

Defining a robust biological prior from Pathway Analysis to drive Network Inference.

Titre: Construction d'un a priori biologique robuste à partir de l'analyse de voies métaboliques pour l'inférence de réseaux.

Marine Jeanmougin^{1,2}, Mickael Guedj² and Christophe Ambroise¹

Abstract:

Inferring genetic networks from gene expression data is one of the most challenging work in the post-genomic era, partly due to the vast space of possible networks and the relatively small amount of data available. In this field, Gaussian Graphical Model (GGM) provides a convenient framework for the discovery of biological networks. In this paper, we propose an original approach for inferring gene regulation networks using a robust biological prior on their structure in order to limit the set of candidate networks.

Pathways, that represent biological knowledge on the regulatory networks, will be used as an informative prior knowledge to drive Network Inference. This approach is based on the selection of a relevant set of genes, called the "molecular signature", associated with a condition of interest (for instance, the genes involved in disease development). In this context, differential expression analysis is a well established strategy. However outcome signatures are often not consistent and show little overlap between studies. Thus, we will dedicate the first part of our work to the improvement of the standard process of biomarker identification to guarantee the robustness and reproducibility of the molecular signature.

Our approach enables to compare the networks inferred between two conditions of interest (for instance case and control networks) and help along the biological interpretation of results. Thus it allows to identify differential regulations that occur in these conditions. We illustrate the proposed approach by applying our method to a study of breast cancer's response to treatment.

Résumé :

L'inférence de réseaux génétiques à partir de données issues de biopuces est un des défis majeurs de l'ère post-génomique, en partie à cause du grand nombre de réseaux possibles et de la quantité relativement faible de données disponibles. Dans ce contexte, la théorie des modèles graphiques gaussiens est un outil efficace pour la reconstruction de réseaux.

A travers ce travail nous proposons une approche d'inférence de réseaux de régulation à partir d'un *a priori* biologique robuste sur la structure des réseaux afin de limiter le nombre de candidats possibles.

Les voies métaboliques, qui rendent compte des connaissances biologiques des réseaux de régulation, nous permettent de définir cet *a priori*. Cette approche est basée sur la sélection d'un ensemble de gènes pertinents, appelé "signature moléculaire", potentiellement associé à un phénotype d'intérêt (par exemple les gènes impliqués dans le développement d'une pathologie). Dans ce contexte, l'analyse différentielle est la stratégie prédominante. Néanmoins les signatures de gènes diffèrent d'une étude à l'autre et la robustesse de telles approches peut être remise en question. Ainsi, la première partie de notre travail consistera en l'amélioration de la stratégie d'identification des gènes les plus

1. Statistics and Genome laboratory UMR CNRS 8071, University of Evry, Evry, France.

E-mail: marine.jeanmougin@genopole.cnrs.fr

2. Department of Biostatistics, Pharnext, Paris, France.

E-mail: mickael.guedj@pharnext.com and E-mail: Christophe.Ambroise@genopole.cnrs.fr

informatifs afin de garantir la robustesse et la reproductibilité de la signature moléculaire.

Notre approche vise à comparer les réseaux inférés dans différentes conditions d'étude et à faciliter l'interprétation biologique des résultats. Ainsi, elle permet de mettre en avant des régulations différentielles entre ces conditions. Nous appliquerons notre méthode à l'étude de la réponse au traitement dans le cancer du sein.

Keywords: Network Inference, Gaussian Graphical Model, ℓ_1 penalization, Prior information, Pathway Analysis

Mots-clés : Inférence de réseaux, Modèle graphique gaussien, Pénalisation ℓ_1 , Information a priori, Analyse de voies métaboliques

AMS 2000 subject classifications: 60K35,

1. Introduction

Advances in Molecular Biology and substantial improvements in microarray technology have led biologists toward high-throughput genomic studies. It has become possible to detect tens of thousands of genes and compare their expression levels between samples in a single experiment. These data describe dynamic changes in gene expression closely related to regulatory events without offering any explanation on how this is all managed by the genome. The use of microarrays to discover differentially expressed genes between two or more conditions (patients *versus* controls for instance) has found many applications. These include the identification of disease biomarkers (i.e. the genes potentially associated with a given disease) that may be important in the diagnosis of the different types or subtypes of diseases. A common challenge faced by the researchers is to translate the identified sets of differentially expressed genes into a better understanding of the underlying biological phenomena.

A wealth of analysis tools are available among which Pathway Analysis and Network Inference are getting especially popular in the field of gene expression data analysis. They both hold great promise to interpret genome-wide profile.

Pathways can be defined as sets of gene products interacting in order to achieve a specific cellular function (cell cycle or apoptosis for instance). Pathway Analysis aims at determining whether the set of differentially expressed genes is “enriched” by a given pathway or cellular function. Pathway Analysis approaches have the advantage of providing a clear and direct biological interpretation. However it can appear restrictive as pathways represent a strong biological prior and results will not deviate from our current knowledge of the cellular mechanisms.

On the other side, the purpose of Network Inference is to predict the presence or absence of edges between a set of genes known to form the vertices of a network, the prediction being based on gene expression levels. Various statistical frameworks have been proposed to solve the network inference problem [7, 33]. Graphical models such as Bayesian network [9] and constraint-based methods are popular frameworks to model the gene regulatory networks. Modelings with the Boolean networks [14, 17, 25, 30] or differential equations have also been investigated. Discovering structures of regulation networks based on large-scale data represents a major challenge in Systems Biology, partly because the space of possible networks is often too large compared to the relatively small amount of data available. Moreover, exploring all possible network topologies is not computationally feasible. However, all the networks are not equally

plausible and the use of biological knowledge may help to limit the set of candidate networks. Little research have already been made to associate biological informative prior to Network Inference in a Bayesian network framework [19] or by the use of dynamic Bayesian networks [3] from time series data. In the field of Machine Learning other kinds of approaches have been proposed based on kernel metric learning [32, 35].

In this paper, we propose a 3-step approach to define a biological prior on network structure to drive the Network Inference. Our intention here is to overcome both Pathway Analysis and Network Inference limitations by integrating these two approaches.

We will dedicate the first part of our work to the improvement of the standard process of biomarker identification to guarantee the robustness and reproducibility of the gene signature. Indeed due to the biological and technical variability within the expression data, robust gene selection is a crucial issue. Most informative genes have to be identified to ensure the inference of a relevant network.

The second step consists in correlating the signature with a priori defined gene sets from biological pathway databases. The resulting metabolic and signaling pathways will be used as an informative prior in the last step.

Finally, a network is inferred with the idea to encourage the inference of an edge between two genes when they are known to participate to the same cellular function. A way to describe the interactions among the genome functional elements is by using conditional independencies and, more concretely, graphical models [16, 23, 34] which provide a convenient setting for inferring gene regulatory networks. We base our approach on this strategy implemented in the R package SIMoNe (Statistical Inference for Modular Network) [5] which enables inference of undirected networks based on partial correlation. In this approach, Chiquet et al presents the first method dedicated to the inference of multiple graphs. In other words, when various experimental conditions are available, they propose to estimate multiple GGMs jointly, in a multi-task framework by coupling the estimation of several networks.

We conduct our global approach on breast cancer data to identify the molecular mechanisms underlying the response to treatment using the cohort from Hess et al. 2006 [11] by comparing the networks inferred under 2 conditions: patients who achieve pathological complete response (pCR) and those who do not (not-pCR).

2. Materials and Methods

The purpose of this work is to provide a global framework for transcriptomic data analysis and in particular for inferring co-expression networks from a relevant biological informative prior based on Pathway Analysis. The co-expression networks are seen as regulation networks for which the direction of the edges and hence causality events are not known. Such analyses require as a preliminary step to select the genes involved in the phenotype of interest that are assumed to be differentially expressed between the conditions to compare. This selection is of great importance since the results of Pathway Analysis and Network Inference directly depend upon it.

Hence, our global approach can be viewed as a 3-step process: i) an improved selection of

differentially expressed genes ii) a Pathway Analysis and iii) a Network Inference strategy using Pathway Analysis results as an informative biological prior. These 3 steps are fully described below and summarized in Figure 1.

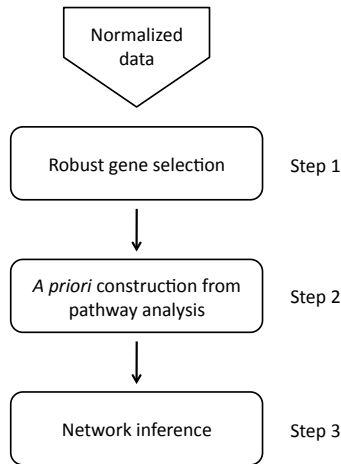


Figure 1. Global approach of expression data analysis. Our approach follows a 3-step process: from the expression data we first perform a robust selection to identify the most informative genes, called the signature. Then, we define the biological prior by conducting a Pathway Analysis on the gene signature. The last step consists in inferring a network driven by the prior defined at the previous step.

Let us first introduce some notations. We consider hereafter a sample of a random vectors $X_{cr} = (X_{cr}^1, \dots, X_{cr}^p)$ where X_{cr}^i is the level of expression observed for gene i , replicate r , under condition c ; The vector X_{cr} is assumed to be Gaussian with positive definite covariance matrix $\Sigma^c = (\Sigma_{ij}^c)_{(i,j) \in \mathcal{P}^2}$ where $\mathcal{P} = \{1, \dots, p\}$. No loss of generality is involved when centering X_{cr} , so we may assume that $X_{cr} \sim \mathcal{N}(\mathbf{0}_p, \Sigma^c)$.

Step1: Gene selection by robust differential analysis

Given two conditions to compare (called condition 1 et condition 2 in the following), a key question in the first step of expression data analysis is to identify a set of genes that shows a differential expression between the two conditions of interest. This consists, for each gene, in testing the null hypothesis (H_0) that the expected values of expression are equal between the two conditions, against the alternative hypothesis (H_1) that they differ. The general model is then given by:

$$\mathbb{E}(X_{cr}^i) = \mu_c^i \quad \text{and} \quad \text{Var}(X_{cr}^i) = (\sigma_c^i)^2.$$

So defined, the null hypothesis to test comes down to:

$$\begin{cases} H_0 : & \mu_1^i = \mu_2^i, \\ H_1 : & \mu_1^i \neq \mu_2^i. \end{cases}$$

At first glance, differential analysis appears easy enough but in practice results from differential analysis show a lack of reproducibility when compared in the literature [8]. In fact, due to the biological and technical noise inherent to high-throughput biological data, methodological aspects are still to investigate to obtain a more robust selection of differentially expressed genes. To do so we propose several methodological improvements to the strategy commonly used, that is generally reduced to the application of a t -test.

If the t -test is certainly the most natural and popular test to assess differential expression, its efficacy in term of variance modeling and power on small sample sizes has been seriously questioned. The recent study from Jeanmougin et al. [12] demonstrates that the Bayesian framework proposed by `limma` [26] outperforms all the usual methods to test for differential expression. Briefly, `limma` has the same interpretation as an ordinary t -statistic except that the standard errors have been moderated across genes:

$$t_{\text{limma}}^i = \frac{\bar{x}_1^i - \bar{x}_2^i}{S^i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where n_1 and n_2 are number of replicates of conditions 1 and 2 and \bar{x}_1^i is the average expression level for gene i and condition 1 (across all possible replicates).

Indeed, posterior variance, S^i , has been substituted into the classical t -statistic in place of the usual variance. Using Bayes rules, this posterior variance becomes a combination of an estimate obtained from a prior distribution and the pooled variance. Including a prior distribution on variances has the effect of borrowing information from the ensemble of the genes to aid with inference about each individual gene. Thus, the posterior values shrink the observed variances towards a common value, that is why it is called “moderated” t -statistic. This approach is implemented in the R package `limma`.

Despite the use of `limma`, such a hypothesis testing strategy to select differentially expressed genes between 2 conditions can be biased by outliers. In this context, re-sampling based approaches are pretty common methods. Here, we use random forest [4] for improving the overall homogeneity of the signature obtained from differential analysis. The algorithm consists in fitting many binary decision trees built using several bootstrap samples and then in combining the predictions from all the trees. Random forest uses both bagging and random variable selection for tree building. Thus, at each node, a given number of input variables are randomly chosen and the best split is calculated only within this subset. The Out-Of-Bag sample (i.e. the set of observations which are not used for building the current tree) is used to estimate the prediction error and then to evaluate variable importance.

We apply Random Forest algorithm (implemented in the R package `randomForest`) to rank the significant genes resulting from `limma` according to variable importance (i.e ability to well discriminate the conditions). This approach enables to clean the set of differentially expressed genes from putative outliers by selecting most important variables. A simple Principal Component Analysis of the samples before and after the application of the Random Forest algorithm clearly show how it increases the separation of the conditions (Figure 2).

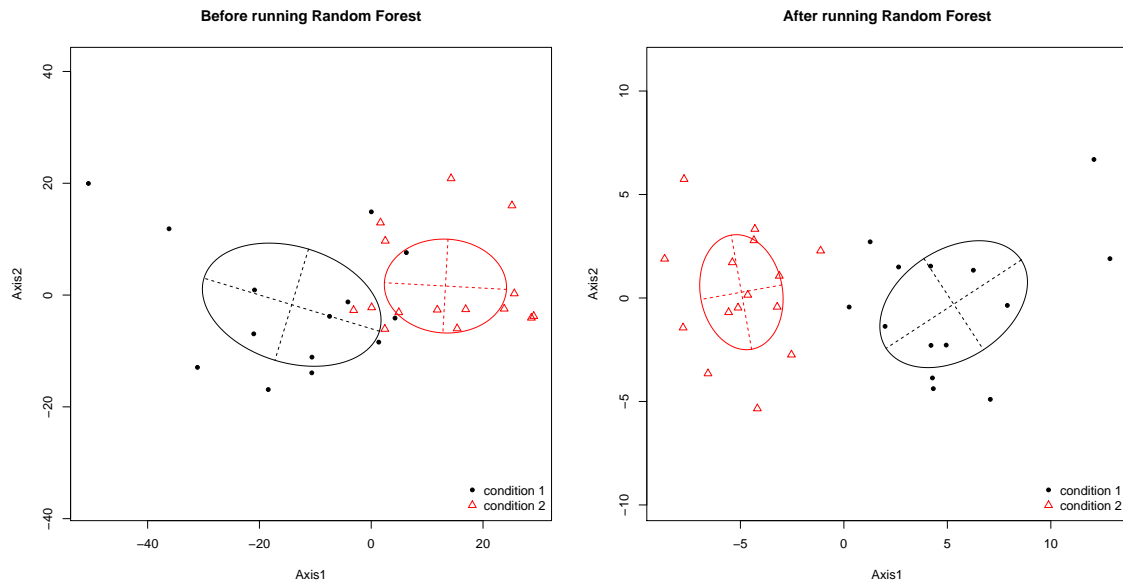


Figure 2. Improving discrimination between conditions using Random Forest. A Principal Component Analysis is performed on conditions of interest. The signature obtained by Random Forest (RF) algorithm enables to better discriminate condition 1 from condition 2 as shown in the figures before and after running the RF.

Finally little agreement between studies in the literature is also due to the heterogeneity inherent to genetic diseases and to sampling variation for moderate sample size. Thus, two differential analyses based on two different studies may lead to different selection of genes, however resulting from the same regulatory process. Motivated by the observation that genes causing the same phenotype are likely to be functionally related [10, 15] (i.e they form some kind of module [21], for instance a multi-protein complex), we explore a network-based approach to identify modules of genes rather than individual genes. We will refer to this approach as “functional partners identification” in the following. One of the strongest manifestation of a functional relation between genes is Protein-Protein Interaction (PPI). Oti et al. [22] demonstrated that exploiting PPIs can lead to novel candidate gene prediction. While many PPI network databases are available, such as Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com) or STRING [27], few studies have actually integrated such knowledge for biomarkers identification. The PPI network was extracted using STRING. The interactions include direct (physical) and indirect (functional) associations derived from various sources on the basis of both experimental evidence for PPIs as well as interactions predicted by comparative genomics and text mining.

STRING uses a scoring system that is intended to reflect the evidence of predicted interactions. In the present study, we add the relevant functional partners with a score of at least 0.9 (which corresponds to a high-confidence network) to the set of differentially expressed genes.

In the following, our final set of genes will be called the "molecular signature" of the conditions

to compare.

Step2: Pathway Analysis to define a robust a priori

Pathway Analysis aims at identifying the cellular processes involved in differentiating the two conditions to compare (cases and controls for instance) in order to enhance the interpretation of the molecular signature generated at step 1.

In practice, Pathway Analysis consists in determining whether the signature is enriched in pathway key actors. Given the M genes measured on a microarray, our signature is defined as a subset of length n of these genes, and a given pathway is defined as another subset of length K of these genes. The probability of having y genes in common between our signature and a pathway is given by the hypergeometric distribution:

$$\mathbb{P}(Y = y) = \frac{\binom{n}{y} \binom{M-n}{K-y}}{\binom{M}{K}}.$$

This approach requires a pre-defined set of pathways to analyze. Several databases exist today and the Kyoto Encyclopedia of Genes and Genomes (KEGG [13]) is often taken as a reference.

The problem with the analysis of pathways is that they do not represent clearly distinct entities and two pathways can involve common genes and hence share a common biological information. Consequently a set of genes can be responsible for the positive results of several pathways. In order to ease the interpretation of Pathway Analysis results, we propose to summarize the set of pathways found significant because of the same genes into a reduced set of "core pathways" (each core pathway represented by a set of pathways). In practice we apply a Hierarchical Clustering algorithm on a binary matrix, that contains the genes in row and the pathways in column, summarizing the membership of each gene to each pathway. Dissimilarity, which accounts for pairwise differences between the pathways, is assessed by using a binary metric, also known as the Jaccard distance. Finally, from this dissimilarity matrix we perform a Hierarchical Agglomerative Clustering using Ward's criterion. Each core pathway defines a gene cluster which will be used in the next step for inferring the regulation network.

Step 3: Network Inference from biological prior knowledge

The last step of the procedure consists in inferring a part of the regulation network using the transcriptomic data from all available conditions and from the prior structure exhibited in the pathway analysis. This last step relies on a Gaussian Graphical Model, taking advantage of the relation existing between conditional independence and the concentration matrix. Variables (i.e. genes) X^i and X^j with $i \neq j$ are independent conditional on all other variables indexed by $\mathcal{P} \setminus \{i, j\}$, if and only if the entry $(\Sigma^{-1})_{ij}$ is zero. The inverse of the covariance matrix $\mathbf{K} = (K_{ij})_{(i,j) \in \mathcal{P}^2} = \Sigma^{-1}$, known as the concentration matrix, thus describes the conditional

independence structure of X . Moreover, each entry K_{ij} , is directly linked to the partial correlation coefficient $r_{ij|\mathcal{P}\setminus\{i,j\}}$ between variables X^i and X^j . The inferred graph is described by the adjacency matrix defined by the nonzero elements of the concentration matrix.

Merging different experimental conditions from transcriptomic data is a common practice in GGM-based inference methods [31]. This process enlarges the number of observations available for inferring interactions. However, GGMs assume that the observed data form an independent and identically distributed (i.i.d.) sample. But assuming that the merged data is drawn from a single Gaussian component is obviously wrong, and is likely to have detrimental side effects in the estimation process.

In this paper, we assume that the distributions of these different conditions have strong commonalities. We thus propose to estimate these graphs jointly. In this context, the multi-task learning provides a convenient framework [5]. This approach allows to overcome the difficulties arising from the scarcity of data in each experimental condition by coupling the estimation problems.

Moreover we take advantage of the previously exhibited gene clusters to put some prior structure on the inferred network. We assume that the edges of the graph defined by the nonzero elements of the concentration matrix are distributed among a set $\mathcal{Q} = \{1, \dots, Q\}$ of given overlapping clusters. For any gene i , the indicator variable Z_{iq} is equal to 1 if $i \in q$ and 0 otherwise, hence describing to which cluster the gene i belongs. The structure of the graph is thus described by the matrix $\mathbf{Z} = (Z_{iq})_{i \in \mathcal{P}, q \in \mathcal{Q}}$.

Thus the group structure over the Q gene clusters and the structure over the 2 conditions lead to estimating the 2 concentration matrices which are the solutions of the following penalized log-likelihood maximization problem:

$$\max_{\mathbf{K}^{(c)}, c \in 1, 2} \sum_{c=1}^2 \mathcal{L}(\mathbf{K}^{(c)}; \mathbf{X}) - \lambda \sum_{\substack{i, j \in \mathcal{P} \\ i \neq j}} \rho_{\mathbf{Z}_i \mathbf{Z}_j} \left\{ \left(\sum_{c=1}^2 [K_{ij}^{(c)}]_+^2 \right)^{1/2} + \left(\sum_{c=1}^2 [K_{ij}^{(c)}]_-^2 \right)^{1/2} \right\}, \quad (1)$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$ and the coefficients of the penalty are defined as:

$$\rho_{\mathbf{Z}_i \mathbf{Z}_j} = \begin{cases} \sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{in}}}, & \text{if } i \neq j, \text{ and } q = \ell, \\ \sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{out}}}, & \text{if } i \neq j, \text{ and } q \neq \ell, \\ 1, & \text{otherwise,} \end{cases} \quad (2)$$

and

$$\mathcal{L}(\mathbf{K}^{(c)}; \mathbf{X}) = \frac{n}{2} \log \det(\mathbf{K}^{(c)}) - \frac{n}{2} \text{Tr}(\mathbf{S}^{(c)} \mathbf{K}) + C,$$

with $\mathbf{S}^{(c)}$ the empirical covariance matrix relative to condition c and C a constant term.

The first part of the criterion $\sum_{c=1}^2 \mathcal{L}(\mathbf{K}^{(c)}; \mathbf{X})$ consists in the sum of the log-likelihoods of the two concentration matrices given the observations. The second part of the criterion is a penalty, which considers two types of edges: edges between two genes belonging to the same core pathway

are penalized with a coefficient $1/\lambda_{in}$ and edges between two genes which are never present together in a pathway are penalized with a coefficient $1/\lambda_{out}$.

The basic form of the penalty has been proposed in [2] where it has been shown via simulations that the knowledge of an existing group structure was indeed improving the estimation of the non-zeros entries of the concentration matrix. The current form of the penalty [5] encourages co-regulation of the same sign across the two conditions. Specifically, co-regulation encompasses up-regulation and down-regulation and the type of regulation is not likely to be inverted across assays: in terms of partial correlations, sign swaps are very unlikely.

3. Application

We apply our approach to the study of the response to chemotherapy in breast cancers. The pathologic Complete Response (pCR), defined as the absence of disease in both the breast and lymph nodes, is used as an early surrogate marker of treatment efficacy. Most studies on response to chemotherapy have considered breast cancer as a single homogeneous entity. However, it is a complex disease that has a molecular and cellular heterogeneity that should not be overlooked in such studies. Perou et al. [24] were the first to report evidence for breast cancer tumor subtypes based upon gene expression profiles. Since the various tumor subgroups were re-evaluated and breast cancers have been classified into five molecular classes: luminal A and B, basal, ErbB2+, and normal-like subgroups [28, 29]. These subtypes differ markedly in prognosis and in the repertoire of therapeutic targets they express [20]. Within these five groups, basal-like breast cancers, quickly became a subtype of interest because of its features. Indeed it is associated with poor prognosis and characterized by its aggressive behavior.

Many reports have demonstrated that basal-like tumors are chemo-sensitive. This particular subtype of breast cancer correlated with significantly higher pCR rates when compared to other subtypes [18]. Nevertheless, basal tumors harboring homogeneous clinical and pathologic features may exhibit highly variable response to chemotherapy. It raises new questions to biologists such as how to explain the differences between patients who achieve a pCR from those that do not (not-pCR), and which are the molecular mechanisms involved in chemo-sensitivity or resistance ?

We demonstrate that our approach based on Pathway Analysis and Network Inference can provides effective tools to address such questions of biological and clinical interests. To do so, we investigate the dataset proposed by Hess et al. [11] including 133 breast cancer samples for which the pCR status is known. Gene expression profiling has been made based on the Affymetrix U133A microarray. Following the methodology described in Sorlie et al. [28], we identify 29 basal tumors from the 133 samples, divided into 15 pCR and 14 not-pCR.

Step 1 - Comparison of the gene expression profiles between the pCR and not-pCR samples yields about 100 genes with statistically significant differences (at a 10^{-3} level). We add 30 key regulatory genes identified from the STRING database. These 130 genes represent the molecular signature of the pCR in our 29 basal tumors.

Step 2 - The Pathway Analysis driven from the molecular signature identified 22 significant

KEGG pathways (at a 5% level). The Hierarchical Clustering of these pathways lead to 6 core pathways that summarize 6 main, distinct and clear biological processes involved in the mechanisms of response to treatment in basal tumors (Figure 3). For instance, one core pathway includes the VEGF, mTOR and IL8 signaling pathways, all associated with angiogenesis, the new blood vessel formation process required for tumor progression. Consequently, such results lead us to relate this core pathway to angiogenesis.

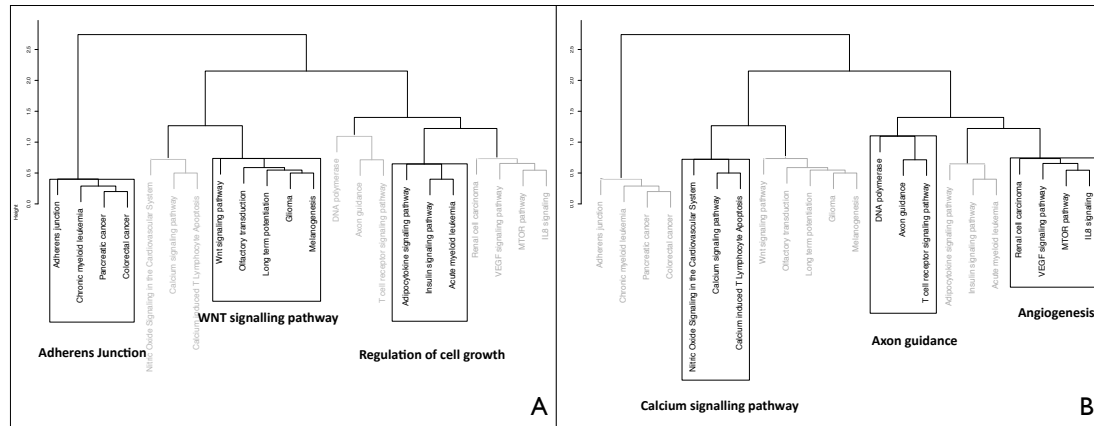


Figure 3. Core pathways. From the Hierarchical Clustering we identify 6 core pathways. Figure 3-A highlights the core pathways related to tumor cell growth and proliferation mechanism and Figure 3-B highlights the core pathways associated with angiogenesis activity.

Step 3 - The 6 core pathways are taken as a biological prior for driving the Network Inference using the R package SIMoNe. A particularly interesting network subpart is shown on Figure 4, which includes many genes known to be associated with breast cancer (for instance: mTOR, WNT, VEGF, AKT1...).

In particular, it has been shown that the CALM3 gene regulates the activity of AKT1 in breast tumors [6]. Our network suggests that this regulation occurs only in the pCR condition, and hence, may be broken for not-pCR tumors.

4. Discussion and conclusion

The ability to infer networks from high-throughput genomic data is one of the most promising applications of Systems Biology. In this paper, we propose a global framework to infer networks on the basis of a biological informative prior over network structures. It has the advantage to reduce the space of possible network structures to investigate, and to propose a more relevant network in order to facilitate the interpretation of its edges. Network Inference is done using the SIMoNe algorithm developed by Chiquet et al. [5] and based on a weighted Lasso criterion in a Gaussian Graphical Model framework.

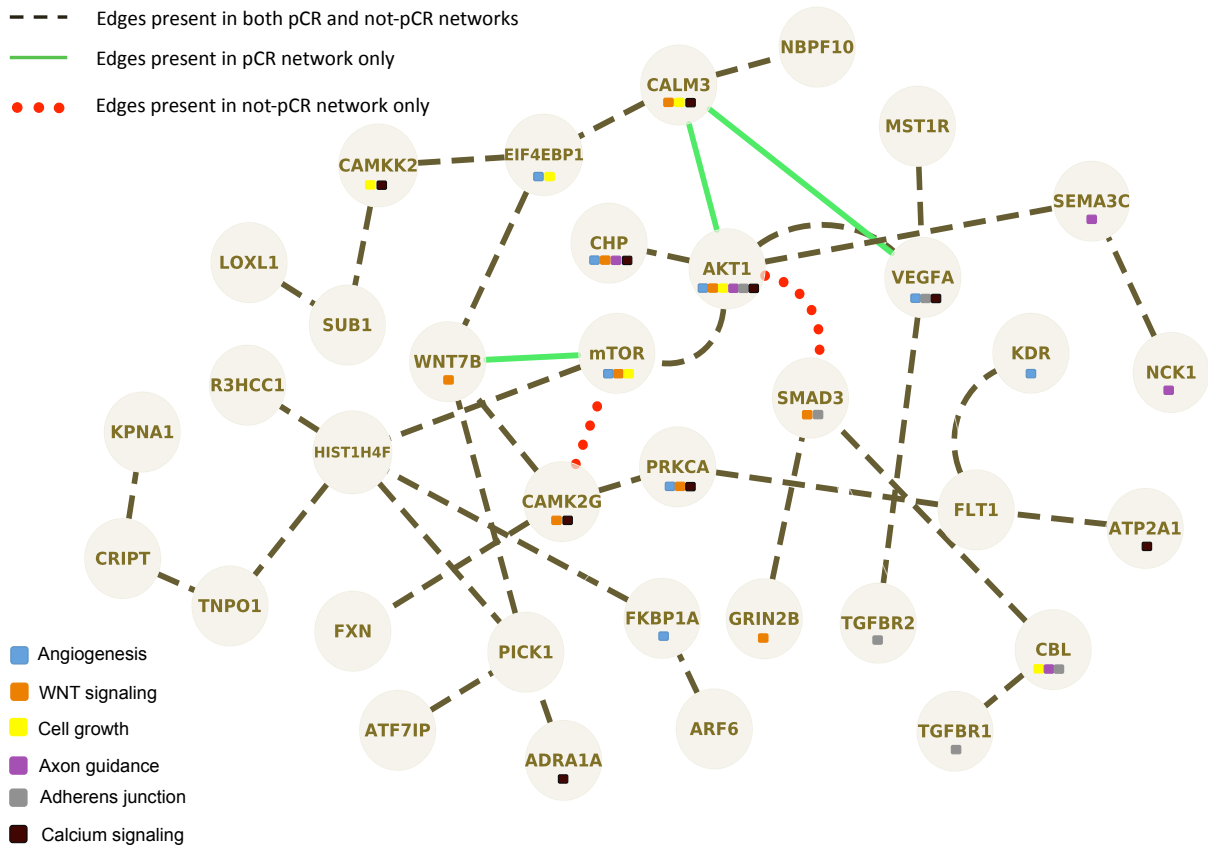


Figure 4. Inferred graph from the Hess cohort. The figure displays an enlarged view of a network subpart. The black dashed edges are present in both pCR and not-pCR networks. In red are the edges inferred only in the not-pCR network and conversely, the green edges are only inferred in the pCR network.

In this context, determining the biological prior to use is not a simple task. To this purpose, we define an informative prior using Pathway Analysis that allows, from molecular signature, to identify which are the main cellular mechanisms associated with a condition of interest. Many public domain repositories exist for storing biological pathways, each based on its own set of conventions. Thus pathways are not consistent among databases and the same pathway can present several definitions that vary greatly in quality and completeness [1]. Pathway Analysis is hence database dependent. However our approach overcomes this limitation in two points: i) the definition of core pathways from the molecular signature makes our analysis less sensitive to the strict definition of pathways found in the databases and ii) it is possible to adjust the weighted penalization according to the confidence we have in each pathway definition.

The core of such an approach, combining both Pathway Analysis and Network Inference, is the selection of the most informative set of genes to discriminate between the two conditions

(i.e. the molecular signature). The relevance of the biological findings at the end of the analysis greatly depends on this signature. In consequence, as a first step of our global Network Inference process, we propose an improved Differential Analysis. In particular, we aim to increase the reproductibility of the signature by removing potential outliers and identifying modules of genes via functional partners identification. It allows to include some well-studied genes (such AKT1) not detected in the differential analysis. However the noise inherent to PPI data is one of the main limitations of this approach. Indeed, only few protein interactions are confirmed by the HPRD (Human Protein Reference Database). These high noise levels reduce the accuracy of functional partners identification. By restricting our study to high-confidence score interactions, we focused on the most biologically relevant interactions. However it introduces a bias into our approach by selecting the most studied and documented proteins. Then, the functional partners identification step is highly dependent on the quality and the amount of data available. The functional partners identification method still needs to be further improved and explored. In particular, we have to develop an evaluation strategy to better assess the contribution of such a method in identifying genes of interest.

Our application on breast cancer data highlights key regulations in cancer progression and response to treatment. The co-expression of AKT1 and CALM3 that occurs specifically in the group responding to chemotherapy is a good example. Such results can be a starting point of more important applications at the biological and clinical levels.

Finally, the regulation of gene expression is a complex process resulting from several regulatory mechanisms occurring at distinct steps of the biological system (miRNA, eQTL, genomic alterations and epigenetic factors for instance) and gene expression data enables us to understand a limited part of the whole system. In this context, the inference of networks from heterogeneous biological data, such as eQTL, is a very promising topic to develop in the near future.

Acknowledgement

We thank Caroline Paccard, Matthieu Bouaziz, Fabrice Glibert, Julien Chiquet, Carene Rizzon, Claudine Devauchelle, Camille Charbonnier and Serguei Nabirotkin for helpful discussions. We also thank Celeste Lebbe and Ilya Chumakov for their support.

Author Contributions

Conceived and implemented the global process: MJ. Wrote the paper: MJ, MG and CA. Co-supervised the study: MG and CA.

References

- [1] Michiel E. Adriaens, Magali Jaillard, Andra Waagmeester, Susan L.M. Coort, Alex R. Pico, and Chris T.A. Evelo. The public road to high-quality curated biological pathways. *Drug Discovery Today*, 13, 2008.
- [2] Christophe Ambroise, Julien Chiquet, and Catherine Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.

- [3] Allister Bernard and Alexander J Hartemink. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput*, pages 459–470, 2005.
- [4] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 10.1023/A:1010933404324.
- [5] Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graph structures. *Statistics and Computing*, 2010.
- [6] Christine M Coticchia, Chetana M Revankar, Tushar B Deb, Robert B Dickson, and Michael D Johnson. Calmodulin modulates akt activity in human breast cancer cell lines. *Breast Cancer Res Treat*, 115(3):545–560, Jun 2009.
- [7] Hidde de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103, 2002.
- [8] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.
- [9] Nir Friedman, Michal Linial, Itach Nachman, and Dana Pe’er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620, 2000.
- [10] T. K B Gandhi, Jun Zhong, Suresh Mathivanan, L. Karthick, K. N. Chandrika, S. Sujatha Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D Boeke, Giovanni Parmigiani, Jörg Schultz, Joel S Bader, and Akhilesh Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, Mar 2006.
- [11] Kenneth R Hess, Keith Anderson, W. Fraser Symmans, Vicente Valero, Nuha Ibrahim, Jaime A Mejia, Daniel Booser, Richard L Theriault, Aman U Buzdar, Peter J Dempsey, Roman Rouzier, Nour Sneige, Jeffrey S Ross, Tatiana Vidaurre, Henry L Gómez, Gabriel N Hortobagyi, and Lajos Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol*, 24(26):4236–4244, Sep 2006.
- [12] Marine Jeanmougin, Aurelien de Reynies, Laetitia Marisa, Caroline Paccard, Gregory Nuel, and Mickael Guedj. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE*, 5(9):e12336, 09 2010.
- [13] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006.
- [14] Stuart A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, March 1969.
- [15] Kasper Lage, E. Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, Yves Moreau, and Søren Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–316, Mar 2007.
- [16] Steffen L. Lauritzen. *Graphical models*. Clarendon Press, 1996.
- [17] Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures, 1998.
- [18] Cornelia Liedtke, Chafika Mazouni, Kenneth R Hess, Fabrice André, Attila Tordai, Jaime A Mejia, W. Fraser Symmans, Ana M Gonzalez-Angulo, Bryan Hennessy, Marjorie Green, Massimo Cristofanilli, Gabriel N Hortobagyi, and Lajos Pusztai. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol*, 26(8):1275–1281, Mar 2008.
- [19] Sach Mukherjee and Terence P Speed. Network inference using informative priors. *Proc Natl Acad Sci U S A*, 105(38):14313–14318, Sep 2008.
- [20] Torsten O Nielsen, Forrest D Hsu, Kristin Jensen, Maggie Cheang, Gamze Karaca, Zhiyuan Hu, Tina Hernandez-Boussard, Chad Livasy, Dave Cowan, Lynn Dressler, Lars A Akslen, Joseph Ragaz, Allen M Gown, C. Blake Gilks, Matt van de Rijn, and Charles M Perou. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res*, 10(16):5367–5374, Aug 2004.
- [21] M. Oti and H. G. Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, Jan 2007.
- [22] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698, Aug 2006.

- [23] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, September 1997.
- [24] Charles M. Perou, Therese Sorlie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Christian A. Rees, Jonathan R. Pollack, Douglas T. Ross, Hilde Johnsen, Lars A. Akslen, Oystein Fluge, Alexander Pergamenschikov and Cheryl Williams, Shirley X. Zhu, Per E. Lonning, Anne-Lise Borresen-Dale, Patrick O. Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000.
- [25] Elisabeth Remy and Paul Ruet. From minimal signed circuits to the dynamics of Boolean regulatory networks. *Bioinformatics*, 24(16):i220–i226, 2008.
- [26] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.
- [27] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, 28(18):3442–3444, Sep 2000.
- [28] Therese Sorlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M Perou, Per E Lønning, Patrick O Brown, Anne-Lise Børresen-Dale, and David Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–8423, Jul 2003.
- [29] Christos Sotiriou, Soek-Ying Neo, Lisa M. McShane, Edward L. Korn, Philip M. Long, Amir Jazaeri, Philippe Martiat, Steve B. Fox, Adrian L. Harris, and Edison T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America*, 100(18):10393–10398, 2003.
- [30] René Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563 – 585, 1973.
- [31] Hiroyuki Toh and Katsuhisa Horimoto. Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics*, 18:287–297, 2002.
- [32] Jean-Philippe Vert and Yoshihiro Yamanishi. Supervised graph inference. *Advances in Neural Information Processing Systems*, pages 1433–1440, 2005.
- [33] Adriano V. Werhli, Marco Grzegorzczak, and Dirk Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.
- [34] Joe Whittaker. *Graphical Models in Applied Multivariate Statistics (Wiley Series in Probability & Statistics)*. John Wiley & Sons, March 1990.
- [35] Yoshihiro Yamanishi, Jean-Philippe Vert, and Minoru Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20 Suppl 1:i363–i370, Aug 2004.