TECHNICAL UNIVERSITY OF CRETE, DEPARTMENT OF ELECTRONIC AND COMPUTER ENGINEERING

Graphical Models in Genomic Networks

Application for Genomic Data Analysis

Kalliopi Kalantzaki,
Prof. M. Zervakis (Supervisor)
Prof. M. Garofalakis
Prof. E. Petrakis

1. CONTENTS

2. ABSTRACT

During the past few years there has been an increasing interest in studying the underlying genetic/proteomic mechanisms behind the discovery of genetic interactions between molecules. This kind of knowledge is of great importance in many scientific areas such as clinical prognosis, diagnosis and treatment. In this context, various methodological approaches have been suggested for the analysis of genetic interactions in terms of predicting the genetic/proteomic associations and for modeling the dependencies among the studied molecules.

This work explores two distinct aspects. In the first approach we estimate the genetic interactions between sets of genes/proteins of interest in which we use two methodologies; the first is a standard technique that relies on partial correlations (PC) while the second is a proposed algorithm based on kernel density estimation (KDE). We compare the two approaches and we highlight KDE as a useful approach for sparse genomic expression data. We also expand the aspect of genetic network construction and we show the importance of including indirect genetic associations for the performance of such algorithmic approaches.

In the second part of this work we estimate the dependencies on the extracted genetic structure, according to above two approaches. We use Gaussian graphical models in order to describe the dependencies among the involved nodes/molecules. For this purpose, we propose a methodology based on KDE that incorporates these associations as Gaussian approximations with non-linear parameters. We apply our methodology on three separate datasets, starting from the prototype organism *Arabidopsis Thaliana* and continuing with complex diseases such oral and breast cancer. The results indicate statistical and biological validity according to various datasets are researches. They also highlight new genetic structures possibly responsible for cancer development.

Lastly, we applied statistical algorithms for applications based on Single-nucleotide polymorphisms (SNPs). We highlighted specific genomic regions on DNA that show statistical significance in osteoarthritis (OA) disease. The results indicate specific genes responsible for OA progression.

3. ACKNOWLEDGMETS

My warmest thanks to my supervisors, Professors of Electronic and Computer Engineering, Technical University of Crete, Professors M. Zervakis and M. Garofalakis for their valuable help and guidance throughout the duration of the thesis. I also thank the committee members, Professors M. Zervakis, M. Garofalakis and E. Petrakis, for their valuable contribution and comments on the presented thesis. Finally, I would like to thank the research staff of the Laboratory of Digital Signal and Image Processing, Department of Electronic and Computer Engineering, Technical University of Crete, namely as: Dr. Bei Ekaterini and MASc Moirogiorgou Konstantina for their continued assistance and cooperation during the thesis.

4. INTRODUCTION

Our environment is a combination of tightly interlinked complex systems at various levels of magnitude. While the exact sciences of physics and chemistry describe our environment from subatomic level up to the molecular level, biology is carrying the burden to deal with an inexact and extremely complex universe that sometimes even seems lawless. Yet biological systems follow "laws" that physicists would rather refer to as "probabilities." By these laws, it is possible to describe biology at different detail levels with a certain precision. The smallest biological detail level is the molecular level of DNA, RNA, proteins, and metabolites. All these molecules are ingredients of a cell, which in turn is a part of a tissue. Different tissues constitute the organs of an organism. Many organisms together form the ecosystem. Additionally, over time these organisms are subjected to evolution, which results in a certain phylogenetic relationship between them.

At all these levels of detail, the relationships between the elements are of great interest. These relationships can be described as networks, in which the elements are the vertices (nodes, points) and the relationships are the edges. Typical biological networks at the molecular level are gene regulation networks, signal transduction networks, protein interaction networks, and metabolic networks. An example of a biological network is given in Figure 1, where is presented the gene-protein interaction network of the prototype organism *Arabidopsis Thaliana*. While parts of all these networks have been modeled since long time, recent technological advances have made

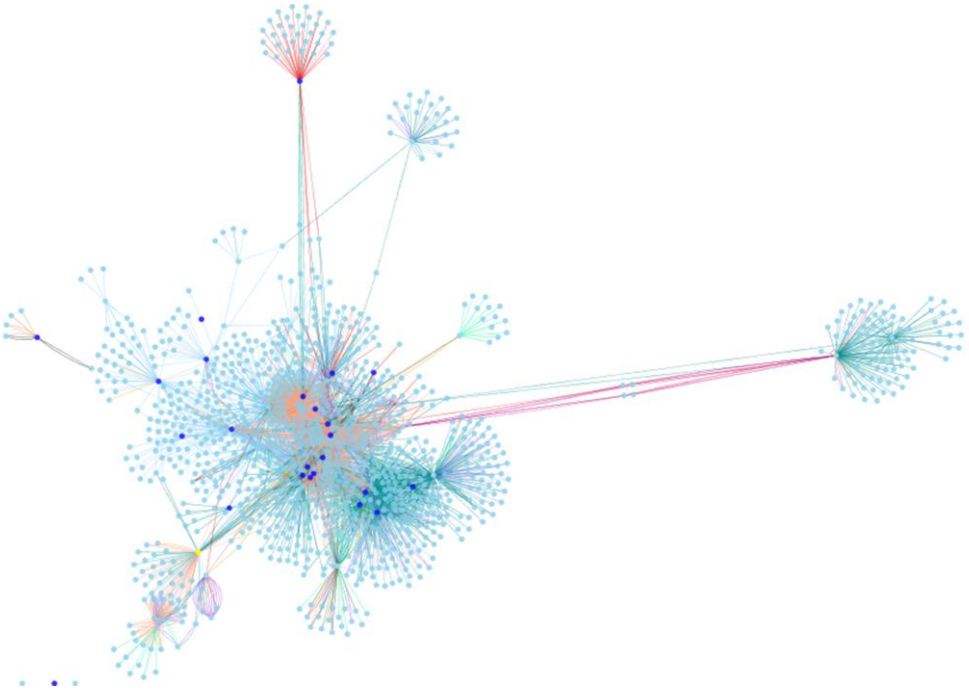


Figure 1: Example of a biological network based on the prototype organism Arabidopsis Thalianna.

it possible to elicit entire networks, or at least large proportions of them.

Biochemical networks have been under investigation for many decades. However, the efforts were until recently limited to the determination of the components of the networks, rather than addressing the design principles of its structure. The fundamental findings about all kinds of networks have also been investigated in biological networks, such as regulation networks, protein interaction networks, and metabolic networks.

Transcriptional regulation networks (or gene interaction networks) are controlling gene expression in cells [1]. The expression of one gene can be controlled by the gene product of another gene. Thus, a directed graph in which the vertices are genes and the directed edges represent control can be used to model these networks. Until recently, only fragments of these networks have been modeled quantitatively, by assigning rate laws to every step. For example, quantitative models containing selected genes have greatly improved the understanding of morphogenesis of prototype organisms [2].

Another type of interaction modeling is the proteomic networks. In this category are included protein to protein interactions that model the enclosed relationships between molecules. (e.g proteins that contribute to the composition of proteomic complexes). Protein interaction networks are generated out of different types of large-scale experimental and computational approaches [3]. The different methods are resulting in significantly different networks, so that we can speak only of a network for a certain organism determined by using a certain method. The protein interaction network of the baker's yeast (*Saccharomyces cerevisiae*) as determined by systematic two-hybrid analyses was found to follow the laws of scale-free networks [4]. Furthermore, it has been shown that the most highly connected proteins in the cell are the most important for its survival. In the network, this corresponds to the vertices with the highest number of connections. In the same network, it has been shown that certain motifs are overrepresented [5].

A multiplicity of mathematical tools has been developed to represent biological regulatory networks with different levels of detail. In the setting of network-structure inference from microarray data, graphical models such as *Bayesian networks* (BNs) represent a commonly used tool to describe the network in a comparatively high level manner. Probabilistic modeling of the involved relationships between the members of the graph outlines the associations' profile. In addition, through probabilistic inference we can identify significant molecules responsible of disease development. Thus, many of the encrypted biological mechanisms can be interpreted so as to reveal the underlying characteristics of a complex organism.

The main focus of this study is the revealing and modeling of inter-relationships between molecules. Using expression data we attempt to estimate the network structure using gene and protein information. The proposed method for network construction is based on Kernel density estimation (KDE) and on Pearson's correlation

(PC), as an attempt to model the nonlinear effect of gene interactions and fill the information loss from the data samples. On the predicted structure, we apply a novel approach based on Gaussian graphical models (GGM) using Kernels to model the genetic dependencies between the nodes. The analysis is applied on four distinct datasets starting from the prototype organism Arabidopsis Thaliana, continuing with the human oral and breast cancer disease and closing with Osteoarthritis. From the analysis we highlight disease-related structures that place important role in disease development, verify and reveal genetic interactions.

In works is organized as follows; there are two parts the theoretical and the application studies. In the first we introduce the basic principles of the methodology and we present the proposed framework and on the second we enclose the experimental applications of our work.

A. THEORETICAL BACKGROUND

In this Section we will present the basic semantics of statistical analysis on genomic datasets and we will continue with the introduction of the Bayesian and Gaussian graphical model analysis. In the last part of this section is presented the proposed framework that was used on the following applications of part B.

5. GRAPHICAL MODELS

Probabilistic graphical models [6] use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space. In a graphical model structure nodes represent random variables (r.v), while (un)directed edges determine the associations between the nodes. There are many types of graphical models such as Boolean network models [7],[8], Bayesian networks [9],[10],[11],[12],[13],[14], Gaussian graphical models [15],[16] , linear and nonlinear models [17], [18] that aim to provide a suitable mathematical representation of stochastic network-like associations and dependence structures.

Many complex systems are characterized by the presence of multiple interrelated aspects, many of which relate to the reasoning task. For example, in a genetic diagnosis application, there are multiple possible genes that the a patient might have, dozens or hundreds of proteins produced by genes, personal characteristics that often form predisposing factors for a disease, and many more matters to consider. These domains can be characterized in terms of a set of random variables, where the value of each defines an important property. For example, a particular disease, such as cancer, may be one variable in our domain that takes two values, i.e. present or absent; a gene/protein may be a variable in our domain, while some of the genetic variables may take continuous values. The set of possible variables and their values is an important design decision and it depends strongly on the questions we may wish to answer about the domain. Figure 2 and Figure 3 illustrate a graphical representation, where the nodes (or ovals) correspond to the variables in our domain and the edges correspond to direct probabilistic interactions between them. The Figure 2 presents a typical Bayesian network where the directed edges imply a direct connection between the nodes while Figure 3 shows a Gaussian network where there is a bidirectional relationship between the involved molecules. In each case, the mathematical representation of these interconnections between the nodes follows a different pattern (Bayesian, Gaussian).

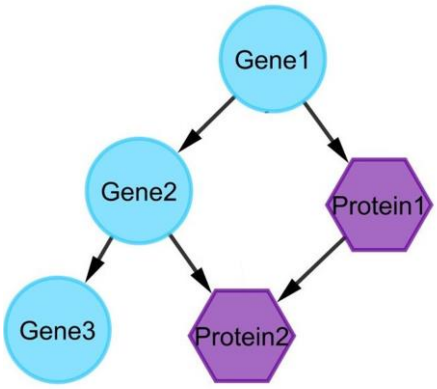


FIGURE 2: Bayesian Directed Network

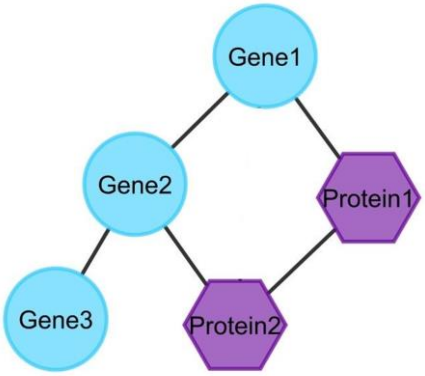


FIGURE 3: Gaussian Undirected Network

In graphical models the goal is to probabilistically predict the values of one or more variables in the set, given the observations about other nodes in the network. For that reason using principled probabilistic reasoning, we need to construct a joint distribution over the space of possible assignments to distribution some set of random variables *X*. With this approach the model allows to answer a broad range of interesting queries of interest, such as to find the posterior distribution of a node X_i based on a specific value x_i, or to find the probability distribution of another node X_j based on the value of X_i.

According to Figure 2, there are many enclosed relationships between genes/proteins that denote dependencies and independencies; with the latter taking the form of *X* is independent of *Y* given $Z(X \perp Y | Z)$. For instance, we can say that $Gene_2 \perp Protein_1 | Gene_1$, $Gene_3 \perp Protein_1$, $Protein_2 | Gene_2$. Also, $Protein_1$ is the child of $Gene_1$ and is depended by its father $Gene_1(Pa_{Protein1} = [Gene_1])$. Similarly, the $Protein_2$ is depended by its parents $Gene_2$ and $Protein_1(Pa_{Protein2} = [Gene_2, Protein_1])$. Exploiting the advantages of conditional independence, we will see that a multi-dimensional joint distribution can be interpreted as a set of (conditional) probability factors:

$$P(Gene_1, Gene_2, Gene_3, Protein_1, Protein_2) = P(Gene_1) * P(Gene_2|Gene_1) * P(Protein_1|Gene_1) * P(Protein_2|Gene_2, Protein_1) * P(Gene_2|Gene_3)$$

This statement is of importance when there is a big number of random variables in a graph that make impossible the modeling of a high-dimensional distribution. In the same context, Figure 3 represents an undirected graph with the same properties, only with the difference that each node is represented as a continuous distribution *f(.)*.

5.1 SEMANTICS

In this section we will provide a formal definition of the graphical models' semantics.

5.1.1 JOINT AND MARGINAL DISTRIBUTIONS

From probability theory and according to Bayes rule, we know that a conditional probability distribution (CPD) of two random variables is defined as $P(X_1|X_2) = \frac{P(X_1,X_2)}{P(X_2)}$, where $P(X_1, X_2)$ is the joint probability distribution. In general, the joint distribution of $X_1 \dots X_n$ is defined as the probability over all random variables that represent the nodes in a graph G and is denoted as $P(X_1, \dots, X_n)$. From the above equation and according to chain rule we take that the joint distribution becomes:

$$P(X_1, \dots, X_n) = P(X_1) * P(X_2|X_1) * \dots * P(X_n|X_1, \dots, X_n). \qquad \text{EQ. 1}$$

On the other hand, marginal probability is the probability over a random variable X_i in the G and is denoted as $P(X_i) = \int_{G \notin X_i} P(X_1, \dots, X_n)$.

5.1.2 CONDITIONAL INDEPENDENCE AND INFLUENCE

Two random variables X_i and X_j are independent if:

I. $P(X_i, X_j) = P(X_i,) * P(X_j)$
II. $P(X_i|X_j) = P(X_i)$
III. $P(X_j|X_i) = P(X_j)$

Similarly, for a set of random variables X_i , X_j and X_i and X_n we say that X_i and X_j are conditionally independent given X_n $(X_i \perp X_j| X_n)$ if one of the following is valid:

I. $P(X_i, X_j|X_n) = P(X_j|X_n) * P(X_i|X_n)$
II. $P(X_i|X_j, X_n) = P(X_i|X_n)$
III. $P(X_j|X_i, X_n) = P(X_j|X_n)$

Looking back in Figure 2, we see that the graph encodes a set of independencies through which the joint distribution can be decomposed as a set of factors. In order to specify those factors we have to find the independencies in the structure.

To specify the conditionally independencies in a graph all possible pathways, through which the influence of a node can be diffused to the other node, have to be examined. If there are no possible pathways through which the influence is diffused, then the nodes are conditionally independent.

Structure (Reasoning Patterns)	Observed nodes	Influence	
		Observed (Yes)	Observed (No)
X→Y	X,Y	X⇄Y	X⇄Y
X←Y	X,Y	X⇄Y	X⇄Y
X→W→Y	W	None ($X \perp Y \mid W$)	X⇄Y
X←W←Y	W	None ($X \perp Y \mid W$)	X⇄Y
X←W→Y	W	None ($X \perp Y \mid W$)	X⇄Y
X→W←Y (v-struct)	W	X⇄Y	None* ($X \perp Y \mid W$)

*If W and none of its descendants are not observed

TABLE 1: Influence diffusion for all possible causal patterns. If there is no influence between the nodes X, Y, then X, Y are conditionally independent.

According to Figure 2 and based on EQ. 1, we can say the joint distribution of the graph G is defined as:

$$P(Gene_1, Gene_2, Protein_1, Gene_3, Protein_2) = P(Gene_1) * P(Gene_2|Gene_1) * P(Protein_1|Gene_2, Gene_1) * P(Gene_3|Gene_2, Gene_1, Protein_1) * P(Protein_2|Gene_2, Gene_1, Gene_3, Protein_2)$$

$$(\boldsymbol{Protein_1} \perp \boldsymbol{Gene_2}|\boldsymbol{Gene_1},) \rightarrow = P(Gene_1) * P(Gene_2|Gene_1) * \boxed{P(Protein_1|Gene_1)} * P(Gene_3|Gene_2, Gene_1, Protein_1) * P(Protein_2|Gene_2, Gene_1, Gene_3, Protein_2)$$

$$(\boldsymbol{Gene_1}, \boldsymbol{Protein_1} \perp \boldsymbol{Gene_3}|\boldsymbol{Gene_2}) \rightarrow = P(Gene_1) * P(Gene_2|Gene_1) * P(Protein_1|Gene_1) * \boxed{P(Gene_3|Gene_2)} * P(Protein_2|Gene_2, Gene_1, Gene_3, Protein_2) =$$

$$(\boldsymbol{Gene_1}, \boldsymbol{Gene_3} \perp \boldsymbol{Protein_2}|\boldsymbol{Gene_2}, \boldsymbol{Protein_1}) \rightarrow = P(Gene_1) * P(Gene_2|Gene_1) * P(Protein_1|Gene_1) * P(Gene_3|Gene_2) * \boxed{P(Protein_2|Gene_2, Protein_2)}$$

$$= P(Gene_1) * P(Protein_1|Gene_1) * P(Gene_2|Gene_1) * P(Gene_3|Gene_2) * P(Protein_2|Gene_2, Protein_2)$$

Thus, we proved that exploiting the conditional independencies in a graph G we can estimate the joint probability distribution as a product of factors that represent the causal dependencies of the nodes with their parents. With this observation we can define more formally the following:

5.1.3 LOCAL INDEPENDENCIES AND CONDITIONAL PROPABILITY

In directed acyclic graph G let $X_1 \dots X_n$ be the nodes represented as random variables. Let $Pa_{X_i}^G$ denote the parents of each node X_i and the $No\text{-}descendantsx_i$ the variables in G that are not descendants ofX_i. Then G encodes the following indecencies:

$$\forall\, X_i: X_i \perp No - descendantsx_i |\, Pa_{X_i}^G$$

Thus, each node X_i is conditionally independent of its no-descendants given its parents. Also, the conditional probability (CPD) of a node $X_i$given is parents $Pa_{X_i}^G$ is defined as:

$$P\left(X_i \middle| Pa_{X_i}^G\right) = \frac{P(X_i, Pa_{X_i}^G)}{P(Pa_{X_i}^G)} = \frac{P\left(Pa_{X_i}^G \middle| X_i\right)P(X_i)}{P(Pa_{X_i}^G)}. \qquad \text{EQ. 2}$$

Also the joint probability distribution is defined as:

$$P(X_1 \dots X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Pa_{X_i}^G\right) \qquad \text{EQ. 3}$$

5.2 GAUSSIAN GRAPHICAL MODELS

The graphical language exploits structure that appears present in many distributions that we want to encode in practice: the property that variables tend to interact directly only with very few others. Distributions that exhibit this type of structure can generally be encoded naturally and compactly using a graphical model. Graphical Models [19], [20] are (un)directed graphs also known as covariance selection models. In typical GGMs the nodes correspond to continuous random variables, modeled by Gaussian distributions. A graph G that is consisted by Gaussian distributions can be represented by a multivariate joint Gaussian distribution. This property, as will be analyzed above, along with many interesting characteristics make the use of GGM's quite interesting. Gaussians are simple subclasses of distributions that make strong assumptions, such as exponential decay of the distribution away of the mean, linearity between the variables [6] . In addition, Gaussian distributions are quite good approximations of real-world distributions which provide a useful tool in modeling real phenomena.

5.2.1 GAUSSIAN DISTRIBUTIONS

A random variable X has a Gaussian distribution $N(m,\sigma^2)$ with mean m and variance σ^2 if it has a probability distribution function (pdf) as $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}}$. On the other hand, a multivariate Gaussian pdf represents a set of random variables $X_1 \dots X_n$and can be characterized by two parameters, an n-dimensional vector of means $\boldsymbol{m}$, a $n \times n$ covariance matrix S with determinant |S| and a pdf:

$$f(x) = \frac{1}{\left((2\pi)^{n/2}\right)|\mathrm{S}|^{1/2}}\, e^{-\frac{1}{2}(x-\boldsymbol{m})^T \mathrm{S}^{-1}(x-\boldsymbol{m})} \qquad \text{EQ. 4}$$

For the validity of the following definition, S must be positive definite. In this case, S is invertible with S^{-1} being the information matrix.

There are specific operations that are needed in a graphical model, such as the computation of marginal, the joint distribution and the conditioning of the distribution on some values $X = x$. Through a multivariate Gaussian distribution these operations are straightforward. A multivariate Gaussian specifies a parallel set of ellipsoidal contours around the $\boldsymbol{m}$ with each contour corresponding to a particular value of the density function. In addition, the extent of each contour is restricted by the covariance matrix as presented in Figure 4, while the joint probability, let's say on X_1, X_2 r.v, is presented as [21]:

$$f(X_1, X_2) = N\left(\begin{pmatrix} m_{X_1} \\ m_{X_2} \end{pmatrix}; \begin{bmatrix} S_{X_1X_1} & S_{X_1X_2} \\ S_{X_2X_1} & S_{X_2X_2} \end{bmatrix}\right) \qquad \text{EQ. 5}$$

Similarly, for a larger number of r.v the joint pdf is represented as a vector of means and through the covariance matrix.

Marginalization is trivial to perform if we observe Figure 4, as it is a factorization over any subset of r.v and can be described solely by the mean and the covariance matrix. More specifically, if $f(X_1, X_2)$ is the joint pdf then the marginal over X_2 is a normal distribution $N(m_{X_2}, S_{X_2X_2})$. Finally, $f(X_1|X_2)$ is normal distribution known as conditional normal and is defined as [22],[23]:

$$f(X_1|X_2) \sim N\left(m_{X_1} + \rho\sigma_{X_1}\frac{x_2 - m_{X_2}}{\sigma_{X_2}}, \sigma_{X_1}\sqrt{1-\rho^2}\right), \rho = \frac{m_{X_1X_2}}{\sigma_{X_2}\sigma_{X_1}}.$$

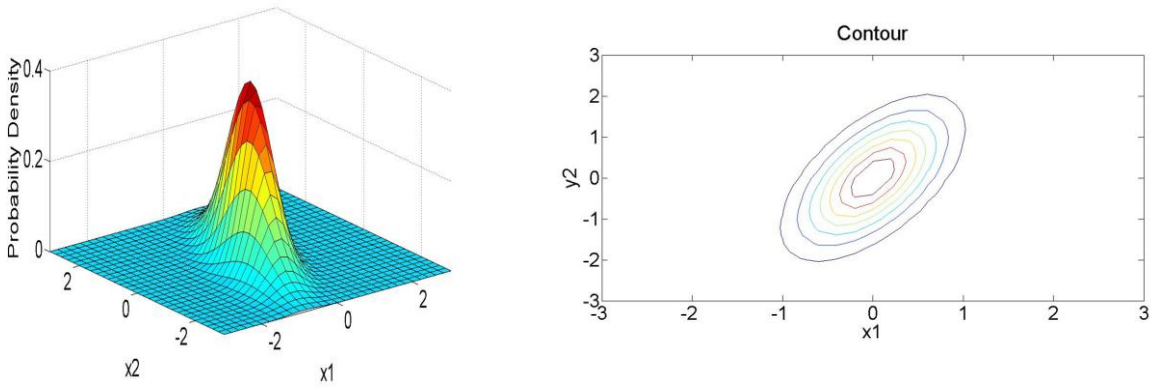


FIGURE 4: Left, joint pdf of a multivariate Gaussian for r.v X_1, X_2. Right, contour representation of the means.

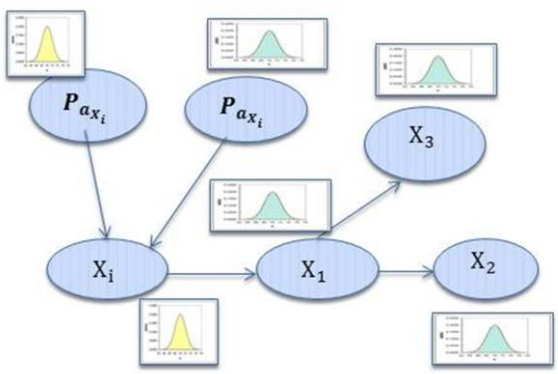


FIGURE 5: Gaussian network. Each node is represented as a Gaussian function, while the CPDs follow Gaussian approximations according to the *Linear Gaussian model*. E.g X_i's CPD is depended by its parents.

Other important properties of a multivariate Gaussian distribution are the rules of independencies. Let $X = X_1 \dots X_n$ have a joint normal distribution $N(m, S)$. Then $X_i$and X_j are independent if and only if $S_{ij} = 0$. Also, the conditionally independency $X_i \perp X_j | X - \{X_i, X_j\}$ holds if and only if $S_{i,j}^{-1} = 0$. The latter property asserts the fact that the information matrix holds the independencies between two r.v, conditioned on all remaining r.vs. Thus, we can say that nodes in a GGM that satisfy $S_{i,j}^{-1} \neq 0$ could be pairwise connected.

5.2.2 LINEAR GAUSSIAN GRAPHICAL MODEL

Linear Gaussian Graphical Model (LGGM)[1] is a classical approach in GGMs that models dependencies between nodes as linear combination of means. In a graph G each node X_i follows a normal distribution depending on its parents as $X_i \sim N(\sum_k w_{ik} x_k, \sigma)$. *N(.)* denotes the normal distribution, whereas the sum extends to all parental nodes of node *i* with x_k denoting the value of node *k*. More formally, let *Y* be a continuous variable with continuous parents $X_1 \dots X_k$, then *Y's* cpd has a linear Gaussian model if there are parameters $b = b_0 \dots b_k$ and σ^2 such that:

$$f(Y|X) = N(b_0 + b^T X; \sigma^2) \qquad \text{EQ. 6}$$

Apparently, LGGM focuses on modeling linear dependencies with parental nodes estimating the mean of a node as a combination of means. In addition, its variance depends only on the experimental data of each node. For instance, according to Figure 2 left, the cpd of $Protein_2$ would be normal, with mean the sum of means of $Protein_1, Gene_2$and variance the variance of $Protein_2$. Apparently, this model captures many interesting dependencies but as we will see we can expand this approach in order to model more complex associations.

5.2.3 GAUSSIAN NETWORKS

A Gaussian network (un)directed graph with a set of nodes $X_1 \dots X_k$ that represent continuous variables. Each node in the network models a Gaussian function while the CPDs of each node are linear Gaussian functions (Figure 5). Thus, from the above

analysis, it is induced that in a Gaussian network the joint pdf is multivariate Gaussian (EQ. 5), while the CPDs are normal pdfs according to EQ. 6. Also, the marginal of each node is $N(m_{X_i}, \sigma^2_{X_i})$ where:

$$m_{X_i} = b_0 + b^T \vec{m}, \quad \sigma^2_{X_i} = \sigma^2 + b^T S b$$

Also, the CPDs $f(X_i|X_j)$ follow the EQ. 6 with:

$$b_0 = m_{X_i} - S_{X_i X_j} S^{-1}_{X_j X_j} m_{X_j}, b = S^{-1}_{X_j X_j} S_{X_i X_j}, \sigma^2 = S_{X_i X_i} - S_{X_i X_j} S^{-1}_{X_j X_j} S_{X_j X_i}$$

The Gaussian networks are generalizations of the graphical models and known as GGMs. They hold all the properties as described in semantics' section with flexible transformations from the network to the joint, marginal and conditional probability distributions with a few parameters to be estimated in each case, the mean and the covariance matrix. Also, through the information matrix we can easily track the conditional dependencies and associations between the nodes. Thus, there is equivalence in representation analog to the Bayesian networks, despite the fact that the latter represent discrete r.vs.

5.3 INFERENCE

Another important graphical model property is the inference. Through the network structure we can use the distributions to answer queries. In particular, the computation of the posterior probability of some variables given evidence is a useful tool that allows the prediction of influence infusion. For example in Figure 2, we might observe that $Gene_1$ is overexpressed and the $Protein_2$ is not expressed and we wish to know how likely it is $Protein_1$ to be overexpressed. Formally written, we want to compute $P(Protein_1\uparrow | Gene_1 \uparrow, Protein_2 \downarrow)$. The inference algorithms work directly on the graph structure and are generally orders of magnitude faster than manipulating the joint distribution explicitly.

The most common query type is already presented in Section 5.2.3 with the conditional probability query as:

$$f(X_i|Z = z) = \frac{f(X_i, z)}{f_{Z=z}(z)} \qquad \text{EQ. 7}$$

These queries allow the prediction of many useful reasoning patterns. However, in a typical graphical model structure there are many parameters that have to be computed in order to find the joint and marginal distributions. In fact, exact inference is an NP-hard problem so we resort to approximation techniques in order to estimate all parameters[6].

A typical GGM structure answers probabilistic queries according to EQ. 7 . In order to estimate this probability we use the variable elimination technique [6], which computes the joint and marginal probabilities eliminating variables X_i, Z from the joint pdf. With

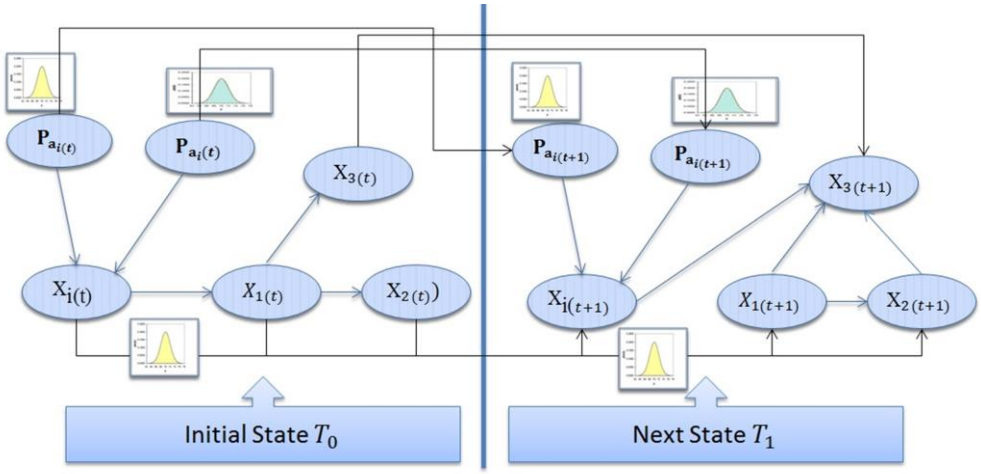


FIGURE 6: Dynamic Graphical Model; between two states each node is connected with itself. The node CPD at T_1 is represented as if having an extra parent (itself from T_0) . The structure does not necessarily remain the same at different time slices.

this approach we find the $f(X_i, z)$ and $f_{Z=z}(z)$ from the joint pdf summing out the variables X_i, Z at each case. Hence, in networks where a large number of parameters is prohibitive, we simply diminish the computation cost using approximation algorithms.

5.4 TEMPORAL BASED MODELS

Probabilistic graphical models specify the above described properties over a set of r.vs and at a specific time stage [24], [25]. The temporal based models are expansions of the graphical models and represent interconnections between different instances of the network at different time stages. The most popular models in this category are the dynamic graphical models as presented in Figure 6, where $X_{i(t)}$ represents the instantiation of the variable X_i at time t and X^t implies all the system variables at state time t. Sequential instances of the network are connected through directed edges that point the next instant of the node.

In order to model the distributions over all time slices $t = 0, \dots, T$ we use the Bayes chain rule as $f(X^{0:T}) = \prod_{t=0}^{T-1} f(X^{t+1}|X^{(0:t)})$. Thus, the distribution over all time slices is the product of cpds for the variables in each time slice given the preceding ones. This property reflects the Markovian assumption which says that $X^{(t+1)} \perp X^{(0:t-1)}|X^{(t)}$, meaning that $X^{(t+1)}$ cannot directly depend on variables $X^{(t')}$ with $t' < t$, but only from the previous state. This means that $f(X^0, \dots X^T) = \prod_{t=0}^{T-1} f(X^{t+1}|X^{(t)})$. From this equation it is clear that structurally there are no dependencies from one state to the other, only direct interactions from each node to itself, at the next state (Figure 6)[6], [26]. This means that in order to compute the joint distribution of all time slices the only information needed is of the previous state.

This property simplifies the dynamic analysis because different time instances the only differentiation compared to the above mentioned properties, is that the node cpds are modeled taking into account an extra parent: itself from the previous time slice.

Based on the above properties of the temporal models we can see the dynamic Gaussian networks (DGN) as extensions of GGMs. In contrast to GGMs that are based on static data, DGNs use time series data for constructing causal relationships among random variables.

In a more compact representation for a biological application, assume that we want to study n genes/proteins and at different time slices. For p microarrays sets and expression levels of n genes/proteins, the data matrix can be summarized as $p \times n$ matrix $X=(X_1,\ldots,X_p)^T$ whose ith row vector $X_i=(x_{i1},\ldots,x_{in})^T$ corresponds to a gene/protein expression level vector measured at time t. Under the concept that the state vector time i depends only by $i-1$ and that each node has the same parents at all states, the joint distribution and conditional probability are composed as [18]:

$$f(X_{11},..X_{pn})=f(X_1)f(X_2|X_1)\ \ldots f\left(X_p\middle|X_{p-1}\right), \qquad \text{EQ. 8}$$

$$f\left(X_i\middle|X_{i-1}\right)=f\left(X_{i_1}\middle|P_{a(i-1),1}\right)*..*f\left(X_{i_n}\middle|P_{a(i-1),n}\right), \qquad \text{EQ. 9}$$

where $P_{a(i-1),\,j}$ are the parents of gene/protein j at time slice $i-1$.

Thus, in DGNs transition between different time slices is modeled as a product of conditional probabilities where the parents of node X_{i-1} are bequeathed to X_i. Profoundly, the pdf at each case is represented as a normal distribution as analyzed in Section of Gaussian Graphical Models.

5.5 KERNEL DENSITY ESTIMATION

Kernel density estimation (KDE) [27], [28], [29], [30] is a non-parametric approach that estimates the probability density function (pdf) of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. In an identically distributed (i.i.d) dataset $X=(x_1,..x_n)$,where x_i denotes the sample i of r.v X, KDE allows the pdf estimation of X as follows:

$$\widehat{f_h(x)}=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right), \qquad \text{EQ. 10}$$

where $K(.)$ is a symmetric kernel function that integrates to one, n is dataset's size and $h>0$ *is* a smoothing parameter. The latter is depended by the samples' standard deviation, the bandwidth that controls the extent of the kernel [29],[31]. Intuitively one wants to choose h as small as the data allow, however there is always a trade-off

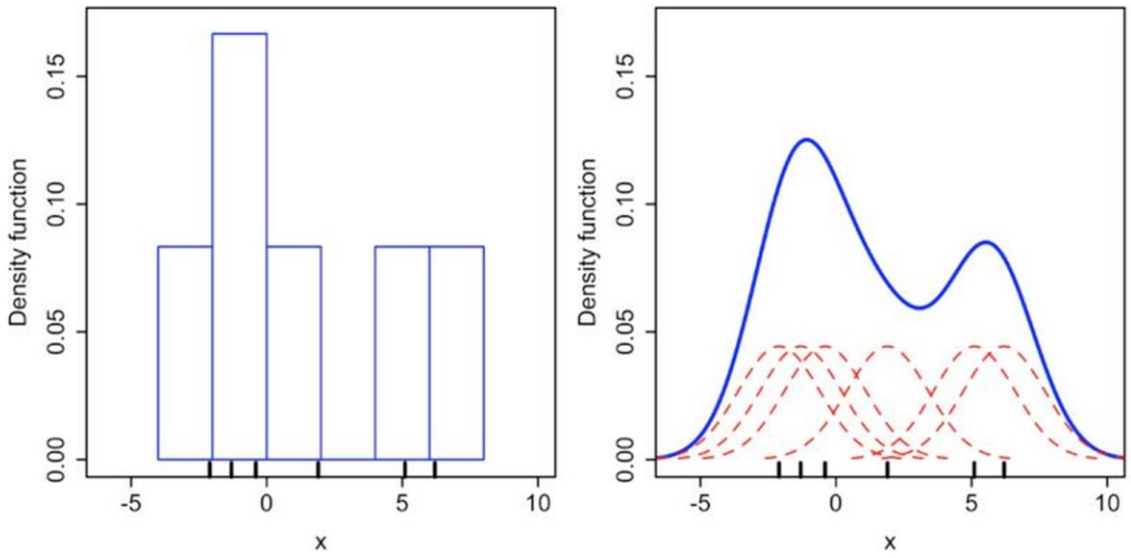


FIGURE 7: Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The 6 individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

between the bias of the estimator and its variance; more on the choice of bandwidth later.

A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others [32]. The Epanechnikov kernel is optimal in a minimum variance sense, though the loss of efficiency is small for the kernels listed previously, and due to its convenient mathematical properties, the normal kernel is often used $K(x) = \phi(x)$, where ϕ is the standard normal density function [33] .

Figure 7 illustrates an example of six data point density estimate. For each of the six points is placed a normal kernel with variance (indicated by the red dashed lines) on each of the data points x_i. The kernels are summed to make the kernel density estimate (solid blue curve). For the histogram, the horizontal axis is divided into sub-intervals or bins which cover the range of the data. In this case, there are 6 bins each of width 2. Whenever a data point (of the kernel) falls inside this interval, it is placed a box of height 1/12. If more than one data point falls inside the same bin, the boxes are stacked on top of each other. The smoothness of the kernel density estimate is evident compared to the discreteness of the histogram, as kernel density estimates converge faster to the true underlying density for continuous random variables

The bandwidth selection of the kernel is a free parameter which exhibits a strong influence on the resulting estimate. Unsuitable bandwidth selection often gives under-smoothed or over-smoothed results. The most common optimality criterion used to select this parameter is the mean integrated squared error:

$$MISE(h) = E \int (\widehat{f_h(x)} - f(x))^2 \, dx \qquad \text{EQ. 11}$$

In Gaussian basis functions with unvaried data where the underlying density being estimated is Gaussian then it can be shown that the optimal choice for h is $h = 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of samples.

6. PROPOSED FRAMEWORK

In an attempt to augment the utilities that graphical models offer, we introduce a new perspective in two separate research areas: in modeling the network's dependencies (cpds) in a GGM network and in revealing the network structure of the involved nodes (genes/proteins).

For this reason, the analysis reported herein is structured in two separate parts. In Section 5.2.2 we consider the analysis of nodes' dependencies through the conditional probability. In this section, we augment the essence of GGMs with a novel algorithm for estimating dependencies between genes/proteins by enforcing a non-linear structure in modeling the parameters of their cpds. We represent conditional probabilities as Gaussian distributions through Kernel density estimation.

In Section 6.2, we study the potential of biological networks and algorithms in revealing the interconnections between the molecules based on experimental data. A generic framework for gene/protein network construction composed of three parts is employed, i.e. network formation based on direct relations, enhancement with indirect interactions and edge orientation. The first two parts referring to network construction are focusing on the partial correlations (PC) and KDE approaches. The third part is enforcing genetic causality according to the Bayesian Information Criterion (BIC). One of the novelties of our framework is the exploitation of not only direct but also indirect genetic interactions. Furthermore, the framework emphasizes the use of the cross correlation metric, as demonstrated in the KDE approach, as well as the exploitation of causality, by means of the BIC criterion.

Figure 8 shows the flow diagram of the framework. The first step of the proposed analysis lies on the discovery of genetic interactions from experimental expression data based on two algorithms, the KDE and PC. For the predicted interactions a direction between nodes is assigned using the BIC. After this step, it is applied the Gaussian modeling in order to find the cpds. If the experimental data are temporal we follow the dynamic Gaussian analysis, otherwise the GGM analysis. In each case, after the second step of the framework we can pose queries through inference to deploy significant structures or molecules in the networks.

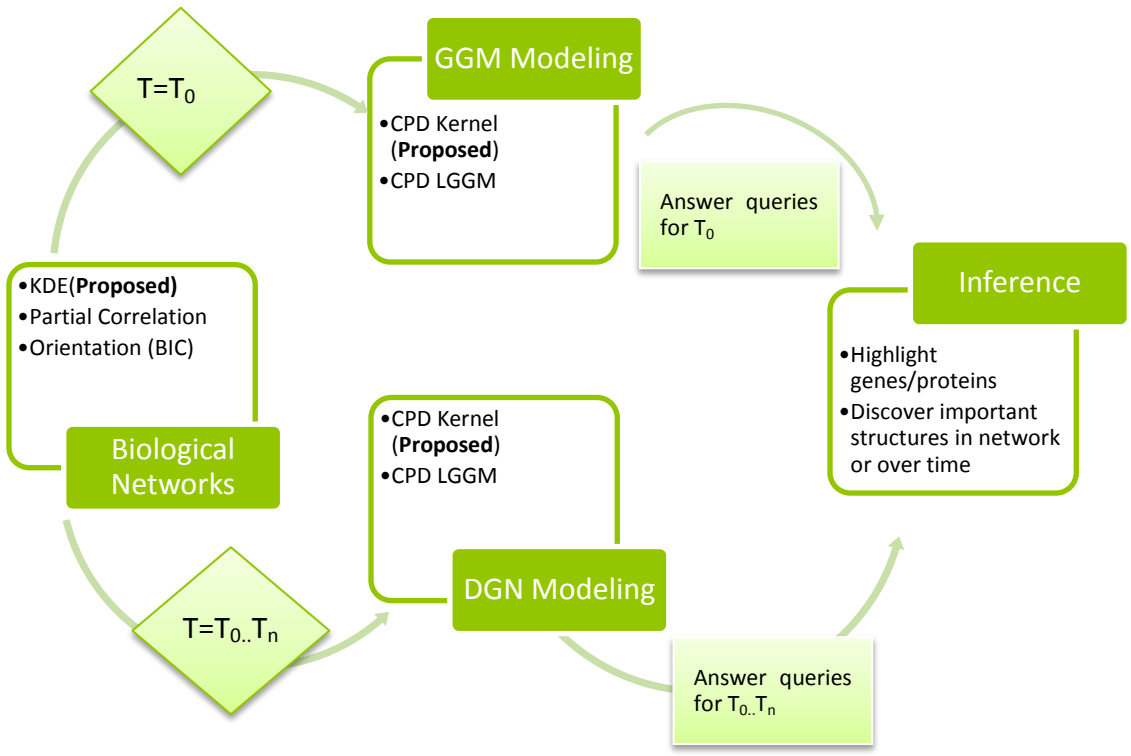


FIGURE 8: Flow Chart for the proposed framework. **1st step**: Revealing the network structure; **2nd step**: we model the dependencies in the produced network through GGM or DGN (for temporal data); **3rd step:** We pose queries in the network to reveal important genes/proteins, alterations through time (for temporal data) or what changes if a gene/protein is (under)expressed.

6.1 STATISTICAL ANALYSIS

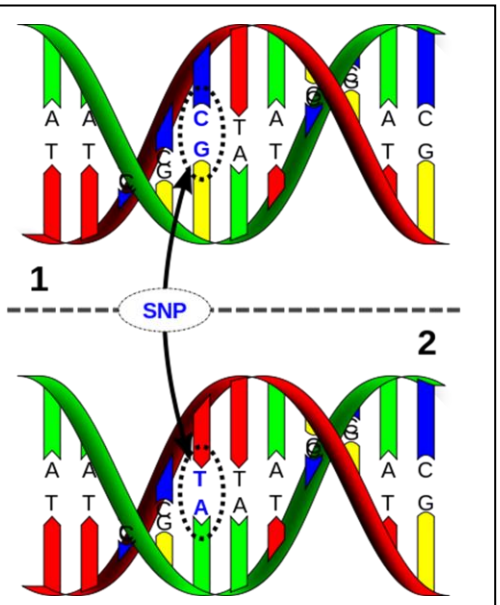


FIGURE 9: DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism).

In this section we will introduce the basic semantics for analyzing differentially expressed genetic datasets. A typical use of the following techniques is the identification of significant genes that are expressed only in a group of interest (e.g healthy and patients). In an attempt to identify such changes between populations, the biologists have deployed a technique that is based on DNA sequence variation, the SNP's. On these data we apply the following statistical analysis so as to select genes responsible for certain diseases.

A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide —A, T, C, G— in the genome differs between members of a biological species or paired chromosomes in a human. For example, two sequenced DNA fragments from different

	A	B	C	D	E	F	G	H	I	EN	EO	EP	EQ	ER	ES	JL	JM	JN
1	Probe Set ID	E100	E101	E102	E103	E104	E105	E106	E107	OR88	OR89	OR91	VA100	VA103	VA105	dbSNP RS	Chromosc	Chromosc
2	AFFX-2315060	AA	BB	AB	AB	AB	AB	AA	AB	AB	AB	BB	AB	AB	BB	rs1699492	20	48440771
3	AFFX-2315061	BB	AB	BB	BB	BB	AB	BB	AA	AB	BB	AB	AA	BB	AB	rs233978	4	1,05E+08
4	AFFX-2315062	AB	BB	NoCall	AA	AB	AB	AB	AB	AB	BB	BB	BB	AB	AB	rs2249922	14	52906081
5	AFFX-2315057	AB	AB	AB	AB	AA	AB	BB	AB	AB	AB	BB	AB	AB	AB	rs7553394	1	21167404
6	AFFX-2315058	AB	AA	AA	AB	AA	AA	AA	AA	AA	AA	AB	AA	AA	AA	rs1782144	16	57996932
7	AFFX-2315059	AA	AA	NoCall	NoCall	AB	AA	AA	NoCall	AB	AB	AB	AB	AA	AB	rs216008	12	2721137
8	AFFX-2315063	AB	AB	AB	AA	AB	AA	AB	AA	AA	AA	AA	AB	AA	AA	rs1254058	7	1,03E+08
9	AFFX-2315064	BB	AA	AB	BB	AB	AB	AB	BB	AB	NoCall	AB	BB	BB	AB	rs2306877	3	4716811
10	AFFX-2315065	BB	NoCall	AA	AB	AB	AB	AB	AB	AB	AB	AA	AA	AB	AA	rs3859360	18	34178190
11	AFFX-2315066	BB	AB	BB	AA	BB	AB	BB	BB	BB	AB	AB	BB	BB	BB	rs1089946	11	78014057
12	AFFX-2315067	AB	AB	BB	AB	AB	BB	AB	AB	BB	BB	BB	AB	BB	BB	rs1682558	14	87714040
13	AFFX-2315068	BB	AB	AA	AB	AA	AA	AB	AA	AB	AA	AA	AA	AA	AB	rs635095	22	21008167
14	AFFX-2315069	AA	AB	AB	AB	AA	AA	AA	AB	AB	AA	AA	AA	AB	BB	rs7683949	4	66746021
15	AFFX-2315070	AB	AB	AB	AB	BB	AB	AA	AA	AB	AB	BB	AB	BB	AB	rs35941	5	53606295
16	AFFX-2315071	BB	AA	BB	AB	AB	AB	AB	AB	AB	BB	BB	BB	AB	AB	rs7192626	16	78425862
17	AFFX-2315072	AA	AA	AA	AB	AA	AA	AA	AA	AA	AB	AB	AA	AA	AA	rs1692909	9	12521826
18	AFFX-2315073	AA	AA	AA	BB	AA	AB	AA	AB	AB	AB	AA	AB	BB	AB	rs1689513	5	22401181
19	AFFX-2315074	BB	AA	BB	AA	AB	BB	BB	AB	AB	AB	BB	BB	BB	BB	rs2472530	11	5153261
20	AFFX-2315075	AB	AB	AB	AB	BB	AB	AA	BB	NoCall	AB	AB	NoCall	NoCall	AB	rs1232722	18	66063259
21	AFFX-2315076	AB	BB	AA	BB	BB	BB	BB	AB	BB	BB	AA	BB	BB	AB	rs4525489	16	6782279
22	AFFX-2315077	BB	NoCall	AB	AB	BB	AB	AB	AA	BB	BB	BB	AB	AB	BB	rs1049283	16	6782303

FIGURE 10 : A SNP dataset. The 1st column represents the SNP ids, the 2nd last the chromosome and the last the chromosomal position. The intermediate columns represent the phenotypes (A, B, NA) for each allele and for each sample (column). Exxx-Oxxx columns concern the patients while the Vxxx the controls.

individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide. In this case we say that there are two alleles. Almost all common SNPs have only two alleles. The genomic distribution of SNPs is not homogenous; SNPs usually occur in non-coding regions more frequently than in coding regions or, in general, where natural selection is acting and fixating the allele of the SNP that constitutes the most favorable genetic adaptation. Other factors, like genetic recombination and mutation rate, can also determine SNP density.

These genetic variations between individuals (particularly in non-coding parts of the genome) are exploited in DNA fingerprinting, which is used in forensic science. Also, these genetic variations underlie differences in our susceptibility to disease. The severity of illness and the way our body responds to treatments are also manifestations of genetic variations. For example, a single base mutation on the on chromosome 7q22 is related to the Osteoarthritis disease [34].

Figure 10 depicts a typical SNP dataset for patients and healthy (controls) and for many different chromosomal positions (case-control study). In practice, an SNP analysis covers thousands genomic regions for all different samples. Each SNP identifies a specific phenotype for each allele (AA, BB) that encodes the nucleotide mismatches as presented in Figure 10. In case where an SNP cannot identify a phenotype, the NoCall is denoted for this sample. In order to identify genomic regions responsible for genetic alterations there is a variety of methodological approaches that isolate chromosome locus or genes to examine if these alterations are related to pathological phenotypes.

6.1.1 QUALITY CONTROL (MAF & HWE)

The first step in a SNP analysis is the quality control where many are excluded due to errors in experimental procedures. For this purpose the minor allele frequency (MAF)

and the Hardly Weinberg equilibrium (HWE) [35] are computed and SNPs that have lower bounds than specific thresholds are excluded.

	AA	*AB*	*BB*	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

FIGURE 11: Genotype distribution for case-control data

In a case-control study as presented in Figure 11, the MAF is computed according to:

- $F_A = \frac{2n_0+n_1}{2n}, F_B = \frac{2n_2+n_1}{2n}$
- MAF=min(F_A, F_B)

If $MAF \geq th$, then the SNP is included for further analysis, otherwise excluded. A typical *th* is 0.1 [34].

The second step in a quality control analysis is the HWD approach. The HWE approach is commonly used as a statistical technique of abnormal allele identification [36]. Given the SNP data both for control and cases we apply the Hardy–Weinberg principle so as to exclude SNPs that do not follow the rule with:

$$f(AA) = p^2 = n_0/n, f(BB) = q^2 = n_2/n, f(AB) = 2pq = n_1/n, \quad \text{EQ. 12}$$

where *f(.)* are the allele frequencies with $p^2 + q^2 = 1$.

We can identify whether the above frequencies have changed compared to the expected Hardy–Weinberg frequencies for a population of *n* as:

$$f'(AA) = np^2, f'(BB) = q^2, f'(AB) = 2pqn \quad \text{EQ. 13}$$

Through a Chi-squared test $x^2 = \sum \frac{(f-f')^2}{f'}$ between the expected frequencies and the observed, we can identify the SNPs that follow the HWE rule. For the produced x^2 values, for 1 degree of freedom, the 5% significance level is 3.84. Thus, if p-values p>0.05 the population follows the HW principle, otherwise the SNP is excluded.

6.1.2 STATISTICAL METRICS

The second step after the quality control is the statistical analysis of the remaining SNPs as an attempt to find differences between control and patients. The main approaches for this purpose is the Hardly Weinberg disequilibrium trend test (HWDTT), the Cochran-Armitage trend test (CATT) and the Odds Ratio (OR) analysis.

6.1.2.1 HWDTT

In the same context with the HWE, the HWDTT [37], [38] is a test that to traces differences between control and cases. Testing with HWE has been used to detect genotype errors or to indicate genetic association. HWE can be tested based on the disequilibrium coefficient $D = \Pr(BB) - \Pr(B)^2$which is based on the difference in disequilibrium coefficients between cases (D_1) and controls (D_0).

Denote the estimators of the genotype frequencies in cases and controls as $\widehat{p_i} = r_i/r, \widehat{q_i} = s_i/s$ for $i = 0, 1, 2$. Then $\widehat{p_B} = \widehat{p_2} + \widehat{p_1}/2$ and $\widehat{q_B} = \widehat{q_2} + \widehat{q_1}/2$ are estimators of the frequencies of allele B in cases (D_1) and controls (D_0):

$$T_{\mathrm{HWDTT}} = \frac{Z^2_{\mathrm{HWDTT}}}{\widehat{Var}\left(Z_{\mathrm{HWDTT}}\right)} = \frac{rsn^3\left[\left(\hat{p}_2 - \hat{p}_B^2\right) - \left(\hat{q}_2 - \hat{q}_B^2\right)\right]^2}{\left\{n - \left(n_2 + n_1/2\right)\right\}^2 \left(n_2 + n_1/2\right)^2}.$$

Under the null hypothesis *H*o, the HWDTT has an asymptotic chi-square distribution with 1df. For low p-values we consider that *H*o is not valid so it is rejected. This means that the examined SNP have phenotype differences between the two examined groups which imply abnormality between control and cases.

6.1.2.2 CATT

Similarly to HWDTT, there is the CATT approach as:

$$T_{\mathrm{CATT}} = \frac{Z^2_{\mathrm{CATT}}}{\widehat{\mathrm{Var}}\left(Z_{\mathrm{CATT}}\right)} = \frac{rsn\left(\hat{p}_B - \hat{q}_B\right)^2}{n\left(n_2 + n_1/4\right) - \left(n_2 + n_1/2\right)^2},$$

The CATT is optimal for the additive model and robust when the underlying genetic model is unknown. Thus, the performance of the CATT could be different under different genetic models. Similarly to HWDTT, if the null hypothesis *H*o is rejected bay low p-values, this indicates abnormality between the studied groups. Thus, the studied SNP is isolated for further examination.

6.1.2.3 OR

The Odd ratio (OR)[39] is a common approach for the case controls analysis. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

- OR=ad/cb or ln(OR)=ln(ad/cb)
- Var(ln(OR))=1/a+1/b+1/c+1/d

The odds ratio can be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

- OR=1 Exposure does not affect odds of outcome
- OR>1 Exposure associated with higher odds of outcome
- OR<1 Exposure associated with lower odds of outcome

However, the OR as a single number is not a representative measure of abnormality identification. For this reason, we use the specified confidence intervals that define the validity of the outcome. The 95% confidence interval (CI) is used to estimate the precision of the OR as[40]:

- CI(a=95%): $\mathrm{OR} \mp z_{1-a/2}\sqrt{\mathrm{Var(OR)}}$

A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR. CI does not report a measure's statistical significance but it is often used as a proxy for the presence of statistical significance if it does not overlap the null value (e.g. OR=1). Nevertheless, it would be inappropriate to interpret an OR with 95% CI that spans the null value as indicating evidence for lack of association between the exposure and outcome. Thus, to find an association between case-controls we cannot only rely on OR and CI.

To overcome this obstacle we use the Mantel-Haenszel Test [40] as a test that identifies homogeneity. This means that we set a null hypothesis *H*o that states both control and cases have the same proportions of alleles. If the hypothesis is rejected, due to low p-values, we have identification of abnormality.

$$E[A] = \frac{r * (n_0 + n_1)}{2 * T}, var[A] = \frac{r * s * (n_0 + n_1 + n_2)}{(T-1)T^2}, T = r + s \qquad \chi^2_{df=1} = \left[\frac{(A - E[A])^2}{\mathrm{var}[A]}\right]$$

The *Ho* is that OR=1, meaning there is no association. For p↑ the *Ho* is retained while for p↓ is rejected. Thus we keep the SNPs that indicate abnormality (statistically significant p-values, $p<10^{-3}$).

6.2 BIOLOGICAL NETWORKS

Biological networks are often described as probabilistic graphs in the context of gene and/or protein sequence analysis in molecular biology. Common approaches to systems biology are based on mathematical representation of biological processes which aim to a deeper understanding of biochemical interactions between genes and genes products.

In recent years the description of genome sequences has resulted in large amounts of gene and protein expression data. The simultaneous examination of thousands genomic units gave a new perspective in the field of bioinformatics as it made possible the study of biological networks. The latest high-throughput microarray technologies allow the simultaneous measurements of expression levels. These technologies have given insight in microbiology since its invention [41] with large amount of data being generated [42]. The extended study of these datasets has provided a new perspective in gene and/or protein network association studies with the network construction from experimental data being a promising approach in modeling functional processes.

Gene regulatory networks (GRN) [43] have provided insight in understanding the working mechanisms of the cell in pathophysiological conditions, as their structure allows the modeling of causal associations. Understanding molecular pathways at the whole-genome level, however, remains a major challenge. Several computational methodologies have been applied to construct biological networks using different data sources [34]. The main focus of networking approaches is to build target-independent networks that describe the pair-wise relations between molecules. Within the last few years, several advanced approaches to address the construction of biological networks from gene-expression data have emerged. These include Pearson's correlation based approaches [44][45][46][13]-[15], clustering and classification algorithms [47] [48][49] [16]-[18]. Although these methods have been successfully used to elucidate the functional relationship between genes and pathways, they are unlikely to directly output the specific gene networks in response to abnormal physiological conditions such as diseases, due to experimental errors and the genetic complexity [34], [17], [8]. Their main drawback is their limited performance when the experimental data is insufficient, especially when the number of the features under examination exceeds the number of samples. This makes the estimation of a network structure a challenging problem due to the uncertainty of calculation of the correlation matrix. The information contained in the expression data is limited by the tissue quality, the experimental design, noise, and measurement errors. These factors negatively affect the estimation of causal relationships in network structure and the derivations of dependencies enclosed between neighbored genes and/or proteins [46].

Similarly to parzen windows [50] as data-interpolation technique that use kernel at sample points to estimate the pdf, Kernel-based models have demonstrated a very competitive computational performance due to their ability to model nonlinear systems and high-dimension data [8]. Using the data distribution in a high-dimensional space, they attempt to interpolate the density of data and, thus, approximate the unknown pdf

of the data model. Support vector machines and relevance vector machines [27] have been applied in prototype organisms and protein-protein networks. In this context, the problem of data scarcity is addressed as a kernel-approximation problem for network estimation. Kernel regression model [28] is also a promising technique for gene and/or protein network analysis with high-throughput genomic data, which could be effectively used for detecting possible altered associations of modules at various disease states. Moreover, in order to identify gene modules associated with diseases or changing conditions, many methods [28] have been developed by integrating gene expression data. A disease-associated active module can be considered as a connected subnetwork or dysfunctional pathway in an interaction network, which has close relationship with a specific disease. Similarly, there exist studies [51] that analyze the underlying mechanisms of differential pathways and molecules responsible for abnormalities of a specific disease.

6.2.1 KDE IN CROSS CORRELATION TEST

Assume that a generic network is developed based on a limited genomic independent identically distributed (i.i.d) dataset $X=(x_1,..x_n)$,where x_i denotes the sample i of gene or protein X. The KDE allows the estimation of probability density $\hat{f}_h$ of X as showed in EQ. 10, where $K(u)=\frac{1}{2\pi}e^{\frac{-1}{2}u^2}$ is a vector of symmetric positive definite Gaussian function, n is dataset's size of the gene or protein X and $h \cong 1.06\sigma n^{-1/5}$ is the bandwidth parameter. The latter is depended by the samples' standard deviation and controls the extent of the kernel [29],[31].

Genes interacting with each other can be linked and organized in a network form. The gene expression over a population provides valuable information on a gene's activity, which can be correlated with other genes as to provide a metric of organization in the network structure. Under the assumption that genes and gene-products share similarities in datasets, the problem of network construction is reduced to the examination of independence between nodes X_i and X_k, through the joint and marginal probabilities:

$$f_h(X_i, X_k) = f_h(X_i) \,.\, f_h(X_k) \qquad \text{EQ. 14}$$

where f_h indicates the probability density estimate of each gene according to EQ. 10 and the right-hand side is computed by point-by-point multiplication. The comparison of the two parts of EQ. 14 can be performed through the cross correlation test, where high correlation indicates independence of the two nodes, thus low connectivity. In contrast, small correlation indicates differences between the two parts of EQ. 14 and dependence of the two nodes, thus demonstrating high interaction between X_i and X_k. The latter justifies the connection between candidate nodes since they share common activity characteristics.

6.2.2 PARTIAL CORRELATION ESTIMATION

Pairwise associations of coexpressed molecules can be modeled by the Pearson's correlation (PC) [20], [19]. Pearson's correlation is a metric that denotes whether two variables share common characteristics. The interaction identification between two vector variables is reduced to estimating the covariance matrix *S.* Each element in S_{ik}, via $S_{ik}=\rho_{ik}\sigma_i\sigma_k$ and $S_{ii}=\sigma_i^2\sigma_i^2$, represents the scalar correlation coefficient ρ_{ik} between nodes X_i and X_k and indicates an association, while $\sigma_i^2\sigma_i^2$ are scalar values that denote the variance of nodes X_i and X_j. A high correlation coefficient between any two genes/proteins may be indicative of either direct interaction, or indirect interaction or regulation by a common gene/protein. However, for the construction of a gene and/or protein association network only the direct interactions are of interest as only these correspond to edges between two nodes (genes) in the resulting graph.

The method of partial correlations [45] measures the correlation between two variables after the common effects of all other variables are removed. An appropriate notion of the strength for these interactions is the partial correlation matrix $\Pi=(\pi_{ik})$. Its coefficients π_{ik}, describe the correlation between genes and/or proteins *i* and *k* conditioned on all remaining genes of the network. This property is reflected in the inverse covariance matrix S^{-1}, with elements:

$$\pi_{ik}=-\frac{S_{ik}^{-1}}{\sqrt{S_{ii}^{-1}S_{kk}^{-1}}}. \qquad \text{EQ. 15}$$

Given the experimental data, the covariance matrix is computed and then it is inverted. Indeed, using EQ. 15 the partial correlations, π_{ik}, can be easily computed. Significantly small values of $|\pi_{ik}|$ indicate conditional independence between *i* and *k* given the remaining variables in graph. On the contrary, high values of $|\pi_{ik}|$ indicate dependence between *i* and *k* which contributes to adding an edge between these nodes.

Despite its straight-forward nature, this approach is only applicable if the sample number in the dataset is larger than the number of genes/proteins. Otherwise, the inversion of *S* is unstable making the estimation of S^{-1} a non-trivial task. To overcome this obstacle we invert *S* using the Moore-Penrose pseudo inverse [52], an approximation of the standard matrix inverse, based on the singular value decomposition (SVD).

6.2.3 EDGE ORIENTATION

Up to this point we have reviewed two approaches in revealing the network structure, thus providing an intuition on whether two nodes interact. Nevertheless, they do not imply anything about directionality, indicating which node is the cause and which is the result. In order to determine the edge orientation for the above networks we have to examine the causality between pairs of nodes. For instance, considering two

nodes we can define two models, i.e. model M_1, where node X_i is the parent of node X_k and model M_2, where node X_k is the parent of node X_i.

Model selection procedures cannot distinguish the above models because their distribution *f(.)* or likelihood is equivalent. In other words, the variation in the level of node X_i causing a variation on node X_k yields the same joint density as the reverse situation [53], [54]:

$$f(X_k|X_i)f(X_i)=f(X_i,X_k)=f(X_k)f(X_i|X_k). \quad \text{EQ. 16}$$

Therefore, the distinction between models M_1 and M_2 is made by inferring direction of causality between nodes using a scoring function, the BIC criterion [53]:

$$BIC=-2\log\hat{L}+K\log N, \quad \text{EQ. 17}$$

where $\hat{L}$ is the maximum likelihood, K is the number of parameters to be estimated in the model, and N is the sample size. A model is better than another if it has a smaller BIC value. Thus, for each edge orientation a BIC score is computed and the edge direction is decided in favor of the lowest BIC value.

For instance, if we assume that an initial direction between 4 nodes is 1→2←3 ←4, we start by computing the BIC score for edge (2-3) including node 1. The process is performed in one direction including node 1 and is repeated for the opposite direction for edge (2-3) including node 4. If the BIC score is smaller in the latter case the direction changes for edge (2-3) and deriving the structure 1 →2→ 3 ←4. Furthermore, the BIC score is recomputed for the edge (3-4) including node 2.

In more complex networks, edges are oriented by splitting the graph structure into smaller subnetworks. For each node, the number of its connected edges is counted. Nodes are then arranged in descending order in terms of the number of connected nodes. A node and all the nodes that are directly connected to it form a subnetwork. For each subnetwork, the BIC score is computed for each edge that connects a pair of nodes, containing all other causative nodes to that pair.

6.3 GRAPHICAL MODELS USING KERNELS

As we described in Section 5.2.2, the GGMs are types of graphical models for representing complex associations among Gaussian random variables. In this context, a gene/protein corresponds to a random variable shown as a node, while gene/protein interactions are shown by directed edges. Consequently, interactions with parental nodes are modeled by the conditional distribution of each gene.

Although graphical models are promising for interaction analysis, their main drawback is their limited performance when the experimental data is insufficient. This problem has two aspects. First, the lack of experimental samples (genes/proteins) when the number of the features under examination has greatly increased. More precisely, in a typical microarray dataset the number of genes exceeds by far the number of sample

points that correspond to a gene. This makes the estimation of a network structure a challenging problem due to the uncertainty of calculation of the correlation matrix [52], [46] . Second, the information contained in expression data is limited by tissue quality, the experimental design, noise, and measurement errors. These factors negatively affect the estimation of causal relationships in network structure and the derivations of dependencies enclosed between neighbored genes/proteins [46].

A common graphical representation scheme is the Gaussian model firstly introduced by Kishino and Waddell [19]. However, there is a critical detail in applying Gaussian modeling. If the number of samples is far smaller than the number of features, then this framework is inefficient. The covariance matrix, embodying the interactions between genes/proteins is often not positive definite, rendering the computation of the partial correlation matrix (Section 5.2.1).

Given these challenges, it becomes obvious that graphical models need additional tools to overcome such obstacles. We propose a new methodology for modeling dynamic Gaussian graphical models from sparse data. More specifically, we focus on the problem of completing the information loss in time varying Gaussian networks through the non-parametric framework of Kernel density estimation [55]. Our approach exploits the idea that Gaussian densities describe sufficiently biological interactions and that neighboring gene/proteins can be described by conditional probabilities as approximations of Gaussians with nonlinear parameters. In addition, due to the fact that Gaussian graphical models are widely known as non-directed graphs, we introduce directions based on Bayesian information criterion. This makes interactions within the graph conceptually more representative to biological processes.

6.3.1 CONDITIONAL PROBABILITY DISTRIBUTION

In the presented analysis we showed the LGGM approach where the dependencies between the nodes are modeled as linear combinations of their parents. This model, despite it's many interesting properties, it also has weaknesses. Firstly, the variance of each node (EQ. 10) is only depended by node's values. This is actual a non-realistic hypothesis because in real applications we expect the variance of each child to be affected by its parents. Also, according to EQ. 10 we clearly see that this modeling considers linear dependencies between the involved nodes. In practice however, we know that physical phenomena enclose non-linear dependencies, thus we would like an enhanced model that would satisfy these requirements.

For these reasons we introduce a new approach based on KDE, as a non-parametric framework, that estimates the cpds of the GGM as Gaussians approximations and solely based on the experimental data. With this proposed methodology we aim to model the Gaussian parameters according to the enclosed relations of the samples. With a KDE estimator we compute the joint and marginal distributions in addition to Gaussian cpd's parameters as the non-linear forms. Hence, the main innovation of this model is that it captures non-linear relationships between molecular units based on expression data. In

addition, there is no information loss. In fact through KDE, the missing data is no longer an obstacle due to estimation from the remaining samples.

In a typical biological application, suppose we have p sets of microarrays and n genes/proteins, where $X_i=(x_{i1},\ldots,x_{ip})^T$ is a p dimensional expression vector obtained for ith gene/protein. Let P_{a_i} be the parents of gene/protein X_i then direct dependencies are encoded as:

$$f(X_i|P_{a_i}) = \frac{f(X_i, P_{a_i})}{f(P_{a_i})} \qquad \text{EQ. 18}$$

In order to model these relations with a coherent mathematical framework based on genomic expressions, we find the joint distributions of EQ. 18 with the Standard Gaussian Kernel (SGK) as follows [27], [28], [29], [30]:

$$\hat{f}_h(x,y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right) K\left(\frac{y-y_i}{h_2}\right) \qquad \text{EQ. 19}$$

From EQ. 18, EQ. 19 we obtain:

$$f(X_i|P_{a_i}) = \frac{\sum_{i=1}^{p} K_{h_1}(x-x_{ij}) K_{h_2}\left(p_{a_i}-p_{a_{ij}}\right)}{\sum_{j=1}^{p} K_{h_2}\left(p_{a_i}-p_{a_{ij}}\right)} \qquad \text{EQ. 20}$$

where $K(.)$ is a Gaussian kernel function described with EQ. 10, p is dataset's size and $h_1=c_1 n^{-1/6}$, $h_2=c_2 n^{-1/6}$ for $c_1,c_2>0$ [29] are the smoothing parameters selected as optimal approximations of Gaussians basis functions [41], [56].

EQ. 20 implies that the conditional density estimate is an asymptotic approximation of Gaussian [41], [55] , [57], [58] $N(\theta_1,\sigma_1^2)$ with $R(K)=\int K(u)^2 du$ and parameters as follows:

$$\theta_1 = \frac{\sigma_K^2}{2\sqrt{c_1 c_2}}\left(c_1^2 f^{(2)}(X_i|P_{a_i}) + c_2^2 f_{(2)}(X_i|P_{a_i}) + 2c_2^2 f_{(1)}(X_i|P_{a_i}) f^{(1)}(X_i|P_{a_i})\right) \qquad \text{EQ. 21}$$

$$\sigma_1^2 = \frac{R(K)^2 f(X_i|P_{a_i})}{c_1 c_2 f(P_{a_i})} \qquad \text{EQ. 22}$$

Hence, EQ. 21 and EQ. 22 encode a Gaussian model that captures non-linear dependencies of network parameters. If a gene/protein has no parents the mean and variance is taken from KDE according to its expression data.

Figure 12 shows in summary the steps for applying the proposed framework from the first step of network estimation to the GGM analysis. From the genetic expression data are found the inter-connections between the molecules. On this structure and using

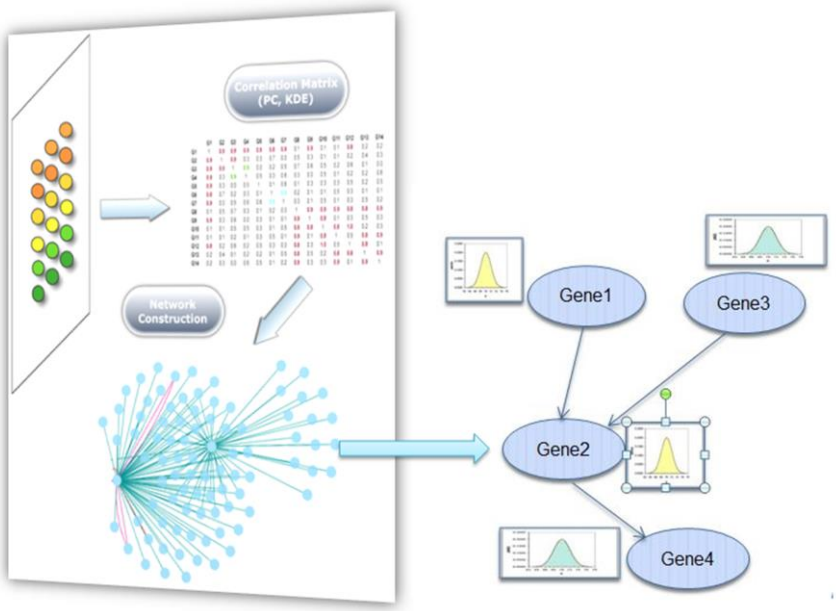


FIGURE 12: Left, using the expression data we estimate the genetic structure through KDE or PC. On the estimated structure we apply the GGM analysis (right). Each node has a cpd estimated from the data samples based on the Gaussian kernel analysis.

the data samples, all nodes are represented as Gaussian approximations. This modeling is computed after applying GGM using Gaussian kernels on the expression data of each node including its parents. In this way, we estimate the dependencies between the molecules as continuous r.v that enclose non-linear associations with their parents.

6.4 EFFECTS OF EXTERNAL GENES

The poor performance in biological network reconstruction is a well-known problem that has been extensively addressed, especially when dealing only with expression data. The problem is attributed to the large number of false-positive predicted interactions and a dominant idea to address, is to characterize the produced associations according to Gene Ontology (GO) terms at a higher level of processes [13], [59]. Other approaches [60], [61] introduce new topological metrics that justify each molecular connectivity and associate it with a biological process. Due to the complicated nature of molecule associations, we propose to accept not only known direct associations between pairs of genes, but also connections that are induced by external molecules [42], which can be identified in various available databases [59]. By exploiting this knowledge we can examine indirect interactions between the studied genes, taking into account all the possible external pathways that connect these molecules. Thus, several initially assigned false-positive edges can be characterized true positive as a result of multiple effects of external molecules.

Other supporting evidence for revisiting the consideration of edges as false positive (FP) is that the actual interactions are either physical or genetic, which may not be direct interactions. Thus, the computed precision may be lower than the actual

performance, since links may be missing in the databases of the known direct interactions. Similarly, the recall presented may be lower than the actual recall, partly because some of the links reported in the databases may be indirect [59] and partly because some presently unsupported edges in the constructed network may find experimental evidence in the near future. Therefore, many unsupported edges may not be necessarily false positives.

In order to compare the performance of the proposed framework with and without external interactions, we employ at the results section the receiver operator characteristic (ROC) and precision-recall curves. For this purpose we consider a ground-truth network that encompasses the available biological knowledge of many public databases and compare it with our network's structure. We use the following notation; TP is the number of edges present in the ground-truth network and in the predicted network; FP is the number of edges not present in the ground-truth network but are included in the predicted network; FN is the number of edges present in the ground-truth network but not in the predicted network; TN is the number of edges not present in the ground-truth network and also not included in the predicted network. The above definitions are graphically illustrated in Figure 13. We consider TP as the existent edges in both networks. Also, when a predicted interaction is verified through indirect associations with external factors (triangle genes) then the predicted association is set as TP. Finally, we consider FN as non-existent in the ground-truth but predicted direct and/or indirect interactions, while TN are edges that are not present in the constructed and ground-truth networks neither as direct nor as indirect connections. With this approach we examine if the predicted interactions are verified as indirect implications through external genes that participate in different pathways.

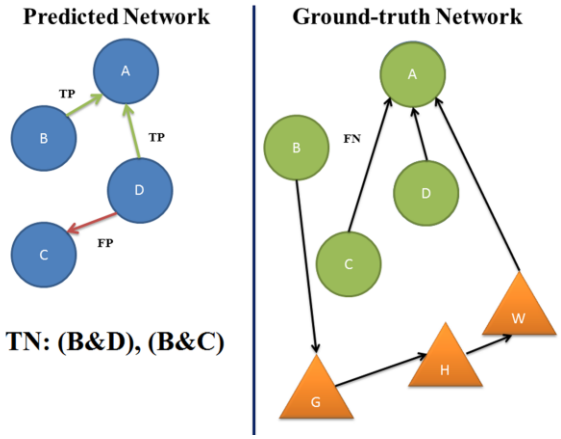


FIGURE 13: Graphical representation of the TP, FP, TN, FN edges according to the existent knowledge of the ground-truth network. External genes are represented with triangles while studied genes with circles. External pathways that indirectly connect studied genes give TP connections.

B. APPLICATIONS

In the following section are presented the applications of the proposed framework. There are four distinct datasets on which we investigate the statistical verification of the methodology. Starting from a prototype organism, the *Arabidopsis Thaliana*, we focus on the network structure estimation with sparse temporal data in addition to gene selection of important molecules through inference. In the same context, we examine the properties of the framework on the human organism, for the breast and oral cancer disease. For the first we attempt to highlight differentially expressed genes between two populations while for the latter we highlight the significance of external genes in the disease development. In the last section, we present the findings of differentially expressed genes on the Osteoarthritis disease based on healthy and patients.

7. ARABIDOPSIS THALIANA RESULTS

Biological networks are often described as probabilistic graphs in the context of gene and protein sequence analysis in molecular biology. Microarrays and proteomics technologies allow the monitoring of expression levels over thousands of biological units over time. Experimental efforts aim to unveiling pairwise interactions. Many graphical models have been introduced in order to discover associations from the expression data analysis. However, the small size of samples compared to the number of observed genes/proteins makes the inference of the network structure quite challenging. In this study we generate gene-protein networks from sparse experimental temporal data using two methods, Partial Correlations and Kernel Density Estimation, in order to capture genetic interactions. Applying KDE method we model the genetic associations as Gaussians approximations, while through the dynamic Gaussian analysis we aim to identify relationships between genes and proteins at different time stages. The statistical results demonstrate important implications of valid biological significance for gene to gene/protein interactions and reveal new indirect relations and dependencies of molecules.

In order to investigate the statistical properties of the proposed framework we start by revealing the network structure using the PC and KDE approaches. After this step, and for each generated network, the conditional probabilities are found based on our proposed algorithm, as well as using the LGGM approach. Finally, through network inference we compute the direct and indirect implications of certain factors in the network and compare with known significant biological relations. We perform comparisons on the inference results based on our algorithm and LGGM. The same framework is applied for different time slices in order to examine time dependencies.

7.1 NETWORK CONSTRUCTION AND DIRECT RELATIONS

The data samples we used for testing concern the developing Arabidopsis thaliana seeds [62], [63], harvested at 5, 7, 9, 11, and 13 days after flowering using Affymetrix ATH1 chips. We isolated the carbohydrate metabolism pathway including 7 'significant' and 6 'unrelated' genes and studies the network associated with this pathway. Genes that encode invertases (At1g35580, At5g22510) or sucrose synthases (At3g43190, At4g02280, At5g20830, At5g37180, and At5g49190), both being important enzymes in the metabolism of sucrose, were designated as 'significant' genes [64]. In order to test our proposed algorithm, we included more than one sucrose synthase genes as internal controls. As 'unrelated' genes we designated six genes that are involved in other biological processes (intracellular traffic, energy, protein destination and storage, disease/defence) in carbohydrate metabolism [62], [65]. These 'unrelated' genes are either expressed in seeds (At1g54050 and At3g17520) or not expressed in seeds (At1g13140, At2g39470, At4g14630, At4g15010) and are identified as biomarkers for

Threshold		Verified Pairs		New Edges		Oriented Edges	
PC	KDE	PC	KDE	PC	KDE	PC	KDE
≥0.1	≤0.1	19/27	1/27	5594	421	192	51
≥0.2	≤0.2	15/27	7/27	4852	1075	181	95
≥0.3	≤0.3	8/27	14/27	4097	1969	159	83
≥0.4	≤0.4	9/27	15/27	3357	2741	140	82
≥0.5	≤0.5	8/27	17/27	2618	3995	165	93
≥0.6	≤0.6	6/27	17/27	1942	5224	133	77
≥0.7	≤0.7	4/27	17/27	1300	5682	124	66
≥0.8	≤0.8	4/27	23/27	753	6100	111	70
≥0.9	≤0.9	0/27	22/27	286	6327	58	60

TABLE 2: Network structure for various thresholds with PC and KDE algorithms

specific organs (flowers, leaves, roots, siliques) in Arabidopsis. Overall, we studied 113 genes and 27 gene-protein pairs, for all stages of growth. Our goal was to verify known gene-protein interactions, direct associations between genes as well as to highlight how the pathway is affected by significant factors.

Table 2 presents the number of verified gene-protein pairs. The first column describes different thresholds on partial correlation set on PC for EQ. 15, while the second column provides the thresholds of absolute difference of EQ. 14 for KDE. The third and fourth columns summarize for both approaches the verified number of gene-protein interactions. The fifth and sixth columns present the number of new edges that have

occurred for each threshold while the two last columns describe the number of edges that changed orientation according to BIC criterion.

The results indicate that as thresholds increase for the inferred networks with the PC algorithm, the graph becomes sparser with less interactions being verified. This is due to the lack of strong partial correlations between molecular units. On the contrary, as thresholds of KDE increase, the correlation also increases implying that genes-proteins are found to be less independent. Thus, more interactions are identified in KDE and the graph becomes more cohesive.

Table 3 shows the verified interactions between genes as well as interactions of proteins. We compared the performance of the two approaches taking into account the existent information on gene-gene and protein-protein interactions from two related databases, namely ATTED-II, the Arabidopsis gene co-expression database [66] and AtPIN, A. thaliana Protein Interaction Network [67]. The former provides 3,321 genes (interacting directly or indirectly), while the latter provides 1,092 protein-protein interactions, when all examined genes are used as input queries for known gene or protein interactions in A. thaliana, respectively. For the examined pathway we retrieved 62 known gene interactions and 729 protein interactions [68]. The high number of protein interactions may be relevant to a number of physically interacting proteins, but also to a number of interacting proteins that are not physically connected.

Threshold		Verified Gene Interactions		Verified Protein Interactions	
PC	KDE	PC	KDE	PC	KDE
≥0.1	≤0.1	58/62	0/62	240/729	46/729
≥0.2	≤0.2	52/62	3/62	212/729	76/729
≥0.3	≤0.3	48/62	6/62	182/729	108/729
≥0.4	≤0.4	44/62	19/62	158/729	148/729
≥0.5	≤0.5	39/62	34/62	130/729	184/729
≥0.6	≤0.6	35/62	47/62	106/729	220/729
≥0.7	≤0.7	28/62	53/62	84/729	236/729
≥0.8	≤0.8	20/62	57/62	60/729	256/729
≥0.9	≤0.9	08/62	60/62	38/729	262/729

TABLE 3: Gene-gene and protein-protein interactions for various thresholds

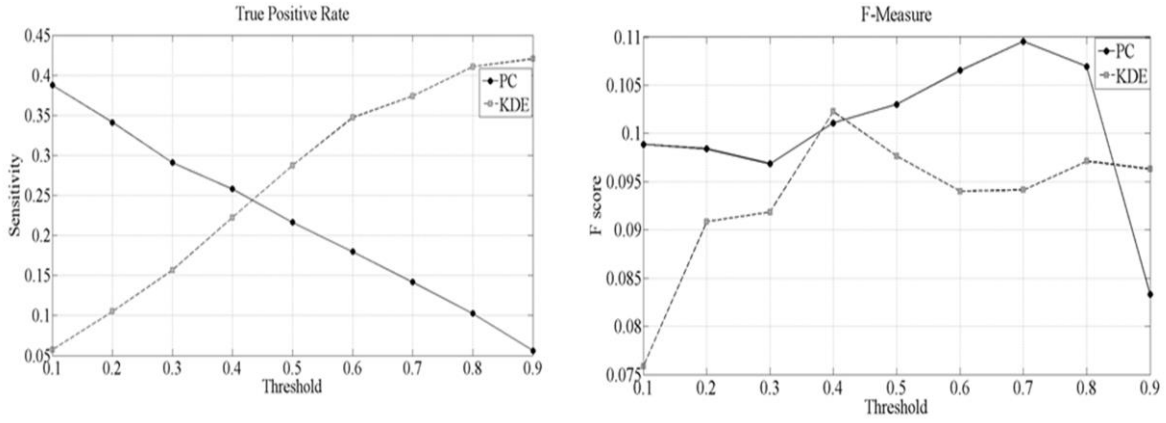


FIGURE 14: Left, True positive rate for the verified gene or/and protein interactions for KDE and PC algorithms. Right, F-measure for different thresholds on KDE and PC algorithms.

Table 2 and Table 3 provide a notion of the identified number of verified interactions. Comparing the performance of two methodologies, KDE appears to behave better in capturing the above biological associations. More precisely, KDE identifies up to 81% of known gene/protein interactions, up to 96% known gene-gene interactions and up to 36% existent protein-protein interactions. These percentages for PC are 70%, 93% and 33%, respectively. Finally, to assess the network reconstruction ability, we counted true positives TP (correctly identified true edges), false positives FP (spurious edges), true negatives TN (correctly identified zero-edges) and false negatives FN (not recognized true edges) edges. Figure 14 (left), summarizes the true positive rate for both algorithms, meaning framework's ability to detect existent interactions.

In order to find the optimal threshold for each algorithm the size of the graph has to be taken into consideration. This is necessitated by the fact that as graph becomes denser, more interactions are generated. Thus the probability of capturing preexistent associations increases. Figure 14 (right), presents for all thresholds the performance of two methodologies according to F-score metric, $F=\frac{2*precision*recall}{precision+recall}$. In conclusion, appropriate threshold for KDE is $th=0.4$ while for PC it is set to $th=0.7$.

From a statistical perspective, many false positive edges were found (leading to low F-score). However, this aspect needs further discussion to reveal its valid implications. The false positive rate of connections becomes large due to the consideration as ground-truth positive of only the direct interactions that have also been biologically confirmed. In practice, the majority of molecules in the neighbor of a gene or protein participate in similar biological processes. It is expected therefore that this neighborhood defines many more direct interactions that have not been established yet. Thus, an alternative consideration would concern the inclusion of all direct connections in the neighborhood of established ones in the definition of ground-truth positive, as a valid assumption that contributes to the determination of relevant false positives and the correct interpretation of the performance metric.

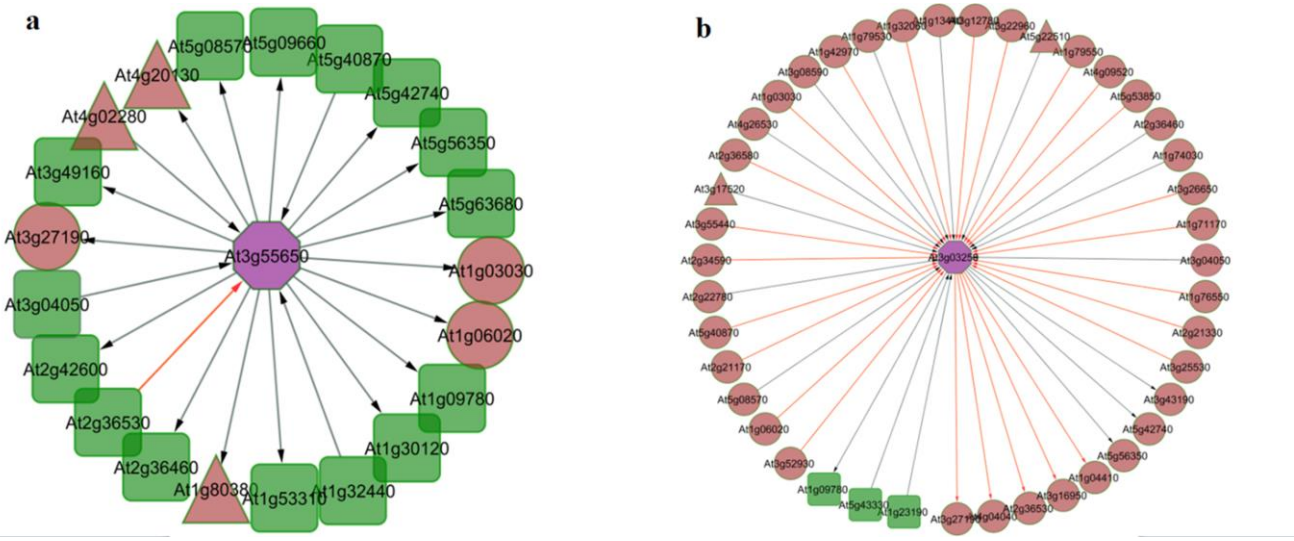


FIGURE 15: Snapshot of direct connected genes/proteins in the KDE (0.4) network for selected (central) genes. Square green molecules indicate true positive relations; brown triangles reflect false positives and brown ellipses denote controversial false positives. Grey and orange edges show genetic and proteomic associations, respectively.

Addressing these aspects in more detail, Figure 15 presents a snapshot of the constructed KDE network where the At3g55650 (a) and At3g03250 (b) genes are directly connected to their neighbors. The central gene is associated with biologically verified genes (green squares), whereas triangle and elliptical elements indicate false positive association. For gene At3g55650, KDE has successfully captured all known biological interactions. However, apart from all known associations of gene At3g03250, KDE derives many interactions considered as false positives (brown ellipses). These interactions deserve closer attention, since some of them could be encountered as true positives because their molecules participate in the same processes as their neighboring genes (green squares). For instance, some interacting molecules may be the result of indirect connections (not physically interacting proteins) with At3g03250 [69]. Similarly, At1g13440, At3g12780, At2g36580 and the At3g03250 in Figure 15 are all identified as 14-3-3 client and binding proteins [70].

Recent studies report that the 14-3-3 proteins interact dynamically with proteins engaged in plant nitrogen and carbon metabolism, and appear to possess a modulatory role in Arabidopsis seed development. They also suggest an important role for 14-3-3 proteins in the homeostatic control of crucial glycolytic intermediates, such as the phosphoenolpyruvate (PEP) [70]. Thus, we speculate that the central molecule At3g03250 might indeed be interacting with At1g13440, At3g12780 and At2g36580 through 14-3-3 proteins that are involved in carbohydrate metabolic process, including glycolysis of developing *A. thaliana* seed. Furthermore, due to the verification of the interaction between the central gene At3g03250 and the At1g09780 gene that encodes a phosphoglycerate mutase isozyme, we expected that the observed interactions with two other isozymes of phosphoglycerate mutase (At3g08590, At4g09520) could also be

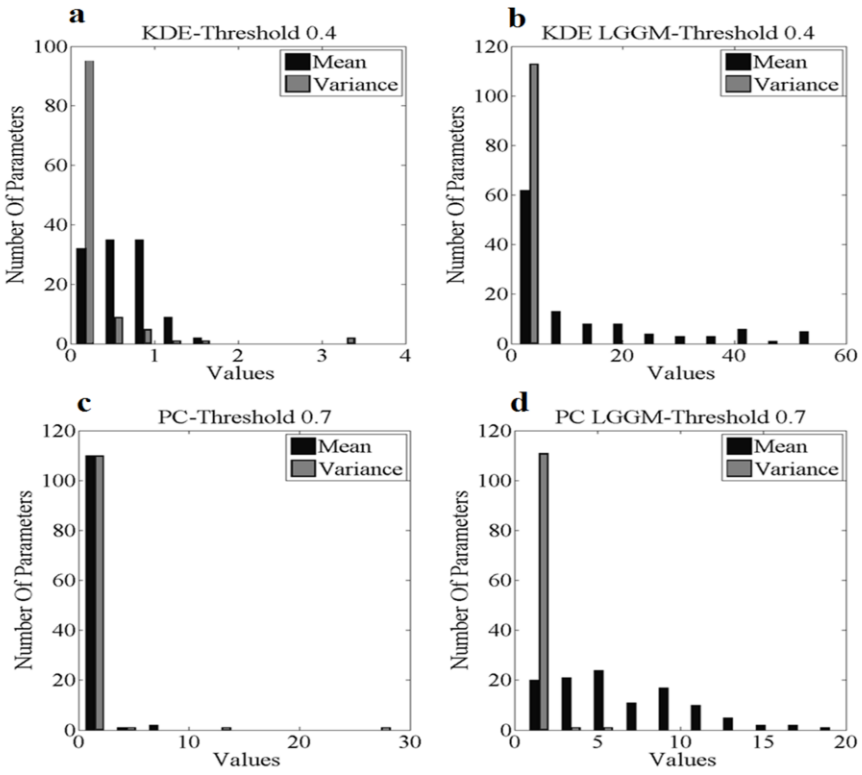


FIGURE 16: Histogram comparison between mean and variances of KDE and PC networks. **(a), (b)** Proposed methodology vs LGGM on KDE network; **(c), (d)** Proposed methodology vs LGMM on PC network. All frameworks show small variances for all nodes. However, with LGGM approach **(c, d)** the means have wider range while with the proposed methodology **(a, b)** means are close to zero similar to the experimental data. Notice that KDE **(a)** also attains very small means as the horizontal scale is much smaller than the other approaches.

considered as true positives, as they all catalyze the same reversible reaction (3-phospho-D-glycerate to 2-phospho-D-glycerate). In this form, the statistically derived interactions could substantiate valid assumptions for biological consideration.

In addition to indirect relations within the pathway, there are many indirect interactions of the selected pathway with external ones that affect the carbohydrate metabolism and need to be taken into consideration. Thus, necessary corrections on biological grounds can drastically decrease the false positive rate and increases the F-score. Nevertheless, we do not examine those interactions because our analysis focuses only on the molecular factors of the carbohydrate pathway. In fact, using ANAP (Arabidopsis Network Analysis Pipeline) [71], an interactive Web tool that contains protein interaction data information from 11 public Arabidopsis databases [72], we constructed a protein interaction network based on our studied genes/gene products as input and confirmed a number of 3,544 edges. This implies that our framework performs well in capturing genetic interactions, since for the proposed threshold of 0.4 the network constructed via KDE consists of approximately 3,000 edges.

7.2 DIRECT AND INDIRECT IMPLICATIONS OF ACTIVATION

The next step in evaluating algorithm's efficiency was the examination of the direct and indirect genetic implications for different time stages. For this purpose, we included in the studied carbohydrate pathway the 13 genes (7 'significant', 6 'unrelated' genes). The basic idea was to select a group of genes whose expression is known to affect the involved genes/proteins in the pathway. For the experimental values of those groups of genes, we predict which genes/proteins seem to be expressed (activated) or under-expressed (inhibited) and if those predicted direct and indirect associations are verified according to the findings of Hajduch et al. (2010) work [62].

Towards this direction, we consider the estimation of the conditional probabilities for the selected networks of PC and KDE. For the generated networks, the mean and variances are compared with the equivalent parameters of the LGGM approach. Figure 16a,b present the histograms of mean and variances for the computed probabilities. For both KDE and PC networks, the conditional Gaussian distributions according to our proposed method fluctuate tightly around low means, while the LGGM approach fluctuates over a wider range. This is due to the fact that conditional dependencies in LGGM are modeled by the sum of parental means, while with our modeling conditional distributions are more depended on the experimental data. This is also conducted by the expression data which show that the values of genes/proteins are low with small deviation.

We now proceed with the verification of expression profiles for the studied genes and proteins for five stages of growth, as presented in Figure 17. For this purpose we isolated the genes At2g01140, At3g03250, At5g52920, At1g73370, At5g47810, At5g56630, At3g55650, At4g29220 and At5g22510, which show different expressions during the five stages of growth. Furthermore, they hold an important role in carbohydrate metabolism and their study is expected to reveal an impact on the genes/proteins involved in the pathway. To determine the impact of expression levels, we pose inference queries conditioned on the observation of each of the above genes. For instance, the probability of gene At3g26650 to be inhibited when At2g01140 gene is activated is summarized as the conditional probability of the first given the expression level of the latter. In order to model the activation and inhibition, we set a threshold as a mean of the experimental values for each gene. We mention here that in our consideration, a gene can either be activated or inhibited. Thus, a gene/protein is considered inhibited if its expression is lower than the experimental mean of this gene/protein in the dataset; otherwise, it is labeled as activated. In order to also account for uncertainties in the close region of the threshold, the computation of the conditional probability takes soft bounds on activation/inhibition, considering the mean plus/minus one standard deviation, respectively. Moreover, the above mentioned genes are observed according to a random value in the range of activation/inhibition.

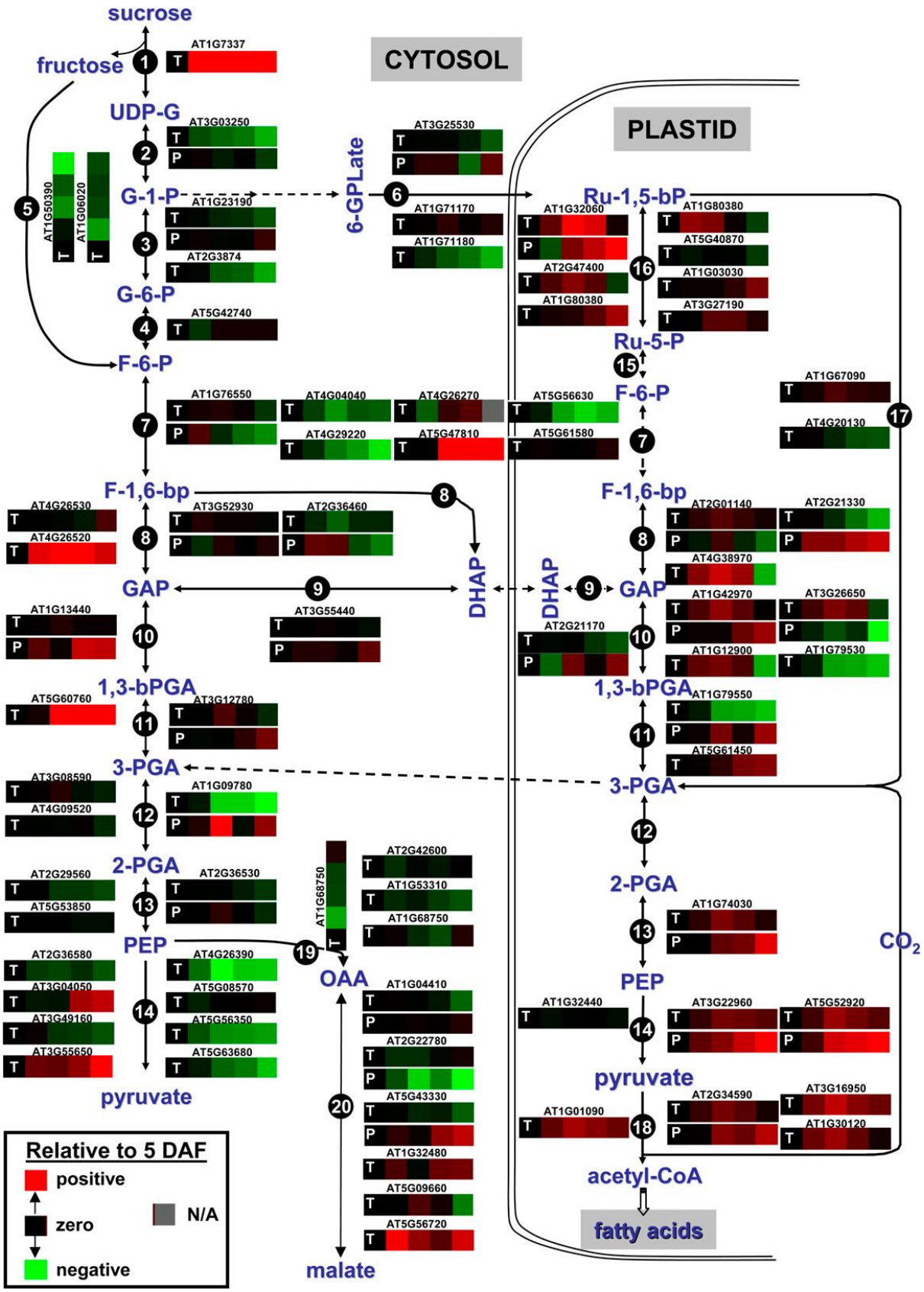


FIGURE 17: Schematic view of carbohydrate metabolism during seed filling of Arabidopsis. Expression (heat) maps of individual protein (P) and transcript (T) expression based on proteomics and microarray experiments as relative value to 5 DAF are shown. Protein/transcript pairs are

under one ATG number. Asterisks in red and bold depict the 9 observed genes. At5g22510 (top) is external to the pathway and considered as 'significant' gene. [With permission of the authors and journal. Originally published in: Hajduch M, Hearne LB, Miernyk JA, Casteel JE, Joshi T, Agrawal GK, Song Z, Zhou M, Xu D, Thelen JJ; Plant Physiol. 2010, 152(4), 2078-2087; doi/10.1104/pp.109.152413; www.plantphysiol.org; "Copyright © American Society of Plant Biologists".]

Table 4 summarizes the predicted expression levels from our proposed method on the KDE network. The first column shows the targeted genes for inference. The rest columns show the predicted expression profiles for the other involved genes/proteins in the pathway for all stages of growth. The presented expression profiles were selected as the most significant with the highest probability to occur. The 'related' gene At5g22510 appears in many cases to have opposite expression compared to At1g73370. Both genes are enzymes that catalyze the sucrose cleavage in plants but the final products of their enzymatic pathways are different in important aspects. Additionally, Table 4 presents the outcome of the predicted associations when At1g73370 and At3g03250 are simultaneously observed as activated and inhibited, respectively. There are many reasons for posing this query. We wanted to examine the robustness of the proposed method for a complex of important genes. Furthermore, queries with opposite genetic behaviors correspond to a more realistic interpretation. Last, we observe that genes At1g73370 and At3g03250 have particular biological significance in *A. thaliana* carbohydrate metabolism. More specifically, At1g73370 encodes a sucrose synthase, which is implicated in sucrose metabolism and is vital for homeostatic regulation between metabolic pathways and sucrose signals. In addition, genes that encode sucrose synthase enzymes are responsive to the action of their own enzyme products [64]. The At3g03250 gene encodes the UDP-glucose pyrophosphorylase, a key enzyme for carbohydrate metabolism that is essential in Arabidopsis [73]. It is reported that the At3g03250 gene is co-regulated with genes implicated in carbohydrate metabolism, late embryogenesis and seed loading [74] .

Our analysis of sparse experimental data in Table 4 allows the generation of gene-protein networks and illustrates three key points focusing on the outcome interactions of the 'significant' genes associated with the KDE method (Table 4). First, we observe that the target genes from the 1nd column interact with genes from other columns, most of which are involved in carbohydrate metabolism. These gene-pairs are indirectly interconnected according to ATTED-II [66]. Second, we highlight new gene-protein interactions between the 'significant' genes and proteins (4th column).We highlight two indicative examples, (i) fructose 1,6-biphosphate aldolase 6 (AtFBA6), which is a key enzyme in glycolysis and gluconeogenesis in plant cytoplasm and may have crucial role in stress and sugar signaling [75], and (ii) plastidial glyceraldehyde 3-phosphate dehydrogenase, A subunit (GAPA) that participates in the reductive carbon cycle and also is involved in response to sucrose stimulus [76]. Third, we reveal new gene-gene (direct or indirect) interactions between the target genes and the genes showed in other columns, including interactions with the seemingly 'unrelated' genes. Interestingly, the

'unrelated' gene At3g17520 has inference significance and is a member of the group 4 late embryogenesis abundant (LEA) protein genes [65]. The presence of their encoded LEA proteins is related to the adaptive response of higher plants caused by adverse conditions to maintain normal metabolism [77]. The observed gene-gene and gene-protein interactions between the various 'significant' genes with LEA gene or GAPA and FBA protein, should be experimentally analyzed in order to find their possible associations or cross-talks between carbohydrate metabolism and other pathways during seed development in *A. thaliana.*

In a related attempt of Hajduch et al. (2010)[62] to examine the behavior of genes and corresponding proteins, the expressions from 2nd to 5th stages of seed development is compared to the corresponding 1st stage (Figure 17). In the associated color map, red regions imply concentration increase compared to 1st stage, green regions indicate decrease, while black regions reflect no change in concentration. Regarding the expression of pairs in Figure 17 (26 pairs), the study of Hajduch et al. (2010) based on linear regression, reveals disagreement with the heat-map in some gene-protein pairs with opposite expression profiles. This effect is also derived from our approach for the gene-protein pairs At2g21330, At3g52930, At3g26650, At2g36460, At1g13440 and At1g76550, as presented in Table 4. For the remaining gene-protein pairs, the predicted expressions from our model attain low probabilities, with one exception of the At2g21170 pair that expresses discordance of expressions in time and is attributed to post transcriptional regulation.

In the last section we compare the performance of our proposed method with LGGM applied on the KDE and PC networks. The genes At2g01140, At3g03250, At5g52920, At1g73370, At5g47810, At5g56630, At3g55650, At4g29220 and At5g22510 are simultaneous observed due to their importance in the involved biological processes into the pathway. Table 4 presents the predicted expressions of our proposed method on

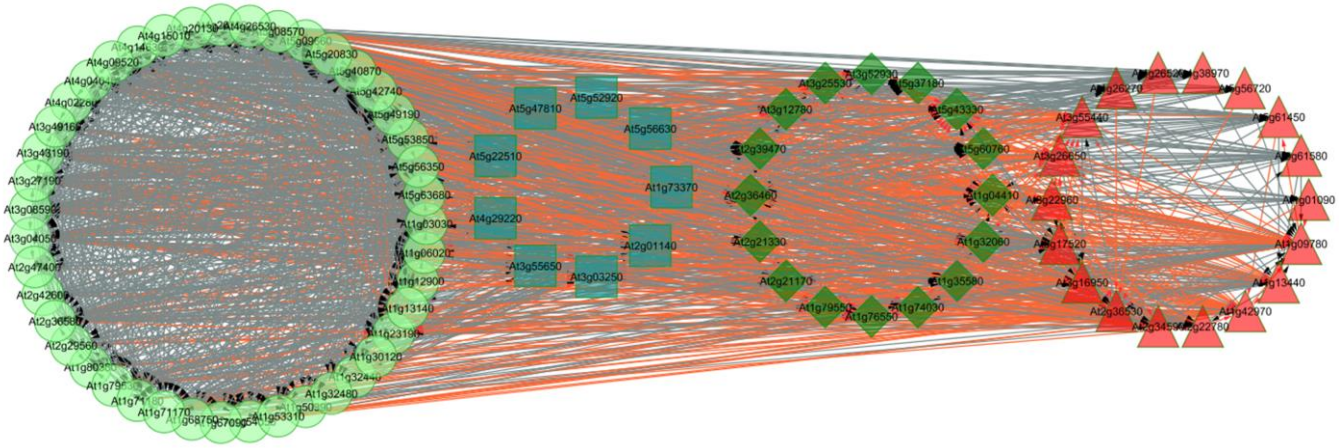


FIGURE 18: Predicted expressions for the nine observed genes (square blue) for all stages of development. Inhibited molecules are marked as green diamonds and activated as red triangles. Genetic interactions are presented as grey and proteomic as orange. [Created by Cytoscape Platform].

KDE network. The predicted outcomes indicate high probabilities and same expression profiles for all stages of development.

Interactions of observed with other genes in the KDE network at different time points. The observed genes are selected based on their inference significance

Observed Genes	Predicted interactions/Day-KDE (0.4)									
	1st Day		2nd Day		3rd Day		4th Day		5th Day	
At1g73370	At2g21330↑	p2225↓(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)
	At3g52930↑	p2173↓(At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)
	At3g26650↑	-	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)
	-	p496↑ (At2g36460)	At2g36460↑	p496↑(At2g36460)	-	p496↑(At2g36460)	At2g36460↓	p496↑(At2g36460)	-	p496↑(At2g36460)
	At2g21170↑	p2392↓(At2g21170)	At2g21170↑	-	At2g21170↑	p2392↓(At2g21170)	At2g21170↑	-	At2g21170↑	p2392↓(At2g21170)
	At2g01140↓	-	At2g01140↓	-	At2g01140↓	-	At2g01140↓	-	-	p2218↓(At2g01140)
	At3g17520a↑	p2322↓(At3g17520)	-	p2322↓(At3g17520)	At3g17520a↑	p2322↓(At3g17520)	-	p2322↓(At3g17520)	At3g17520a↑	p2322↓(At3g17520)
	-	p1877↓(At3g22960)	-	p1877↓ (At3g22960)	-	p1877↓ (At3g22960)	-	p1877↓(At3g22960)	-	p1877↓(At3g22960)
	-	-	-	p512↓(At1g13440)	At1g13440↓	-	-	p512↓(At1g13440)	At1g13440↑	-
	-	-	At1g76550↓	p175↓(At1g76550)	At1g76550↓	-	At1g76550↓	p175↓(At1g76550)	At1g76550↓	-
	At3g25530↑	-	-	-	At3g25530↓	-	At3g25530↓	-	At3g25530	-
	At1g42970↑	-	-	-	At1g42970↑	-	-	-	At1g42970↑	-
	At3g12780↑	-	-	-	At3g12780↑	-	At3g12780↑	-	At3g12780↑	-
	At5g47810↑	-	At5g47810↑	-	At5g47810↑	-	At5g47810↑	-	-	-
	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-
	At5g60760↑	-		-	At5g60760↑	-	At5g60760↑	-	At5g60760↑	-
	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-
	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-
	At1g50390↑	-	-	-	At1g50390↓	-	At1g50390↓	-	At1g50390↓	-
	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-
	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-
At5g22510	At2g21330↓	p2225↓(At2g21330)	At2g21330↑	p2225↓(At2g21330)	-	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)
	At3g52930↑	-	At3g52930↑	-	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)
	At3g26650↓	-	At3g26650↓	-	At3g26650↓	-	-	-	-	-
	At2g36460↓	p496↑(At2g36460)	At2g36460↑	p496↑(At2g36460)	-	p496↑(At2g36460)	At2g36460↑	p496↑(At2g36460)	-	p496↑(At2g36460)
	At2g21170↑	-	At2g21170↓	-	At2g21170↓	-	At2g21170↓	-	At2g21170↓	-
	At2g01140↓	-	At2g01140↓	-	At2g01140↓	p2218↓ (At2g01140)	At2g01140↓	-	-	p2218↓(At2g01140)
	At3g17520a↑	-	-	-	At3g17520a↑	-	-	-	At3g17520a↑	-
	-	p512↑(At1g13440)	-	-	-	p512↓(At1g13440)	-	-	At1g13440↑	p512↓ (At1g13440)
	At1g42970↑	-	-	-	At1g42970↑	-	-	-	At1g42970↑	-
	At3g12780↑	-	-	-	At3g12780↑	-	At3g12780↑	p2124↓ (At3g12780)	At3g12780↑	p2124↓ (At3g12780)
	At5g60760↑	-	-	-	-	-	-	-	-	-
	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-
	At4g02280↑	-	At4g02280↑	-	-	-	At4g02280↑	-	At4g02280↑	-
	-	-		-	At1g32060↓	-	At1g32060↓	-	At1g32060↓	-
	-	-	At1g32060↓	p2035↓(At1g74030)	-	-	-	p2035↓(At1g74030)	-	p2035↓(At1g74030)
	At4g38970↓	-	-	-	-	-	At4g38970↓	-	At4g38970↓	-
	At1g73370↑	-	-	-	-	-	At1g73370↑	-	At1g73370↑	-
			At1g73370↓							
At1g73370 At3g03250	At2g21330↓	p2225↑(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)
	At3g52930↓	-	At3g52930↓	-	At3g52930↓	p2173↑(At3g52930)	At3g52930↓	p2173↓(At3g52930)	At3g52930↓	p2173↓(At3g52930)
	At3g26650↓	-	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)
	-	p496↓(At2g36460)	-	-	-	p496↓(At2g36460)	At2g36460↑	p496↓(At2g36460)	-	p496↓(At2g36460)
	At2g21170↓	p2392↑(At2g21170)	At2g21170↓	-	At2g21170↓	p2392↑(At2g21170)	At2g21170↓	-	At2g21170↓	p2392↑(At2g21170)
	At2g01140↑	-	At2g01140↑	p2218↑(At2g01140)	At2g01140↑	p2218↓(At2g01140)	-	p2218↑(At2g01140)		
	At3g17520a↓	p23222↑(At3g17520)	-	p23222↑(At3g17520)	At3g17520a↓	p23222↑(At3g17520)	-	p23222↑(At3g17520)	At3g17520a↓	p23222↑(At3g17520)
	-	-	-	p512↑(At1g13440)	At1g13440↑	-	-	p512↑(At1g13440)	At1g13440↓	p512↑(At1g13440)
	At1g42970↓	-	-	-	At1g42970↓	-	-	-	At1g42970↓	-
	At3g12780↓	-	-	-	At3g12780↓	-	At3g12780↓	-	At3g12780↓	-
	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-
	-	-	At1g76550↓	p175↑(At1g76550)	At1g76550↓	-	At1g76550↓	p175↑(At1g76550)	At1g76550↓	p175↓(At1g76550)
	At1g50390↓	-	At1g50390↓	-	At1g50390↓	-	At1g50390↓	-	-	-

At1g09780↓ - - - At5g49190↑ -	At1g09780↓ - - - At5g49190↑ -	At1g09780↓ - At5g63680↓ At5g49190↑	At1g09780↓ - At5g63680↓ - At5g49190↑ -	At1g09780↓ - At5g63680↓ - At5g49190↑

TABLE 4 : Predicted interactions from inference

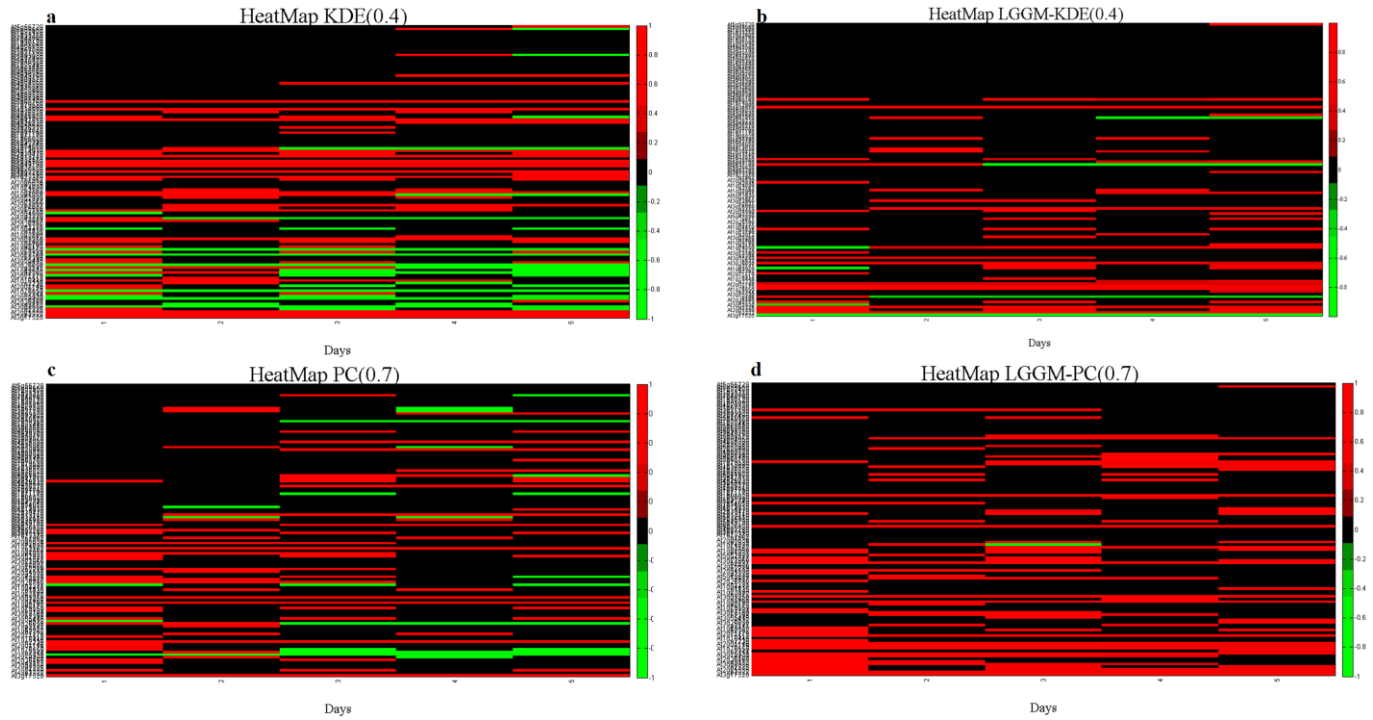


FIGURE 19: Heat maps of the predicted expressions for all stages of seed development. White regions imply inhibition, grey regions imply activation, and black regions reflect none predicted expression. **a)** Proposed method on KDE network; **b)** LGGM method on KDE network; **c)** Proposed method on PC network; **d)** LGGM method on PC network.

Figure 19a and Figure 19b show how the heat-maps of the predicted expression profiles for all 5 days applying our proposed method on the networks produced by KDE and PC. Figure 19c and Figure 19d present the results after applying the LGGM approach on the respective networks. Regions on the heat maps marked as red represent the predicted activation while green regions the predicted inhibition. The intensities for both cases reflect signed probabilities with the inhibited predictions set as negative values. All predicted outcomes were chosen with probabilities higher than 0.4 while black regions in the heat maps indicate cases with probabilities smaller than the above threshold. All cases represent predicted results of the observation of the nine above mentioned genes. Clearly, the proposed method enables both networks (constructed by KDE and PC) to achieve higher numbers of predicted interactions compared to the LGGM approach. Moreover, while our model captures inhibited in addition to activated expressions, the LGGM approach fails in identifying expressions with high probabilities for both types of networks. This illustrates another aspect of the proposed method as it predicts expression of genes for both activation and inhibition with high probabilities.

In order to validate the above observations we compare the 4 derived heat-maps with the results of Hajduch et al. (2010). More specifically, we compute the precision of our outcome in relation to the results of Figure 17 and set each predicted expression as true (false) positive if it agrees (disagrees) with the corresponding prediction of Figure

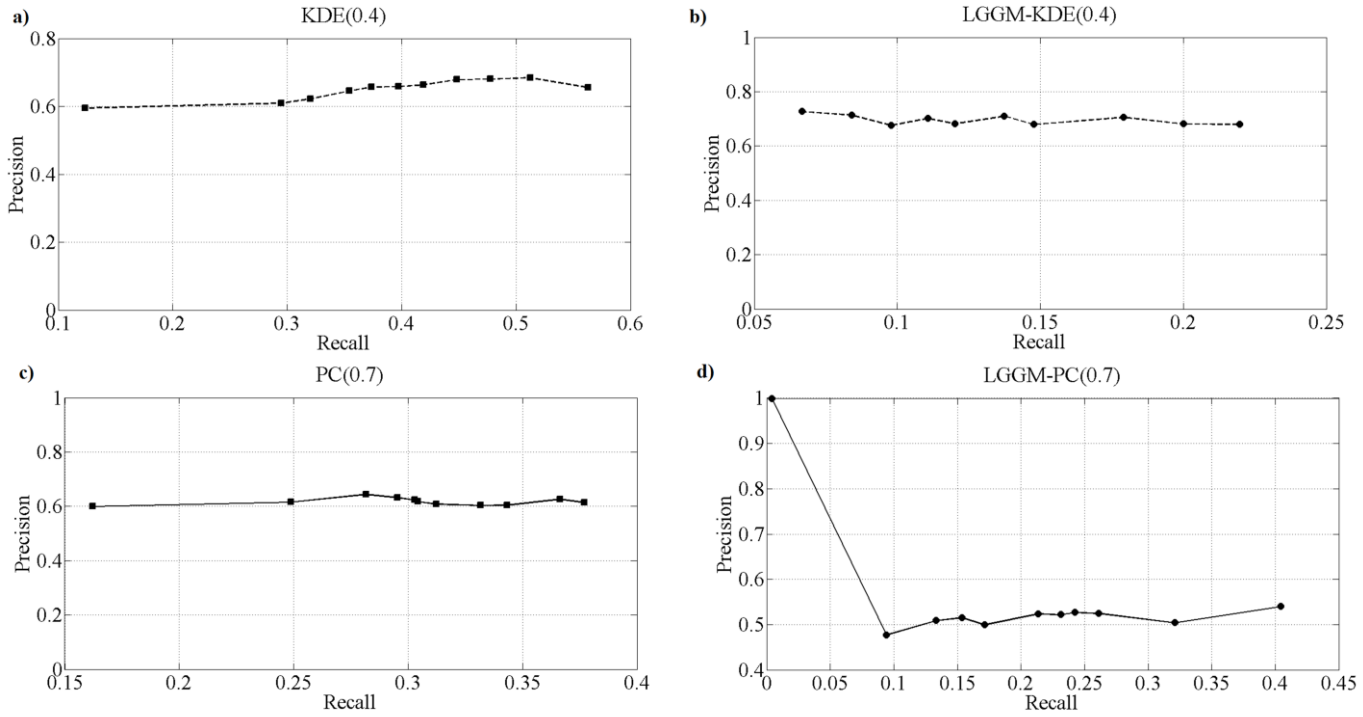


FIGURE 20: Precision-recall curves for different levels of probability according to the het-maps. The rightmost point in each figure reflects probability over 0.4 while the leftmost point is for probability 1; increase in probability is for 0.5. **a)** Proposed method on KDE network; **b)** LGGM method on KDE network; **c)** Proposed method on PC network; **d)** LGGM method on PC network

17. However, the true positive relations are difficult to define in the entire dataset. In our consideration, they involve only the direct connections in Figure 17 to the observed genes/proteins, in addition to their neighboring molecules. The main rationale for this assumption is that adjacent molecules are expected to engage in similar biological processes and interact with the observed ones.

Figure 20 describes the precision-recall curves on KDE and PC networks for different levels of probability, higher than 0.75. Right-most points indicate probability higher than 0.6, while left-most points reflect probability bound equal to 1. Notice that none framework can reach precision level higher than 0.75. This is an expected outcome because many of the involved genes considered as true positives do not show discriminant expressions (activated/inhibited). In essence, we cannot identify these interactions because the experimental data do not clearly reflect information on activation/inhibition. The proposed method applied on the KDE network (Figure 20a) reaches higher levels of precision and recall that other approaches, with higher probabilities. It can identify successfully up to 40% of the true positives with probability 1 and as the subset of the predicted interactions is augmented (lower thresholds on probability), the precision also rises reaching its highest score at probability 0.5. This implies that the proposed framework is able to discriminate the expression profiles giving high scores to true positives. In contrast, while other approaches show similar

precision scores they fail in reaching high recall. This implies that the amount of predicted interactions is far less than the expected number of true positives, which also increases the number of false negatives. Overall, our method outperforms others in revealing predicted temporal expressions in terms of recall and precision, in addition to discriminating activated and inhibited genes/proteins. Interestingly, while the LGGM-KDE (Figure 20b) approach has low recall, it has higher precision levels compared to LGGM-PC (Figure 20c) and PC (Figure 20d). This suggests that the network structure based on KDE has successfully captured many biological interactions verifying the superiority of KDE over PC.

7.3 CONCLUSION

Clearly, the KDE approach models quite well the verified associations between the participating genes/proteins, as the majority of the genes/proteins are located close to the processes of the carbohydrate metabolism pathway. On the contrary, the PC approach appears to capture less of those associations. Thus, our results indicate that KDE performs better on the of network construction. This supports the aforementioned statement that KDE is resilient in modeling the genetics associations with sparse experimental data. Considering the modeling of conditional dependencies, both heat maps and precision curves prove that genetic associations enclose more complex dependencies, whereas linear Gaussian approaches lack the ability to model such relations. Ongoing research is under investigation in an attempt to reveal the potential benefits of this methodology on human cancer, so as to highlight important gene profiles of the participant genes in dysregulated pathways in cancer diseases (oral, breast) .

Perhaps the most important contribution of this study is the provision of a different perspective in revealing the identity of genetic interactions. More specifically, network construction studies have proven that extracting direct genetic interactions is a far more complex problem than the one studied with simplifications in the different layers of genetic information, explaining the poor results obtained on precision especially in complex organisms. The direct interactions are to a large extent unknown, especially if we take into account all the possible pathways that affect groups of genes. In addition, the available knowledge of direct interactions is established under specific conditions, which also seem to change when abnormalities happen. These issues imply the need to reexamine the generation mechanism for expression profiles in relation to underlying genetic factors and their direct or indirect relevance in specific pathways. In this direction, our approach enables the verification of relations in the expression profiles from the underlying interactions, and can be used as a first step in studying whether indirect effects of important genetic molecules verify to a good extent the expression profiles of genes involved in the pathway.

At this point we need to address an important issue associated with the thresholds on the bounds of activation and inhibition. The consideration of the sample means from

experimental data is a naïve approach, since it is drastically influenced by the frequency of activations/inhibitions in the data; if one gene is far more often activated than inhibited, the bounds of inhibition cannot be computed reliably. Thus, the mean is not always an acceptable and robust threshold. In our study, this was not an obstacle due to the scarcity of the data, but in datasets with a respectable amount of samples this has to be examined carefully, since inappropriate thresholds may give different results in terms of activation and inhibition. Subsequently, this would affect the precision in identifying activated and inhibited expressions.

8. BREAST CANCER DATASET

In this chapter we apply the proposed methodology to the human organism and especially on breast cancer disease. Using existing knowledge on molecular pathway organization, we apply the proposed methodology as described in Section 6.2.1 to find differences, such as up and down regulations, between different populations. Based on a previous study [78], we will attempt to examine if the predicted up and down regulations are consistent with the results of Koumakis et al. approach. For this reason we will follow a similar concept with Koumakis et al. analysis so as to have the same base of discussion.

More specifically, we focus only on two populations which are grouped as estrogen responsive positive (ER+) and estrogen responsive negative (ER-). Most breast cancer (BC) cases are estrogen responsive, a series of growth promoting pathways are activated, for example, ErbB signaling GRN. In an effort to reveal the underlying regulatory mechanisms that govern BC patients' treatment responses we applied the presented methodology on a set of three independent gene-expression studies targeting the ER phenotypic status of the respective patients, i.e., ER+ vs. ER- . The details of the gene-expression data from the three studies are: GSE7390 (the GEO-Gene Expression Omnibus study code), 286 patients; GSE2990, 183 patients; GSE3494, 247 patients;[78].

We isolated 2 pathways for these purposes which are signal transduction pathways associated with cancer. These pathways describe how molecules interact, thus based on these graphs we will take for granted the basic genetic structure. Among many pathways responsible for cancer development we only focused on the most known for BRCA: the ErbB (Kegg code: hsa04012), mTOR (Kegg code: hsa04150) according to KEGG database. Figure 21 shows the ErbB and mTOR pathways with the total involved molecules. Totally, from the selected pathways 121 genes including their closely related human genes, such as the three RAS genes (Hras, Nras, Kras) or encoded protein family members (e.g. Crk and CrkL) and protein isoforms (e.g. AKT1, AKT2 and AKT3), were studied.

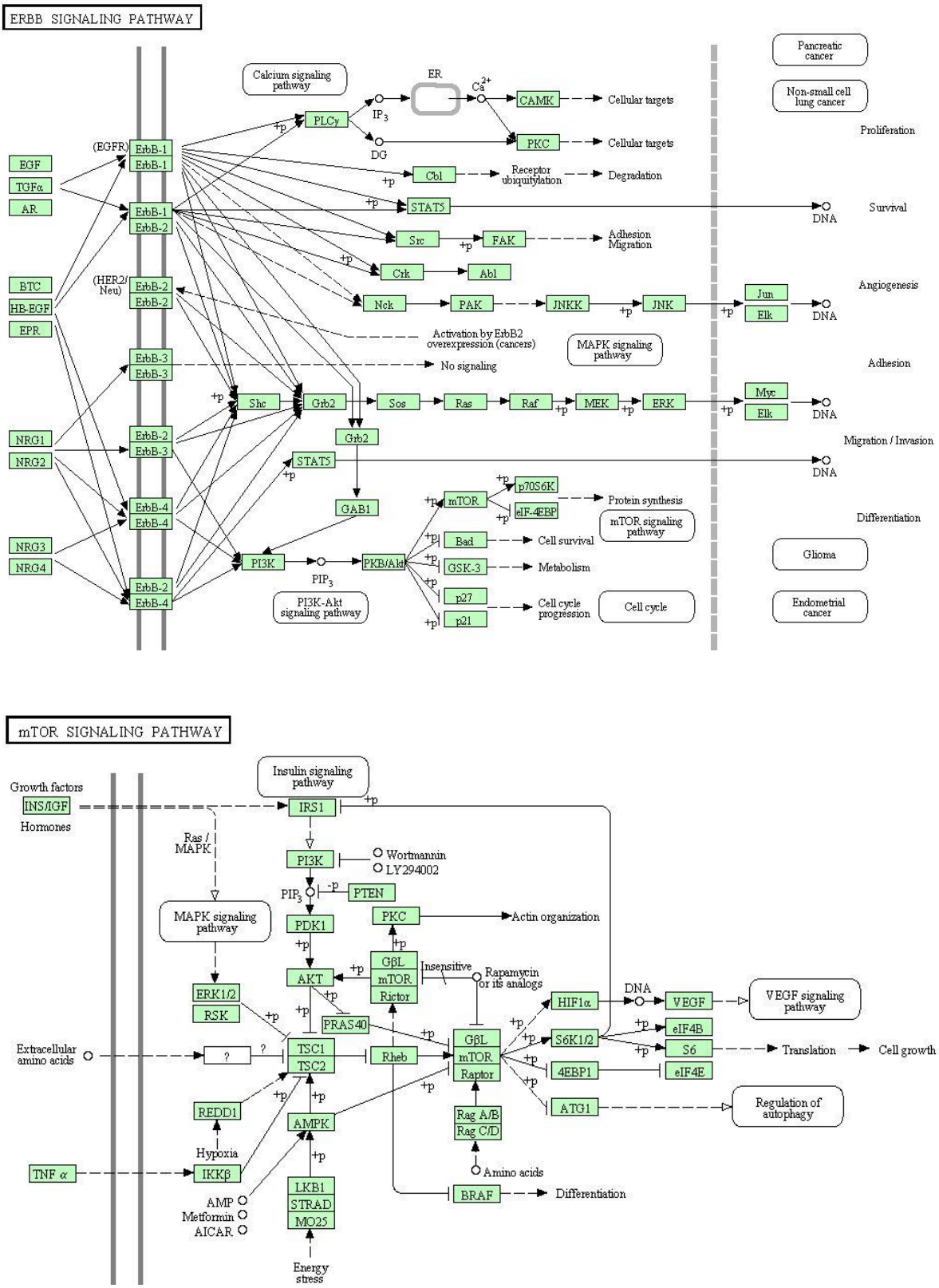


FIGURE 21: ErbB signaling and mTOR pathways

8.1 EXPERIMENTAL SETUP

Combining the knowledge from three different datasets –GSE7390, GSE2990 and GSE3494- we made a distinct separation between the ER+ and ER- samples. Totally, 123 ER- and 44- ER+ samples were isolated from all datasets and based on these observations we estimated the cpd for all 121 genes in the above mentioned pathways. Also, each dataset contains a multiple number of probes that correspond to the same gene. These id probes are sample repetitions for all genes using a different probe. For this reason we only took into consideration one randomly selected probe id that corresponds to each gene.

In order to estimate probabilities of up and down regulation we must first define specific expression thresholds for all genes. Expression values higher than these thresholds correspond to up regulation while values smaller values correspond to down regulation. For this purpose we followed the approach as analyzed in Koumakis et al. [78], where for each gene is computed a threshold according to the frequency of the samples. Thus, for gene we took one threshold as the mean of the computed thresholds of all datasets after applying the Koumakis et al algorithm.

Probe	GSE2990	GSE3494	GSE7390
211550_at	6,373277	3,224526	3,461737
209951_s_at	5,679941	5,813863	7,246689
201466_s_at	6,187776	7,829451	9,916724
211665_s_at	7,785601	5,866292	8,233298
212240_s_at	9,72953	9,090286	10,38258
205015_s_at	6,242149	2,681882	3,299597
216551_x_at	7,642446	6,956974	7,8291
206794_at	4,89411	4,938971	6,30431
202123_s_at	10,07357	8,44693	10,37818
202431_s_at	8,599652	8,505663	7,589204
209956_s_at	4,488093	5,469424	4,622889

	ER	ER+	ER+
event,rfs		0	1
time,rfs		2,580821918	0,58082192
event,dmfs		0	1
time,dmfs		2,583561644	0,58356164
ID_REF		GSM65316	GSM65317
1007_s_at		12,123249	12,123692
1053_at		6,574572	6,920157
117_at		7,163689	7,710973
121_at		9,192505	9,251458
1255_g_at		4,390527	4,368292
1294_at		7,784241	8,375649

TABLE 5: **(LEFT)** Computed thresholds for each dataset and for all genes. **(Right)** The expression values for each gene grouped as ER+ and ER- patients.

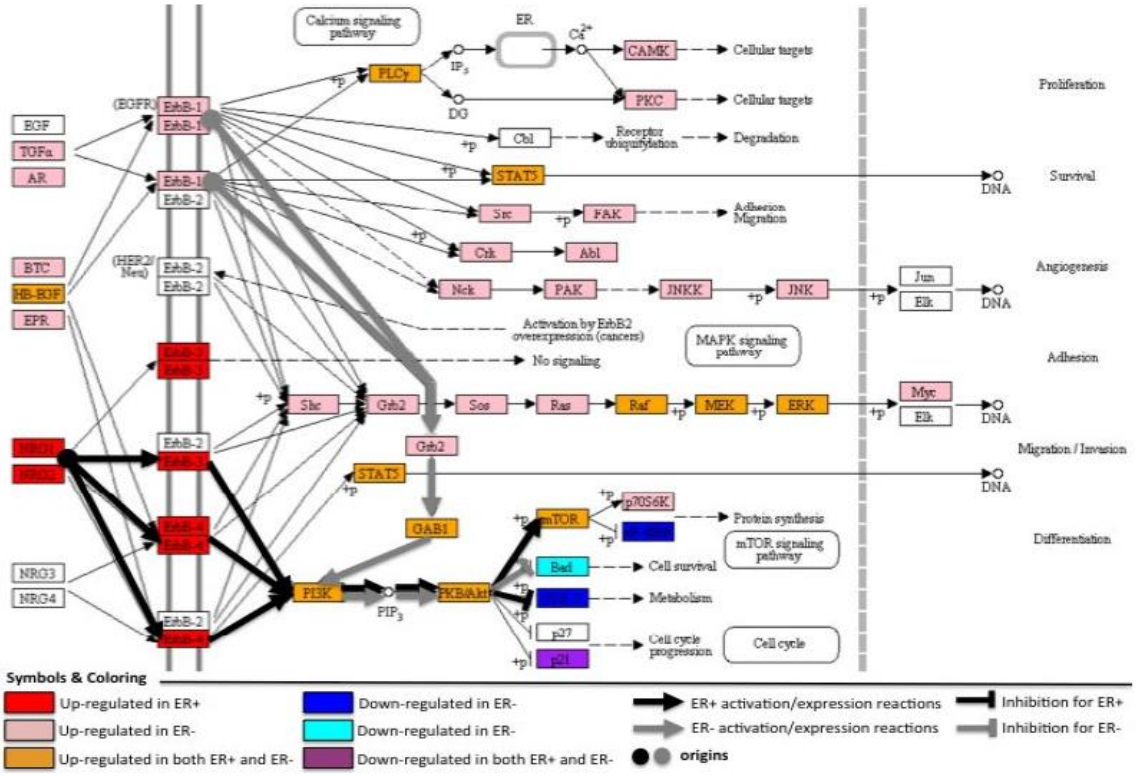
8.2 INFERENCE QUERIES-RESULTS

The query selection was based on the findings of Koumakis et al research. According to Table 6 there are obvious differences between ER+ and ER- patients as far the up and down regulations are concerned. The results are summarized as [78]:

- The ER- path originates from TGFα (transforming growth factor, alpha), AR (amphiregulin), BTC (betacellulin), and EPR (epiregulin) epidermal growth factors that activate both ErbB-1 and ErbB-2 receptors; then, the two receptors initiate the path GRB2 → GAB1 → PI3K → PKB/Akt that guides to the activation of

mTOR that activates p70S6K which signals “protein synthesis”, and inhibits BAD which signals “cell survival”;

- The ER+ path originates from the neuregulins NRG1, NRG2 (neuregulin1,2) that that bind and activate ErbB-3 and ErbB-4 followed by the PI3K → PKB/Akt activation reaction which is also part of the ER- path. But now, PKB/Akt acts just as an inhibitor of GSK-3 and blocking of “Metabolism”. Moreover, PKB/Akt activates mTOR, which now acts as an inhibitor of EIF-4EBP with the result of blocking “protein synthesis”. According to the recent biomedical literature the aforementioned results are quite relevant to the estrogen-receptor status - we focused our exploration on the mechanisms underlying the resistance to pure estrogen antagonist



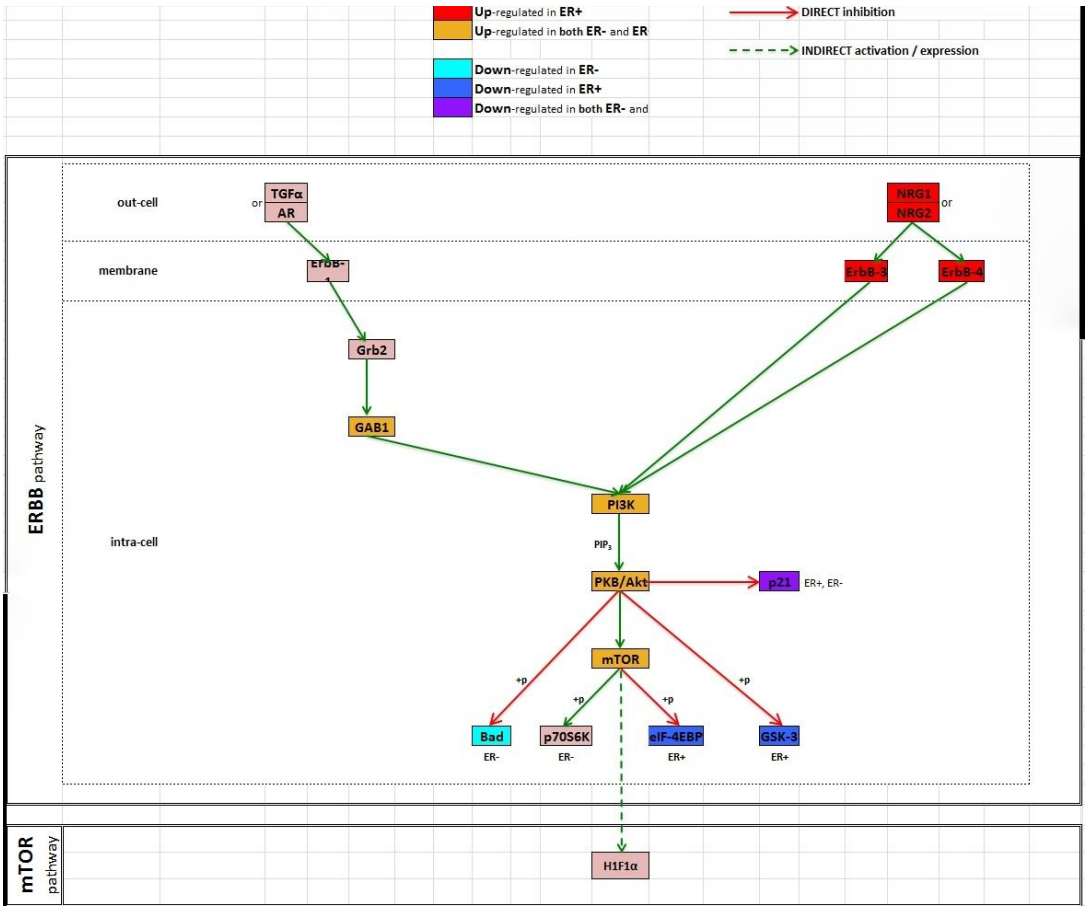


TABLE 6: Activation or Inhibition for the ER+ and ER- patients for the ErbB signaling pathway. Results of Koumakis et al research.

Based on the above findings we aim to validate whether these results are verified following our analysis. For this reason, we organize the posed queries into four groups of observed genes:

1. Which molecules are activated or inhibited, when EGF is observed
2. Which molecules are activated or inhibited, when **pink** genes are activated in the ER- patients (**TGFa,AR,ErbB1, ErbB2, Grb2**)
3. Which molecules are activated or inhibited, when **red** genes are up regulated activated in the ER+ patients (**NRG1,NRG2,ErbB-3, ErbB-4**)
4. Which molecules are activated or inhibited, when **brown** genes are activated both in the ER+ and ER- patients (**GAB1,mTOR,PI3K,PKB/Akt**)

For instance, if we want to examine whether EGF is responsible for differences between ER+ and ER- patients, we compute the probability of activation or inhibition for all the other molecules in the pathway. If the produced probability is higher than 0.5, then we consider EGF as an important factor that affects the signaling in the pathway. If the equivalent queries result to differences between ER+ and ER-, we consider EGF as an important factor for signaling alterations. Finally, for each of the above queries are included : a) the closely related human genes, and/or b) the family members, and/or c) the isoforms.

Observed gene	Activation ER+ ER-	Inhibition ER+ ER-	Activation ER-	Inhibition ER-	Activation ER+	Inhibition ER+
Koumakis et al study	**GAB1; PI3K (**PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIK3R5**); PKB/Akt (AKT1,** AKT2, AKT3**); MTOR**	**CDKN1A**	TGFα/AREG; EGFR; GRB2; p70S6K (RPS6KB1, **RPS6KB2); HIF1A**	BAD	**NRG1/ NRG2; ERBB3; ERBB4**	**GSK3B; EIF4EBP1**
EGF	EGFR (EGF/AREG); PIK3CA, PIK3CB, PIK3CD, PIK3R1, PIK3R5; AKT1, AKT2, AKT3; MTOR; RPS6KB1; BAD; GSK3B; GAB1;CDKN1B, CDKN1A; NRG1; HIF1A	RPSK6KA2; ERBB3; ERBB4; NRG2; BTC	PIK3CG	PIK3CB, PIK3R2, PIK3R3	PIK3CB, PIK3R2, PIK3R3	PIK3CG
GAB1-PI3K-PKB/Akt-mTOR (Brown)	EIF4B, EIF4E; RPS6KB1, PRS6KB2, RPS6KA1, RPS6KA3, RPS6KA6; GRB2; EGFR; TGFA; AREG; EREG; NRG1	ERBB3, ERBB4; BAD; GSK3B; CDK1A, CDK1B; HIF1A	NRG2; EIF4E2, EIFEBP1; ERBB2; BTC	-	-	ERBB2; EIF4EBP1, EIF4B4E2; BTC; NRG2
TGFα/AREG-ERBB1-GRB2 (Pink)	PIK3CA, PIK3CG, PIK3CD, PIK3R1, PIK3R3; PIK3R5; AKT2, AKT3; MTOR; RPS6KB1; RPS6KA1, RPS6KA3, RPS6KA6; BAD; GSK3B, CDKN1B, CDKN1A; BTC; NRG1, NRG2, HBEGF; HIF1A	RPS6KA2; EREG	-	AKT1; PIK3CB, PIK3R2	AKT1; PIK3CB, PIK3R2	-
NRG1/NRG2-ERBB3-ERBB4 (Red)	PIK3CA, PIK3R5, PIK3R1, PIK3R2; AKT1, AKT2, AKT3; MTOR; RPS6KB1, RPS6KA1, RPS6KA3, RPS6KA6; BAD; GSK3B; CDKN1B;TGFA, AREG; EREG; EGFR, ERBB2; GRB2; GAB1; EIF4B, EIF4BP1; EIF4E, EIF4E2; HIF1A	RPS6KA2	HBEGF ; PIK3CD, PIK3CG	PIK3CB, PIK3R3; BTC	PIK3CB, PIK3R3; BTC	HBEGF; PIK3CD, PIK3CG

TABLE 7: Experimental results of the proposed analysis. The 1st column describes the posed queries, while the other columns the main regulation profile for ER+ and ER- patients.

Ideally, we expect to predict that the groups in Table 6 have the same behavior in ER- and ER+ patients a. This means, for example, that we expect to find the pink genes as activated only in ER- patients or that BAD is inhibited only in the ER-. Table 7 describes the results after applying the proposed methodology in Section 6.3. The first column shows the posed query while the other columns the main genes that formulate the expression profile of the ER+ and ER- patients. The red indications in the table define the common genes between the proposed methodology and Koumakis et al analysis.

Results focusing at PI3K pathway of ErbB signaling as presented in Table 6 show that some regulation activation/inhibition profiles are in concordance with the observations of Koumakis et al analysis (bolded genes).

Also, there are many genes that are both activated in ER+ and ER- and show that are common factors that affect the expression in the two populations. Specific differences, such as BTC, NRG1 and NRG2 which are placed at the begging of the pathway have different profile between the two populations. This may be indicative of therapeutic uses as these genes seem to be responsible for the activation of specific molecules in PI3K signaling pathway.

In general, it has shown that the PI3K pathway is activated in a diversity of malignancies including breast, ovarian, endometrium, and other tumor types. Moreover, the influence of PI3Ks proteins in oncogenesis has been validated by several studies indicating that aberrations in this pathway are potential causes of cell transformation and, more significant, that PI3K pathway inhibition causes tumor regression. Given that PI3K/Akt drives proliferation as well as tumor cell survival, it is perhaps expected that tumor cells endeavor to maintain constitutive activation of this pathway. In human breast cancer, the PI3K signaling network is affected at different levels. More than 70% of breast tumors have molecular alterations in at least one component of this pathway. PIK3CA mutations, and mutations or other aberrations at the level of PDK1, AKT1, AKT2, and p70S6kinase are some of the known mechanisms that activate the pathway [79][80].

Furthermore, the proposed methodology provides a useful tool for the discovery of new activated or inhibited molecules that may reveal novel mechanisms of PI3K pathway activation in BC subtypes. Perhaps the most important outcome of this research is that there are distinct expression profiles in ER+ and ER- patients. More precisely, apart from the common activated genes (brown), all the other observations indicate that activated genes in ER+, are inhibited in ER-.

It is well established the major importance of estrogen and progesterone receptors (ER and PR, respectively) in the development and progression of breast cancer, as well as the association of ER/PR reduced expression with poor response to antiestrogen therapy and worse prognosis [80]. A bidirectional cross talk, where the PI3K pathway affects the levels and activity of ER, and the endogenous membrane ER can stimulate growth factor receptors (GFRs) and PI3K/AKT pathway [80], indicate the significance of

the above observations. Recent studies provide also evidence that the frequency and type of PI3K pathway aberrations vary among the different breast cancer subtypes, such ER+/ER- status [80][81][82]. As referred by Hernandez-Aya και Gonzalez-Angulo (2011) each molecular aberration may have a different clinical impact depending on the breast cancer molecular background, the presence of other aberrations, and the treatments received. The authors also denote that the BC genetic heterogeneity and likely different cell origin for each tumor subtype make necessary an irrespective analysis of the PI3K pathway alterations by tumor subtype [80].

Another important aspect of our study is the ability to focus to specific down and upstream effectors of the PI3K signaling pathway, as they comprise potential targets for drug development in BC. Nowadays, there are many agents that inhibit the network at various levels and used alone or in combination with chemotherapy, radiation, or other targeted therapies are developed and being evaluated continuously in preclinical and clinical trials, such as PI3K selective inhibitors, AKT inhibitors, Rapamycin analogs, Dual PI3K/mTOR, and mTOR kinase inhibitors [80].

Thus, as presented here, it is important to recognize the impact of specific signals (activation/inhibition) through the PI3K pathway, and to identify key effectors of the entire PI3K pathway in the different subtypes of breast cancer. This may be very useful for a targeted therapy at different levels of PI3K pathway.

9. ORAL CANCER DATASET

Biological networks in living organisms can be seen as the ultimate means of understanding the underlying mechanisms in complex diseases, such as oral cancer. During the last decade, many algorithms based on high-throughput genomic data have been developed to unravel the complexity of gene network construction and their progression in time. However, the small size of samples compared to the number of observed genes makes the inference of the network structure quite challenging. In this study, we propose a framework for constructing and analyzing gene- networks from sparse experimental temporal data and investigate its potential in oral cancer. We use two network models based on Partial Correlations and Kernel Density Estimation, in order to capture the genetic interactions. Using this network construction framework on real clinical data of tissue and blood at different time stages, we identify common disease-related structures that may decipher the association between disease state and biological processes in oral cancer. The analysis emphasizes an altered MET (hepatocyte growth factor receptor) network during oral cancer progression. In addition, we demonstrate that the functional changes of gene interactions during oral cancer progression might be particularly useful for patient categorization at the time of diagnosis and/or at follow-up periods.

9.1 STATISTICAL RESULTS

In order to investigate the statistical properties of the proposed methodology, we apply the PC and KDE approaches to reveal network structure from gene expression data. In the previous Section 7[83][82][81][80][79], our framework was applied on the prototype organism *Arabidopsis thaliana* on developing seeds harvested at 5, 7, 9, 11, and 13 days after flowering. This analysis gave a clear advantage for KDE over PC in revealing gene-gene and gene and/or protein associations. In this study, we examine the biological performance on the human organism for the oral cancer disease. We compare the performance of both algorithms and investigate the biological implications of our results.

9.1.1 DIRECT INTERACTIONS

Table 8 presents the number of gene interactions on blood samples, for the first follow-up. Accordingly, Table 9 presents the gene associations on tissue samples. Both tables present the performance of PC and KDE. The first column lists different thresholds on partial correlation set on PC for EQ. 15, while the second column provides the thresholds of similarity of EQ. 14 for KDE. Columns 3 to 6 summarize the verified numbers of direct and indirect gene to gene interactions for both approaches. Columns 7 and 8 present the number of new edges that have occurred for each threshold,

Threshold		Verified Gene Interactions				New Edges		Oriented Edges	
PC	KDE	PC		KDE		PC	KDE	PC	KDE
≥0.1	≤0.6	**1167**	(42/63)	**1**	(0/63)	3166	1	279	1
≥0.15	≤0.7	**957**	(34/63)	**75**	(4/63)	2551	108	185	70
≥0.175	≤0.75	**848**	(30/63)	**129**	(5/63)	2234	202	181	85
≥0.2	≤0.8	**738**	(27/63)	**167**	(7/63)	1968	347	172	92
≥0.3	≤0.85	**394**	(17/63)	**423**	(18/63)	1068	1081	187	158
≥0.4	≤0.875	**181**	(6/63)	**711**	(33/63)	474	1678	157	204
≥0.5	≤0.9	**71**	(4/63)	**1225**	(54/63)	172	2813	67	321

TABLE 8: Gene-Gene Interactions for the first follow-up on blood samples. Bold columns of PC and KDE indicate the gene interactions considering the external genes

respectively, while the last two columns describe the number of edges that changed orientation according to the BIC criterion.

We compared the performance of the two approaches, taking into account existing information on molecular interactions from the BioGRID (Biological General Repository for Interaction Datasets) public database (version 3.2.95), an interaction repository with data from model organisms and humans. BioGRID is a database that archives and provides both genetic and protein interactions from humans (150,273 protein and 1,622 gene interaction data) curated from high-throughput datasets as well as individual focused studies, as derived from over 19,000 primary publications [84]. For the 115 selected genes (110 oral cancer related genes and five control genes) BioGRID database derived 3,380 direct and indirect interactions (65 genetic and 3,315 protein interactions; accessed on December 2012) among them and at most three external genes. Notice that the currently available information provided 63 direct interactions between the examined molecules. In addition, we validated all new interactions created from our network-construction framework using HIPPIE (Human Integrated Protein-Protein Interaction rEference) and we discovered that only these 63 direct interactions have protein interaction annotations in the current human interactome reference [85].

Thus, the goal of our study at this stage was to examine how many of these available associations can be verified from expression data. The results for the inferred networks with PC algorithm indicate that, as thresholds increase, the graph becomes sparser with fewer interactions verified. This is due to the lack of strong partial correlations between molecular units. However, as the thresholds of KDE increase, correlation also increases. This implies that genes are found to be less independent, more interactions are identified and the graph becomes more cohesive.

The two approaches reveal that the molecules under examination do not present high association. This is deduced by the extracted interactions for the various thresholds. For PC at high thresholds there are only few strong associations; for KDE at lower thresholds of similarity there is some indication of dependence. However, for these thresholds the actual number of intense associations is small. The above

Threshold		Verified Gene Interactions				New Edges		Oriented Edges	
PC	KDE	PC		KDE		PC	KDE	PC	KDE
≥0.1	≤0.6	**1060**	41/63	**41**	(0/63)	3005	95	222	65
≥0.15	≤0.7	**854**	33/63	**79**	(0/63)	2379	164	189	83
≥0.175	≤0.75	**735**	29/63	**145**	(1/63)	2069	330	185	116
≥0.2	≤0.8	**627**	25/63	**287**	(4/63)	1797	590	183	183
≥0.3	≤0.85	**316**	14/63	**616**	(21/63)	927	1278	151	186
≥0.4	≤0.875	**122**	3/63	**934**	(34/63)	397	1979	92	231
≥0.5	≤0.9	**41**	1/63	**1328**	(50/63)	135	3000	46	210

TABLE 9: Gene-Gene Interactions on tissue samples. Bold columns of PC and KDE indicate the gene interactions considering the external genes.

observation indicates that molecules from various pathways are not likely to directly interact. This is also verified by the small number of the direct genetic interactions. Thus, in addition to direct interactions, it would be of great interest to take into consideration the external influence of additional molecules, for which we expect indirect associations with the 115 genes under examination.

Table 8 and Table 9 provide the numbers of verified direct interactions. Comparing the performance of the two methodologies, KDE appears to behave better in capturing the biological associations. More precisely, KDE identifies up to 86% of known genetic direct interactions for the blood constructed network and up to 79% of known direct interactions for the tissue network. These percentages for PC are 66% and 65%, respectively. To further reinforce this statement, we present in Figure 22 and Figure

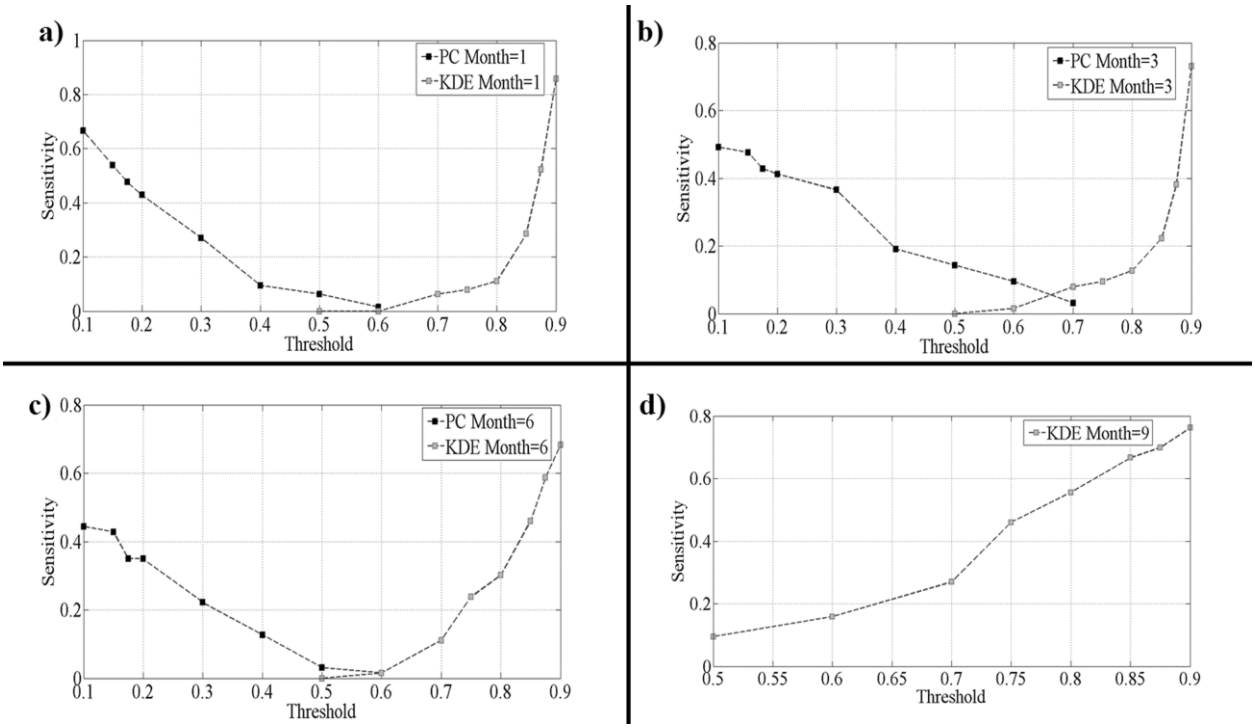


FIGURE 22: True positive rate for all time stages on blood samples.

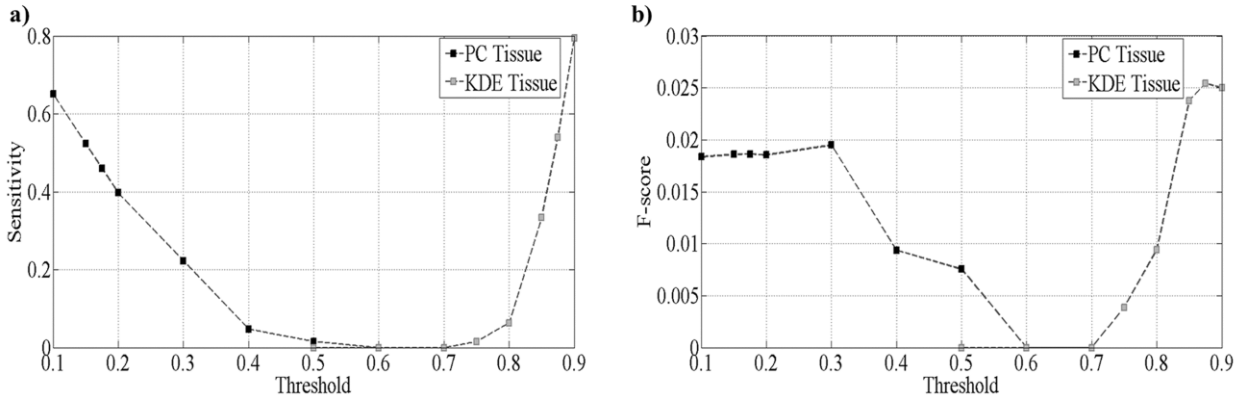


FIGURE 23: **(a)** True positive rate and **(b)** F-score metric on tissue samples.

23a, the true positive rate for the networks constructed from all monthly follow-ups, in addition to the tissue network. These figures (Figure 22 and Figure 23a) justify KDE's superiority in detecting existing interactions over PC for all oral cancer stages. Surprisingly, for the last follow-up (Figure 22d), PC was unable to generate a reliable network due to small patients' attendance at that time. In fact, for this stage, PC gave for all different thresholds almost 6500 new connections due to instabilities of the correlation matrix inversion.

To assess the network reconstruction ability, we counted true positives-TP (correctly identified true edges), false positives-FP (spurious edges), true negatives-TN (correctly identified zero-edges) and false negatives-FN (not recognized true edges) edges. In order to specify the optimal threshold for each algorithm, the size of the graph has to be taken into consideration. This is necessitated by the fact that as the graph becomes denser, more interactions are generated. Thus, the probability of capturing pre-existing associations increases. Figure 23b, Figure 24 present the performance of the two methodologies for all thresholds, according to the F-score metric $\left(F=\frac{2*precision*recall}{precision+recall}\right)$. For each temporal instant, the F-score analysis derives the thresholds 0.7, 0.9, 0.75 and 0.75 for KDE and 0.5, 0.6 and 0.15 for PC, respectively. We note that the 4th instant does not provide a reliable network for PC. Similarly, the appropriate thresholds for both algorithms on the tissue network are 0.88 and 0.3, respectively (Figure 23b).

From a statistical perspective, many false positive edges were found (leading to low F-score). However, this aspect needs further discussion to reveal its valid implications. The false positive rate of connections becomes large due to the fact that we consider only the direct interactions that have been biologically confirmed. In practice, the majority of molecules participate in a variety of biological processes. As a consequence, they affect (or, are affected) by many external factors participating in pathways that connect indirectly with the molecules under examination. Therefore, we expect that external factors define many more interactions that have not been established yet. This inclusion of direct connections through external pathways is a valid assumption that

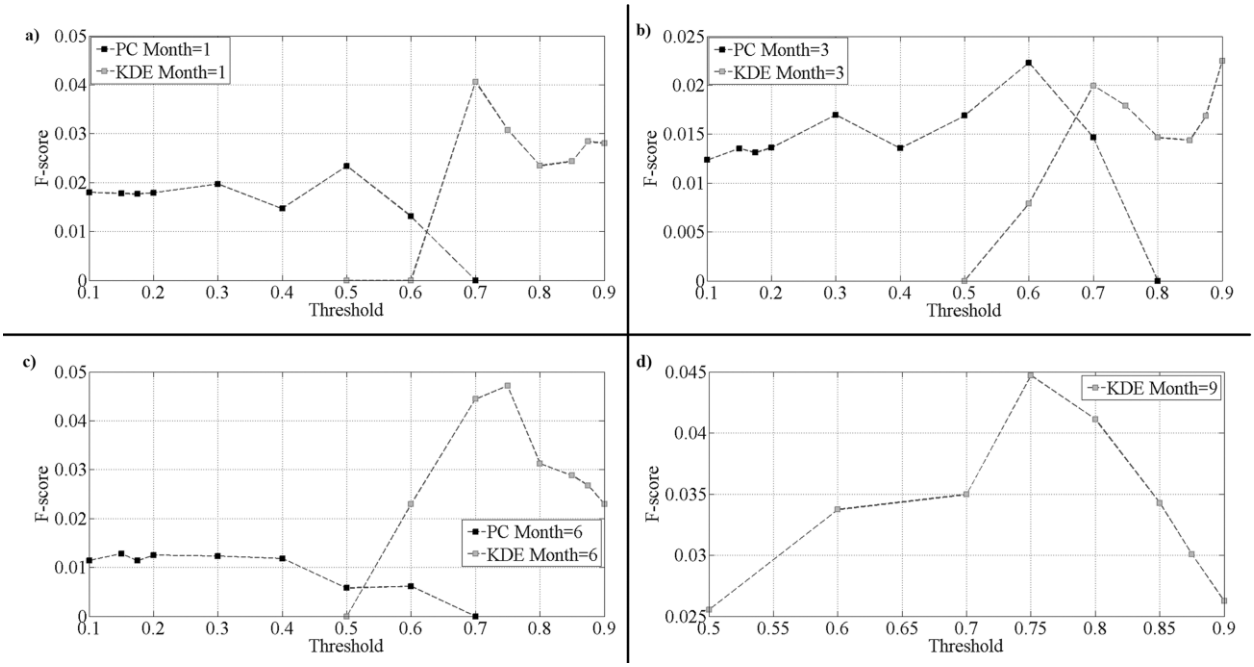


FIGURE 24: F-score metric for all follow-ups on blood samples.

contributes to the consideration of relevant false positives and the correct interpretation of the performance metric.

9.1.2 INDIRECT INTERACTIONS USING EXTERNAL GENES

In Section 6.4 we stressed the need to examine the indirect associations through external genes that connect molecules in other pathways, to justify interactions between the analyzed genes. However, it is too expensive to validate the full set of predictions experimentally [42]. During the last decade, interaction databases have grown exponentially. More than 230 web-accessible biological pathway and network databases have been created. In order to integrate molecular interactions and other types of high-throughput data from different public databases towards automatically building biological networks, we used BioNetBuilder [59] which is an open- source client-server Cytoscape plug-in and offers a user-friendly interface to create biological networks integrated from several databases. For the studied genes, BioNetBuilder retrieved more than 300.000 interactions with more than 25.000 genes from the following databases: (BIND, 11631); (BioGrid, 24313); (DIP, 1387); (IntAct, 20201); (Interologger, 24136); (KEGG, 112230); (MINT, 11411); (MPPI, 469); (Prolinks, 136770). The resulting network through extensive consideration of available biological knowledge is considered as the ground-truth, against which we compare our analysis.

Columns 3 and 5 in Table 8 and Table 9 present the results according to the above analysis for the blood and tissue samples, respectively. According to ground-truth network, apart from the 63 direct edges there are 1558 indirect implications; these result when a maximum of three external genes is considered (Figure 13). Furthermore, from the 115 analyzed genes, there are 22 uncharacterized, for which BioNetBuilder

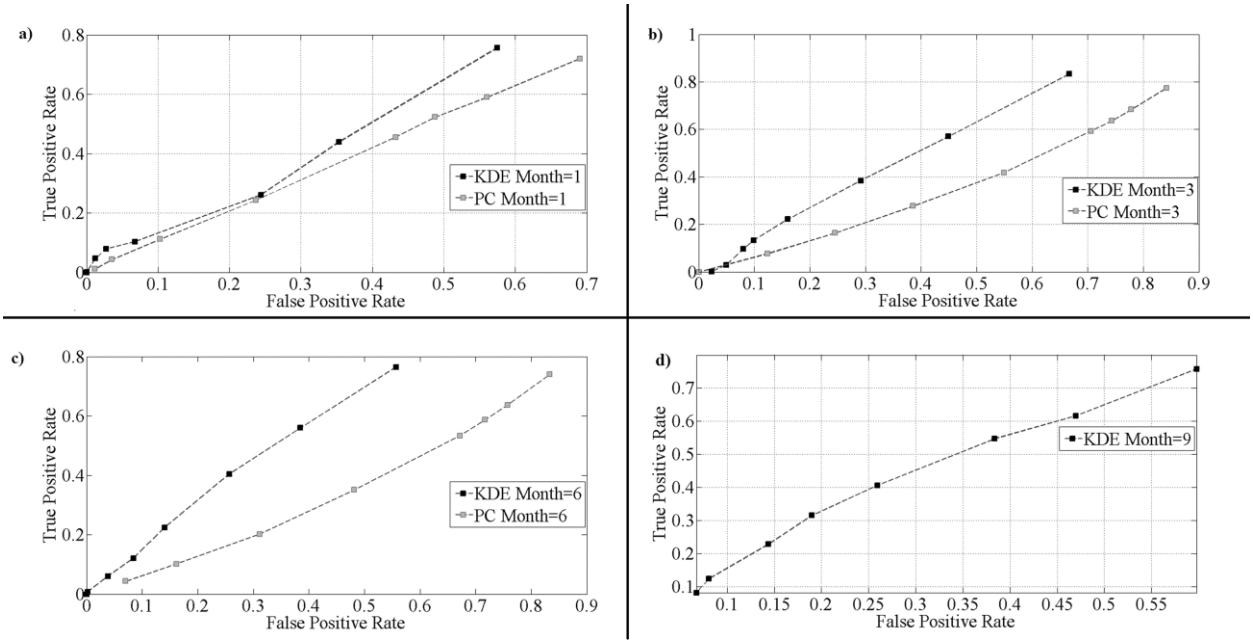


FIGURE 25: ROC comparison between KDE and PC for the blood samples **(a-d)**. For the ninth month, PC cannot derive a reliable network structure. Apparently KDE results in larger AUC for all cases.

resulted in no associations; in our framework we did not take into account edges connecting these genes. Notice that, for this reason, the number of new edges (columns 3, 4) in Table 9 differs from Table 8, which include all edges. In order to specify the number of TN associations we found all the possible interactions between the 115 studied genes and from this set we omitted the TP interactions (direct, indirect). In total, the set of TN associations comprised of 2657 edges.

Figure 25 presents the ROC curves for the blood samples associated with the 4 follow-ups (Figure 25a-d), while Figure 26a presents the ROC curve for the tissue network. For all listed cases, KDE outperforms PC as the area under the curve (AUC) is larger compared to PC. Furthermore, both algorithms show improvement in performance after taking the external genetic influence into consideration. In fact, the equivalent plots of precision and recall, Figure 27a-d and Figure 27b, show significant improvement for all studied cases. The diagrams show the levels of precision comparing the initial approach based on the 63 direct interactions, with the proposed idea based on the 1558 indirect external interactions. In fact, the latter approach considers many more edges for which there exists an indirect pathway through external molecules. Considering these edges, precision greatly improves for all network cases, reaching quite high levels, to support of the conclusion that expression data enclose dependencies from a variety of sources. Therefore, when dealing with expression data, direct associations obtained from statistical analysis should be interpreted as possible indirect influences of external factors and not as spurious edges. In fact, the MET-CD44 interaction that was found as TP external association is also verified by the updated HIPPIE version as direct association.

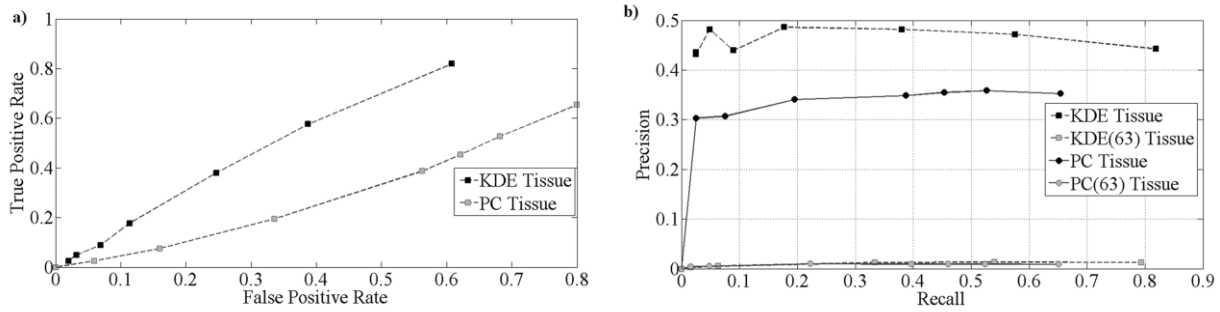


FIGURE 26: **(a)** ROC curve comparison between KDE and PC for the tissue network. KDE outperforms PC as it covers larger AUC; **(b)** Precision vs Recall comparison between KDE and PC for the tissue network. The precision has been significantly improved from the initial approach, which considers only 63 interconnections as TP.

9.2 BIOLOGICAL DISCUSSION

After the basic gene structure, we first analyze the global organization of the gene network by examining the major gene clusters. Groups of genes that are densely connected to each other in the network may represent functional modules in which the genes are highly related in function and/or cooperate in some biological processes. We performed k-means cluster analysis [86] on the primary gene expression data and recovered five major clusters (Figure 29). As shown in Figure 29, the content and structure of blood and tissue networks based on gene interactions is different at the

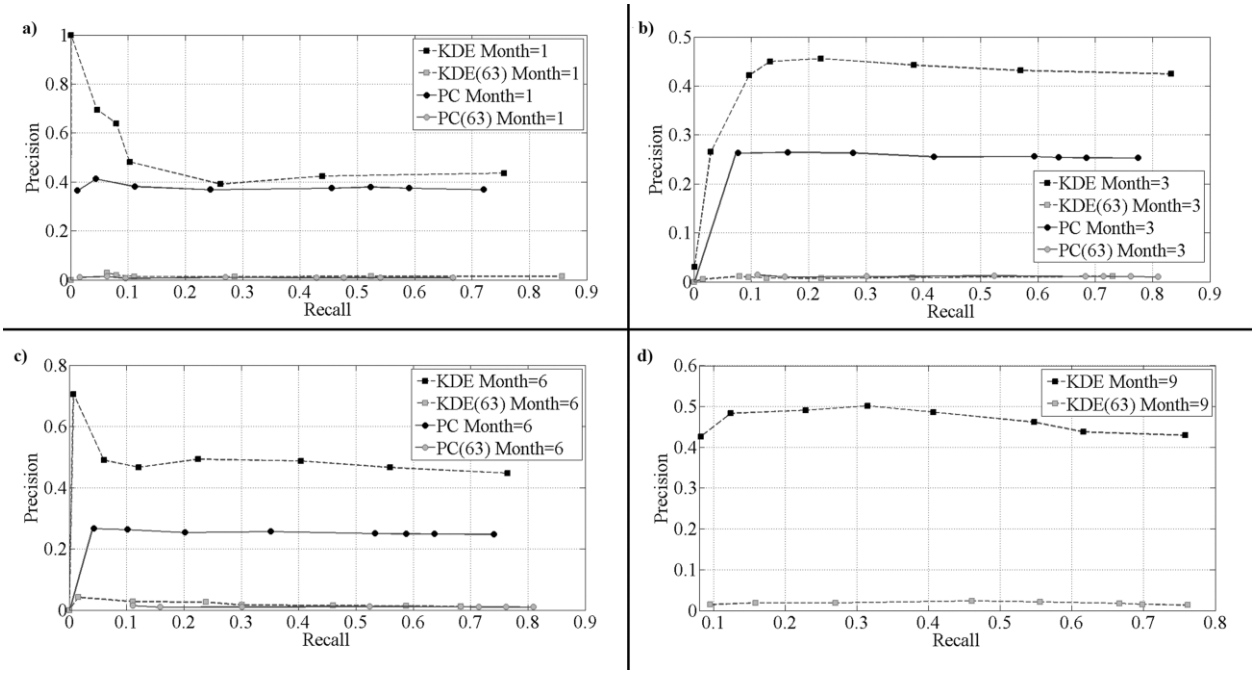


FIGURE 27 : Precision vs Recall comparison between KDE and PC for all cases on blood samples **(a-d)**. KDE(63) and PC(63) represent the networks considering as TP the set of 63 direct interactions, while KDE and PC curves represent the performance considering as TP all direct and indirect edges. KDE outperforms PC reaching higher levels of precision and recall for all periods. For the ninth month, PC could not result in a reliable network structure.

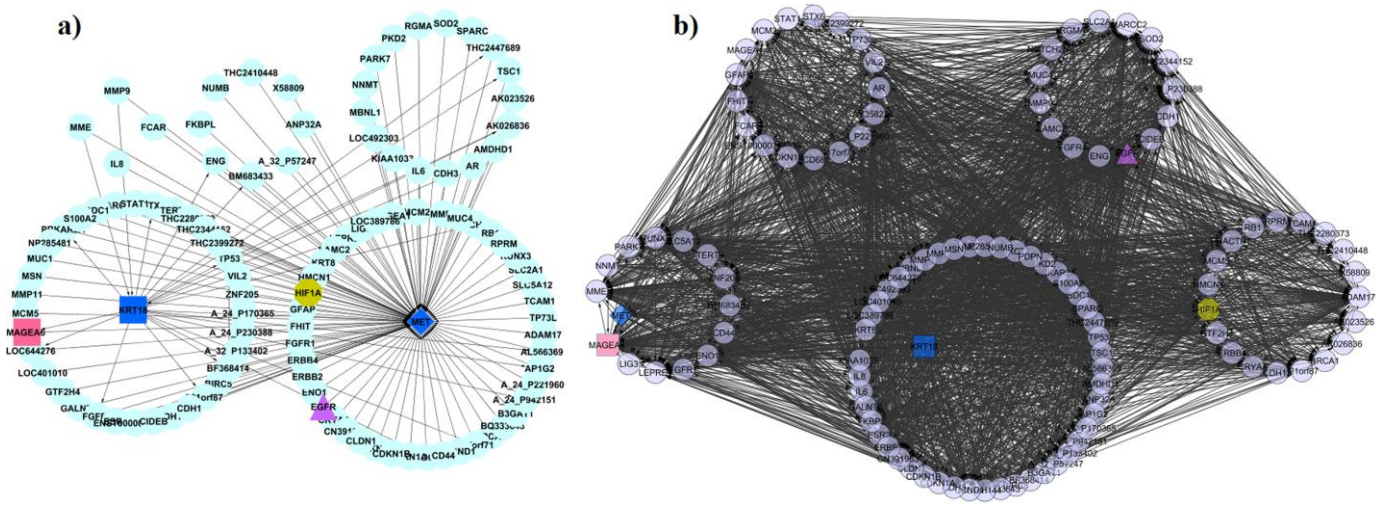


FIGURE 28 : **(a)** K-means clustering on the blood samples from the first follow-up and **(b)** on tissue samples. Five clusters were identified as presented in the grouped areas in both networks. Highlighted are the common "intersection genes" *MET, EGFR, HIF1A* and *MAGEA6.*

first visit to the doctor. For the remaining time stages, the network on blood samples preserves a similar structure, with small variations among the peripheral genes. To explore whether the selected genes share specific functional features, we performed GO enrichment analysis using WebGestalt [87]. The genes in the same cluster are densely connected with each other (Figure 28 : b), and GO analysis indicates that these five gene-clusters are enriched in certain GO annotation terms (Supplementary Table II).

The enriched GO terms support the current knowledge about the multiple functional roles of the implicated genes in oral cancer as well as in the disease progression [88], [89]. Regardless of GO terms in the category of biological process, we found that cell proliferation (adj$P$=6.94X10^{-9}), regulation of cell proliferation (adj$P$=2.58X10^{-7}), and regulation of cell cycle (adj$P$=4.48X10^{-7}) are significantly enriched in these gene clusters of both blood and tissue samples, as well as in blood follow up samples (Figure 28a, b; Supplementary Figs. 1a, b, 2a, b; Supplementary Table II). Overall, each cluster is dominated by distinct GO terms, a number of which is also present in other clusters, however with varying statistical significance. More importantly, the enrichment significance of specific GO terms varies between blood and tissue samples and/or time stages (e.g. cell cycle, regulation of cell cycle, regulation of apoptosis, positive regulation of locomotion), accompanied by the reorganization of many genes (e.g. *TP53, EGFR, MET, HIF1A, CDH1, MMP2, MMP9, MMP11, CD44)* in these five clusters (Supplementary Table I, II). Furthermore, OSCC development depends on the accumulation of multiple genetic changes. During the multistep process of oral tumorigenesis, the normal functions of proto-oncogenes and tumor suppressor genes are modified, thus affecting regulation of cell cycle, cell proliferation and death, DNA repair, cell differentiation and immunity [84], [89], cellular processes which are reflected by the above enriched GO terms but also by less significant GO terms (Supplementary Figs. 1a, b, 2a, b; Supplementary Table II).

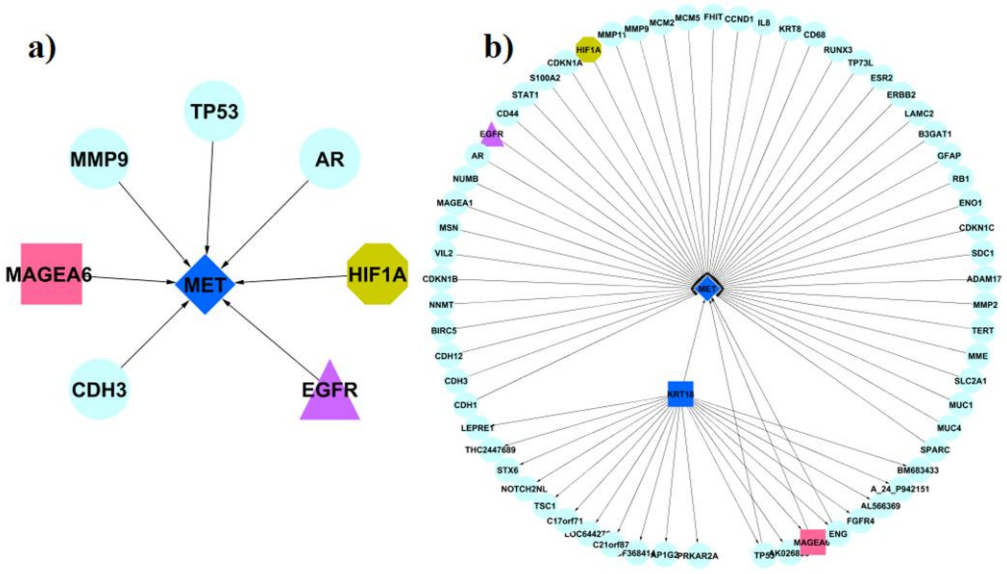


FIGURE 29: Network intersection for blood and tissue. The left part **(a)** presents the intersection of all time stages for the blood samples; The Right part **(b)** presents the network intersection between the tissue and the blood from the first follow-up. Highlighted are the common "intersection genes" *MET, EGFR, HIF1A* and *MAGEA6.*

To investigate the biological meaning of the proposed framework we found the intersections of the blood based networks for all stages. The appropriate thresholds for each stage were chosen according to the F-score metric, as is presented in Section 9.1.1. Figure 29 presents the "intersection genes", i.e. genes that are induced by the network intersection for all time stages for the blood samples (Figure 29a, Supplementary Table IIIc), as well as the intersection of tissue network with the blood from the first time stage (Figure 29b, Supplementary Table IIIc). Figure 29a,b depict a number of oncogenes (e.g. *EGFR*), tumor suppressor genes (e.g. *TP53*), transcription factors (e.g. *RUNX3, HIF1A, AR*) and other important molecules in many aspects of multistep tumor development (e.g. *KRT8, KRT18, MMP9, MMP11, CDH1, CDH3, MAGEA6, ENO1, CDKN1C, SDC1, LEPRE1*) and highlight the central role of the proto-oncogene *MET* (hepatocyte growth factor receptor) on both tissue and blood/follow-up samples. The MET gene product is a proto-oncogenic receptor tyrosine kinase and its activation elicits cell proliferation, cell scattering, survival, invasion, and angiogenesis. MET deregulation promotes tumor formation, growth, progression and metastasis as well as resistance to therapy. Due to its key role in cancer development and progression, it is also a potential candidate for therapeutic intervention [88], [90].

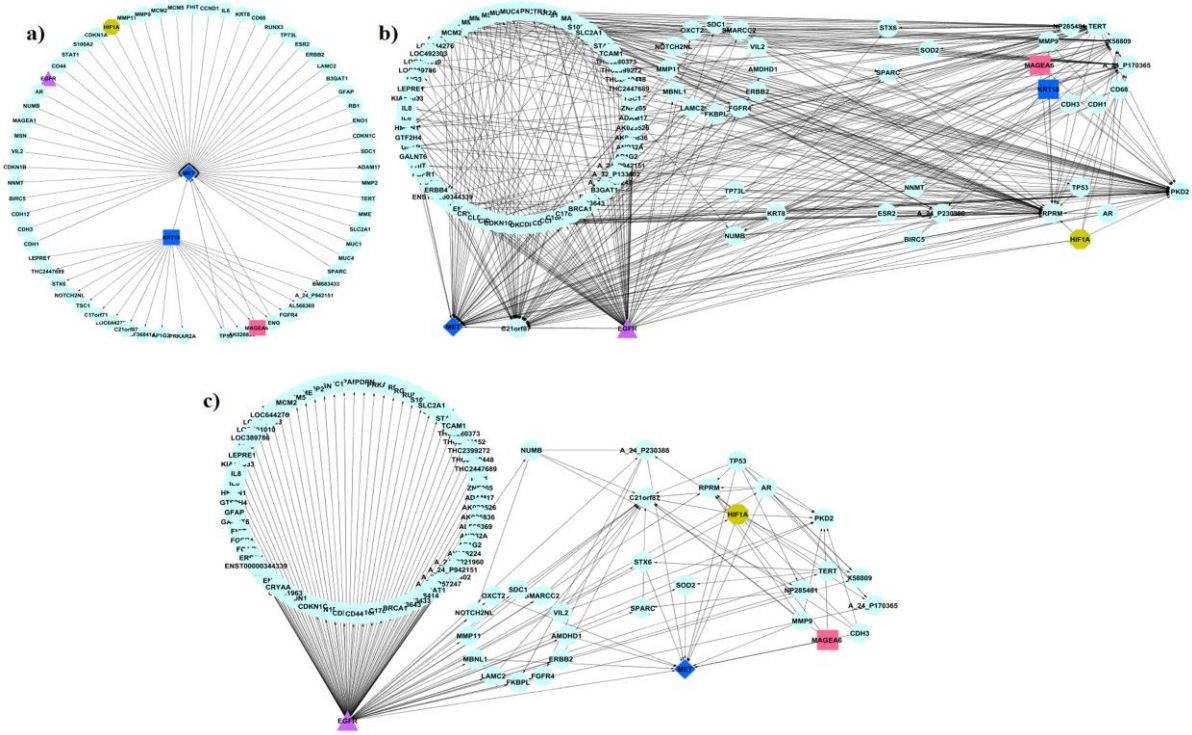


FIGURE 30 : Network intersection for sequential time stages on blood samples; **(a)** Network intersection between the 1st and 3rd month; **(b)** Network intersection between the 3rd and 6th month; **(c)** Network intersection between the 6th and 9th month. Highlighted are the common "intersection genes" *MET, EGFR, HIF1A and MAGEA6.*

In an attempt to investigate the network structure at different stages in the oral cancer disease, we compute the intersections of sequential follow-ups. Figure 30 presents the networks in groups which have a common degree. The degree as a topology metric shows the number of links that a node has with another in the network. Especially in biological networks, the degree is an important metric that highlights genes with high connectivity playing an important role in disease development [61]. The temporal development of the network is demonstrated in Figure 30, presenting the sequential network-intersections of the blood samples.

By examining the important molecules in Figure 29 we conclude on the following:

i. *For the temporal development of network intersection of the blood samples* (Figure 29a, Supplementary Table IIIa, IIIc):

- all genes have extremely low degree (1 or 2 or 4) on blood samples at the first time slice, with the exception of MET (109)
- all genes have higher degrees at the 3rd and 6th-month follow-up and acquire their highest degree (107 or 114) at 9th-month follow-up
- MET degree is slightly higher (114) at the 3rd and 6th month follow-up and very low (13) at 9th month follow-up, through the loss of genes [61], [91] and possibly as a result of disease-causing mutations [92].

ii. *For the intersection of tissue with the blood network for the first time-stage* (Figure 29b; Supplementary Table IIIa, IIIc):

- all genes have extremely low degree (1 or 2) on blood samples at the first time-slice, with the exception of KRT18 (24) and MET (109)
- all genes have higher degrees (39 to 114) with the exception of MET that has lower degree (38) on tissue samples (similar to blood follow-up samples [61], [91], [93]).

iii. *For both network intersections* (Figure 29, Figure 30; Supplementary Table IIIa, IIIc):

- MET interacts with EGFR (epidermal growth factor receptor), HIF1A (transcriptional factor hypoxia inducible factor 1α) and MAGEA6 (melanoma antigen family A, 6)
- MET interacts with important molecules that are not or loosely connected themselves
- MET interactions with its neighboring molecules appear to change drastically in time and among tissue and blood samples compared to all other disease genes; this is demonstrated more clearly during disease progression, in particular at the last stage considered.

It is obvious from the above results that MET plays a crucial role in oral cancer. Further emphasizing on the biological effects of the above intersections, the proposed dynamic networks exemplify the following issues: (a) MET interacts consistently with EGFR, HIF1A and MAGEA6 at both tissue and blood samples and during OSCC progression (Figure 29, Figure 30). Despite the known MET/EGFR association in cancer [94], the existence of the MET/HIF1A and MET/MAGEA6 associations remain unknown. However, previous studies [95], [96], [97], [98] [44]-[47], referring to the functional role of these molecules in cancer and to their involvement in OSCC further support their potential interaction with MET and their relevance to oral cancer initiation and progression. For example, the MAGEA6 gene product (*MAGEA6* is expressed in OSCC) has been reported to bind to p53 tumor suppressor (TP53) and impair its function causing decreased apoptosis and increased cell growth [95]. Furthermore, the transcriptional activation of *MET* proto-oncogene during hypoxia via HIF1-mediated cascade could possibly explain the MET overexpression reported in OSCC specimens [96]. In addition, the EGFR increased expression and its ligand (i.e., transforming growth factor alpha) can play a critical role in oral tumor development and progression; it is recently reported that both EGFR and MET mediate cellular responses in partly redundant and partly complementary ways [88], [97]. This counter-balancing activity of MET and EGFR pathways may also be viewed as a potential target for oral cancer therapeutic intervention [98]. (b) MET interacts with *EGFR* oncogene and *TP53* tumor suppressor gene at blood samples from all disease stages, partly supporting the

existence of a large complex consisting of many oncogenes, tumor suppressors, and DNA repair proteins [99]. (c) MET loses many of its interactions through the loss of genes [61], [91] and possibly as a result of disease-causing mutations; deranged protein-DNA interactions, disruptions of protein-protein interactions due to protein misfolding, new undesirable protein interactions or pathogen-host protein interactions are examples of the impact of such disease-causing mutations [92].

Finally, from k-means clustering we infer that MET clusters together with a few other molecules on both tissue and blood samples; the clustered molecules are often different at different time stages (Supplementary Table IIIb, IIIc). This aspect illustrates the contribution of complex signaling pathways in the activation or repression of specific biological processes, which are indicative of tumor initiation, promotion and progression and result in genetic alterations [100]. This also highlights a dynamically functional reprogramming of a number of implicated genes and especially MET.

According to a recent study [61] , *MET* could be characterized as "broker" gene, i.e., a disease gene that holds a crucial position in the network topology as broker interacting with many neighboring molecules that are less or not connected with each other. Cai et al. (2010) [61] suggest that disease genes are found in especially vulnerable positions in networks, which is a reason of identifiable disease phenotypes accompanied by their disorganization [61]. MET appears as a highly connected hub molecule in a central position at the onset of cancer initiation; following disease progression, it is dynamically reorganized and takes a peripheral position in the constructed network. This central position has been reported in cancer, where disease genes tend to encode hubs, although in other pathologies disease genes reside at the periphery of the networks [92]. In addition, recent studies [101] support that hub proteins displaying modified modularity in the human interactome (like MET) could be useful markers for predicting oral cancer outcome.

We suggest that MET is a key molecule with unique network-topological features, which are in agreement with its biological role as proto-oncogene, so that it may be considered vital for oral cancer.

Our findings also support the claim that the networks of molecular interaction provide information about the alterations of gene-gene/gene product and/or gene product-gene product interactions in a complex disease, such as oral cancer. The consideration of the *in vivo* MET cellular network at a specific disease state might be an important guide for screening patients at the time of diagnosis, for predicting oral cancer progression and for deciding on effective treatment plan.

Although this study attempts a coupling of the mathematical or computational model to experimental data, the small sample size remains a limiting parameter in estimating the network structure. Furthermore, even though it offers potential grounds for biological validation, many predicted outcomes of this analysis are difficult to be

validated for clinical use due to the extensive simulation procedures needed for this purpose.

9.3 CONCLUSION

Clearly, the KDE approach models quite well the verified direct and indirect associations among the participating genes in oral cancer. On the contrary, the PC approach appears to capture fewer of these associations. Thus, our results indicate that KDE performs better on the network construction. In addition, while PC fails in modeling genetic interactions with sparse data, KDE due to sample estimation succeeds in capturing biological interactions. This supports the aforementioned statement that KDE is resilient in modeling the genetic associations with sparse experimental data.

Perhaps the most important contribution of this study is that it gives a different perspective in revealing genetic interactions as a result of multiple genetic factors. Within this framework we proved that external factors that participate in different pathways affect the genetic expression. Thus, when statistical analysis gives a large amount of typically false edges, indirect pathways should be examined. Moreover, we focused on the edge interpretation as existing or not, solely based on expression data. In fact, due to the analyzed obstacles many studies resort to characterizing the predicted edges as TP according to the biological process they participate. This gives an advantage in boosting our framework's performance but it introduces generality in justifying the genetic association.

From the biological knowledge point of view, the proposed framework of analysis provides strong evidence on the importance of MET. More specifically, it suggests an initial central role of this molecule. This is modified to peripheral with time and disease progression, while other significant genes like EGFR take the central role(s). It appears that the activation of the MET network occurs earlier than the EGFR network, at the onset of the disease. Overall, the specific interplay of HIF1-MET, MET-EGFR and MET-MAGEA6 and their associated signaling cascades may denote key mechanisms of oral cancer initiation and progression and may carry therapeutic implications. The provided MET network is not only validated by known interactions but also offer predictive value of new interactions that should be further considered experimentally.

Supplementary information on our work can be found on http://www.display.tuc.gr/kalan.osccstudy/.

10. OSTEOARTHRITIS

Osteoarthritis (OA)[102] is the most prevalent form of chronic joint disease and accounts for substantial morbidity and disability, particularly among older people. It is characterized by loss of joint homeostasis. The articular cartilage cannot maintain its integrity and is progressively damaged, the subchondral bone envelope is thickened changing loads in the bone-cartilage biomechanical unit, the synovium shows signs of inflammation and bony spurs (osteophytes) appear at the edges of the bone. Its etiology is multifactorial with a signify cant genetic component as shown by twin and family studies. Due to the severity of the disease, a variety of clinical SNP studies have been conducted in an attempt to reveal genetic factors related to the disease.

In this study we analyze 270 samples from 125 healthy and 145 OA patients. For each sample an approximate number of 250.000 SNPs were taken each at different genomic position. Each sample was given two possible polymorphisms A or B, thus for each allele there were 3 combinations: AA, AB and BB, as presented in Figure 10. Also, if no polymorphism was found NoCall was assigned as an SNP value.

Our goal was to isolate SNPs suspicious for genomic OA alterations. We applied the methodology of Section 6.1, starting by the quality control. SNPs with $\text{Pvalue}_{MAF}<0.1$ and $\text{Pvalue}_{HWE}<0.05$ were excluded from further analysis. From this procedure remained approximately 143.000 SNPs for which was made a statistical analysis with the HWDTT, CATT and OR.

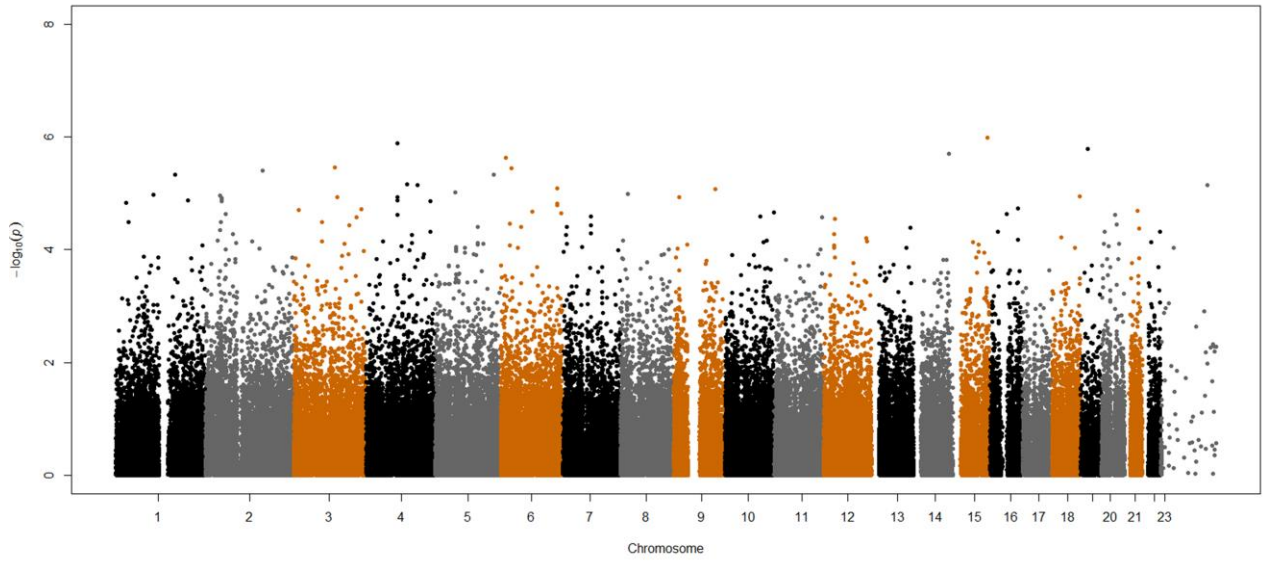


FIGURE 31: Manhattan plot for the HWDTT approach. Strongest associations have higher logarithmic p-values.

SNP	Chromosome	Position	ID	P-value	-log
rs732781	15	95855819	SNP_A-2090364	1,03E-06	5,986111
rs7699990	4	86222796	SNP_A-2311457	1,28E-06	5,894131
rs1167155	19	20866028	SNP_A-4236388	1,64E-06	5,783849
rs2297243	14	97022900	SNP_A-1883903	1,98E-06	5,70281
rs9472528	6	12781353	SNP_A-2197710	2,37E-06	5,624608
rs923349	3	112225934	SNP_A-2008988	3,49E-06	5,457447
rs2148008	6	28855831	SNP_A-1911873	3,63E-06	5,439791
rs298247	2	157233542	SNP_A-1965504	4,02E-06	5,396175
rs1963837	5	162064605	SNP_A-2148208	4,64E-06	5,33361
rs4657364	1	164532325	SNP_A-2147897	4,68E-06	5,329721
rs6855671	4	114502911	SNP_A-2207888	6,87E-06	5,162978
rs3775652	4	143049980	SNP_A-4210588	7,09E-06	5,149593
rs3585091	X	127969409	SNP_A-4223428	7,15E-06	5,145925
rs2294682	6	155740704	SNP_A-4212085	8,11E-06	5,090912
rs4541999	9	113907759	SNP_A-4225931	8,48E-06	5,071667
rs404820	5	54513456	SNP_A-2090775	9,51E-06	5,021718
rs1709273	8	20390409	SNP_A-2092887	1,04E-05	4,984296
rs1701527	1	105486489	SNP_A-2057837	1,07E-05	4,968753
rs376535	2	40411717	SNP_A-1788212	1,09E-05	4,96125
rs1260793	18	76651887	SNP_A-1803755	1,15E-05	4,940169
rs1700946	4	86035461	SNP_A-1951838	1,16E-05	4,934018
rs1051137	3	118188090	SNP_A-1974117	1,18E-05	4,927565
rs770205	9	15870719	SNP_A-2306830	1,18E-05	4,926812
rs7572190	2	42889233	SNP_A-1948826	1,19E-05	4,922801

SNP	Chromosome	Position	ID	P-value	-log
rs751207	10	73887615	SNP_A-4214196	1,17E-13	12,93298
rs2880301	13	20100534	SNP_A-2257575	4,61E-13	12,33654
rs2932174	13	20065041	SNP_A-4224554	7,46E-13	12,12736
rs1910130	18	50975302	SNP_A-2226169	8,08E-12	11,09246
rs12000384	9	37857067	SNP_A-4227798	1,17E-11	10,93086
rs8713	7	116199797	SNP_A-4193124	2,41E-11	10,61843
rs17116635	10	105945842	SNP_A-4196050	2,76E-11	10,55977
rs3213031	10	62540547	SNP_A-2140146	3,21E-11	10,49401
rs5968981	X	86002218	SNP_A-4238831	3,76E-11	10,42488
rs17124913	12	51092019	SNP_A-1951520	6,83E-11	10,16548
rs4300294	10	65927516	SNP_A-4206370	7,30E-11	10,13691
rs16965398	15	53352551	SNP_A-1780521	9,33E-11	10,03019
rs7117211	11	13248647	SNP_A-4214713	1,69E-10	9,772934
rs13417546	2	12316126	SNP_A-2248383	1,80E-10	9,743869
rs936634	18	49371058	SNP_A-2061321	2,03E-10	9,692372
rs1431068	14	34596172	SNP_A-4222588	2,92E-10	9,535072
rs41498646	4	82441654	SNP_A-1957916	3,10E-10	9,509267
rs16973258	15	82252329	SNP_A-2097697	3,52E-10	9,45347
rs41382145	6	125494570	SNP_A-2199476	3,89E-10	9,409542
rs4415869	13	31076750	SNP_A-4226493	4,19E-10	9,377399
rs4673740	2	214206077	SNP_A-1825696	4,29E-10	9,367537
rs16941272	15	88658044	SNP_A-2303250	4,67E-10	9,33107
rs1925925	X	69625447	SNP_A-1910772	4,85E-10	9,314367
rs2581654	18	72929289	SNP_A-4200239	6,06E-10	9,217455

FIGURE 32: HWDTT (LEFT) and CATT (right) analysis for the 142.000 SNPs. The red rows indicate the 12 first SNPs with low P-values,

Figure 32 depicts the results of the HWDTT and CATT sorted by the lowest p-values. SNPs that reject the initial *Ho* hypothesis (p-value<0.001) are candidate for further analysis. In order to graphically present the tendency of the SNPs we show the Manhattan plot for these approaches. A Manhattan plot is a type of scatter plot, usually used to display data with a large number of data-points - many of non-zero amplitude, and with a distribution of higher-magnitude values, for instance in genome-wide association studies (GWAS). In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative logarithm of the association *P*-value for each polymorphism displayed on the Y-axis. Strongest associations have the smallest *P*-values (e.g., 10^{-6}) so their negative logarithms will be the greatest low logarithmic p-values (Figure 31 and Figure 34).

Similarly, Figure 33 and Figure 36 present the results of OR analysis. Comparing the three different approaches we can see that CATT highlights a big number of suspicious SNPs ($-\log(pvalues) \geq 3$) while OR and HWDTT a smaller number. In order to see the common SNPs between the approaches we enlist Figure 35. For the first 100 ranked SNP positions we see that only CATT and OR have highlighted common SNPs. Thus, we propose OR and CATT as the most efficient approaches in SNP analysis for $p<10^{-3}$ and $p<10^{-10}$respectively. The results from those approaches will be given for further clinical examination in order to verify whether those genomic positions are responsible for OA disease.

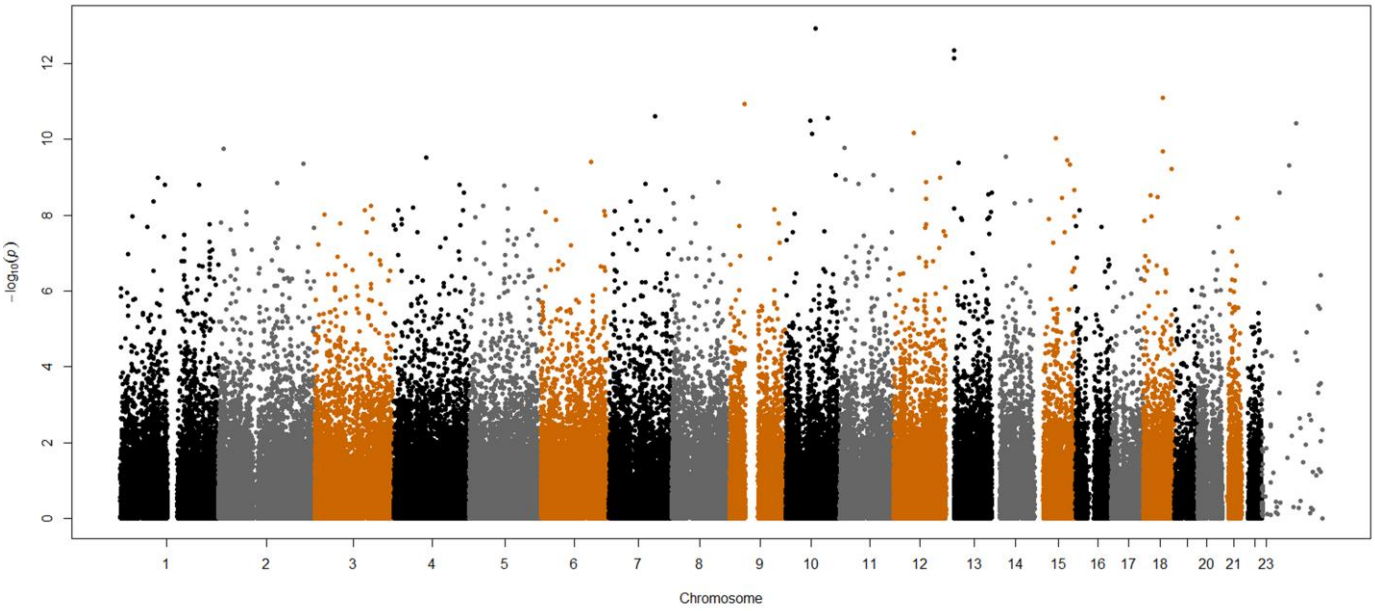


FIGURE 34: Manhattan plot for the CATT approach. Strongest associations have higher logarithmic p-values.

SNP	Chromosome	Position	ID	OR	CI-min	CI-max	P-value	-Log
rs3213031	10	62540547	SNP_A-2140146	0,334049	-1,5624	-0,63053	2,51E-06	5,600268
rs8713	7	116199797	SNP_A-4193124	0,388903	-1,34659	-0,54226	3,38E-06	5,470831
rs2880301	13	20100534	SNP_A-2257575	3,224719	0,655234	1,686457	4,14E-06	5,383352
rs5968981	X	86002218	SNP_A-4238831	5,359137	0,896483	2,461123	4,22E-06	5,375094
rs1431068	14	34596172	SNP_A-4222588	2,43304	0,502616	1,275668	5,60E-06	5,252011
rs10501580	11	84599002	SNP_A-2012840	2,840232	0,576401	1,51137	7,87E-06	5,103842
rs12000384	9	37857067	SNP_A-4227798	4,34	0,779805	2,155944	8,03E-06	5,09542
rs4471519	12	83761458	SNP_A-2192548	5,196262	0,852957	2,442922	8,63E-06	5,063888
rs41382145	6	125494570	SNP_A-2199476	0,368573	-1,44758	-0,54865	9,50E-06	5,022472
rs41485553	7	93183990	SNP_A-4230855	0,435445	-1,20267	-0,4601	9,56E-06	5,019684
rs2690039	1	200137649	SNP_A-1979786	0,415231	-1,27155	-0,48629	9,96E-06	5,001605
rs2581654	18	72929289	SNP_A-4200239	6,0033	0,90338	2,681239	1,08E-05	4,966952
rs751207	10	73887615	SNP_A-4214196	2,83356	0,561833	1,521234	1,39E-05	4,855933
rs1910130	18	50975302	SNP_A-2226169	0,296296	-1,78717	-0,64562	1,43E-05	4,845176
rs10735465	12	83785172	SNP_A-1798041	5,043062	0,820139	2,415888	1,44E-05	4,841326
rs13417546	2	12316126	SNP_A-2248383	0,237688	-2,13458	-0,73901	1,79E-05	4,747549
rs4300294	10	65927516	SNP_A-4206370	0,239778	-2,12247	-0,7336	1,83E-05	4,73852
rs2134151	5	169460902	SNP_A-1921089	0,434067	-1,22075	-0,44836	1,89E-05	4,72467
rs1925925	X	69625447	SNP_A-1910772	3,542408	0,655844	1,87377	2,09E-05	4,68084
rs10184605	2	147116314	SNP_A-2189691	2,378226	0,460361	1,272348	2,44E-05	4,611886
rs11223027	11	132226747	SNP_A-2181564	2,369231	0,455988	1,269143	2,71E-05	4,567016
rs1059214	4	166263460	SNP_A-4210672	2,657119	0,511586	1,442899	2,85E-05	4,544669
rs10134160	14	96747986	SNP_A-2025294	2,431284	0,464788	1,312051	3,11E-05	4,506572
rs12657135	5	87725387	SNP_A-1818153	3,13949	0,587577	1,700544	3,21E-05	4,493939
rs7322718	13	106461753	SNP_A-2215580	0,437087	-1,22194	-0,43331	3,25E-05	4,488454
rs2932174	13	20065041	SNP_A-4224554	0,36411	-1,49976	-0,52083	3,65E-05	4,4377
rs16941272	15	88658044	SNP_A-2303250	0,230326	-2,21491	-0,72161	3,67E-05	4,435833
rs17188012	13	37015242	SNP_A-2126255	2,532252	0,479985	1,378233	3,97E-05	4,401438
rs4869383	5	97014421	SNP_A-4211112	0,336976	-1,62221	-0,55327	4,04E-05	4,393126
rs11466229	2	70722763	SNP_A-2033552	3,682504	0,64757	1,959616	4,08E-05	4,389483

FIGURE 33: OR analysis for the 142.000 SNPs. The red rows indicate the 12 first SNPs with low P-values

OR-CATT			OR-HWDTT	CATT-HWDTT
OR Rank	SNP	CATT Rank		
1	rs3213031	8	#N/A	#N/A
2	rs8713	6	#N/A	#N/A
3	rs2880301	2	#N/A	#N/A
4	rs5968981	9	#N/A	#N/A
5	rs1431068	16	#N/A	#N/A
6	rs10501580	25	#N/A	#N/A
7	rs12000384	5	#N/A	#N/A
8	rs4471519	31	#N/A	#N/A
9	rs41382145	19	#N/A	#N/A
10	rs41485553	34	#N/A	#N/A
11	rs2690039	36	#N/A	#N/A
12	rs2581654	24	#N/A	#N/A
13	rs751207	1	#N/A	#N/A
14	rs1910130	4	#N/A	#N/A
15	rs10735465	51	#N/A	#N/A
16	rs13417546	14	#N/A	#N/A
17	rs4300294	11	#N/A	#N/A
18	rs2134151	39	#N/A	#N/A
19	rs1925925	23	#N/A	#N/A
20	rs10184605	32	#N/A	#N/A
21	rs11223027	42	#N/A	#N/A
22	rs1059214	35	#N/A	#N/A
23	rs10134160	52	#N/A	#N/A
24	rs12657135	38	#N/A	#N/A
25	rs7322718	46	#N/A	#N/A
26	rs2932174	3	#N/A	#N/A
27	rs16941272	22	#N/A	#N/A
28	rs17188012	81	#N/A	#N/A
29	rs4869383	61	#N/A	#N/A
30	rs11466229	99	#N/A	#N/A
31	rs9806603	82	#N/A	#N/A
32	rs17124913	10	#N/A	#N/A
33	rs6790201	57	#N/A	#N/A
34	rs4415869	20	#N/A	#N/A
35	rs12868114	78	#N/A	#N/A
36	rs2895889	97	#N/A	#N/A
37	rs17169323	67	#N/A	#N/A
38	rs4303351	87	#N/A	#N/A
39	rs2504923	74	#N/A	#N/A
40	rs7928807	29		
41	rs16913048	62		
42	rs2538071	54		
43	rs2718215	98		
44	rs2352740	63		
45	rs7516521	37		
46	rs33917597	76		
47	rs936634	15		

CATT-Chi-Square		
CATT Rank	SNP	Chi-Square
35	rs6998440	24
40	rs4784429	6
64	rs2037021	84
68	rs2049475	42

FIGURE 35 : Common SNPs between the 3 approaches. OR and CATT have the biggest number of common SNPs between the first 100 ranked positions. The 1st and 3rd columns present the ranked SNP position for each analysis.

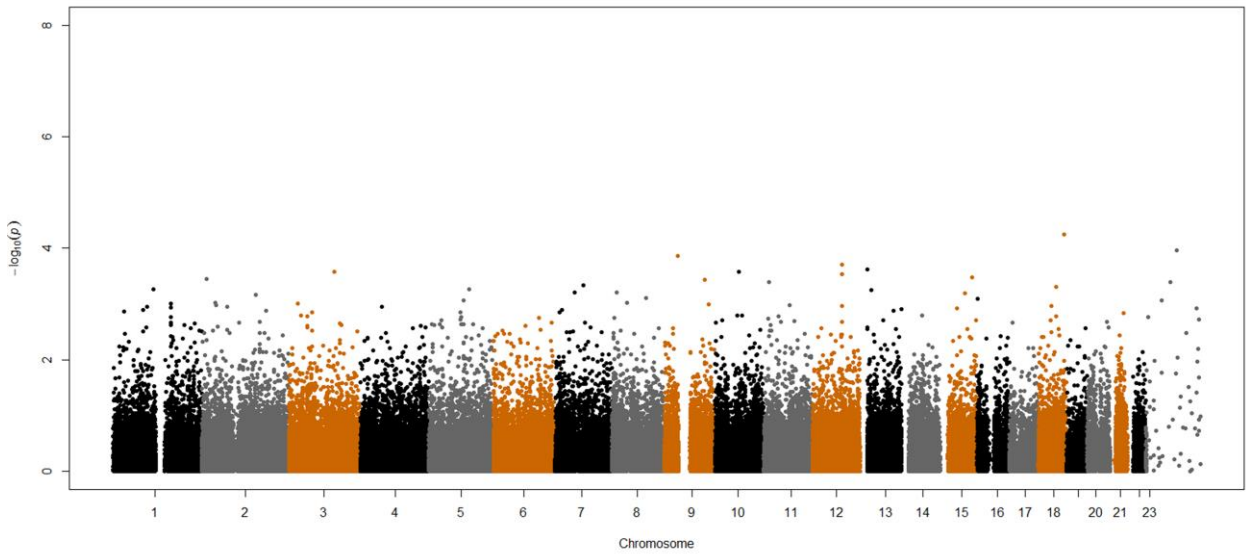


FIGURE 36: Manhattan plot for the OR approach. Strongest associations have higher logarithmic p-values.

11. APPENDIX

Table of Abbreviations	
CPD	Conditional probability distribution
PDF	Probability distribution function
N(.)	Normal distribution
LGGM	Linear Gaussian Graphical Model
DGN	Dynamic Gaussian networks
PC	Partial correlation
KDE	Kernel density estimation
BIC	Bayesian Information Criterion
GRN	Gene regulatory network
IID	Independent identically distributed
SVD	Singular value decomposition
GO	Gene Ontology
FP	False positive
TP	True positive
TN	True negative
FN	False negative
ROC	Receiver operator characteristic
SGK	Standard Gaussian Kernel
SNP	Single-nucleotide polymorphism
MAF	Minor allele frequency
HWE	Hardly Weinberg equilibrium
HWDTT	Hardly Weinberg disequilibrium trend test
CATT	Cochran-Armitage trend test
OR	Odds Ratio

CI	Confidence interval
PEP	Phosphoenolpyruvate
ANAP	Arabidopsis Network Analysis Pipeline
AtFBA6	Fructose 1,6-biphosphate aldolase 6
LEA	Late embryogenesis abundant
ER+, ER-	Estrogen responsive positive, negative
BRCA	Breast cancer
GSE7390	GEO-Gene Expression Omnibus study code
TGFα	Transforming growth factor, alpha
AR	Amphiregulin
BTC	Betacellulin
EPR	Epiregulin
NRG1,2	Neuregulin1,2
MET	Mepatocyte growth factor receptor
BioGrid	Biological General Repository for Interaction Datasets)
HIPPIE	Human Integrated Protein-Protein Interaction rEference
AUC	Area under the curve
EGFR	Epidermal growth factor receptor
HIF1A	Transcriptional factor hypoxia inducible factor 1α
MEGEA6	Melanoma antigen family A, 6
OA	Osteoarthritis
GWAS	Genome-wide association studies

12. BIBLIOGRAPHY

[1] A. V Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks.," *Bioinformatics (Oxford, England)*, vol. 22, no. 20, pp. 2523–2531, Jul. 2006.

[2] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz., "Dynamic control of positional information in the early Drosophila embryo," *Nature*, vol. 430, pp. 368–371, 2004.

[3] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein, "Relating three-dimensional structures to protein networks provides evolutionary insights.," *Science (New York, N.Y.)*, vol. 314, no. 5807, pp. 1938–41, Dec. 2006.

[4] A. Presser, M. B. Elowitz, M. Kellis, and R. Kishony, "The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 3, pp. 950–4, Jan. 2008.

[5] S. Wuchty, Z. N. Oltvai, and A.-L. Barabasi., "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nature Genetics*, vol. 35, pp. 176–179, 2003.

[6] D. Koller and N. Friedman, *Probabilistic Graphical Models Principles and Techniques*. The MIT Press, Cambridge Massachusetts, London, England.

[7] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks.," *Bioinformatics (Oxford, England)*, vol. 18, no. 2, pp. 261–274, Oct. 2002.

[8] S. Huang, "Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery.," *Journal of molecular medicine (Berlin, Germany)*, vol. 77, no. 6, pp. 469–480, Jun. 1999.

[9] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data.," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, no. 3–4, pp. 601–620, Jan. 2000.

[10] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks.," *Proceedings / IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference*, vol. 2, pp. 104–13, Jan. 2003.

[11] N. Friedman, N. Iftach, and D. Peér, "Learning Bayesian Network Structure from Massive Datasets : The ' Sparse Candidate ' Algorithm," in *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 1999, pp. 206–215.

[12] J. Hu, J. Wan, L. Hackler, D. J. Zack, and J. Qian, "Computational analysis of tissue-specific gene networks: application to murine retinal functional studies.," *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. 2289–2297, Jul. 2010.

[13] F. Browne, H. Wang, H. Zheng, and F. Azuaje, "A knowledge-driven probabilistic framework for the prediction of protein-protein interaction networks.," *Computers in biology and medicine*, vol. 40, no. 3, pp. 306–317, Jan. 2010.

[14] S. Bulashevska, A. Bulashevska, and R. Eils, "Bayesian statistical modelling of human protein interaction network incorporating protein disorder information.," *BMC bioinformatics*, vol. 11, no. 46, Jan. 2010.

[15] C. Borgelt and R. Kruse, "Possibilistic Graphical Models," pp. 1–17.

[16] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks.," *Bioinformatics (Oxford, England)*, vol. 21, no. 6, pp. 754–764, Sep. 2005.

[17] X. Deng, H. Geng, and H. Ali, "EXAMINE: a computational approach to reconstructing gene regulatory networks.," *Bio Systems*, vol. 81, no. 2, pp. 125–136, Aug. 2005.

[18] Z. Wang, X. Liu, Y. Liu, J. Liang, and V. Vinciotti, "An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series.," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 6, no. 3, pp. 410–419, Jul. 2009.

[19] X. Wu, Y. Ye, and R. K. Subramanian, "Interactive Analysis of Gene Interactions Using Graphical Gaussian Model," in *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, 2003, pp. 63–69.

[20] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 196–212, Jul. 2004.

[21] S. L. Miller and D. G. Childers, *Probability and Random Processes With Applications to Signal Processing and Communications*. Elsevier Academic Press Publications, 2004, pp. 1–529.

[22] S. L. Miller and D. G. Childers, *Probability and Random Processes With Applications to Signal Processing and Communications*. Elsevier Academic Press Publications, 2004, pp. 1–529.

[23] Wikiedia, "Conditional Normal Distribution," 2010. [Online]. Available: http://www.edegan.com/wiki/index.php/Conditional_Normal_Distribution.

[24] S. Y. Kim, S. Imoto, and S. Miyano, "Inferring gene networks from time series microarray data using dynamic Bayesian networks.," *Briefings in bioinformatics*, vol. 4, no. 3, pp. 228–235, Sep. 2003.

[25] K. Murphy and S. Mian, "Modelling Gene Expression Data using Dynamic Bayesian Networks."

[26] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.," *Bioinformatics (Oxford, England)*, vol. 21, no. 1, pp. 71–79, Aug. 2005.

[27] C.-C. Wu, S. Asgharzadeh, T. J. Triche, and D. Z. D'Argenio, "Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning.," *Bioinformatics (Oxford, England)*, vol. 26, no. 6, pp. 807–813, Feb. 2010.

[28] Y.-Q. Qiu, S. Zhang, X.-S. Zhang, and L. Chen, "Detecting disease associated modules and prioritizing active genes based on high throughput data.," *BMC bioinformatics*, vol. 11, no. 26, Jan. 2010.

[29] H. Wang, D. Mirota, and G. D. Hager, "A generalized Kernel Consensus-based robust estimator.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 178–184, Jan. 2010.

[30] R. Boscolo, H. Pan, and V. P. Roychowdhury, "Independent component analysis based on nonparametric density estimation.," *IEEE transactions on neural networks /a publication of the IEEE Neural Networks Council*, vol. 15, no. 1, pp. 55–65, Jan. 2004.

[31] E. Barnard, "Maximum Leave-one-out Likelihood for Kernel Density Estimation," in *Proc. of the 21st annual Symposium of the Pattern Recognition Association*, 2010, pp. 19–24.

[32] Wikipedia, "Kernel functions in common use," 2012. .

[33] Wikipedia, "Kernel density estimation," 2010. .

[34] K. Wang, M. Narayanan, H. Zhong, M. Tompa, E. E. Schadt, and J. Zhu, "Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases.," *PLoS computational biology*, vol. 5, no. 12, p. e1000616, Dec. 2009.

[35] G. Y. Zou and A. Donner, "The Merits of Testing Hardy-Weinberg Equilibrium in the Analysis of Unmatched Case-Control Data : A Cautionary Note," *Annals of Human Genetics*, vol. 70, no. 6, pp. 923–933, Nov. 2006.

[36] K. Song and R. C. Elston, "A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies.," *Statistics in medicine*, vol. 25, no. 1, pp. 105–126, Nov. 2006.

[37] G. Zheng, K. Song, and R. C. Elston, "Adaptive two-stage analysis of genetic association in case-control designs.," *Human heredity*, vol. 63, no. 3–4, pp. 175–86, Jan. 2007.

[38] D. W. Fardo, K. D. Becker, L. Bertram, R. E. Tanzi, and C. Lange, "Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium.," *European journal of human genetics : EJHG*, vol. 17, no. 12, pp. 1676–1682, Jun. 2009.

[39] G. K. Chen and D. C. Thomas, "Using biological knowledge to discover higher order interactions in genetic association studies.," *Genetic epidemiology*, vol. 34, no. 8, pp. 863–878, Sep. 2010.

[40] L. Wittgenstein, "Unit 4 Categorical Data Analysis," PubHlth, 2012, pp. 1–87.

[41] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray.," *Science (New York, N.Y.)*, vol. 270, no. 5235, pp. 467–470, Oct. 1995.

[42] N. R. Clark, R. Dannenfelser, C. M. Tan, M. E. Komosinski, and A. Ma'ayan, "Sets2Networks: network inference from repeated observations of sets.," *BMC systems biology*, vol. 6, no. 1, p. 89, Jul. 2012.

[43] N. Noman and H. Iba, "Inferring gene regulatory networks using differential evolution with local search heuristics.," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 4, no. 4, pp. 634–647, Oct. 2007.

[44] D. Ucar, I. Neuhaus, P. Ross-MacDonald, C. Tilford, S. Parthasarathy, N. Siemers, and R.-R. Ji, "Construction of a reference gene association network from multiple profiling data: application to data analysis.," *Bioinformatics (Oxford, England)*, vol. 23, no. 20, pp. 2716–2724, Aug. 2007.

[45] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks.," *Bioinformatics (Oxford, England)*, vol. 21, no. 6, pp. 754–764, Sep. 2005.

[46] B. F. Wong, C. K. Carter, and R. Kohn, "Efficient estimation of covariance selection models," *Biometrika*, vol. 90, no. 4, pp. 809–830, Dec. 2003.

[47] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data.," *Bioinformatics (Oxford, England)*, vol. 18 Suppl.1, pp. S145–S154, Mar. 2002.

[48] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert, "Classification of microarray data using gene networks.," *BMC bioinformatics*, vol. 8, no. 35, Feb. 2007.

[49] A. Benso, S. Di Carlo, and G. Politano, "A cDNA microarray gene expression data classifier for clinical diagnostics based on graph theory.," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 8, no. 3, pp. 577–91, May 2011.

[50] E. Parzen, "On Estimation of a Probability Density Function and Mode." Stanford University, 1961.

[51] W. Jiang, X. Li, S. Rao, L. Wang, L. Du, C. Li, C. Wu, H. Wang, Y. Wang, and B. Yang, "Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements.," *BMC systems biology*, vol. 2, no. 72, Aug. 2008.

[52] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks.," *Bioinformatics (Oxford, England)*, vol. 21, no. 6, pp. 754–64, Mar. 2005.

[53] E. Chaibub Neto, C. T. Ferrara, A. D. Attie, and B. S. Yandell, "Inferring causal phenotype networks from segregating populations.," *Genetics*, vol. 179, no. 2, pp. 1089–1100, Apr. 2008.

[54] X.-W. Chen, G. Anantha, and X. Wang, "An effective structure learning method for constructing gene networks.," *Bioinformatics (Oxford, England)*, vol. 22, no. 11, pp. 1367–1374, Mar. 2006.

[55] B. E. Hansen, "Nonparametric Conditional Density Estimation," no. November, 2004.

[56] K. Yu, A. K. Ally, and D. J. Hand, "Kernel quantile-based estimation of," vol. 12, no. 4, pp. 15–32, 2010.

[57] J. Hayya, D. Armstrong, and N. Gressis, "A NOTE ON THE RATIO OF TWO NORMALLY DISTRIBUTED VARIABLES," *The Institute of Management Sciences*, vol. 21, no. 11, pp. 1338–1341, Jul. 1975.

[58] S. Y. Shatskikha, "MULTIVARIATE CAUCHY DISTRIBUTIONS AS LOCALLY GAUSSIAN DISTRIBUTIONS," *Journal of Mathematical Sciences*, vol. 78, no. 1, pp. 102–108, Jan. 1996.

[59] F. M. Alakwaa, N. H. Solouma, and Y. M. Kadah, "Construction of gene regulatory networks using biclustering and Bayesian networks.," *Theoretical biology & medical modelling*, vol. 8, no. 39, Oct. 2011.

[60] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, "Increasing confidence of protein interactomes using network topological metrics.," *Bioinformatics (Oxford, England)*, vol. 22, no. 16, pp. 1998–2004, Jul. 2006.

[61] J. J. Cai, E. Borenstein, and D. a Petrov, "Broker genes in human disease.," *Genome biology and evolution*, vol. 2, pp. 815–25, Oct. 2010.

[62] M. Hajduch, L. B. Hearne, J. a Miernyk, J. E. Casteel, T. Joshi, G. K. Agrawal, Z. Song, M. Zhou, D. Xu, and J. J. Thelen, "Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression.," *Plant physiology*, vol. 152, no. 4, pp. 2078–2087, Apr. 2010.

[63] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann, "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana.," *Genome biology*, vol. 5, no. 11, p. R92, Jan. 2004.

[64] K. Koch, "Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development.," *Current opinion in plant biology*, vol. 7, no. 3, pp. 235–46, Jun. 2004.

[65] E. H. P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, "The Arabidopsis Information Resource (TAIR): improved gene annotation and new Tools," *Nucleic. Acids. Res*, vol. 40, no. D, pp. 1202–1210, 2012.

[66] T. Obayashi, S. Hayashi, M. Saeki, H. Ohta, K. Kinoshita, and and K. K. T. Obayashi, S. Hayashi, M. Saeki, H. Ohta, "ATTED-II provides coexpressed gene networks for Arabidopsis," *Nucleic. Acids. Res*, vol. 37, no. Database, pp. D987–991, Jan. 2009.

[67] and M. C. S.-F. M. M. Brandão, L. L. Dantas, "AtPIN: Arabidopsis thaliana protein interaction network," *BMC. Bioinformatics*, vol. 10, p. 454, 2009.

[68] G. Liu, L. Wong, and H. N. Chua, "Complex discovery from weighted PPI networks.," *Bioinformatics (Oxford, England)*, vol. 25, no. 15, pp. 1891–7, Aug. 2009.

[69] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein, "Relating three-dimensional structures to protein networks provides evolutionary insights.," *Science (New York, N.Y.)*, vol. 314, no. 5807, pp. 1938–41, Dec. 2006.

[70] S. Swatek, K, K. Graham, K. Agrawal, G, and J. J. Thelen, "The 14-3-3 isoforms chi and epsilon differentially bind client proteins from developing Arabidopsis seed," *J Proteome Res*, vol. 10, no. 9, pp. 4076–4087, Sep. 2011.

[71] C. Wang, A. Marshall, D. Zhang, and Z. a Wilson, "ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis.," *Plant physiology*, vol. 158, no. 4, pp. 1523–1533, Apr. 2012.

[72] and Z. A. W. C. Wang, A. Marshall, D. Zhang, "ANAP: an integrated knowledge base for Arabidopsis protein interaction network Analysis," *Plant. Physiology*, vol. 158, no. 4, pp. 1523–1533, 2012.

[73] M. Meng, M. Geisler, H. Johansson, J. Harholt, H. V Scheller, E. J. Mellerowicz, and L. a Kleczkowski, "UDP-glucose pyrophosphorylase is not rate limiting, but is essential in Arabidopsis.," *Plant & cell physiology*, vol. 50, no. 5, pp. 998–1011, May 2009.

[74] M. Meng, M. Geisler, H. Johansson, J. Harholt, H. V Scheller, E. J. Mellerowicz, and L. a Kleczkowski, "UDP-glucose pyrophosphorylase is not rate limiting, but is essential in Arabidopsis.," *Plant & cell physiology*, vol. 50, no. 5, pp. 998–1011, May 2009.

[75] W. Lu, X. Tang, Y. Huo, R. Xu, S. Qi, J. Huang, C. Zheng, and C. Wu, "Identification and characterization of fructose 1,6-bisphosphate aldolase genes in Arabidopsis reveal a gene family with diverse responses to abiotic stresses.," *Gene*, vol. 503, no. 1, pp. 65–74, Jul. 2012.

[76] J. Muñoz-Bertomeu, B. Cascales-Miñana, J. M. Mulet, E. Baroja-Fernández, J. Pozueta-Romero, J. M. Kuhn, J. Segura, and R. Ros, "Plastidial glyceraldehyde-3-phosphate dehydrogenase deficiency leads to altered root development and affects the sugar and amino acid balance in Arabidopsis.," *Plant physiology*, vol. 151, no. 2, pp. 541–58, Oct. 2009.

[77] S. Hong-Bo, L. Zong-Suo, and S. Ming-An, "LEA proteins in higher plants: structure, function, gene expression and regulation.," *Colloids and surfaces. B, Biointerfaces*, vol. 45, no. 3–4, pp. 131–135, Jul. 2005.

[78] L. Koumakis, V. Moustakis, M. Zervakis, M. E. Kafetzopoulos, and D. Potamias, "Coupling Regulatory Networks and Microarays : Revealing Molecular Regulations of Breast Cancer," *Artificial Intelligence: Theories and Applications. Lecture Notes in Computer Science*, vol. 7297, pp. 239–246, 2012.

[79] N. E. Hynes and G. MacDonald, "ErbB receptors and signaling pathways in cancer.," *Current opinion in cell biology*, vol. 21, no. 2, pp. 177–84, Apr. 2009.

[80] L. F. Hernandez-Aya and A. M. Gonzalez-Angulo, "Targeting the phosphatidylinositol 3-kinase signaling pathway in breast cancer.," *The oncologist*, vol. 16, no. 4, pp. 404–14, Jan. 2011.

[81] E. López-Knowles, S. a O'Toole, C. M. McNeil, E. K. a Millar, M. R. Qiu, P. Crea, R. J. Daly, E. a Musgrove, and R. L. Sutherland, "PI3K pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality.," *International journal of cancer. Journal international du cancer*, vol. 126, no. 5, pp. 1121–31, Mar. 2010.

[82] K. Stemke-hale, A. M. Gonzalez-angulo, A. Lluch, R. M. Neve, W. Kuo, M. Davies, M. Carey, Z. Hu, Y. Guan, W. F. Symmans, L. Pusztai, L. K. Nolden, H. Horlings, M. Hung, M. J. Van De Vijver, V. Valero, and J. W. Gray, "NIH Public Access," *Cancer Research*, vol. 68, no. 15, pp. 6084–6091, Aug. 2009.

[83] K. D. Kalantzaki, E. S. Bei, M. Garofalakis, and M. Zervakis, "Biological Interaction Networks Based on Sparse Temporal Expansion of Graphical Models," in *Proc. 12th IEEE International Conference on BioInformatics and BioEngineering*, 2012, pp. 460–465.

[84] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers, "The BioGRID Interaction Database: 2011 update.," *Nucleic acids research*, vol. 39, no. Database issue, pp. D698–D704, Nov. 2011.

[85] M. H. Schaefer, J.-F. Fontaine, A. Vinayagam, P. Porras, E. E. Wanker, and M. a Andrade-Navarro, "HIPPIE: Integrating protein interaction networks with experiment based quality scores.," *PloS one*, vol. 7, no. 2, p. e31826, Feb. 2012.

[86] P. J. Lisboa, T. a Etchells, I. H. Jarman, and S. J. Chambers, "Finding reproducible cluster partitions for the k-means algorithm," *BMC Bioinformatics*, vol. 14 Suppl.1, no. S8, Sep. 2013.

[87] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts.," *Nucleic acids research*, vol. 33, no. Web Server issue, pp. W741–W748, Apr. 2005.

[88] S. Choi and J. N. Myers, "Molecular Pathogenesis of Oral Squamous Cell Carcinoma: Implications for Therapy," *Journal of Dental Research*, vol. 87, no. 1, pp. 14–32, Jan. 2008.

[89] S. D. da Silva, A. Ferlito, R. P. Takes, R. H. Brakenhoff, M. D. Valentin, J. a Woolgar, C. R. Bradford, J. P. Rodrigo, A. Rinaldo, M. P. Hier, and L. P. Kowalski, "Advances and applications of oral cancer basic research.," *Oral oncology*, vol. 47, no. 9, pp. 783–791, Sep. 2011.

[90] C. M. Stellrecht and V. Gandhi, "MET receptor tyrosine kinase as a therapeutic anticancer target.," *Cancer letters*, vol. 280, no. 1, pp. 1–14, Oct. 2009.

[91] J. P. A. Bánkfalvi, M. Krassort, I. B Buchwalow, A. Végh , E. Felszeghy, "Gains and losses of adhesion molecules (CD44, E-cadherin, and β-catenin) during oral carcinogenesis and tumour progression," *Journal of Pathology*, vol. 198, no. 3, pp. 343–351, Jul. 2002.

[92] M. W. Gonzalez and M. G. Kann, "Chapter 4: protein interactions and disease.," *PLoS computational biology*, vol. 8, no. 12, p. e1002819, Dec. 2012.

[93] Y. Fang, W. Benjamin, M. Sun, and K. Ramani, "Global geometric affinity for revealing high fidelity protein interaction network.," *PloS one*, vol. 6, no. 5, p. e19349, Jan. 2011.

[94] K. L. Mueller, Z.-Q. Yang, R. Haddad, S. P. Ethier, and J. L. Boerner, "EGFR/Met association regulates EGFR TKI resistance in breast cancer.," *Journal of molecular signaling*, vol. 5, no. 8, Jul. 2010.

[95] U. D. Müller-Richter, A. Dowejko, T. Reuther, J. Kleinheinz, E. T. Reichert, and O. Driemel, "Analysis of expression profiles of MAGE-A antigens in oral squamous cell carcinoma cell lines," *Head & Face Medicine*, vol. 5, no. 10, Apr. 2009.

[96] S. Pennacchietti, P. Michieli, M. Galluzzo, M. Mazzone, S. Giordano, and P. M. Comoglio, "Hypoxia promotes invasive growth by transcriptional activation of the met protooncogene.," *Cancer cell*, vol. 3, no. 4, pp. 347–361, Apr. 2003.

[97] I. J. Brusevold, M. Aasrum, M. Bryne, and T. Christoffersen, "Migration induced by epidermal and hepatocyte growth factors in oral squamous carcinoma cells in vitro: role of MEK/ERK, p38 and PI-3 kinase/Akt.," *Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology*, vol. 41, no. 7, pp. 547–558, Feb. 2012.

[98] H. Xu, P. L. Stabile, T. C. Gubish, E. W. Gooding, R. J. Grandis, and M. J. Siegfried, "Dual blockade of EGFR and c-Met abrogates redundant signaling and proliferation in head and neck carcinoma cells," *Clinical Cancer Research*, vol. 17, no. 13, pp. 4425–4438, Jul. 2011.

[99] R. Raftogianis and A. Godwin, "Impact of Protein Interaction Technologies on Cancer Biology and Pharmacogenetics," in *Protein–Protein Interactions: A Molecular Cloning Manual (Cold Spring Harbor Laboratory Press)*, no. Chapter 3, 2002, pp. 15–68.

[100] J. A. Gentles and D. Gallahan, "Meeting Report: 'Systems Biology: Confronting the Complexity of Cancer'," *Cancer Research*, vol. 71, no. 18, pp. 5961–5964, Sep. 2011.

[101] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[102] P. Bernstein, C. Sticht, A. Jacobi, C. Liebers, S. Manthey, and M. Stiehler, "Expression pattern differences between osteoarthritic chondrocytes and mesenchymal stem cells during chondrogenic differentiation.," *Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society*, vol. 18, no. 12, pp. 1596–1607, Sep. 2010.