35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Hierarchical clustering combining numerical and biological similarities for gene expression data classification*

Mattia Bosio, Philippe Salembier, Pau Bellot, and Albert Oliveras-Vergès[1]

Abstract— **High throughput data analysis is a challenging problem due to the vast amount of available data. A major concern is to develop algorithms that provide accurate numerical predictions and biologically relevant results. A wide variety of tools exist in the literature using biological knowledge to evaluate analysis results. Only recently, some works have included biological knowledge inside the analysis process improving the prediction results.**

In this work, a knowledge integration scheme is proposed to improve the microarray classification results from [3]. Biological knowledge is used to infer biological similarity which is combined with the classical numerical similarity. The resulting similarity measure is used in a hierarchical clustering process producing new features called metagenes. The goal of the numerical and biological similarities integration is to produce metagenes involving more useful and significant gene signatures.

The proposed algorithm has been tested on 7 publicly available datasets. The results have been compared with the state of the art method. The knowledge inclusion has proven beneficial both for the predictive ability, improving the results repeatability, and for the biological relevance after evaluating the produced signatures with two gene list analysis tools.

I. Introduction

The analysis of high throughput genomic data such as gene expression or methylation data, is currently a very active field of research. A common issue when dealing with this type of information is the need to extract reliable knowledge from the extremely high amount of available data [1]. This is a primary concern in the processes of hypothesis formulation and knowledge generation. For this reason, a plethora of analysis tools have been developed to help the interpretation task and to infer relationships between the gene signatures and biological knowledge databases [12], [7].

In parallel to the development of interpretation supporting tools, in the last years, the inclusion of biological knowledge inside data analysis frameworks has gained importance [1]. Biological knowledge has been used, for example, to identify biologically relevant activated pathways by integrating Gene Ontology (GO) in the analysis process [10]. Moreover, biological knowledge is also used in tools like Hanalyzer [4] to identify gene-to-gene relationships and facilitate the data analysis. In all cases, the knowledge inclusion led to more interpretable results and an easier hypothesis generation from a biological viewpoint.

*Financed by the Fundació privada CELLEX and the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya".

[1]All authors are from Department of Signal Theory and Communications, Technical University of Catalonia UPC, Campus Nord, Jordi Girona 1-3, 08034 Barcelona, SPAIN. mattia.bosio@upc.edu

In this work, prior biological knowledge is used to improve classifiers for high throughput data. The aim is to utilize the biological knowledge to develop algorithms useful not only for data analysis but also for prediction. The classification approach is based on the algorithm presented in [3], which showed very good predictive properties. The algorithm [3] works exclusively with numerical data and relies on a two-step approach. In the first step, a hierarchical clustering is applied over the data to create an extended feature space and the second step takes care of the feature selection. The hierarchical clustering generates a binary tree and a set of new features called metagenes, one for each node of the tree. Metagenes have proved to be useful for classification because they summarize the common behavior of related genes and, in this process, they may filter out residual noise.

The algorithm proposed here modifies the hierarchical clustering to include prior biological knowledge to define the similarity between genes. A similar idea has been implemented in Hanalyzer [4], where pairwise gene similarity is defined as a combination of numerical similarity and of knowledge similarity to infer gene regulatory networks. The difference with our hierarchical clustering is that, in [4], the pairwise similarities are used only once and a threshold is applied. In our case, the similarities are used to infer a complete hierarchical structure and to produce new features. The aim is to generate metagenes able to help both in the noise reduction as in [3], and in the data interpretation by summarizing genes with related biological functions. The proposed algorithm has been introduced in the framework presented in [3] to classify public microarray datasets and the results have been compared with the original algorithm.

This paper is organized as follows: in Section II, the knowledge inclusion algorithm is detailed. The experimental protocol is presented in Section III. The results are discussed in Section IV and the conclusions are drawn in Section V.

II. Materials and methods

This section explains the knowledge integration algorithm, the adopted knowledge database and the computational details to generate the metagenes.

A. The knowledge database

To integrate some prior biological knowledge in the clustering process, the first step is to define the source of information. Many sources are available to evaluate analysis results from a biological perspective. Here a subset of the Molecular Signature Database (MSigDB) [12] has been selected. It is a

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

collection of annotated gene sets provided by the Gene Set Enrichment Analysis (GSEA) software [12].

The MSigDB is composed of six gene set collections varying from manually *curated gene sets*, to *motif sets* or *Gene Ontology terms* related sets. We have chosen the C2, C3 and C5 collections. The C2 gene sets are curated from online pathway databases, publications in PubMed and knowledge of domain experts. C3 are motif gene sets based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes. Finally, C5 consists of gene sets sharing the same GO term.

The database is publicly available and can be represented as a binary matrix $\mathbf{M}$ whose rows are the different genes, while the columns represent the MSigDB gene sets. The information from MSigDB C2, C3 and C5 gene sets are represented in a knowledge matrix $\mathbf{M}$ composed of 22680 unique gene identifiers and 5607 MSigDB gene sets. It is used as knowledge database for the clustering process.

B. Knowledge integration in the clustering

The clustering process is outlined here. It is an iterative process that merges pairs of similar genes or metagenes, and the merging result is considered as a new feature called metagene. If the dataset is composed of p genes $\mathbf{g_i}$, the clustering can be described by the following pseudo code:

Define the active set $\mathbf{F} = \{\mathbf{g_i}\}, i = 1 \rightarrow p$
For $k = 1 \rightarrow p - 1$

1) Calculate the pairwise similarity of each feature pair $S(\mathbf{f_i}, \mathbf{f_j})$
2) Choose the feature pair with highest similarity
3) Build a metagene merging the two joined features: $\mathbf{m_k} = G(\mathbf{f_i}, \mathbf{f_j})$ and add it to F.
4) Remove the two features $\{\mathbf{f_i}, \mathbf{f_j}\}$ from the active set:

End

In [3], the similarity is defined as the correlation between two gene expressions, normalized between -1 and 1. A metagene is defined as the first principal component of the local Principal Component Analysis (PCA) over the two features to be merged [6]. Starting from the second merging step, metagenes and genes are considered as features. As a result, the similarity must be calculated for all genes and metagenes. This point needs to be considered when dealing with the knowledge database since it initially defines binary attributes for genes and does not consider linear combinations of genes.

In this work, the similarity has been modified to include the prior biological knowledge information. For each feature pair $(\mathbf{f_i}, \mathbf{f_j})$, two quantities are calculated: $S_n(\mathbf{f_i}, \mathbf{f_j})$ which is the numerical similarity as in [3] and $S_k(\mathbf{f_i}, \mathbf{f_j})$ the knowledge similarity. The final pairwise similarity is defined as the average of the two partial similarities:

$$S(\mathbf{f_i}, \mathbf{f_j}) = \frac{1}{2}(S_n(\mathbf{f_i}, \mathbf{f_j}) + S_k(\mathbf{f_i}, \mathbf{f_j})) \quad (1)$$

The average has been chosen for its simplicity and because it provided good results in [4]. The knowledge similarity is a Noisy-OR over the attributes as in [4], [5] and defined by:

$$S_k(\mathbf{f_i}, \mathbf{f_j}) = 1 - \prod_k ((1 - r_k)(\mathbf{M_{(i,k)}}\mathbf{M_{(j,k)}})) \quad (2)$$

The Noisy-OR choice is motivated by the results reported in [4] for finding relevant gene relations. It measures the integrated likelihood of a relationship for a feature pair as a product of an array of independent experts [5]. Here, the experts are the attributes and r_k indicates the attribute reliability: $r_k = 1 - \exp(\hat{r}_k)$, where $\hat{r}_k$ is the "consensus reliability estimate" [5] which is learned over the available knowledge matrix $\mathbf{M}$. r_k is bounded between 0 and 1. Eq.(2) allows us to compute the knowledge similarity of two genes or metagenes. As metagenes are obtained as linear combinations of genes, the attribute vector for a metagene is also a linear combination of the gene attribute vectors. The coefficients are set to be ≥ 0 and the final result is a real valued vector bounded between 0 and 1. The formulation of Eq.(2) is an adapted version of the Noisy-OR of [4] to allow working with real values and not only with binary values.

III. Experimental protocol

The experimental protocol to evaluate the benefits of the biological knowledge inclusion is here described. The seven publicly available datasets from the Micro Array Quality Control study, MAQC [11], have been analyzed. They are high quality datasets, analyzed by a wide variety of state of the art algorithms, whose results are available in [11].

The MAQC datasets have been analyzed with a 50 run Monte Carlo simulation as in [3]. In each iteration, a metagene set is built and a classifier up to five features is trained on a training dataset and later evaluated on an independent test dataset. The prediction ability is measured in terms of Matthews Correlation Coefficient (MCC) as advised in [11], [9] since it is more informative than the prediction error rate.

Beside the prediction ability, an additional evaluation is performed studying the difference in the selected gene signatures between including or not the biological knowledge (both with 50 Monte Carlo runs). The aim is to see if the biological knowledge helps in selecting genes which are good for classification and also useful for biological interpretation. The biological usefulness assessment is an extremely complicated task. It is related to the specific problem under study and depends on the scientist' experience. Nevertheless, an established practice in the literature is to evaluate the different gene signatures with automatic analysis tools, for example to find enriched functions or to find genes related to an investigation topic from the literature. Here, all the genes used by the classifiers in the Monte Carlo study are included in the gene signature. In case of metagene, all the genes it summarizes have been added to the gene signature.

Two analysis tools have been used. The first one uses GSEA resources [12][1]. For each gene signature, it calculates an output p-value for each one of the selected MSigDB gene sets [12]. The p-values are calculated as hypergeometric

[1] web: http://www.broadinstitute.org/gsea/msigdb/annotate.jsp

TABLE I
MCC STATISTICS FROM THE MONTE CARLO SIMULATION.

Endpoint		Mean MCC	Std.
A	**COR**	0.278	0.055
	BIO	0.266	0.046
C	**COR**	0.797	0.025
	BIO	0.776	0.011
D	**COR**	0.315	0.085
	BIO	0.297	0.076
E	**COR**	0.773	0.019
	BIO	0.779	0.014
F	**COR**	0.249	0.045
	BIO	0.259	0.034
G	**COR**	0.162	0.042
	BIO	0.154	0.034
H	**COR**	0.863	0.014
	BIO	0.866	0.012

Averages: $\mathbf{\Delta_{Mean}} = -0.008$, **Std**(COR) / **Std**(BIO) = 1.378

distributions of overlapping genes between the analyzed gene signature and the MSigDB gene set. A low p-value indicates a high probability that the MSigDB gene set is represented in the gene signature and therefore that genes used for classification have something in common from a biological viewpoint (function, position, disease, etc.).

The second tool is Biograph [7], it quantifies relationships between individual genes and a key term (e.g. the studied disease). Biograph analyzes individually the genes of the signature and quantifies their relationship with the key term based on a knowledge database. The output score is proportional to the gene key-term relationship.

The algorithm is implemented in Matlab and the tree-construction has a n^2 processing time growth, where n is the feature number, taking about 30" for 10^4 genes. This time could be reduced by an optimized parallel implementation.

IV. RESULTS

This section presents first the results in terms of predictive ability and then, the collected gene lists are compared to assess their biological usefulness.

Table I reports the statistics of MCC values obtained during the Monte Carlo study. The columns define the classified endpoint, the clustering method, the mean and standard deviation of MCC values. The rows are hierarchically organized depending on the classified dataset (A to H) and on the clustering method: **COR** for the clustering exclusively defined with correlation [3], **BIO** for the clustering integrating biological knowledge. Two overall values are also reported: the global mean MCC difference between the **COR** and **BIO** cases, and the ratio between the standard deviations.

As can be observed, the mean MCC values are very similar for all datasets. An average MCC difference of 0.008 is observed over the seven datasets. The MCC results are slightly better with the **BIO** clustering. However, this difference is not significant when compared to the MCC standard deviations. It can be concluded that both methods have equal mean performances.

The analysis of the standard deviations offers additional insights about the changes induced by the knowledge integrated clustering. The standard deviation is consistently smaller for the **BIO** case, on average 1.378 times smaller than the **COR** case (equivalently the variance is 1.899 times smaller). Using the **BIO** clustering is therefore beneficial for classification since it preserves the mean performance and reduces the results variance. This is a good feature concerning the results robustness and repeatability. This behavior in MCC values is maintained if the error rate is studied (not included here for space reasons): the mean difference is 0.54%, while the variance ratio equals 1.585.

A summary of the gene lists comparison is shown in Fig.2 where a graphical representation for the analysis with GSEA and Biograph is presented for all the datasets except H[2]. The data are organized in six blocks, one for each dataset. Each block has two bar plots: on the right side, the 5 most relevant MSigDB gene sets from the GSEA analysis are shown; the reported scores correspond to $-\log(p)$, where p is the calculated p-value. The white bars represent the results for the **BIO** case, while the black ones are for the **COR** case. On the left side, the scores of the 5 genes selected by Biograph as the most relevant with respect to the query key term are reported. The query key term for the Biograph analysis depends on each dataset and is related to the studied phenomenon: A dataset: lung neoplasms; C dataset: liver neoplasms; D and E datasets: malignant breast neoplasms; F dataset: Multiple Myeloma, G dataset: Survival Analysis.

In the majority of the GSEA results presented in Fig. 2, the gene signatures from the **BIO** case involve MSigDB gene sets with more significant p-values than the signatures from the **COR** case. This happens for the A, C, D and F datasets. For the E dataset, the p-values are substantially equivalent. Moreover, they are not as high as in other datasets. By contrast, the G dataset leads to results in which the **COR** gene signature is more significantly enriched with MSigDB gene sets than the **BIO** case. This may be connected to the fact the number of genes of the **COR** signature is about three times that of the **BIO** signature. Therefore, genes of the **COR** signature have a higher probability of being annotated in MSigDB gene sets. Globally, it can be concluded that the gene signatures obtained with the **BIO** clustering are enriched with more significant MSigDB gene sets. Therefore, the **BIO** clustering is a better method to find genes with common biological functions or relationships.

The Biograph results offer insights about the datasets when the GSEA scores are equal, like the E dataset. In this case, the GSEA scores are low, showing how none of the gene signatures significantly overlaps with known MSigDB gene sets, but this does not mean that these gene signatures are useless. From Biograph analysis, the ESR1 gene is the most important in both signatures. This is because the samples are classified by the Estrogen Receptor status, which is the ESR1 gene [8]. In the E datasets, the ESR1 gene alone is enough to correctly classify the majority of samples, thus reducing the need for additional genes. Therefore the possibility to have significant overlaps with MSigDB gene sets is small, but both gene signatures identify the key gene for classification.

[2]H results are not shown because **COR** and **BIO** have the same output: all classifiers are based on the same gene signature composed of XSIT and EIF1AY genes, both actively involved in sex determination.

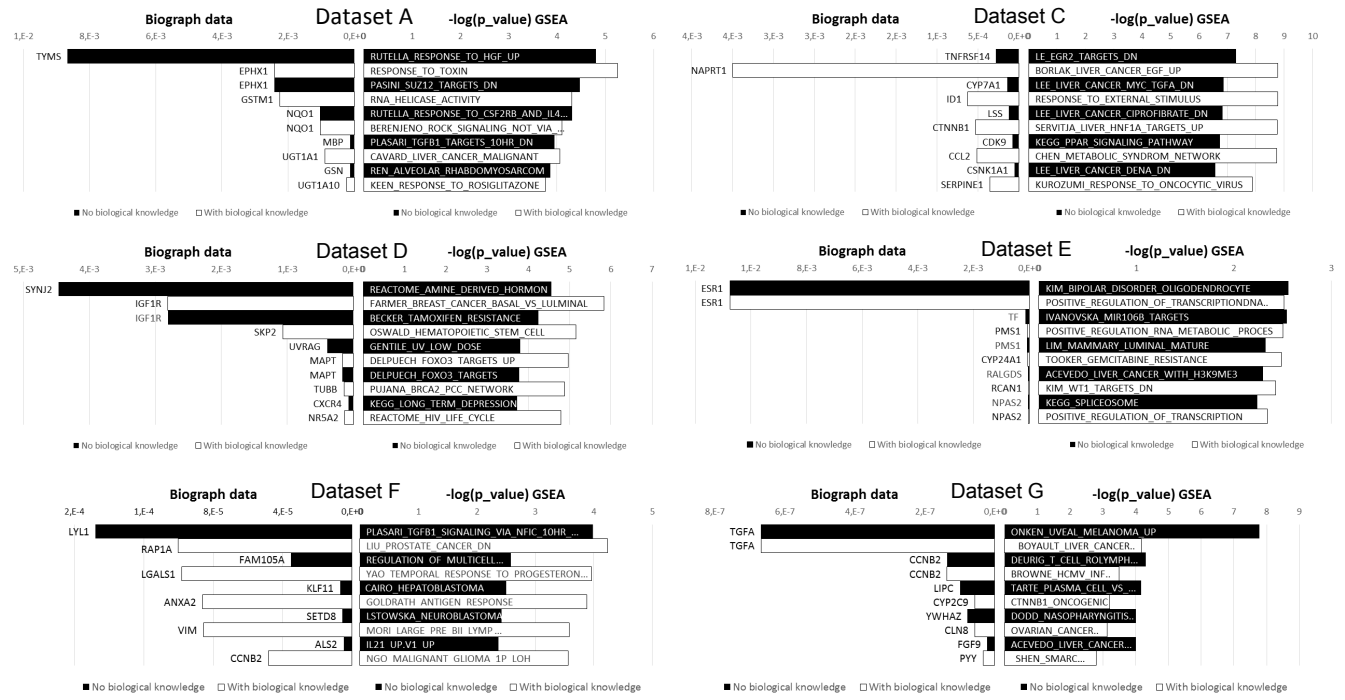

Fig. 1. Experimental results over the MAQC datasets.

In the C and F, the **BIO** clustering chooses genes that are more related to the key term. Sometimes, the **COR** clustering includes a gene with higher score like in A, D and F case, but this behavior is not observed for the remaining genes in the list. The G dataset does not show significant differences because both methods chose the same individual genes as basis for the classifier. More than 80% of the classifiers are based on a two-gene signature composed of CHML gene (6th in both lists) and either TGFA or CCNB2, the first two genes in both Biograph lists. Both lists are based on the same relevant genes and none of the two methods was able to produce metagenes with better predictive abilities.

From these experimental results, it can be concluded that, in average, the inclusion of biological knowledge in the clustering process helps to have more stable prediction performances without losing predictive power. Moreover, the proposed method is able to find gene signatures with more significantly represented MSigDB gene sets from GSEA analysis. Finally, this approach favors the identification of gene groups more correlated with the studied phenomenon.

V. Conclusions

A new method introducing biological knowledge in hierarchical clustering and in the generation of metagene has been proposed. It has been tested on public datasets and its performances have been compared to state of the art method [3]. It showed interesting results in terms of both predictive potential and gene signature significance making of it an interesting tool for genomics data analysis.

The proposed method is very flexible and can be applied to other genomics data like methylation data or RNA-seq. Furthermore its modularity makes it suitable for any kind of knowledge database, once it can be represented with a binary matrix. Finally, future work will focus on the possibility to define improved similarity metrics and combination rules (beyond the average used here) [4], [2] which can further improve the clustering process and the biological relevance of the found gene signatures.

References

[1] G. Alterovitz and M. Ramoni, “Knowledge-Based Bioinformatics From analysis to interpretation.”, Wiley, 2010.

[2] S. Boriah et al., “Similarity Measures for Categorical Data: A Comparative Evaluation”, SIAM Int. Conference on Data Mining, 2008.

[3] M. Bosio et al., “Gene expression data classification combining hierarchical representation and efficient feature selection.” Journal of Biological Systems, Special Issue on Genomic Signal Processing, to be published.

[4] S.M. Leach et al., “Biomedical discovery acceleration, with applications to craniofacial development.” PLoS computational biology, vol. 5, 2009.

[5] S. Leach et al., “Assessing and combining reliability of protein interaction sources.” Pac Symp Biocomput, 433, 2007.

[6] A. B. Lee, B. Nadler, and L. Wasserman, ‘Treelets - an adaptive multi-scale basis for sparse unordered data,” Annals of Applied Statistics, no. 2, 435-471, 2008.

[7] A.M.L. Liekens et al., “BioGraph: Unsupervised Biomedical Knowledge Discovery via Automated Hypothesis Generation,” Genome Biology 12:R57, 2011.

[8] G.B Livshyts et al. “ESR1 gene allelic polymorphism analysis in population of Ukraine.” Cytology and Genetics, Volume 46, Issue 4, pp 220-226, 2012.

[9] B. W. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme.” Biochimica et Biophysica Acta, vol.405, no. 2, pp. 442-451, 1975.

[10] A. Rao and A.O. Hero, “Biological pathway inference using manifold embedding.” IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 5992-5995, 2011.

[11] L. Shi et al., “The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.” Nature biotechnology, vol. 28, pp. 827–838, 2010.

[12] A. Subramanian et al., “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.” Proc. Natl. Acad. Sci. USA 102, 15545-15550, 2005.