Intégration de données hétérogènes pour l'identification de signatures moleculaires : une approche par score-local

Marine Jeanmougin [1] & Mickael Guedj [2] & Christophe Ambroise [3]

[1] *Immunité et Cancer, Institut Curie - 26 rue d'Ulm, Paris - marine.jeanmougin@curie.fr*
[2] *Département de Bioinformatique et Biostatistique - Pharnext, 11 rue des Peupliers, 92130 Issy-les-Moulineaux - mickael.guedj@pharnext.com*
[3] *Laboratoire de Mathématiques et Modélisation d'Evry (UMR 8071), Université d'Evry-Val-d'Essonne, 23 Bd. de France, 91037 Evry Cedex - christophe.ambroise@genopole.cnrs.fr*

Résumé. Au cours de la dernière décennie, les progrès en Biologie Moléculaire ont favorisé l'essor de techniques d'investigation à haut-débit. En particulier, l'étude du transcriptome à travers les puces à ADN ou les nouvelles technologies de séquençage, a permis des avancées majeures dans les sciences du vivant et la recherche médicale. Dans ces travaux, nous nous intéressons au problème de sélection d'un ensemble de gènes d'intérêt, aussi appelés "signature moléculaire". De telles signatures sont utilisées en recherche médicale, et en particulier en oncologie, pour le diagnostic et le pronostic ainsi que pour l'identification de nouvelles cibles thérapeutiques.
Afin de pallier les limites des méthodes classiques de sélection de gènes qui s'avèrent peu reproductibles, nous présentons un nouvel outil, DiAMS (DIsease Associated Modules Selection), dédié à l'identification de modules enrichis en gènes significativement associés à la maladie. DiAMS repose sur une extension du score-local et permet l'intégration de données d'expressions et de données d'interactions protéiques. Dans cet exposé, nous détaillerons les différents principes de cette approche et proposerons une stratégie de simulation afin d'évaluer les performances de notre méthode, en terme de puissance, de taux d'erreur de type I et de reproductibilité. DiAMS sera ensuite intégré dans un pipeline d'analyse que nous appliquerons à l'étude de la rechute métastatique dans le cancer du sein.

Mots-clés. Transcriptome, Information a priori, Intégration de données hétérogènes, Score-local, Cancer du sein.

Abstract. During the last decade, an incredible number of statistical tools have emerged for studying the transcriptome. A key motivating factor is the selection of genes, often referred to as the "molecular signature", whose combination is characteristic of a biological condition. Signatures give rise to new clinical opportunities, for understanding disease predispositions, improving diagnostics or prognostics, and providing new therapeutic targets as well as individualized treatment regimens. Their identification has become a topic of much interest in medical research, with several applications emerging, particularly in the field of Oncology. However, it turns out that the signatures resulting from classical tools proposed in the literature suffer from a lack of reproducibility and are not statistically generalizable to new cases. A major statistical issue in high-throughput transcriptome experiments is how to select relevant and robust signatures given the large number of genes under study. We focus on robust gene selection through differential analysis approaches. We present a new approach, DiAMS (Disease Associated Modules Selection), that aims at improving the robustness of signatures across studies. The proposed methodology integrates both Protein-Protein Interactions (PPI) and gene expression data in a local-score approach. In this talk, we will present the global approach of DiAMS, for the selection of modules significantly enriched in disease associated genes. We will also introduce the algorithm for module ranking and how to assess the significance of modules by Monte-Carlo simulations. Finally, we evaluate the performance of DiAMS in terms of power, false-positive rate and reproducibility.

Keywords. Transcriptome, Prior knowledge, Integration of heterogeneous data, Local-score, Breast cancer.

1 Background

During the last decade, an incredible number of statistical tools have emerged for studying the transcriptome. A key motivating factor is the selection of genes, often referred to as the "molecular signature". A major statistical issue in high-throughput transcriptome experiments is how to select relevant and robust signatures given the large number of genes under study. We present a new approach, DiAMS (Disease Associated Modules Selection), that aims at improving the robustness of signatures across studies. The proposed methodology integrates both Protein-Protein Interactions (PPI) and gene expression data in a local-score approach.

2 Extended version of the local-score approach for module discovery

The local-score statistic is a matter of interest in biological sequence analysis. It found many applications in pattern identification to locate transmembrane or hydrophobic segments, DNA-binding domains as well as regions of concentrated charges. The literature on the subject of local-score includes, but is not limited to [Brendel et al., 1992], [Karlin and Brendel, 1992] or [Guedj et al., 2006].
We propose to extend the local-score statistics to the discovery of high-scoring modules of genes in a PPI network. Let us consider here that we have enumerated all the possible modules of the network in a list called $\mathcal{M}$. Obviously, it is not possible in large-scale networks and we dedicate the section 2.2.1 to the development of an alternative approach. We denote W_g, the score of a given gene g. The local-score is thus defined as the value of the highest scoring module (*i.e.* the module whose sum of gene score is maximal):

$$\mathrm{L} = \max_{M \in \mathcal{M}} \left(\sum_{g \in M} W_g \right).$$

Note that a module is maximal in the sense that it can not be extended or shortened without reducing the local-score statistic.

This definition of the local score restricts our search to the highest scoring module. However, the next highest scoring modules may be potentially interesting for the study. We therefore rank all modules of the initial network, such that the kth best module is defined as the module with the kth best local-score denoted L_k such as $\mathrm{L}_1 > ... > \mathrm{L}_m$, and identify significant ones. Such an approach will probably yield to the identification of overlapping modules. For instance, the second best module will likely include or be contained in the first highest scoring module. To avoid such situations that provide limited information, we look at disjoint modules. Thus, once the best module has been identified, each gene included in it is thus removed from the remaining modules.

2.1 Module scoring

The local-score statistic relies on gene scores, denoted W_g, that reflects the association of a given gene to the phenotype of interest. We define the scoring function as follows: $W_g = Z_g - \delta$, such as Z_g is the individual score of each gene g and δ a constant specified in the following paragraph. In this work, we derive the individual score Z_g of gene g from its p-value, denoted p_g, resulting from a statistical test such as `limma` from [Smyth, 2004]. Given that a high score Z_g should denote a high chance of association with phenotypes of interest, the p-values need a transformation such as $Z_g = -log_{10}(p_g)$, to be used as an individual score for each gene.
A constraint of the strategy is to have expected negative individual scores, *i.e.* $\mathbb{E}(W_g) \leq 0$, otherwise the module with the highest score would easily span the entire network. Consequently, a constant δ must be subtracted to obtain corrected scores. Genes with a score higher than δ will improve the cumulative score of a given module whereas genes with a score below the threshold will penalize it. We set the value of δ equal to the significance level $\alpha = 0.05$.

2.2 Disease associated modules selection

In the present subsection we detail the global strategy, to search for functional modules presenting unexpected accumulations of genes associated to a phenotype of interest in a PPI network.

2.2.1 Input parameters

The first input parameter that is passed to `DiAMS` is a PPI network. The main issue when working with biological networks lies in the impossibility of exploring the huge space of possible gene subnetworks. Here, we propose a strategy which allows the entire network to be screened without constraints on module sizes by converting the network into a tree structure using a clustering algorithm. This is driven by the observation that biological graphs are globally sparse but locally dense, *i.e.* there exist groups of vertices, called communities, highly connected within them but with few links to other vertices. Therefore, by applying a strategy of clustering which enables to obtain a hierarchical community structure we are able to capture much information about the network topology. The main advantage is that the hierarchical structure renders it relatively easy to go through it instead of exploring all possible subnetworks. Thus the preliminary step of our approach is to convert the network structure into a relevant tree structure. For this purpose, we use the approach of [Pons and Latapy, 2004], named `walktrap`. The authors employed a random walk strategy through the network for detecting dense modules, introducing a similarity measure based on short walks, which is used to define a distance matrix between any two genes (or nodes) of the network. According to Ward's criterion, they are able to

infer a tree structure. A module is no longer defined as a subnetwork but as a subtree of the hierarchical structure (see Figure 1). In analogy with the definition of the local-score for network, we define it for a hierarchical community structure, denoted $\mathcal{H}$, as follows:

$$L = \max_{H \subseteq \mathcal{H}} \left(\sum_{g \in H} W_g \right),$$

such as H is included in $\mathcal{H}$ if H is a subtree of $\mathcal{H}$, *i.e.* H can be obtained from $\mathcal{H}$ by deleting nodes in $\mathcal{H}$.

The second parameter that has to be passed to the method is a vector of scores Z_g, that quantifies for each gene its association to the disease. In this study the scoring function is related to the differential expression of the gene such as significant genes, *i.e.* those that are significantly differentially expressed, have a higher score than non-significant genes. However, other kinds of scoring approaches may also be suitable as well as high scores, denoting a strong association to the disease.

2.2.2 Module ranking through a local-score strategy

Once both the tree structure and the score vector have been defined, we search for accumulation of high-scoring genes in the tree. The strategy for the selection of significant modules can be described in the following three-step algorithm:

1. ***Initialization*** - The first step consists of enumerating modules of the tree in a list and assigning them a score, which is defined as the sum of individual scores, denoted W_g, of all the genes that constitute it, see Figure 1.

2. ***Module ranking*** - The second step involves an iterative local-score algorithm: (i) the highest-scoring module is identified (ii) then, it is removed from the list of modules. Steps (i) and (ii) are then repeatedly applied until all disjoint modules have been enumerated. Thus, we obtain a ranked list of m modules and their respective local-scores $\mathrm{L}_1, ..., \mathrm{L}_m$ such as $\mathrm{L}_1 > ... > \mathrm{L}_m$ with the ith best module being disjoint from the preceding $i\mathrm{th} - 1$ best modules.

3. ***Module significance assessment*** - Given $\mathrm{L}_1, ..., \mathrm{L}_m$, the last step proposes a way to select a set of modules significantly enriched in disease associated genes. The global significance of each module is assessed via Monte-Carlo simulations. Through this permutation procedure we obtain a p-value for each module and are able to make a conclusion about its significance of at a given level.

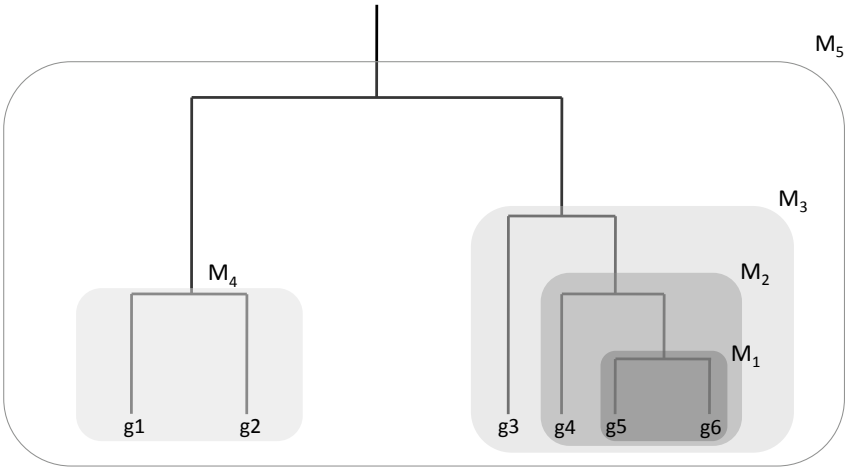

Figure 1: **Module description.**

A module is defined as a subtree of the hierarchical structure. Leaves, i.e. genes, are also considered as modules. Thus, in this figure we count eleven modules: six modules of size one and five modules of size greater than one. For instance, the module M_3 is composed of four genes. Its score is the sum of each individual gene score, W_{g3}, W_{g4}, W_{g5} and W_{g6}.

To evaluate the significance of modules we derive their distributions under the null hypothesis of no accumulation of high-scoring genes, using Monte-Carlo permutations.

3 Evaluation strategy

In this section we detail the strategy adopted to evaluate `DiAMS`. Each evaluation criterion, namely the power, the type-I error rate and the reproducibility, are compared with our modular strategy and its individual scoring counterpart, `limma`. Here, we perform the simulations under a Gaussian model, *i.e.* for data produced by a microarray experiment.

Power Study Recent results from [Gandhi et al., 2006, Lage et al., 2007] or [Oti and Brunner, 2007], which have motivated the development of `DiAMS`, suggest that genes involved in the molecular mechanisms of genetic diseases interact together in functional modules. Therefore, to evaluate our approach, we designed a simulation study under this hypothesis of a modular activity of genes. Firstly, it involves randomly sampling significant modules in the tree structure. Secondly, we simulate a gene expression matrix. The genes belonging to non-significant modules are simulated under the null hypothesis of equality across the mean expression levels for both conditions: $\mu_g^{(1)} = \mu_g^{(2)}$, while genes of significant modules are

simulated under H_1, such as $\mu_g^{(2)} = \mu_g^{(1)} + \Delta$, with Δ in $\{0.5, 0.75, 1, 1.25, 1.5, 2, 3\}$. The p-values obtained from Monte-Carlo permutations are then adjusted using the Benjamini-Hochberg procedure to control the FDR criterion at a level of 5%.

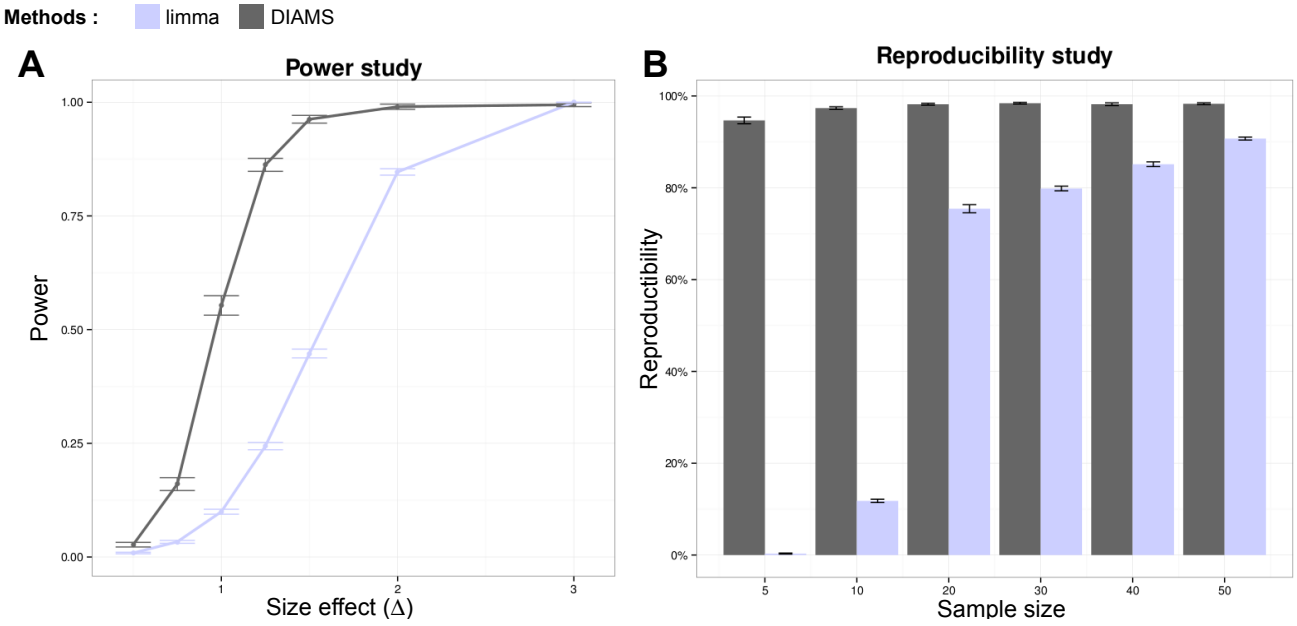

Figure 2: **Power and reproducibility results.**

A *- The mean of power values over the* $1,000$ *simulations and its* 95% *confidence interval are calculated at a* 0.05 *FDR level for the* `DiAMS` *method (in dark gray) and the* `limma` *statistic (in light gray) and displayed according to* Δ*, the difference of mean expression levels under* H_0 *and* H_1*.* ***B*** *- This barplot displays the results of the reproducibility analysis for which we compute the mean of the overlap between a signature of reference and signatures of subsampled expression matrices over* $10,000$ *simulations. We represent the* 95% *confidence interval for each sample size.*

The Figure 2-A illustrates the results of the power analysis for both `DiAMS` and `limma`. As expected, the curve describing the statistical power converges to 1 with increasing values of Δ. For $\Delta = 0.5$, it appears that the power is very similar for both approaches , although `DiAMS` is slightly more powerful. For all values of Δ in $\{0.75, 1, 1.25, 1.5, 2\}$, we observe large differences in power between the two approaches with `DiAMS` outperforming `limma`.

We also consider a scenario where genes are simulated independently under H_1, *i.e.* without assuming a modular activity. The power values obtained are identical for both methods, due to the fact that the p-values of individual genes are exactly the same as

those resulting from `limma`. At worst, if the hypothesis of a functional relationship between disease genes is wrong, the power results are equivalent to `limma`.

False-Positive Rate Using the same simulation strategy as described in the previous subsection, we assess the false-positive rate. A statistical test conducted at a significance level of 0.05 should control the false-positive rate at 5%. Thus, by simulating an entire dataset under H_0, *i.e.* $\forall g : \Delta = 0$, we evaluate the proportion of genes spuriously selected as significant. Both the `limma` and `DiAMS` false-positive rates are estimated for various sample sizes ranging from 5 to 50 samples per condition.

It appears that the rates are similar for both approaches and they lie within the 95% confidence interval (data not shown). For each sample size, `limma` and `DiAMS` meet the theoretical false-positive rate.

Reproducibility study Next, we examined the agreement between signatures using a subsampling procedure. As described in the power study, we simulated modules under H_1 as well as the corresponding expression matrix and compute a signature of reference. Then, we randomly subsampled the replicates of the initial matrix with replacement and estimate the signature again. The reproducibility is calculated as the overlap between the reference signature and the signature of subsampled expression matrices. This procedure is performed for various subsample sizes from an initial dataset containing 50 samples for two conditions.

The reproducibility results are averaged over $10,000$ simulations and displayed in Figure 2-B. For the larger sample size, the initial matrix has been re-sampled with replacement. Even if the sample size is the same, meaning that the noise added to the initial dataset is relatively low, the percentage of reproducibility for `limma` is only 90% while `DiAMS` almost reaches 100%. All the results displayed in Figure 2-B show that `limma` is very sensitive to the noise in data while `DiAMS` results appear to be more consistent. This is especially true for small sample sizes, for which the reproducibility of the signature is about 95% with the `DiAMS` approach while the percentage is almost null (0.3%) with the `limma` selection method. The gap remains very large for the other sample sizes and `DiAMS` clearly provides significantly better results than `limma` in terms of reproducibility.

4 Conclusion

We developed a network-based approach named DiAMS for the selection of gene signatures. We demonstrated through simulations that, under the assumption of a modular activity of genes, DiAMS is more efficient in terms of power and reproducibility than the moderated t-statistic strategy used in limma. We also applied this method to study

the metastatic relapse of Estrogen Receptor negative breast cancers and demonstrated the relevance of signatures obtained using DiAMS, by highlighting relevant biological phenomena. In addition, such an approach facilitates the ease of the interpretation of the resulting signature by providing information on molecular mechanisms through the extraction of PPI subnetworks.

Bibliographie

References

[Brendel et al., 1992] Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E., and Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proceedings of the National Academy of Sciences*, 89(6):2002–2006.

[Gandhi et al., 2006] Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293.

[Guedj et al., 2006] Guedj, M., Robelin, D., Hoebeke, M., Lamarine, M., Wojcik, J., and Nuel, G. (2006). Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. *Stat Appl Genet Mol Biol*, 5(1):Article22.

[Karlin and Brendel, 1992] Karlin, S. and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257(5066):39–49.

[Lage et al., 2007] Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316.

[Oti and Brunner, 2007] Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11.

[Pons and Latapy, 2004] Pons, P. and Latapy, M. (2004). Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293.

[Smyth, 2004] Smyth, G. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1).