# A module-based approach for evaluating differential genome-wide expression profiles

José Flávio S. Dias Júnior
Evandro Chagas Institute
PPGCC, Federal University of Pará
Belém, PA, Brazil
Email: joseflavio@iec.gov.br

Ronnie Alves
Instituto Tecnológico Vale
PPGCC, Federal University of Pará
Belém, PA, Brazil
Email: ronnie.alves@itv.org

Thérèse Commes
Institut de Biologie Computationnelle
Universite Montpellier
Montpellier, France
Email: commes@univ-montp2.fr

*Abstract*—**Transcription is the process of making an RNA copy of a gene sequence. Next, this copy (mRNA) is then translated into proteins. Proteins dictates the expected behavior inside the cells and are required for the structure, function, and regulation of the body's tissues and organs. Together, transcription and translation are known as gene expression. *Transcriptograms* are basically defined as "images" of gene expression data of genomes, by generating expression profiles for transcriptomes. They allow to assess cell metabolism, being capable of discriminating the stage the cell is going through at a given instant, as well as pointing metabolic changes in altered cellular states as compared to a control state, independently of the transcriptome profilling protocol. Though, they cannot highlight differential expression profiles. We present a new possibility of RNA-Seq data analysis using *Transcriptograms* for discovering module-based differential expression profiles. We demonstrate its practical application while obtaining more specific gene signatures as well as functional annotations, closely related to biomedical context. Moreover, these signatures are also enriched by survival cancer analysis.**

## I. Introduction

In multicellular organisms, nearly every cell contains the same genome and thus the same genes. However, not every gene is transcriptionally active in every cell. These variations underlie the wide range of physical, biochemical, and developmental differences seen among various cells and tissues and may play a role in the difference between health and disease. Thus, by studying transcriptomes, researchers hope to determine when and where genes are turned on or off in various types of cells and tissues.

When the genome-wide transcriptional profile of heterogeneous samples is measured under different physiological states, any observed differences are strongly confounded by differences in cell type compositions between samples. Recent studies suggest that the microenvironment of a tissue may change under different physiological states and can contribute to the etiology of diverse diseases [1], [2], [3].

Unsupervised mixture models have been developed to explore gene expression profiles. However, they require prior knowledge of either the cell type frequencies within a given tissue, or the in vitro gene expression profiles of each component cell type. In reality, this information can be difficult to obtain and presents a major drawback for these kinds of approaches [4].

Analysis of trancriptomes often cluster together genes by their co-expression, or co-variation in time, which implies that these cluster definitions depend on the stage the cell is going through or on the protocol used to produce the assessed sample. *Transcriptograms* are basically defined as gene expression profiles, generating expression data for transcriptomes. The idea of the method is to consider averages of expression data over neighboring genes disposed on a line. In one hand, this procedure targets a global assessment of expression data of whole genomes. On the other hand, it requires the definition of gene neighborhood when disposed on a line, which is not straightforward.

In [5] it was introduced a method for sorting a list of genes using the computational physics method known as Monte Carlo, called Cost Function Method (CFM). The genome ordering of CFM defines a mathematical metric that correlates the distance between two genes on the list with their mutual influence.

The traditional method of *transcriptogram* analysis is to explore gene expression profiles based on a seriated protein-protein interaction network. It can not detect differential genome-wide signatures between group of samples. In order to complement and expand the analytical possibilities of the *transcriptogram*, it is presented a new module-based strategy to evaluate differential genome-wide expression profiles.

## II. Material and Methods

### A. RNA-Seq Data

A RNA-Seq data corresponding to a real experimental study [6] was selected to explore differential gene expression profiles, which is accessible on-line in Gene Expression Omnibus (GEO) database through the GEO Series accession number (GSE48173). The data has 72 samples Illumina HiSeq 2000 (Homo sapiens), classified as follows: 43 Acute Myeloid Leukemia (AML), 12 Acute Lymphoblastic Leukemia (ALL) and 17 (Healthy).

### B. Transcriptograms

The *Transcriptograms* take into account the notion that a pair of genes must correlate with the probability that their protein products are associated as well. In protein-protein interaction (PPI) network maps, nodes represent proteins and

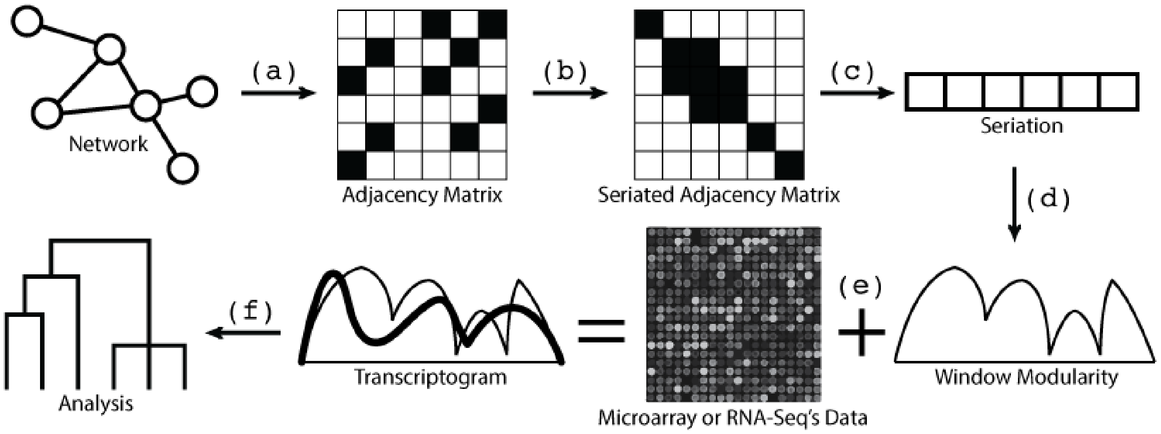

Fig. 1. Pipeline for *Transcriptograms* data analysis. (a) The graph that represents the PPI network is transformed into an adjacency matrix. (b) The adjacency matrix is optimally adjusted by a seriation algorithm (CFM or Claritate). (c) The seriation is extracted from the adjacency matrix. (d) The window modularity is calculated based on network seriation and its modules are evidenced by the peaks. (e) The expression data (RNA-Seq or Microarray) are adjusted along the window modularity, resulting in the *Transcriptogram*. (f) A module-based analysis for evaluating differential genome-wide expression profiles.

edges represent a physical interaction between two proteins. The edges are non directed, as it cannot be said which protein binds the other, that is, which partner functionally influences the other [7]. As the interactions, in which a given protein participates, are likely to correlate with the protein's functional properties, protein interaction maps are frequently utilized to uncover in a systematic fashion the potential biological role of proteins of unknown functional classification [2], [8], [9], [3].

Once having the protein-protein interaction network of a genome, the next step is to devise a seriation scheme in such way that this sort (of protein-protein network modules) correlates to functional gene expression profiles. Modules are present in the form of black dots agglomerations around the adjacency matrix diagonal, as can be seen for example in Figure 2. This visual analysis is good but not enough to identify interactive modules. The modules identification uses the measure called window modularity [10].

Thus, the *Transcriptogram* is, essentially, a 1-dimensional projection of the expression profiles over the seriated PPI network. An overall scheme of the pipeline to use *Transcriptograms* for RNA-Seq Data Analysis is depicted in (Figure 1). In this work we present a new possibility of RNA-Seq data analysis using *Transcriptograms* for discovering module-based differential expression profiles. We demonstrate how the proposed approach yields insights into functional annotations over a cancer-related RNA-Seq data, which can not be detected by single-gene analysis (ranked gene lists)

### C. Seriation over a high-quality human binary PPI

*Transcriptograms* have been devised solely based on seriation over PPI obtained from the STRING database [11]. Though, it can map to a more large catalog of genes, it does not imply in having curated information. Therefore, in this particular study we make use of the human PPI network introduced in [12], named HI-II-14, corresponding to a systematic map of 14,000 high-quality human binary PPI. The map also uncovers significant inter-connectivity between known and candidate cancer gene products, providing unbiased evidence for an expanded functional cancer landscape.

In this work we also introduce the method Claritate to build *Transcriptograms*, being a promising alternative to CFM. The Claritate uses a strategy of spatial proportion of the actual distances between the proteins in the ordered list in relation to the virtual minimum distance observed in the graph representative of the PPI network. The first step relies on building an ordinary matrix having all the minimum distances among all protein-protein interactions. This matrix is calculated through the utilization of the Floyd-Warshall algorithm over the associated adjacency matrix [13]. A metaheuristic process of protein selection takes place to calculate the proportional distances. Thus, the distances between the proteins in the ordered list are adjusted and subjected to a metric of acceptance, called *dispersion*, which guides the movement of relationships (black dots of the adjacency matrix) for the main diagonal, forming the protein modules. A similar strategy has been used in [14] to measure the quality of clustering solutions. Claritate also make uses of the scale-free property of biological networks [2], focusing on a pre-selection of potential *hubs* and sorting them according to its excentricity [15], resulting into the first seriation. Figure 2 shows the *Saccharomyces* PPI network, used in Rybarczyk-Filho et al. (2011) along with its corresponding seriation results (CFM and Claritate). It can be observed, clearly, a significant improvement of Claritate in the detection of network modules (black bloxes along the diagonal).

### D. Calculation of Differentially Expressed Genes (DEG)

DEG were calculated along with the R package *GeneSelector* [16]. Ranked gene lists were enumerated using the function *RankingWelchT()*, which provide gene rankings based on Welch $t$ statistic.

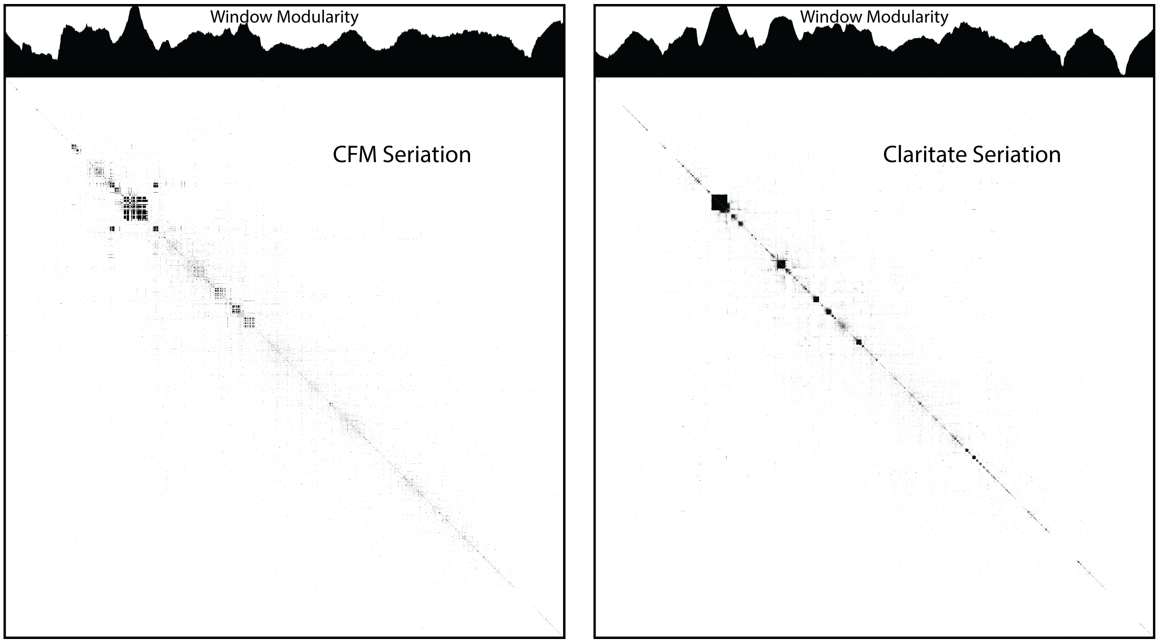

Fig. 2. *Saccharomyces cerevisiae* PPI network. Comparison of window modularity and PPI adjacency matrix seriated by the CFM and the Claritate. The window modularity is a representation of the matrix for identification of modules through the peaks and valleys.

### E. Functional enrichment analysis

Ranked gene lists are usually evaluated over biological databases such as Gene Ontology to verify whether the gene list is enriched (covered) by terms enregistred in the annotation database. For example, given a list of genes being up-regulated under certain conditions, an enrichment analysis will search for GO terms that are over-represented (or under-represented) using annotations for that gene set. In this work we have used the AmiGO/Term Enrichment Service [17] of the Gene Ontology Consortium (GO) [18].

### F. RNA-Seq Data Analysis using Transcriptograms

We have developed a pipeline for RNA-Seq data analysis using *Transcriptograms* that can be run directly from terminal code line, using a set of shell scripts (associated to data analysis steps). The pipeline covers all steps (Figure 1), including other facilities for exploration and visualization of *Transcriptograms*. The pipeline has been implemented in JAVA, R and C++ programming languages, being available on https://github.com/joseflaviojr/transcriptograma/wiki.

## III. Results and Discussion

The differential gene expression profiles for two groups of camparison were carefully evaluated using Welch's unequal variances *t*-test. The first (i) group is related to ALL patients versus Healthy ones, and the second (ii) group is related to AML patients versus Healthy ones. Next, we make use of the HI-II-14 seriated network to highlight patterns of differential gene expression between patients in both groups. Thus, gene expression variance is arranged following the optimal order calculated by the seriation procedure. We can observe co-expressed patterns which is closer related to the notion of complex regulatory gene networks. Moreover, these patterns correlates to functional modules from the seriated protein-protein interaction. Module-based differential expression profiles are identified automatically by closer inspection of peaks and valleys of gene variance, where the lowest gene expression variances are used as modules' frontiers (Figure 3). Without loss of generality, the remaining discussions is focused on the second group (AML vs Healthy).

In classical differential expression analysis the main goal is the selection of a ranked list of genes that might be potential markers to differentiate the control group from the target one. Next, functional enrichment analysis takes place to find biological soundness of the discovered gene signatures. Rather than focus on high ranked genes that might be poorly annotated, the proposed approach takes into account all genes in the experiment, not only that ones resulting from an arbitrary threshold. Although, one can easily select ranked genes either i) locally, by searching highest ranked genes intra-modules; or ii) globally, by searching highest ranked genes inter-modules. The availability of so many different gene ranking methods and the lack of consensus in the community, with respect to the limitations and capabilities of all of them, opens a clear space for systematic studies to better evaluate the current methods with relevant and objective criteria [16]. Nonetheless, the choice of an unique ranking method is not recommended, and thus module-based gene rankings could be a potential alternative for the "choosing gene lists dilemma".

As an example, if we search for functional annotation using a gene list solely based on the top-100 DEG (AML vs Healthy), so a single-gene analysis, only eight biological functions are retrieved from Gene Ontology database:

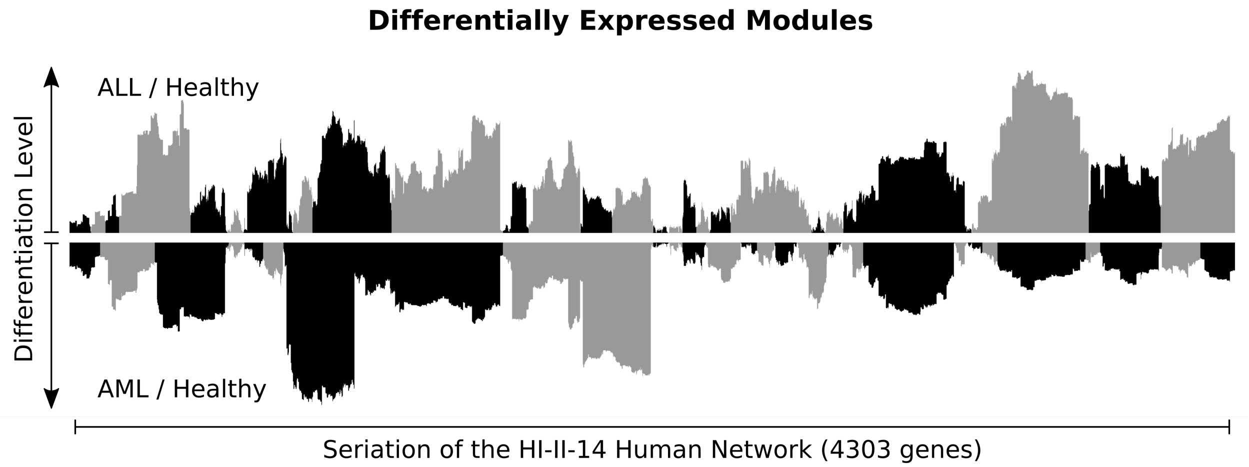

Fig. 3. Module-based differential expression profiles estimated from the *Claritate* seriation of the human network PPI (HI-II-14) over two groups of patients: i) ALL vs Healthy and ii) AML vs Healthy. The y-axis corresponds to the differentiation level of each gene between Healthy and Unhealthy samples.

- BP GO:0045930 - negative regulation of mitotic cell cycle (p-value $= 6.02 \times 10^{-4}$)
- BP GO:0000075 - cell cycle checkpoint $(6.02 \times 10^{-4})$
- BP GO:2000785 - regulation of autophagosome assembly $(6.75 \times 10^{-4})$
- BP GO:0044088 - regulation of vacuole organization $(8.75 \times 10^{-4})$
- BP GO:2000786 - positive regulation of autophagosome assembly $(8.75 \times 10^{-4})$
- MF GO:0005515 - protein binding $(2.44 \times 10^{-11})$
- MF GO:0005488 - binding $(1.00 \times 10^{-5})$
- MF GO:0016308 - 1-phosphatidylinositol-4-phosphate 5-kinase activity $(9.22 \times 10^{-4})$

Conversely, if we search for functional annotations using a gene list based on the top-100 DEG (AML vs Healthy), resulting from the top-4 modules (4 x 25 genes) showing higher differentiation, then seventeen biological functions are retrieved. Thus, two-fold more than the single-gene analysis:

- BP GO:0001776 - **leukocyte homeostasis** (p-value $= 7.51 \times 10^{-4}$)
- BP GO:0035821 - modification of morphology or physiology of other organism $(7.78 \times 10^{-4})$
- BP GO:0002513 - tolerance induction to self antigen $(7.78 \times 10^{-4})$
- BP GO:0002260 - **lymphocyte homeostasis** $(8.18 \times 10^{-4})$
- BP GO:0032945 - negative regulation of mononuclear cell proliferation $(8.18 \times 10^{-4})$
- BP GO:0050672 - negative regulation of lymphocyte proliferation $(8.18 \times 10^{-4})$
- BP GO:0001782 - **B cell homeostasis** $(8.18 \times 10^{-4})$
- BP GO:0070664 - negative regulation of leukocyte proliferation $(8.18 \times 10^{-4})$
- BP GO:0051817 - modification of morphology or physiology of other organism involved in symbiotic interaction $(8.18 \times 10^{-4})$
- BP GO:0018107 - peptidyl-threonine phosphorylation $(8.18 \times 10^{-4})$
- BP GO:0050869 - negative regulation of B cell activation $(8.18 \times 10^{-4})$
- BP GO:1901841 - regulation of high voltage-gated calcium channel activity $(8.18 \times 10^{-4})$
- BP GO:0010799 - regulation of peptidyl-threonine phosphorylation $(8.18 \times 10^{-4})$
- BP GO:0018210 - peptidyl-threonine modification $(8.18 \times 10^{-4})$
- BP GO:0007435 - salivary gland morphogenesis $(8.27 \times 10^{-4})$
- MF GO:0005515 - protein binding $(1.09 \times 10^{-16})$
- MF GO:0005488 - binding $(8.00 \times 10^{-7})$

In fact, the proposed strategy was able to highlight annotations (in bold) tightly related to the experimental study, being more sensitive to the biomedical context than the classical single-gene methods. Moreover, the seriation enriches the DEG result through an integrated procedure, exploring correlation between functional affinity (PPI network) and co-expression patterns (transcriptome).

We can say that the example above employs module-based DEG search locally. However, as metioned previously, it is also possible to enumerate a DEG list globally. A total of 1055 GO terms (BP: 737, CC: 210 and MF: 108) were retrieved by taking into account the same comparison group (AML vs Healthy). Next, we present the top-15 annotations (BP) in accordance to the highest module-based DEG profiles. Note that they are more general annotations which is indeed explained by the inclusion of lowest differentially expressed modules. Though, there is one annotation (in bold) that is closely related to the biomedical context.

- GO:0035556 - intracellular signal transduction
- GO:0007049 - cell cycle
- GO:0002376 - immune system process
- GO:0000209 - protein polyubiquitination
- GO:1901685 - glutathione derivative metabolic process
- GO:0016070 - RNA metabolic process

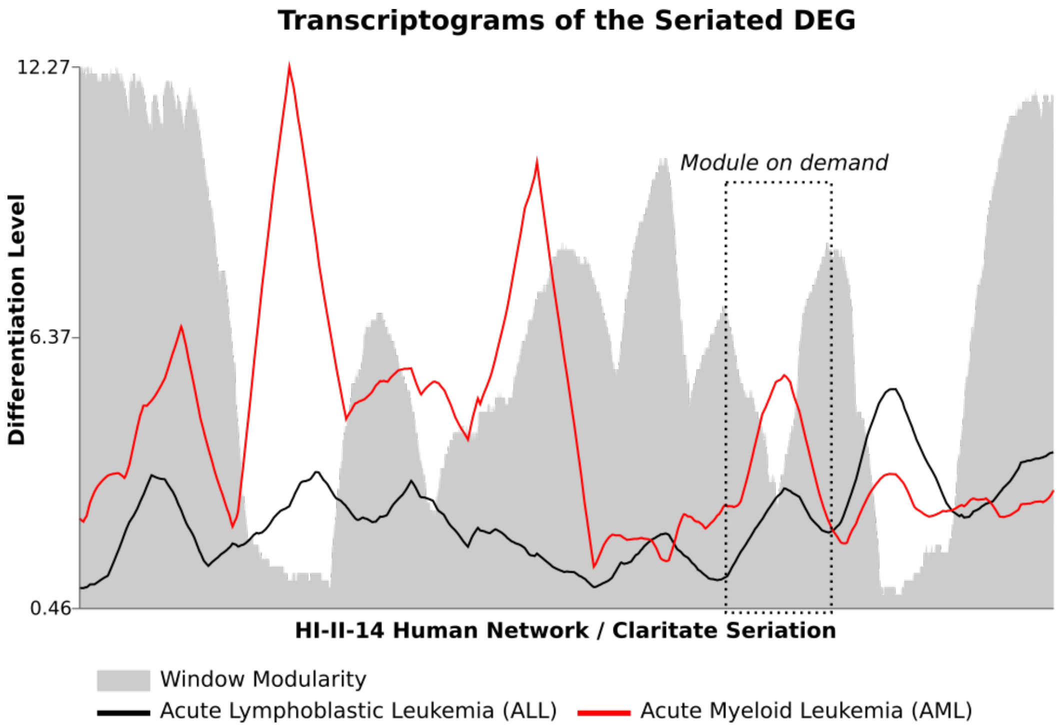

Fig. 4. Transcriptograms of the differential expression profiles estimated from the *Claritate* seriation of the human network PPI (HI-II-14) over two groups (patients) of comparison: i) ALL vs Healthy (black) and ii) AML vs Healthy (red). The background image (gray) is the window modularity. There is an example of module on demand, a highlighted gene group for meta-analisys comparison.

- GO:0010467 - gene expression
- GO:0015031 - protein transport
- GO:0044260 - cellular macromolecule metabolic process
- GO:0044403 - symbiosis, encompassing mutualism through parasitism
- GO:0090174 - organelle membrane fusion
- GO:0008380 - RNA splicing
- GO:0060370 - **susceptibility to T cell mediated cytotoxicity**
- GO:0051534 - negative regulation of NFAT protein import into nucleus
- GO:0030422 - production of siRNA involved in RNA interference

In [19], a new method for Gene Set Enrichment Analysis (GSEA) were introduced. The GSEA method relies on the idea of exploring groups of genes by the notiong of gene sets. Gene sets are defined based on prior biological knowledge, such as, published information about signaling pathways or co-expression in previous experiments. The main goal is to determine whether members of a gene set $S$ tend to occur toward the top (or bottom) of the gene list $L$, in which case the ranked gene list is correlated with the phenotypic class distiction, for instance (AML vs Healthy). GSEA does not explore functional association between proteins, for instance PPI networks. Taking again the same comparison group, a total of 225 GO terms were retrieve. We present bellow the most enriched 15 BP terms. As it can be observed, GSEA is less sensitive to the biomedical context than the proposed strategy.

- GO:0006508 - proteolysis
- GO:0007088 - regulation of mitotic nuclear division
- GO:0032940 - secretion by cell
- GO:0007610 - behavior
- GO:0044257 - cellular protein catabolic process
- GO:0006512 - obsolete ubiquitin cycle
- GO:0030163 - protein catabolic process
- GO:0051248 - negative regulation of protein metabolic process
- GO:0009628 - response to abiotic stimulus
- GO:0045045 - obsolete secretory pathway
- GO:0007626 - locomotory behavior
- GO:0043285 - biopolymer catabolic process
- GO:0045184 - establishment of protein localization
- GO:0009967 - positive regulation of signal transduction
- GO:0051641 - cellular localization

In Figure 4 it can be observed all module-based differential expression profiles for both comparison groups. Note that these experiments are of different phenotypic classes, and thus having distinct modules (size and variance). In order to make these experiments comparable, we must calculate its corresponding *Transcriptograms*. Note, there are some group of genes that are required together in both differention scenarios, but a more strong signal is highlighted on the AML patients (red curve). One my also be interested in the evaluation of module-based meta analysis of differential gene expression signatures by the inspection of “modules on demand” (Figure 4, dot-rectangle).

Finally, a cancer survival analysis using the top-100 DEG

(AML vs Healthy), resulting from the top-4 module-based differential expression profiles (4 x 25 genes), were evaluated though the PPISURV tool (http://bioprofiling.de). Interestingly, 37 genes were identified as positive to survival in diffuse large B cell lymphoma data set (GSE10846). As an example, the *UBE2R2* gene is directly associated to *UBE21* gene by PPI together with *DTX3L* gene (http://biograph.be/concept/graph/C1150669/C1421283). The *UBE21* is one the selected candidate endogenous control genes in normal hematopoietic cells on Leucégène RNA-seq data [6]. The causes of diffuse large B-cell lymphoma are not well understood. Usually it arises from normal B cells, but it can also represent a malignant transformation of other types of lymphoma or leukemia. There are also other positive and negative feedbacks of survival retrieved by PPISURV for other types of cancer (Breast: 15 genes and Lung: 21 genes). Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules [3].

All data and the sequence of commands of this analysis are available on https://github.com/joseflaviojr/transcriptograma/tree/master/UseCase-Leukemia.

## IV. Conclusion

In classical transcriptome data analysis genes are clustered either by their co-expression, or co-variation in time, assuming that these cluster definitions are closed linked to the stage the cell is going through or on the protocol used to assessed sample(s). *Transcriptograms* are independent of the profilling protocol and, this is due to the fact that seriation identifies network modules directly over the PPI to further explore gene expression profiles. Module-based differential expression profiles using *Transcriptograms* is a promising strategy, obtaining more specific gene signatures as well as functional annotations closely related to biomedical context. Moreover, cancer survival analysis over the discovered signatures are enriched to a positive class of genes.

## References

[1] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis, "Analysis of next-generation genomic data in cancer: accomplishments and challenges," *Human Molecular Genetics*, vol. 19, no. R2, pp. R188–R196, 2010. [Online]. Available: http://hmg.oxfordjournals.org/content/19/R2/R188.abstract

[2] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. [Online]. Available: http://dx.doi.org/10.1126/science.286.5439.509

[3] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007. [Online]. Available: http://www.pnas.org/content/104/21/8685.abstract

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8 – 17, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2001037014000464

[5] J. L. L. Rybarczyk-Filho, M. A. Castro, R. J. Dalmolin, J. C. Moreira, L. G. Brunnet, and R. M. de Almeida, "Towards a genome-wide transcriptogram: the Saccharomyces cerevisiae case." *Nucleic acids research*, vol. 39, no. 8, pp. 3005–3016, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq1269

[6] T. Macrae, T. Sargeant, S. Lemieux, J. Hébert, E. Deneault, and G. Sauvageau, "RNA-Seq Reveals Spliceosome and Proteasome Genes as Most Consistent Transcripts in Human Cancer Cells." *PloS one*, vol. 8, no. 9, pp. e72 884+, Sep. 2013. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0072884

[7] M. Vidal, M. E. Cusick, and A.-L. Barabási, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986 – 998, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867411001309

[8] P. J. Hernandez and T. Abel, "The role of protein synthesis in memory consolidation: progress amid decades of debate." *Neurobiology of learning and memory*, vol. 89, no. 3, pp. 293–311, Mar. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.nlm.2007.09.010

[9] S. hyung Yook, Z. N. Oltvai, and A. lászló Barabási, "Functional and topological characterization of protein interaction networks," *Proteomics*, vol. 4, pp. 928–942, 2004.

[10] F. Kuentzer, A. Pereira, A. Amory, G. Perrone, S. da Silva, J. Dinis, and R. de Almeida, "Optimization and analysis of seriation algorithm for ordering protein networks," in *Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on*, Nov 2014, pp. 231–237.

[11] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "STRING 8–a global view on proteins and their functional interactions in 630 organisms." *Nucleic acids research*, vol. 37, no. Database issue, pp. D412–D416, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn760

[12] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A.-R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J.-C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A.-L. Barabási, L. M. Iakoucheva, P. Aloy, J. D. L. Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal, "A proteome-scale map of the human interactome network," *Cell*, vol. 159, no. 5, pp. 1212 – 1226, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867414014226

[13] R. W. Floyd, "Algorithm 97: Shortest path," *Communications of the ACM*, vol. 5, no. 6, pp. 345–, Jun. 1962. [Online]. Available: http://doi.acm.org/10.1145/367766.368168

[14] S. E. Schaeffer, "Survey: Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cosrev.2007.05.001

[15] P. Hage, "Eccentricity and centrality in networks," *Social Networks*, vol. 17, no. 1, pp. 57–63, Jan. 1995. [Online]. Available: http://dx.doi.org/10.1016/0378-8733(94)00248-9

[16] A.-L. Boulesteix and M. Slawski, "Stability and aggregation of ranked gene lists," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 556–568, Sep. 2009. [Online]. Available: http://dx.doi.org/10.1093/bib/bbp034

[17] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, AmiGO Hub, and Web Presence Working Group, "AmiGO: online access to ontology and annotation data." *Bioinformatics (Oxford, England)*, vol. 25, no. 2, pp. 288–289, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btn615

[18] T. G. O. Consortium, "Gene ontology consortium: going forward," *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015. [Online]. Available: http://nar.oxfordjournals.org/content/43/D1/D1049.abstract

[19] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005. [Online]. Available: http://www.pnas.org/content/102/43/15545.abstract