Nonparametric Network Design and Analysis of Disease Genes in Oral Cancer Progression

K. Kalantzaki[1], E. S. Bei[1], K. P. Exarchos[2], M. Zervakis[1] *Member, IEEE*, M. Garofalakis[1]*Member,IEEE*, D. I. Fotiadis[2] *Senior Member, IEEE*

***Abstract*— 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 identified common disease-related structures that may decipher the association between disease state and biological processes in oral cancer. Our study 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.**

***Index Terms*— Kernel density estimation, partial correlation, gene network construction, oral cancer**

I. INTRODUCTION

Oral carcinogenesis is a multistep process implicating numerous genetic events such as changes of oncogenes and tumour suppressor genes[1], justifying that oral cancer can be conceptualized by means of networks of molecular interactions. Moreover, functional associations between different genes within molecular networks that organize specific biological processes could reveal possible aetiopathological mechanisms of various diseases, including oral cancer. A variety of high-throughput experimental data, such as DNA microarray andChIP-chip technology allow the simultaneous measurements of expression levels and generate large datasets associated with various cellular procedures [2].

The extended study of such datasets has provided a new perspective in gene-gene network association studies, with the network construction from experimental data being a promising approach in modeling functional processes. Gene regulatory networks (GRN) [3] 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. Thus, the study of GRNs from diseased tissue is crucial to understanding complex cancer phenotypes and inventing effective therapeutic regimens [4].

Several computational methodologies have been applied to construct biological networks using different data sources [5]. 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 linear and nonlinear models [6], [7], Boolean network models [8], Bayesian networks[9], [10], Pearson's correlation-based approaches[11], [12], clustering and classification algorithms[13]-[15]. 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 [5], [6], [16]. 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 in thecomputationof 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 [12].

Kernel-based models have demonstrated very competitive computational performance due to their ability to model nonlinear systems and high-dimension data [19]. Support vector machines and relevance vector machines [17] 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 [18] is also a promising

[1]K. Kalantzaki, E. S. Bei, M. Garofalakis and M. Zervakis are with the Department of Electronic and Computer Engineering, TUC, Chania 73100, Greece (kkalantzaki@isc.tuc.gr, abei@isc.tuc.gr, michalis@display.tuc.gr minos@acm.org).

[2]K. P. Exarchos and D. I. Fotiadis are with the Department of Materials Science and Engineering, University of Ioannina, Ioannina, 45110, Greece (kexarcho@gmail.com, fotiadis@cc.uoi.gr).

technique for gene 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, several methods [18] 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 [19] that analyze the underlying mechanisms of differential pathways and molecules responsible for abnormalities of a specific disease.

The analysis reported herein is an effort in revealing and modeling the inter-relationships of molecules in oral cancer that participate in many different pathways incriminated for this disease. The proposed method for network construction is based on Kernel density estimation (KDE), as an attempt to model the nonlinear effect of gene interactions and fill the information loss from the data samples. The novelty of our approach relates to the use of kernels for identifying the genetic dependencies in the network structure, instead of using estimators as in previous studies. Our framework is applied on experimental blood data of oral cancer patients received from four successive follow-ups. The goal is to reveal the differences in network structure between different time stages, in addition to conspicuous genes that play a central role in all stages of the disease. Furthermore, by applying our framework to tissue dataset we successively identify common disease related structures which are known as molecular modules responsible for oral cancer. Importantly, our framework can be extended to other data types, besides gene expression. According to ROC curves, KDE outperforms PC with the area under the curve (AUC) reaching 83% and 77%, respectively. Our study exemplifies the central role of *MET* proto-oncogene and MET network in oral cancer, as well as its interactions with EGFR, HIF1A and MAGEA6, which may be of importance in oral cancer progression.

II. Methods and Procedures

A. Proposed Framework

A generic framework for gene 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 partial correlations (PC) and KDE approaches, but any other network construction method can be applied. 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 interaction. 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. The application of our framework to specific datasets provides insight into its effectiveness and reliability, as presented in the results section.

1) Partial Correlation Estimation

Pairwise associations of coexpressed molecules can be modeled through Pearson's correlation. The interaction identification between two variables is reduced to estimating the covariance matrix S. Each element in S_{ik}, via $S_{ik}=\rho_{ik}\sigma_i\sigma_k$ and $S_{ii}=\sigma_i^2$, represents the correlation coefficient ρ_{ik} between nodes X_i and X_k and indicates an association, while σ_i^2 denotes the variance of node X_i. A high correlation coefficient between any two genes may be indicative of either direct interaction, or indirect interaction or regulation by a common gene. However, for the construction of a gene 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 [20] 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 i and k conditioned on all remaining genes of the network. This property is reflected in the inverse covariance matrix S^{-1}, with elements:

$$\pi_{ik}=-\frac{S_{ik}^{-1}}{\sqrt{S_{ii}^{-1}S_{kk}^{-1}}} . \qquad (1)$$

Given the experimental data, the covariance matrix is computed and then it is inverted. Indeed, using Eqn. (1) 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, suggesting the addition of an edge between the nodes.

Despite its straightforward 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 [11], an approximation of the standard matrix inverse, based on the singular value decomposition (SVD).

2) Kernel Density Estimation

Kernel density estimation[17], [18], [21], [22] is a non-parametric approach that estimates the probability density function (pdf) of a random variable. 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 X. KDE allows the estimation of X as follows:

$$\hat{f}_h=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) , \qquad (2)$$

where $K(u)=\frac{1}{2\pi}e^{\frac{-1}{2}u^2}$ is a symmetric positive definite Gaussian function, n is dataset's size of the gene X and $h>0$ is a smoothing parameter, the bandwidth that controls the extent of the kernel [21].

Genes interacting with each other can be expressed as a

network. The gene expression data provides valuable information on a gene's activity, which can be represented by weights of genes and interactions in the network. Formally, an interaction network with weighted nodes and weighted edges can be expressed as *G=(V, E)*, where node set *V* represents genes, edge set *E* represents interactions. Under the assumption that gene 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 Pearson's cross correlation test:

$$f_h(X_i,X_k)=f_h(X_i)*f_h(X_k)\ . \quad (3)$$

The smaller the absolute difference between the two sides of the equation, the more independent the corresponding nodes are. In contrast, high absolute difference indicates dependence between X_i and X_k, and, thus, connection betweenthe candidate nodes. This means that $X_i$and X_k, share common information characteristics that imply interaction. Reducing this scheme to correlations tests, the more correlated the two sides of the above equation, the more independent genes $X_i$and $X_k$are. Otherwise, the candidate genes share dependencies which implies association.

3) Effects of External Genes

The poor performance of 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 it, is to characterize the produced associations according to Gene Ontology (GO) terms at a higher level of processes [23], [24]. Other approaches [25], [26] 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 [27], which can be identified in various available databases [24]. 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 [24] 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 necessarily be false positives.

In order to compare the performance of the proposed framework with and without external interactions, we employ the receiver operator characteristic (ROC) and precision-recall curves, as described in Davis and Goadrich (2006) [28]. 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 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 Fig. 1. 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.

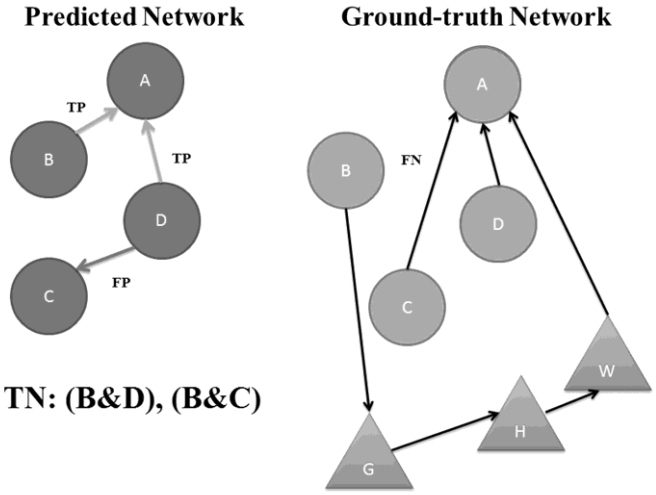

Fig. 1**.**Graphical representation of the TP, FP, TN, FN edges according to the existent knowledge of the ground-truth network. External genes are represented with triangles while studied genes with circles. External pathways that indirectly connect studied genes give TP connections.

4) 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, these approaches do not imply anything about directionality, indicating which node is the cause and which is the effect. 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 distributions*f(.)* or likelihoodsare 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 [29], [30]:

$$f(X_k|X_i)f(X_i)=f(X_i,X_k)=f(X_k)f(X_i|X_k). \quad (4).$$

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 [29]:

$$BIC=-2\log\hat{L}+K\log N, \quad (5)$$

where $\hat{L}$ is the maximum likelihood, K the number of parameters to be estimated in the model, and N denotesthe 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.

BIC also offers the advantage of introducing larger penalty for the number of parameters, resulting in highly reliable edge directions [31].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.

B. Oral Cancer Dataset

The framework is applied on a dataset related to oral cancer [32] which includes 86 subjects [NeoMark project, FP7-ICT-2007-224483, ICT enabled prediction of cancer reoccurrence - D6.1: Research protocol] that have been enrolled from three major clinical centers residing in Italy (University Hospital of Parma and National Cancer Center Regina Elena) and Spain (MD Anderson Cancer Center). All patients have been diagnosed with oral squamous cell carcinoma (OSCC) and have reached remission, after successful therapeutic intervention (i.e. surgery, chemo/radiotherapy). Thereafter, gene expression data were collected both from the primary tumor as well as from circulating blood cells, at the baseline state of the patient. The clinical data contain standard measurements and laboratory markers from the patient's medical record as well as pathology and risk factor data referring to the organism as a whole. More detailed clinical information has been recently described by Exarchos et al. [32].

The oral cancer dataset consists of a) tissue genomic data from diagnosis (86 samples), b) blood genomic data (23 samples), and c) blood genomic follow-up data (maximum 23 samples) [32]. Four different time stages were analyzed, corresponding to the first, third, sixth, and ninth months after the initial diagnosis. Notice that, during the follow-up periods the sample numbers were reduced (from 23 at 1st to 10 at 3rd, 13 at 6th, and 4 at 9th month). The network constructed from blood samples is considered for all temporal stages, and is also compared with the one produced by the tissue samples at the point of the first diagnosis.

The proposed framework is applied on both blood and tissue datasets. In order to discover functional associations between genes in the constructed networks and possible network changes during disease progression from a large amount of genomic expression data, we generated gene interaction networks by entering specific genes as input in the graphical model. These genes are selected based on literature and particularly on a set of previously implicated genes for oral cancer [NeoMark project, FP7-ICT-2007-224483, ICT enabled prediction of cancer reoccurrence], [32]. The final gene list consists of 110 genes that are related to oral cancer disease and 5 control genes that are not related to oral cancer (Supplementary (S) Table I). Control genes are included in this list in order to test the algorithm correctness (*ERBB4*) and as positive (*FGFR1*) or negative (*BRCA1, MBNL1, PARK7*) reference for oral cancer (STable I). The estimated network structure from blood samples is compared with all temporal stages, as well as with the network produced from tissue samples at the first diagnosis.

In order to perform functional enrichment tests of the selected genes of each molecular network, we used WebGestalt (WEB-based GEneSeTAnaLysisToolkit) for Gene Ontology (GO) term analysis. WebGestalt (Version 2.0) applies the hypergeometric test for the enrichment of GO terms in the selected genes, and the Benjamini& Hochberg (BH) method for the multiple test adjustment (adj$P$) [33].

III. RESULTS

A. Statistical Results

In order to investigate the statistical properties of the proposed methodology, we apply PC and KDE approaches to reveal network structure from gene expression data. In a previous work [34],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. For pdf estimation we follow the Parzen approach with a Gaussian kernel function, since the histogram of genes in the preprocessing stage approximate Gaussian characteristics. 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.

1) Direct Interactions

Table I presents the number of gene interactions on blood samples, for the first follow-up. Accordingly, Table II presents the gene associations on tissue samples. Both tables present the number of TP edges that PC and KDE identified among a set of experimentally known genetic interactions[35]**.** The first column describes different thresholds *th* for partial correlation set on PC for Eqn. (1) ($\pi_{ik}\geq th$), while the second column provides the thresholds of correlation r between the two members of Eqn. (3) for KDE ($r\leq th$). The 3 to 4 columns summarize the verified numbers of direct and indirect gene to gene interactions for both approaches. The fifth and sixth columns present the number of new edges that have occurred

TABLE I
GENE-GENE INTERACTIONS FOR THE FIRST FOLLOW-UP ON BLOOD SAMPLES[A]

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

[a]Bold columns of PC and KDE indicate the gene interactions considering the external genes. Threshold *th* is defined as $\pi_{ik} \geq th$ for PC and $r \leq th$ for KDE.

TABLE IIGENE-GENE INTERACTIONS ON TISSUE SAMPLES[A]

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

[a]Bold columns of PC and KDE indicate the gene interactions considering the external genes. Threshold *th* is defined as $\pi_{ik} \geq th$ for PC and $r \leq th$ for KDE.

for each threshold, 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 [36]. 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 [35].

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 observation indicates that molecules from various pathways are not likely to directly interact. This is also verified by the small number of

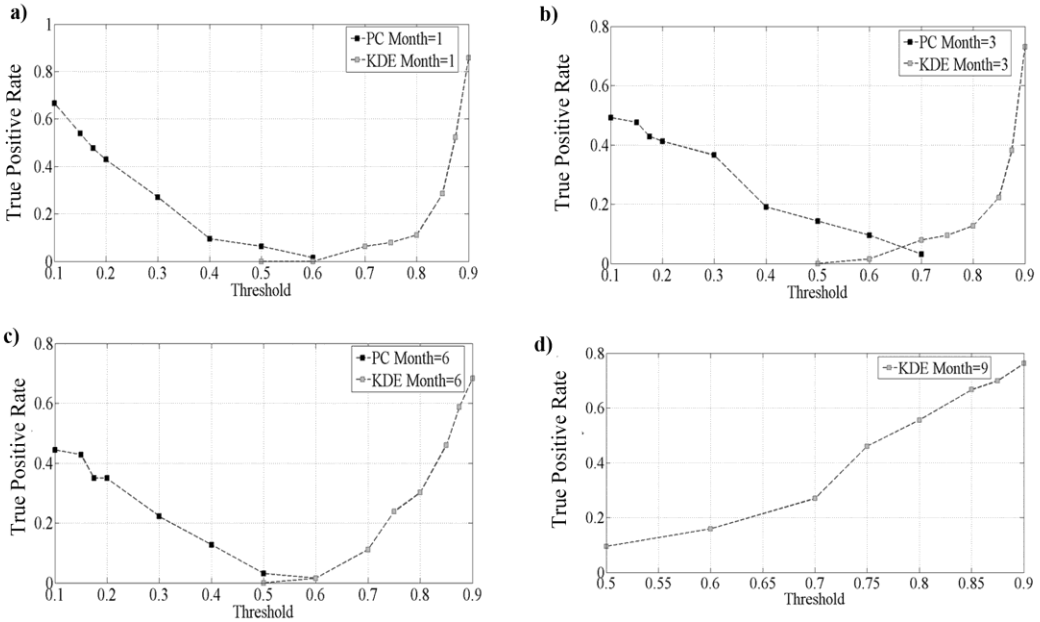

Fig. 2.True positive rate for all time stages on blood samples.

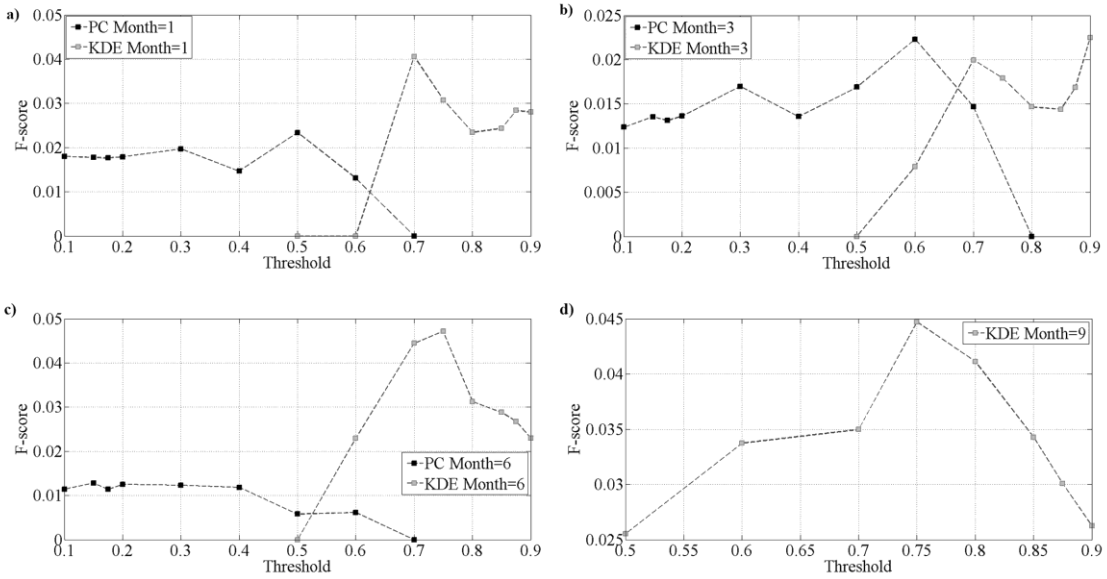

Fig. 3.F-score metric for all follow-ups on blood samples.

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.

Tables I and II 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 Fig.2 the true positive rate for the networks constructed from all monthly follow-ups. This figure (Fig. 2) justify KDE's superiority in detecting existent interactions over PC for all oral cancer stages. Surprisingly, for the last follow-up (Fig. 2d), 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-existingassociations increases. Fig. 3 presents the performance of the two methodologies for all thresholds, according to the F-score metric [28]:

$$F=\frac{2*precision*recall}{precision+recall}. \quad (6)$$

For each temporal instant, the F-score analysis derives the thresholds 0.7, 0.9, 0.75 and 0.75for 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.

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

2) Indirect Interactions using External Genes

In section II.A.3 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 [27]. 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[24] 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,

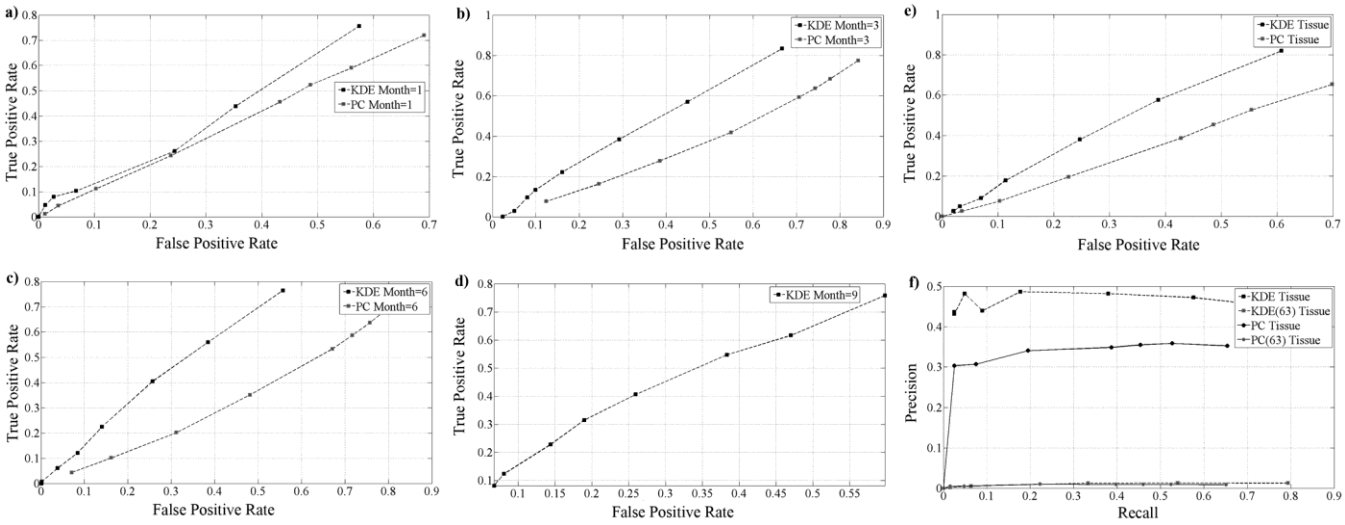

Fig. 4.ROC comparison between KDE and PC for the blood samples **(a-d)**. For the ninth month, PC cannot derive a reliable network structure. Apparently KDE results in larger AUC for all cases. **(e)** ROC curve comparison between KDE and PC for the tissue network. KDE outperforms PC as it covers larger AUC; **(f)** 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.

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.

Bolded columns in Tables I and II 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 (Fig. 1). Furthermore, from the 115 analyzed genes, there are 22 uncharacterized, for which BioNetBuilder resulted in none association; in our framework we did not take into account edges connecting these genes. Notice that for this reason the number of new edges (columns 5, 6) in Table II differs from Table I, 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.

Fig. 4 presents the ROC curves for the blood samples associated with the 4 follow-ups (Fig. 4a-4d), while Fig.4e 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, Figs. 5a-5d and 4f, 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.

B. 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 [37] on the primary gene expression data and recovered five major clusters (Fig. 6a, b). As shown in Figs. 6a and b, the content and structure of blood and tissue networks based on gene interactions are different at the 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[33]. The genes in the same cluster are densely connected with each other (Fig. 6b), and GO analysis indicates that these five gene-clusters are enriched in certain GO annotation terms (STable 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 [1], [38]. Regardless of GO terms in the category of biological process, we found that cell proliferation (adj$P$=6.94×10^{-9}), regulation of cell proliferation (adj$P$=2.58×10^{-7}), and regulation of cell cycle (adj$P$=4.48×10^{-7}) are significantly enriched in these gene clusters of both blood and tissue samples, as well as in blood follow up samples (Fig. 6a, b; SFigs. 1a, b, 2a, b; STable

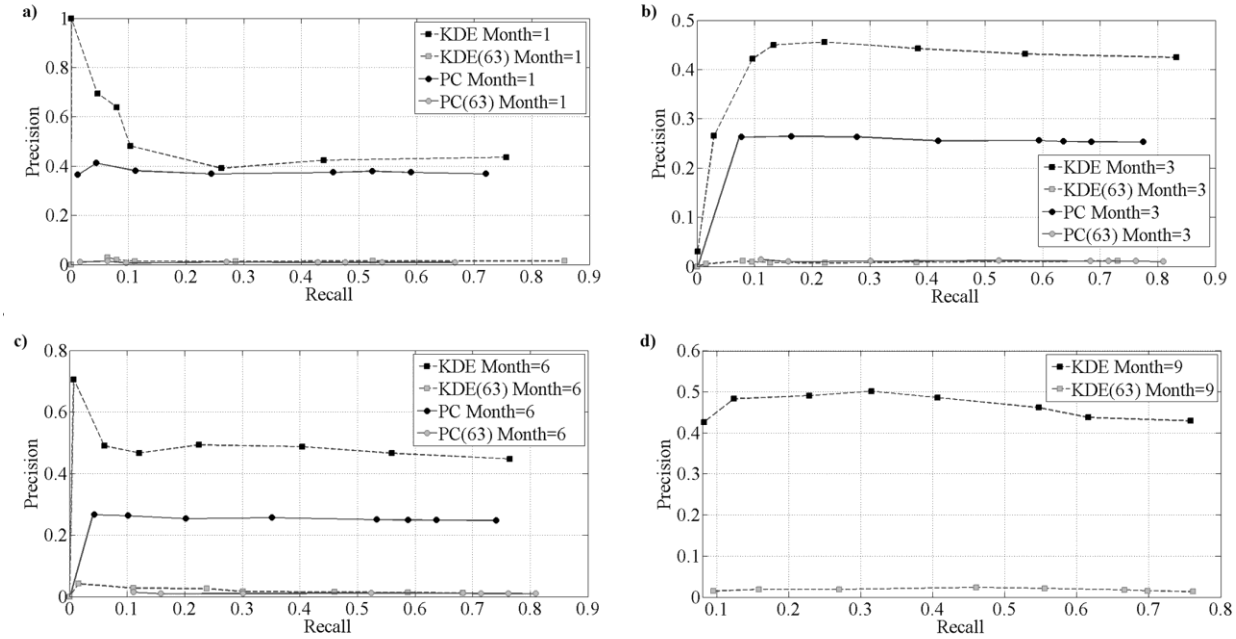

Fig. 5.Precision vs Recall comparison between KDE and PC for all cases on blood samples **(a-d)**. KDE(63) and PC(63) represent the networks considering as TP the set of 63 direct interactions, while KDE and PC curves represent the performance considering as TP all direct and indirect edges. KDE outperforms PC reaching higher levels of precision and recall for all periods. For the ninth month, PC could not result in a reliable network structure.

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 (STable 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 [36], [38], cellular processes which are reflected by the above enriched GO terms but also by less significant GO terms (SFigs. 1a, b, 2a, b; STable II).

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 III.A.1. Fig. 7 presents the "intersection genes", i.e. genes that are induced by the network intersection for all time stages for the blood samples (Fig. 7a, STableIIIc), as well as the intersection of tissue network with the blood from the first time stage (Fig. 7b, STableIIIc). Figs. 7a and 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* on both tissue and blood/follow-up samples. The MET gene product, hepatocyte growth factor receptor,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 [1], [39].

In an attempt to investigate the network structure at different stages in the oral cancer disease, we compute the intersections of sequential follow-ups. Fig. 8 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 [26]. The temporal development of the network is demonstrated in Fig.8, presenting the sequential network-intersections of the blood samples.

By examining the important molecules in Fig. 7 we conclude on the following:

i. *For the temporal development of network intersection of the blood samples* (Fig. 7a, STableIIIa, 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

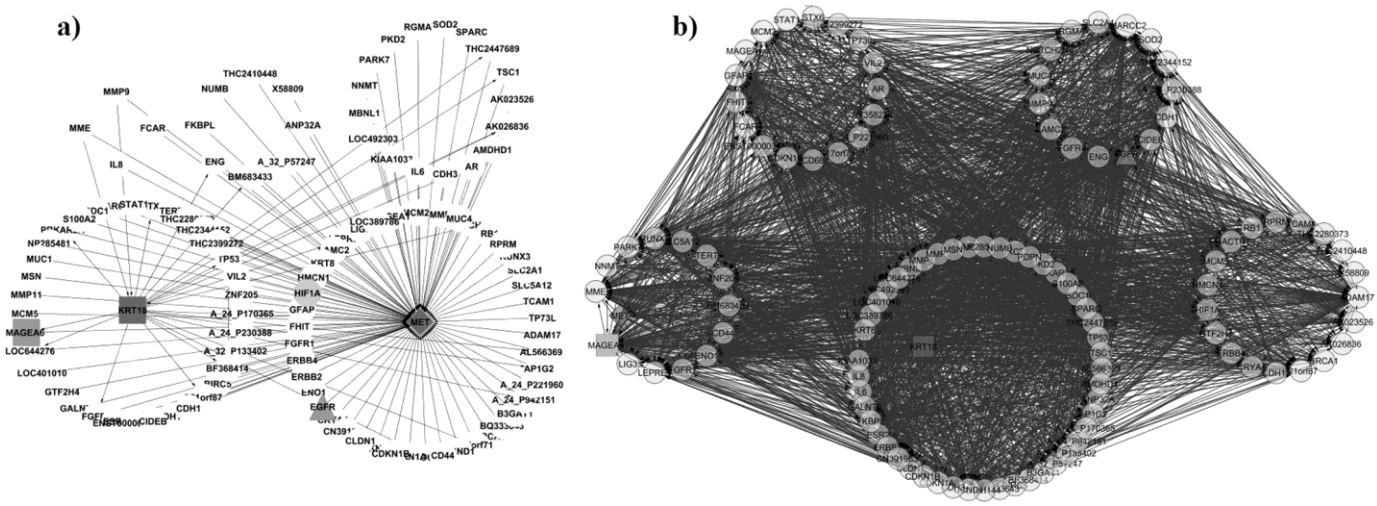

Fig. 6.**(a)** K-means clustering on the blood samples from the first follow-up and **(b)** on tissue samples. Five clusters were identified as presented in the grouped areas in both networks. Highlighted are the common "intersection" genes *MET, EGFR, HIF1A* and *MAGEA6*.

- 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 [26], [40] and possibly as a result of disease-causing mutations [41].

ii. *For the intersection of tissue with the blood network for the first time-stage*(Fig.7b; STableIIIa, 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 [26], [40], [42]).

iii. For both network intersections (Figs.7, 8; STableIIIa, 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
- METinteractions with its neighboring moleculesappear 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 (Figs. 7, 8). Despite the known MET/EGFR association in cancer [43], the existence of the MET/HIF1A and MET/MAGEA6 associations remain unknown. However, previous studies [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* isexpressed in OSCC) has been reported to bind to p53 tumor suppressor (TP53) and impair its function causing decreased apoptosis and increased cell growth [44]. Furthermore, the transcriptional activation of *MET* proto-oncogene during hypoxia via HIF1-mediated cascade could possibly explain the MET overexpression reported in OSCC specimens [45]. 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 [1], [46]. This counter-balancing activity of MET and EGFR pathways may also be viewed as a potential target for oral cancer therapeutic intervention [47]. (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 [48]. (c) MET loses many of its interactions through the loss of genes [26], [40] 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 [41].

Finally, from k-means clusteringwe 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 (STableIIIb, 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 [49]. This also highlights a dynamically functional reprogramming of a number of implicated genes and especially MET.

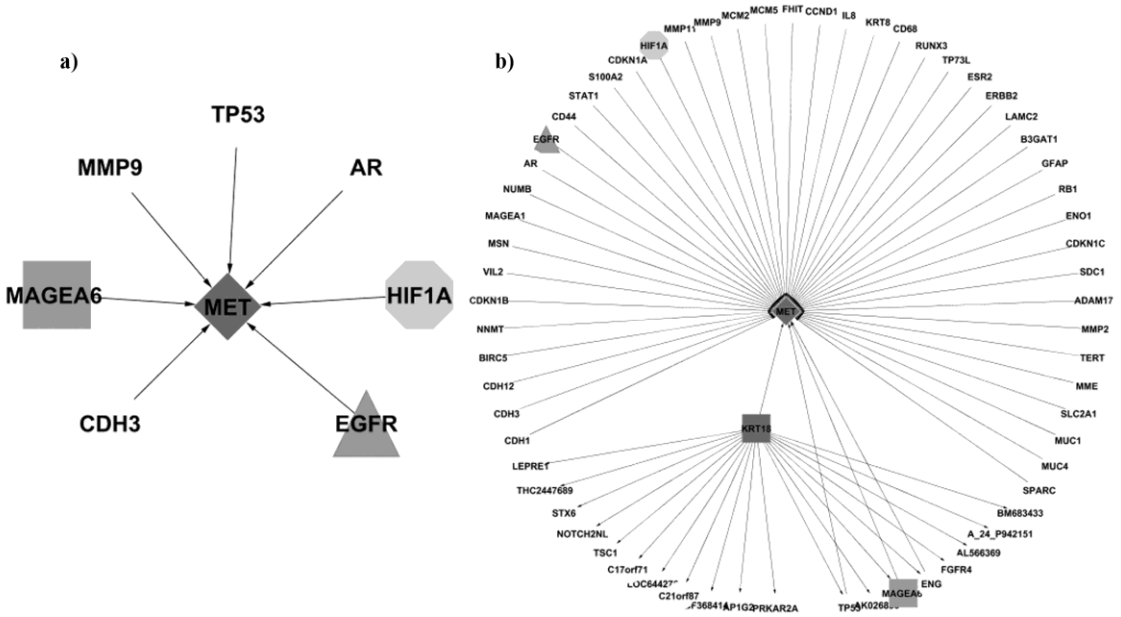

Fig. 7.Network intersection for blood and tissue. The left part **(a)** presents the intersection of all time stages for the blood samples; The Right part **(b)** presents the network intersection between the tissue and the blood from the first follow-up. Highlighted are the common "intersection" genes *MET, EGFR, HIF1A* and *MAGEA6*.

According to a recent study [26], *MET* could be characterizedas "broker" gene, i.e. a disease gene that holds acrucial 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) suggest that disease genes are found in especially vulnerable positions in networks, which is a reason of identifiable disease phenotypes accompanied by their disorganization [26].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 incancer, where disease genes tend to encode hubs, although in other pathologies disease genes reside at the periphery of the networks [41]. In addition, recent studies [50]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 akey 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* METcellular 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.

IV. 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 existenting 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

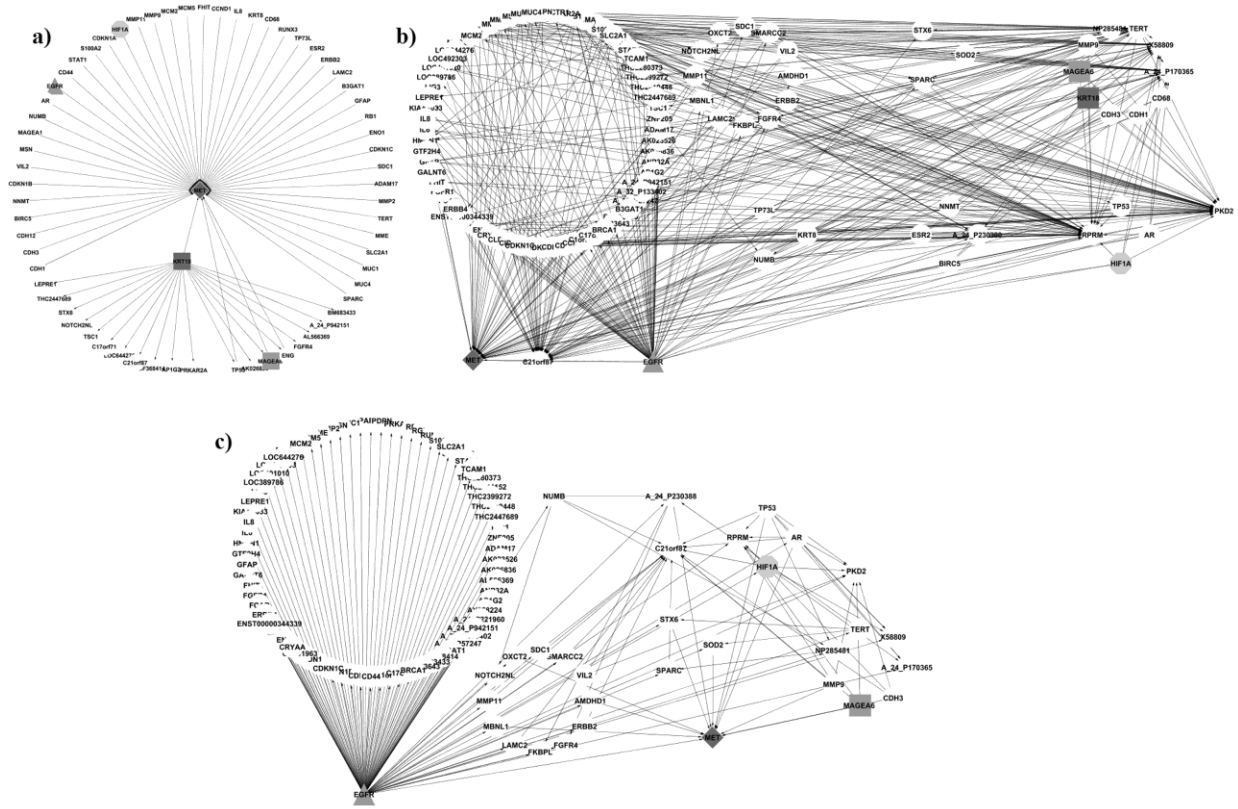

Fig. 8.Network intersection for sequential time stages on blood samples; **(a)** Network intersection between the 1st and 3rd month; **(b)** Network intersection between the 3rd and 6th month; **(c)** Network intersection between the 6th and 9th month. Highlighted are the common "intersection" genes *MET, EGFR, HIF1A* and *MAGEA6*.

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.

APPENDIX

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

ACKNOWLEDGMENT

This work is partly supported by the "OASYS" project funded by the NSRF 2007-13 of the Greek Ministry of Development, and by the "YPERThEN" project, which are funded by the INTERREG programs. It is also partly funded by the European Commission NeoMark project (FP7-ICT-2007-224483 – ICT enabled prediction of cancer reoccurrence).

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] P. B. Madhamshettiwar, S. R. Maetschke, M. J. Davis, A. Reverter, and M. a Ragan, "Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets.," *Genome medicine*, vol. 4, no. 41, May 2012.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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. 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 2006, pp. 233–240.

[29] 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.

[30] 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.

[31] Z. Liu, B. Malone, and C. Yuan, "Empirical evaluation of scoring functions for Bayesian network model selection.," *BMC bioinformatics*, vol. 13 Suppl 1, no. Suppl 15, p. S14, Jan. 2012.

[32] K. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "A multiscale and multiparametric approach for modeling the progression of oral cancer.," *BMC medical informatics and decision making*, vol. 12, no. 136, Nov. 2012.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.