] in order to classify genes into groups. It is based on the following assumption: genes with similar expression profiles are part of the same

biological function or are regulated by common factors. Recently, co-clustering approaches have been proposed to compute groups of objects associated to a set of attributes [3, 4]. However, clustering methods do not take into account genes that act in regulatory networks: genes may coordinate the same actions while having very different expression profiles, and consequently, not be clustered together. ***(3) Pattern mining methods:*** [7, 8] were among the first approaches used for biological databases based on association rule algorithms. For instance, [9] extracts association rules of the type: G1,G2 => cancer. These approaches allow for the detection of groups of genes whose expressions are specific to a phenotype. However, these approaches do not take the gene values into account but only Boolean values (over/under-expressed).

In this proposal, we apply a ***sequential pattern mining*** method. For instance, the pattern $S_{75\%}$=<($G_1$)($G_2$ $G_3$)> means '*for 75% of the studied DNA microarrays, the gene $G_1$ is less expressed than the genes $G_2$ and $G_3$ which have a similar expression*'. Such a pattern introduces the order relationship between the gene expressions. As traditional sequential pattern mining algorithms are not efficient for biological data (size and complexity), we develop a specific algorithm DSPAB [10]. In order to reduce the search space and to get more relevant results (from a biological point of view) we integrate the domain knowledge during the mining step. We reduce the set of patterns to discriminant patterns, frequent in a class and unfrequent in the complementary class (e.g., healthy controls vs. patients). Extracted patterns can thus be used as a signature.

### *2.3. Visualization*

The structure used to visualize networks or graphs is node-link diagrams. [11] shows that the graph readability depends on its size (number of nodes) and its density. Visualising large graphs is an important research problem with many applications in different domains (such as social networks [12]). In our context, we must visualise a large number of data like these traditional applications. However our user-friendly tool takes the specificity of the data (sequential patterns in biomedical domain) into account. We propose a clustering method to group similar patterns, based on S2MP [13]. We design a visualization tool to support the navigation (see Figure 1). This interface supports the discovery process and helps users to focus on smaller sets of meaningful patterns. A presentation of the tool is available on our Web site[3].

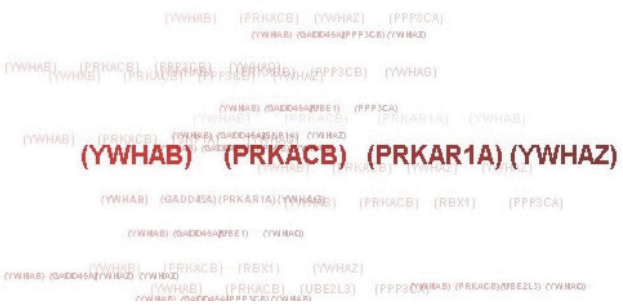


**Figure 1.** ***Clouds visualization*** (inspired from tag clouds which are a visual representation of the more frequent keywords used in a Web site). A cloud is a group of patterns. The centre is at the forefront. The higher the similarity of a pattern to the centre, the nearest in the cloud.

[3] http://www.lirmm.fr/tatoo/spip.php?page=prototypes.

### *2.4. Interpretation*

Our objective is to ease interpretation of experts by allowing them to associate patterns to domain resources. A first step consists in integrating available online knowledge bases as proposed by [14]. When a user selects a pattern, we display information about associated genes in the GO and KEGG systems. For example, to help the expert identifying relations between genes and diseases, we query the KEGG's Web services in order to find all diseases, which are associated with the genes in a pathway. Another step consists in finding the right documents at the right time as suggested by [15], i.e., the best publications in the bibliography databases. For example, we identify in Pubmed the ten most relevant publications to analyze a pattern according to various criteria (type of article, genes involved in the signature, etc.).

## 3. Experiments

Our complete methodology has been applied to decipher mechanisms of brain ageing and associated pathologies (Alzheimer's and Parkinson's diseases).

**Case study:** Ageing is the primary risk factor in neurodegenerative disorders. We have analyzed the transcriptome of the temporal cortex of Microcebus murinus. It is a relevant primate model of Alzheimer's disease studies because as it ages, it shows similar lesions (amyloid plaques) observed in the human brain affected by Alzheimer's disease. We have used human Affymetrix microarrays HG 133 Plus 2.00. Primates have been divided in 3 age groups: 5 young adults, 7 healthy and aged and 2 sick and aged.

**Sequential pattern mining:** We have extracted discriminant sequential patterns (between 100 and 185,240) for various supports in DSPAB (minimal number of individuals for which a pattern is present), i.e., frequent for a biological class (young adults) and not frequent for the complementary class (aged animals).

**Visualization:** The biological experts involved in our project have used our interface (GUI) to analyse the results of the data mining phase. For example, they have observed the sequence $S_{75}$=<(MRVI1)(PGAP1)(PLA2R1)(A2M)(GSK3B)>, which means that for 75% of the DNA microarrays, gene MRVI1 is less expressed than gene PGAP1, etc. Interestingly, those proteins might be involved in signalling or metabolism, and some of them interfere with Alzheimer's disease cellular events.

**Interpretation:** After the gene identification phase, biologists have investigated complementary information on PubMed. For each pattern (composed of n genes), we have looked for texts in PubMed associated with 1, 2 or n genes of the pattern. This process was reiterated with synonyms of these genes found in GO. This provided two types of analysis to the experts: *validation* (identification of patterns which contain genes related in the texts) and *research of innovations* (identification of patterns which contain genes that are not linked in the text or in recent texts). In our first experiments based on two genes (operator AND) with its synonyms (operator OR), 73% of PubMed queries return less than 15 documents. Then experts can manually analyse these publications.

## 4. Discussion

By obtaining knowledge from transcriptomic data that showed biological significance we pave the way for promising research both in terms of computer science and biology. We have extracted ***sequential patterns***, i.e., correlations between genes, which can be used as a signature of a specific trait. As these patterns are new material for biologists, ***visualization*** and ***interpretation*** tools are necessary. To overcome the huge number of extracted patterns, we have applied a clustering algorithm to group them and proposed a method of visualization based on tag clouds. Information from GO and KEGG have also provided in order to help interpretation of results. As most of the knowledge is available in the literature, we have also proposed a simple process to retrieve relevant texts from PubMed. We will improve this process with literature-based discovery [5] which is a set of methods for automatically generating hypotheses for scientific research by finding overlooked implicit connections in the literature. We have applied this methodology to help deciphering mechanisms of brain ageing and associated pathologies and some relevant patterns have been discovered. If each step of this methodology can be improved, we will generalize this process to other types of data mining techniques to offer a relevant framework for transcriptomic analysis. Moreover, we will consider other types of massive data such as genomic data. Finally, the interest of sequential patterns for prediction tasks will be demonstrated in a future work.

## References

[1] Hoerndli, F. et al. (2005) Functional genomics meets neurodegenerative disorders. Part II: Application and data integration. *Progress in Neurobiology* 76:169–188.
[2] Eisen, M., Spellman, P., Brown, P., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA* 85(25):14863–14868.
[3] Madeira, S., Oliveira, A. (2004) Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(1):24–45.
[4] Pensa, R., Boulicaut, J.F. (2008) Constrained Co-clustering of Gene Expression Data. In *Proceedings of the 2008 SIAM International Conference on Data Mining*.
[5] Tusher, G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing data analysis radiation response. In *Proceedings of the National Academy of Sciences of the United States of America* 98:5116–5121.
[6] Butte, A., Chen, R. (2006) Finding Disease-Related Genomic Experiments Within an International Repository: First Steps in Translational Bioinformatics. *AMIA Annual Symposium Proceedings 2006*, 106–110.
[7] Pan, F., Cong, G., Tung, A., Yang, J., Zaki, M. (2003) Carpenter: Finding closed patterns in long biological datasets. In *Proceedings of KDD'03*, 637–642.
[8] Rioult, F. (2004) Mining strong emerging patterns in wide SAGE data. In *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, Pisa, Italy, 127–138.
[9] Xu, X., Cong, G., Ooi, B., Ta, K., Tung, A. (2004) Semantic mining and analysis of gene expression data. In *Proceedings 2004 VLDB Conference* 1261–1264.
[10] Salle, P., Bringay, S., Teisseire., M. (2009) Mining Discriminant Sequential Patterns for Aging Brain. In Proceedings of the Conference on Artificial Intelligence in Medicine, July 2009 (To appear).
[11] Lee, B., Plaisant, C., Parr, C.S., Fekete, J.-D., Henry, N. (2006) Task taxonomy for graph visualization. In *Proceedings of BELIV'06*, 82–86.
[12] Scott, J.P. (2000) *Social Network Analysis: A Handbook*. Sage Publications Ltd.
[13] Saneifar, H., Bringay, S., Laurent, A., Teisseire, M. (2008) S2MP: Similarity Measure for Sequential Patterns. In *Proceeding of AusDM'2008*, 95–104.
[14] Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., Tarczy-Hornoch, P. (2007) Data integration and genomic medicine. *Journal of Biomedical Informatics* 40(1):5–16.
[15] Demner-Fushman, D., Hauser, S., Humphrey, S. et al. (2006) MEDLINE as a Source of Just-in-Time Answers to Clinical Questions. *AMIA Annual Symposium Proceedings 2006*, 190–194.