Integrating 'omics' data sets and biological knowledge: Multiple Factor Analysis as a powerful strategy

Marie de Tayrac[1], Sébastien Lê[2], Marc Aubry[1], François Husson[2], and Jean Mosser[1]

[1] CNRS UMR 6061 Génétique et Développement
Université de Rennes 1, Groupe Oncogénomique
2 avenue du Pr. Léon Bernard - CS 34317
35043 RENNES Cedex, France
(e-mail: marie.de-tayrac@univ-rennes1.fr)

[2] CNRS UMR 6625 Mathématiques appliquées
5 rue de Saint-Brieuc - CS 84215
35042 RENNES Cedex, France
(e-mail: Sebastien.Le@agrocampus-rennes.fr)

Abstract. The huge amount of data provided by genome-scale technologies makes discernible biological meanings difficult to access. Here, we report a powerful integrative method to combine genome-wide scale data sets and biological knowledge. Multiple Factor Analysis (MFA) is used to investigate jointly large observation data sets from different 'omic' areas enriched with biological annotations. This multifactorial method is suitable for a wide range of biological investigations and offers a comprehensive view of the datasets structures and associated knowledge.
Keywords: Multiple Factor Analysis, Genomic, Transcriptomic, Functionnal Annotation, Integrative Analysis, Gliomas.

1 Introduction

High throughput technologies provide an unprecedented amount of data leading to new interpretation challenges in biology. Indeed, scientists are facing a lack of strategies to identify the genes and the gene products involved in different biological processes of interest. This becomes particularly true in cancer studies where genes causative roles are sustained by high level of complexity. During the last few years, many efforts have been made to tackle this problem [Joyce and Palsson, 2006]. Notwithstanding, it remains difficult to obtain from such data a concise visualization of the biological mechanisms involved in the situation under study. Moreover, providing an easy access to the worldwide scientific knowledge, is another challenge.

Here, we are interested in the global understanding of cancer studies using jointly two levels of investigation: the genome structure and its expression (transcriptome). Chromosomal locus copy number alterations are detected by the use of array-based Comparative Genomic Hybridization (array-CGH).

Microarrays allow the monitoring of the expression of potentially all genes within a cell or tissue sample. These two technics provide continuous measurements on a genome-scale level including transcripts abundance and relative genomic copy number variations. To integrate the biological knowledge into a computational method it has to be formalized. The *de facto* standard used in molecular biology is the Gene Ontology (GO)[1]. GO adresses the need for consistent descriptions of gene products and the representation of the functional informations related to them. It is a controlled vocabulary of about 20,000 terms organized in three independent hierarchies for cellular components, molecular functions, and biological processes. Most of the model organisms and annotation databases use GO terms to describe the gene products or have been mapped to the Gene Ontology [Consortium, 2001]. These informations are publicly available and it is easy to retrieve the GO terms associated with a gene or a set of genes [Khatri and Draghici, 2005].

Gathering all these heterogeneous informations potentialize a reliable biological study but necessitates accurate management and exploration methods. For that reason we propose to handle the integration of 'omics' data and biological knowledge by the use of Multiple Factor Analysis (MFA) [Escofier and Pagès, 1988]. We illustrate our method with the analysis of glioma genomic and transcriptomic signatures and show that it allows a good separation of the different glioma subtypes. It also identifies robust regulatory mechanisms implicated in glioblastomagenesis.

2 Methods

2.1 Recovering Biological Data

We used two publicly available data sets from the Gene Expression Omnibus (GEO) database[2] (GSE1991 and GSE2223). These experiments on glial tumors are taken from two studies of Bredel *et al.* in 2005. They allow the comparison of alterations measured at the genome level [Bredel *et al.*, 2005b] and at the transcriptome level [Bredel *et al.*, 2005a]. Among studied gliomas, we selected those from the four main types defined by the standard World Health Organization (WHO) classification (O: oligodendrogliomas, A: astrocytomas, OA: mixed oligo-astrocytomas and GBM: glioblastomas). We retrieved the corresponding measurements for subsets of genes highlighted in the Bredel *et al.* studies. Ratios of the two channels intensities were log_2 tranformed and mean centered per array. ANOVA ($p \leq 0.2$) was used to keep only potential meaningful genes. We then build two matrices of continuous variables. X_1 contains the data for the expression study. X_2 contains those from the genome investigation. Each matrix contains a description of 43 tumor samples (5A, 8O, 6OA, 24GBM).

[1] http://www.geneontology.org/
[2] http://www.ncbi.nlm.nih.gov/geo/

For these two data sets, the conversion from the array probes ID to suitable identifiers[3] was made using the array probe description provided by GEO and manual search. The functional annotations of the corresponding genes were extracted from the Gene Ontology Annotation (GOA) database[4]. We have restricted these annotations to the GO biological process (BP) terms only. We used the *true path rule*[5] to associate each gene with all the GO terms subsumed by its annotated terms. Data are then divided into four parts: samples description (WHO), normalized signal intensities from expression microarrays (expr), normalized signal intensities from CGH arrays (CGH), and gene annotations.

2.2 Multiple Factor Analysis

Multiple Factor Analysis (MFA) [Escofier and Pagès, 1988] is dedicated to the exploration of multiple tables simultaneously. The same set of individuals is described by several groups of variables. It is then possible to analyze different points of view taking them equally into account.

Combining expression microarrays with CGH arrays. The first issue of this study was to combine two different points of view, one provided by expression microarrays and the other provided by CGH arrays, but also to balance their part in the construction of a compromise. We consider the merged data set: $X = [X_1, X_2, ..., X_J]$ where each X_j corresponds to an 'omic' data table. First of all, separate analysis are performed by principal components analysis (PCA) on each group j of variables. Second, a global analysis is carried out: each variable belonging to a group j is weighted by $1/\lambda_1^j$, where λ_1^j denotes the first eigenvalue of the matrix of variance-covariance associated with each data table X_j. The rationale of the scaling is that information that is common to the data tables emerges. Besides no data table can by itself generate the first dimension of the global analysis: the first dimension's variance of each data table is then equal to one. In this way, MFA provides a balanced representation of each individual according to the joint data table X, but also a partial representation of each individual according to each of the group j of variables. Partial representations of a same sample are all the more close that they do express the same information.

Finally, MFA provides a representation of each matrix of variables that allows the visualization of specific and common structures. Consequently, it is possible to get an overall picture of the common structure emerging from the X_j.

[3] Extracted from the HUGO Gene Nomenclature Comittee (HGNC) database: http://www.gene.ucl.ac.uk/nomenclature/

[4] http://www.ebi.ac.uk/GOA/

[5] http://www.geneontology.org/GO.usage.shtml#truePathRule

Adding biological processes annotations. The second issue of this study was to superimpose biological knowledge on the structure extracted from 'omic' investigations. We thus make full use of one of the main features of MFA which is its ability to take in charge supplementary data tables. We then made groups of genes in terms of BP annotations. This was done separately for genes implicated on genome alteration and for genes with expression changes. A group of genes involved in the same biological pathway can be viewed as a module by reference to the so called 'modular biology' [Ge *et al.*, 2003]. A module is described by the genomic or the transcriptomic state of all its constituting genes. The module matrices (X_{BPi}) were built by grouping the data of all the genes annotated by a same BP term. Most of these modules are extended entities that are hierarchically imbricated due to ontological structure. A merged data table of all the modules is constructed: X_{BP}=[X_{BP1}, X_{BP2},...,X_{BPI}].

The X_{BPi} data tables are projected in the common space built out of the X_j matrices. In that space, two modules are all the most close that they have common structure.

3 Results

3.1 First step: joining 'omic' different points of view

We focus on the first two main components created by MFA and visualize the corresponding individual factor map, see Figure 1. This factorial map shows a relatively well-defined partition of tumors by WHO classification (Figure 1a). This is particularly true along the first component which underlines a good separation from glioblastomas (GBM) to lower grade gliomas (O, A, OA). Partial representations (Figure 1b) show that this structure is jointly sustained by genome and transcriptome variations. Indeed, the projections on axe 1 of the two partial mean individuals (CGH and expr) are very close among each subtype. On axe 2, all mean individuals for expression point of view (expr) are located around the origin. It is not the case for genomic partial representations (CGH). The second dimension is therefore specific to the genomic point of view and is not shared by the expressional one. Axe 2 also provides a partition of the histological subtypes and particularly stresses differences between astrocytomas (A) and oligodendrogliomas (O). The manual examination of the genes sustaining this component underlines genomic status modifications of genes located on 1p and 19q positions. These allelic alterations of chromosomes 1 and 19 are frequently reported as important events in gliomas [Smith *et al.*, 1999] and especially in oligodendrogliomas [Reifenberger *et al.*, 1994]. Indeed, it is reported that these chromosomal aberrations patterns vary according to the categories of glial neoplasms and could be marks of malignant progression [Bigner *et al.*, 1999].

However, the interpretation of such emphasized structures remains somehow difficult when only associated gene IDs are accessible. For that reason,

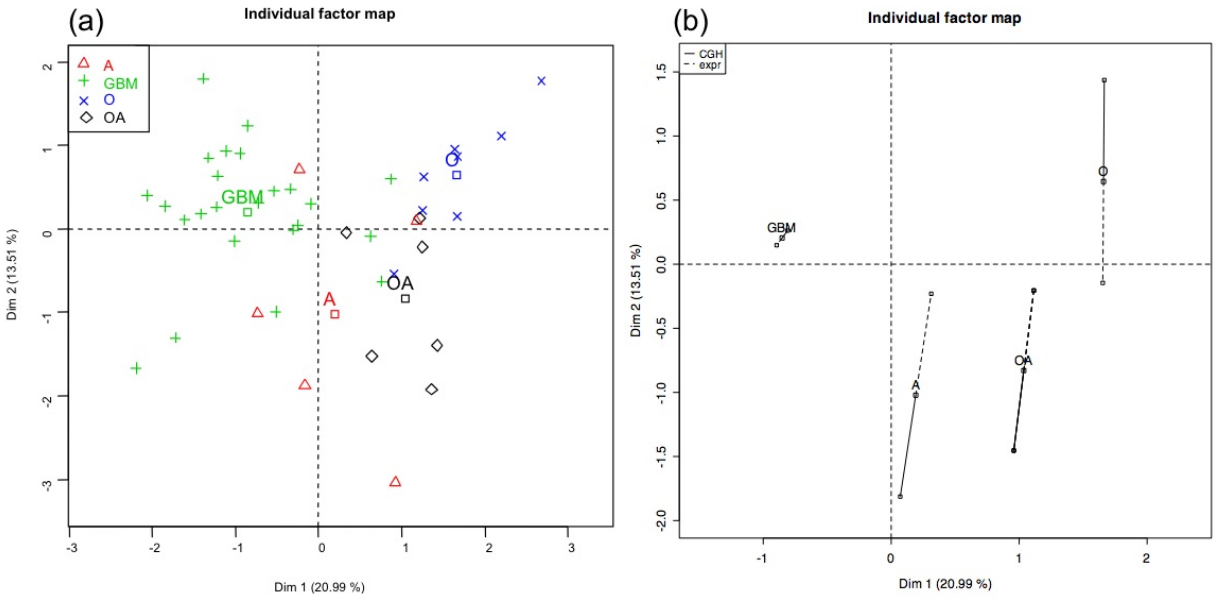

Fig. 1. Individual Factor Map. The scatter plot created with the first two main dimensions of MFA is provided. (a) Individual representation: for each glioma subtype, all tumor samples and mean individual are displayed ; (b) Partial representation : for each subtype, plots of the partial mean individuals (CGH and expr) are linked to the common mean individual extracted from MFA.

providing gene annotations in a corresponding plot is necessary to settle a concise way of understanding these results.

3.2 Second step: Integrating biological knowledge

The second part of our approach is dedicated to the interpretation of the structures pointed out with the joint analyses of genomic and transcriptomic data sets. The formalized biological knowledge becomes accessible with the projection of gene modules on the factor map created by MFA (Figure 2). This graph provides a typology of the modules and highlights shared dimensions between BP terms and tumor groups. The coordinates of the annotation groups provides a direct measure of the links between modules and the corresponding factors (i.e., glioblastomagenesis for axe 1 and precursor cell types (O and A) for axe 2). It is thus possible to found modules highly linked to glioblastoma phenotype and others linked to malignant progression.

Based on the GO terms associated with glioblastomagenesis (axe 1), three main pathways are highly represented. The first one supports the invasive behavior of GBM cells with GO terms like 'localization of cell'

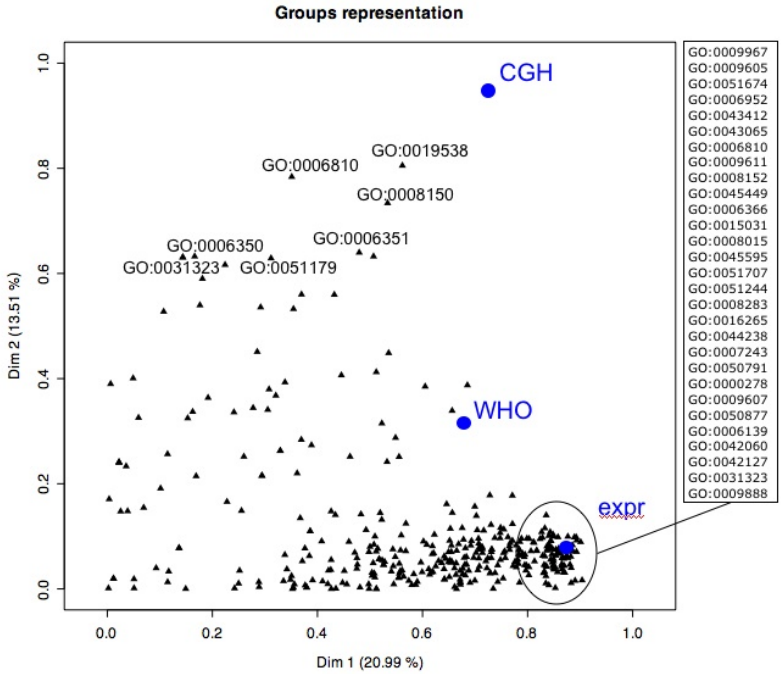

Fig. 2. Groups representation. Each group of variables is projected on the factor map created by MFA: active groups (points) and modules (triangles) are plotted. Only GO identifiers representing modules highly linked to the dimensions 1 and 2 are displayed. To facilitate the interpretation of the plot on axe 1, interesting GO identifiers are shown on a separate box. The qualitative group WHO classification is also shown.

(GO:0051674), 'cell proliferation' (GO: 0008283), and 'regulation of cell proliferation' (GO:0008284). The second one is related to the cell cycle with 'positive regulation of apoptosis' (GO:0043065), 'death'(GO:0016265), and 'mitotic cell cycle' (GO:0000278). Eventually, a third one indicates a link with the response to a stimulus, particularly stress and defense: 'defense response' (GO:0006952), 'response to wounding' (GO:0009611), and 'wound healing' (GO:0042060). These annotations underline the major characteristics of glioblastomas among other malignant gliomas: a rapid progression accompanied by an intense angiogenesis, sustained by cell cycle dysfunctions.

The factor delineating astrocytic tumors from oligendroglial ones (axe 2) is mainly associated with modules related to transport and to transcription processes. Among these biological processes, 'protein metabolism' (GO:0019538), 'transport' (GO:0006810), and 'transcription DNA-dependent' (GO:0006350) annotate 18 genes of those located on 1p or 19q. The homogeneity and coherence of these modules associated with targeted damages of the genome struc-

ture appear as potential cumulative events. They are distinctive features between tumorigenic cell precursors and could therefore constitute reliable markers for glioma diagnostic.

4 Conclusion

When investigating complex diseases such as pleomorphic cancer, it seems necessary to take into account, as far as possible, all the informative experiments available. To tackle this challenging task, we propose to use MFA in such way that it becomes possible to combine data sets coming from different 'omic' areas and to integrate biological knowledge to these data.

Our approach is divided into two parts. In a first step, MFA is used to jointly analyse the structure emerging by the separate molecular levels investigated. The common structures are underlined and graphic outputs are provided such that biological meaning becomes easily retrievable. Partial representations allow the visualization of each 'omic' specific point of view and improve the understanding of the biological bases involved in the situation under study. In a second step, the capacity of MFA to manage supplementary data tables is used to integrate GO annotations. Gene modules are created as groups of illustrative variables and are projected on the space previously created by MFA.

We validated our approach on a complex setting which is the gliomagenesis. The integration by MFA of measured expression changes and genomic locus copy number alterations gives very good insights into the molecular bases of these malignant primary brain tumors. Briefly, a typology of the tumors by WHO classification subtypes was extracted from transcriptomic and genomic modifications. Two main axes holding important parts of tumoral variability were defined: the first was sustained by a cumulative effect of the two molecular levels of investigation, and the second arised only from genomic damages. Relevant mechanisms involved in cancer have been identified and more precisely some well defined in glioblastomagenesis.

Our approach is suitable for a wide range of biological investigations needing a comprehensive view of the datasets structures and an integration of their associated knowledge. Futhermore, one major advantage of this method is not to be bound to any specific experimental design nor to any type of annotation.

References

[Bigner *et al.*, 1999]S. H. Bigner, M. R. Matthews, B. K. Rasheed, R. N. Wiltshire, H. S. Friedman, A. H. Friedman, T. T. Stenzel, D. M. Dawes, R. E. McLendon, and D. D. Bigner. Molecular genetic aspects of oligodendrogliomas including analysis by comparative genomic hybridization. *Am J Pathol*, 155(2):375–386, Aug 1999.

[Bredel *et al.*, 2005a]Markus Bredel, Claudia Bredel, Dejan Juric, Griffith R Harsh, Hannes Vogel, Lawrence D Recht, and Branimir I Sikic. Functional network analysis reveals extended gliomagenesis pathway maps and three novel myc-interacting genes in human gliomas. *Cancer Res*, 65(19):8679–8689, Oct 2005.

[Bredel *et al.*, 2005b]Markus Bredel, Claudia Bredel, Dejan Juric, Griffith R Harsh, Hannes Vogel, Lawrence D Recht, and Branimir I Sikic. High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*, 65(10):4088–4096, May 2005.

[Consortium, 2001]Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33, 2001. 1088-9051 (Print) Journal Article.

[Escofier and Pagès, 1988]B. Escofier and J. Pagès. *Analyses factorielles simples et multiples. Objectifs méthodes et interprétation.* Dunod, Paris, 1988.

[Ge *et al.*, 2003]Hui Ge, Albertha J M Walhout, and Marc Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19(10):551–560, Oct 2003.

[Joyce and Palsson, 2006]Andrew R Joyce and Bernhard O Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7(3):198–210, Mar 2006.

[Khatri and Draghici, 2005]P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–95, 2005. 1367-4803 (Print) Journal Article.

[Reifenberger *et al.*, 1994]J. Reifenberger, G. Reifenberger, L. Liu, C. D. James, W. Wechsler, and V. P. Collins. Molecular genetic analysis of oligodendroglial tumors shows preferential allelic deletions on 19q and 1p. *Am J Pathol*, 145(5):1175–1190, Nov 1994.

[Smith *et al.*, 1999]J. S. Smith, B. Alderete, Y. Minn, T. J. Borell, A. Perry, G. Mohapatra, S. M. Hosek, D. Kimmel, J. O'Fallon, A. Yates, B. G. Feuerstein, P. C. Burger, B. W. Scheithauer, and R. B. Jenkins. Localization of common deletion regions on 1p and 19q in human gliomas and their association with histological subtype. *Oncogene*, 18(28):4144–4152, Jul 1999.