35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Identification of altered MET network in Oral Cancer Progression based on Nonparametric Network Design*

K. Kalantzaki, E. S. Bei, K. P. Exarchos, M. Zervakis *Member, IEEE* and D. I. Fotiadis, *Member, IEEE,* M. Garofalakis *Member, IEEE*

***Abstract*—Oral cancer is characterized by multiple genetic events such as alterations of a number of oncogenes and tumour suppressor genes. The aim of this study is to identify genes and their functional interactions that may play a crucial role on a specific disease-state, especially during oral cancer progression. We examine gene interaction networks on blood genomic data, obtained from twenty three oral cancer patients at four different time stages. We generate the gene-gene networks from sparse experimental temporal data using two methods, Partial Correlations and Kernel Density Estimation, in order to capture genetic interactions. The network study reveals an altered MET (hepatocyte growth factor receptor) network during oral cancer progression, which is further analyzed in relation to other studies.**

I. Introduction

Biological processes organizing functional associations between different genes are central in understanding the biological mechanisms of several diseases, including oral cancer [1]. A variety of high-throughput experimental data, such as DNA microarray, ChIP-chip technology allow the simultaneous measurements of expression levels. These technologies have given thorough insight in complex molecular events in healthy and disease states. The extended study of related 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.

Several computational methodologies have been applied to construct biological networks using different data sources [2]. The main focus of networking approaches is to build target-independent networks that describe the pair-wise relations between molecules. Recent studies include Bayesian networks [3], Pearson's correlation-based approaches [4]. Although these methods have been successfully used to elucidate the functional relationship between genes and pathways, they are unlikely to directly indicate the specific gene networks in response to abnormal physiological conditions such as diseases, due to experimental errors and the inherent genetic complexity [2-4].

The analysis reported herein is an effort of revealing and modeling the inter-relationships of molecules in oral cancer that participate in many different pathways incriminated for this disease. The proposed method (in section II) for network construction is based on Kernel density estimation denoted as KDE, as an attempt to model the nonlinear effect of gene interactions and to fill the information loss from the data samples. Our framework is applied on experimental blood data of oral cancer patients received from four successive follow-ups in section III. The goal is to reveal the network structure and differences between different time slices, in addition to conspicuous genes that play central role in all stages of the disease.

*Research supported by "YPERThEN" project, which is funded by the EU and funds from Greece and Cyprus, and by "OASYS" project funded by the NSRF 2007-13 of the Greek Ministry of Development.

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).

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

II. Methodology

A. Partial Correlation

Pair-wise associations of co-expressed molecules can be modeled by 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. The method of partial correlations (PC) [4] 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 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, S^{-1}, with elements:

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

Given the experimental data, the covariance matrix is computed and then it is inverted. Indeed, using (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 which contributes to adding an edge between these nodes.

However, this approach is only applicable if the sample number in dataset is larger than the number of genes/proteins. Otherwise, the inversion of S is unstable making the estimation of S a non-trivial task. To overcome this obstacle we invert S through Moore-Penrose pseudo inverse [4], an approximation of the standard matrix inverse, based on the singular value decomposition (SVD).

B. Kernel Density Estimation

Kernel density estimation [5], is a non-parametric framework that estimates the probability density function

(pdf) of a random variable. Assume that a generic network is developed based on a limited genomic i.i.d dataset $X=(x_1,..x_n)$, where x_i denotes the sample i of gene X. The KDE allows the estimation of X as follows:

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

where $K(.)$ is a symmetric positive definite Gaussian function $K(u)=\frac{1}{2\pi}e^{\frac{-1}{2}u^2}$, n is dataset's size of the gene X and $h=1.47\sigma n^{1/6}$ is the optimal Gaussian bandwidth parameter, with σ standard sample deviation [5].

Genes interacting with each other can be expressed as a 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 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 two members of (3), the more independent the corresponding nodes are. In contrast, high absolute difference indicates dependence between X_i and X_k, thus connection between 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 members of (3) the more independent are genes X_i and X_k. Otherwise, the candidate genes share dependencies which implies association.

C. Edge Orientation

Determination of edge direction networks is made based on Bayesian Information Criterion (BIC) [6]. For each node, the number of edges connecting to it is counted. Nodes directly connected to it form a sub-network. For each sub-network, BIC score is computed for each edge that connects a pair of nodes, containing all other causative nodes to that pair that do not form a cycle. We evaluate the BIC score comparing the two possible orientations for the edge under examination. Finally, the edge is oriented in favor of the direction with the lowest BIC value.

III. Results and Discussion

In order to investigate the statistical properties of the proposed methodology, we apply PC and KDE approaches to reveal the network structure from gene expression data. In a previous work [7] the analysis was performed on the prototype organism *Arabidopsis thaliana* on developing seeds. 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 meaning of the results.

A. Oral Cancer Dataset

The analysis is performed on the oral cancer dataset [8] of 23 patients that have been enrolled from three major clinical centers (University Hospital of Parma, National Cancer Center Regina Elena and MD Anderson Cancer Center). Gene expression data were collected from circulating blood cells, at the baseline state of the patient, and from monthly follow-ups. Totally, were analyzed four different time stages corresponding to the first, third, sixth and ninth month after the initial diagnosis. The number of blood samples per patient changed depending on repeatability during the follow-ups. In particular, the total sample size was 23, 10, 13 and 4, respectively, for the above monthly follow-ups.

B. Direct Interactions

Table I presents the number of molecular interactions on blood samples, for the first follow-up with the PC and KDE algorithms. The first column describes different thresholds of the strongest partial correlations set on PC for (1), while the second column provides the thresholds of strongest similarity of (3) for KDE. The third and fourth columns summarize the verified number of gene-gene/gene products interactions for both approaches, respectively. 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.

We compared the performance of two approaches, also taking into account existing information on gene-gene and gene-gene product interactions from BioGRID, an interaction repository database. The currently available information provided 63 interconnections between the examined molecules. 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 less interactions being verified. This is due to the lack of strong partial correlations between molecular units. However, as 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.

Table I provides 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 86% of known genetic interactions for the blood constructed network while PC up to 66%. However, table I shows that many false positive edges are found as the number of predicted interactions is far larger than 63, leading to low precision. This aspect is further addressed next.

TABLE I. Gene-gene interactions for the 1st follow-up on blood samples

Threshold		Verified Interactions		New Edges		Oriented Edges	
PC	***KDE***	***PC***	***KDE***	***PC***	***KDE***	***PC***	***KDE***
≥0.1	≤0.6	42/63	0/63	4607	1	279	1
≥0.15	≤0.7	34/63	4/63	3763	134	185	70
≥0.175	≤0.75	30/63	5/63	3336	262	181	85
≥0.2	≤0.8	27/63	7/63	2953	535	172	92
≥0.3	≤0.85	17/63	18/63	1662	1417	187	158
≥0.4	≤0.875	6/63	33/63	754	2261	157	204
≥0.5	≤0.9	4/63	54/63	279	3790	67	321

C. Performance through External Genes

The poor performance in biological network reconstruct ion is a well-known problem that has been extensively addressed, especially when it is dealt only with expression data [3-6]. The problem is focused on the large number of false-positive predicted interactions due to the consideration of only direct interactions that have been biologically confirmed. In this section, we consider not only known direct associations between pairs of genes, but we also acceptconnections that are induced by external molecules, which can be identified in various available databases [9]. We compare the performance of both algorithms taking into account the expression data and the available knowledge of associations from various databases. In this way, we can examine indirect interactions between the studied genes taking into account all the possible external pathways that connect these molecules. Hence, several initially assigned false-positive edges could be characterized true-positive as a result of multiple effects of external molecules. Additionally, we used HIPPIE database for the validation of all new interactions generated by our network framework; we found that only these 63 interactions have protein annotations in the human interactome reference [10].

In order to integrate molecular interactions from different public databases we used BioNetBuilder [9], 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 110 oral cancer disease genes and five disease unrelated genes (e.g. *PARK7*) the BioNetBuilder retrieved more than 300,000 interactions with more than 25.000 genes. This produced network through extensive consideration of available biological knowledge is considered as the ground-truth, against which we compare our analysis.

To determine the performance of the proposed algorithm we used the receiver operator characteristic (ROC) and precision recall curves. 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. Between the two networks, we consider TP as the existent edges in both networks. Also, when a predicted interaction is verified through indirect associations with external factors (apart from the 115 genes) then the predicted association is set as TP. Finally, we consider FN as the 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.

Table II presents the results according to the above analysis for all thresholds of Table I. According to ground-truth network, apart from the 63 direct edges there are 1558 indirect implications; these result from considering a maximum of three external genes. Furthermore, from the 115 analyzed genes, there are 22 association; in our analysis we did not take into account edges connecting these genes. For this reason the number of new edges (columns 1, 2) differs from those in table I. 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). Totally, the set of TN associations was composed of 2657 edges.

Fig.1 presents the ROC curves for the blood samples associated with the three follow-ups (fig.1a-1c). For the ninth month PC could not give a reliable structure thus was omitted. For all listed cases, both methods show monotonic performance over the parameters. For the optimal bandwidth h in eqn (2), KDE outperforms PC as it covers larger area under the curve (AUC) compared to PC. For non-optimal h, KDE is expected to give lower precision. Furthermore, both algorithms show improvement in performance after taking the external genetic influence into consideration. In fact, precision and recall (fig.1d-1f) 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. With this consideration the precision is highly improving for all network cases, in support of the conclusion that expression data enclose dependencies from a variety of sources. Thereafter, when dealing with expression data direct associations that come from statistical analysis should be interpreted as a result of indirect influence of external factors and not as spurious edges. We note that the precision of KDE can be greatly improved by evaluating a larger number of indirect interactions. Another future enhancement is to include experimental protein data. By combining both gene and protein data the method can yield more precise results. However, at this stage the results of our framework can be validated because of the small number of the "newly discovered" gene/protein interactions related to the network intersection of all time-slices, as illustrated in fig.2a.

D. Biological Interpretation

As shown in fig. 2a, MET is obviously a central molecule that interacts with a number of critical molecules for tumor development and progression, including oncogenes (EGFR, epidermal growth factor receptor), suppressor genes (TP53, tumor protein p53), and transcription factors (HIF1A, hypoxia inducible factor 1 alpha subunit). Fig. 2b depicts the degree distribution of these molecules in each constructed network for all four disease stages. From these identified MET interactors, only the MET/EGFR associations have been reported before [10]. We conclude that we recover known disease genes and provide potential associations that could be linked to oral carcinogenesis.

TABLE II. GENE INTERACTIONS FOR THE 1ST FOLLOW-UP ON BLOOD SAMPLES CONSIDERING THE EXTERNAL GENES

New Edges		**TP**		**FP**		**TN**		**FN**	
PC	***KDE***	***PC***	***KDE***	***PC***	***KDE***	***PC***	***KDE***	***PC***	***KDE***
3166	1	1167	1	1999	0	895	2657	454	1620
2551	108	957	75	1594	33	1250	2627	664	1546
2234	202	848	129	1386	73	1458	2588	773	1492
1968	347	738	167	1230	180	1614	2484	883	1454
1068	1081	394	423	674	658	2170	2038	1227	1198
474	1678	181	711	293	967	2551	1771	1440	910
172	2813	71	1225	101	1588	2743	1176	1550	396

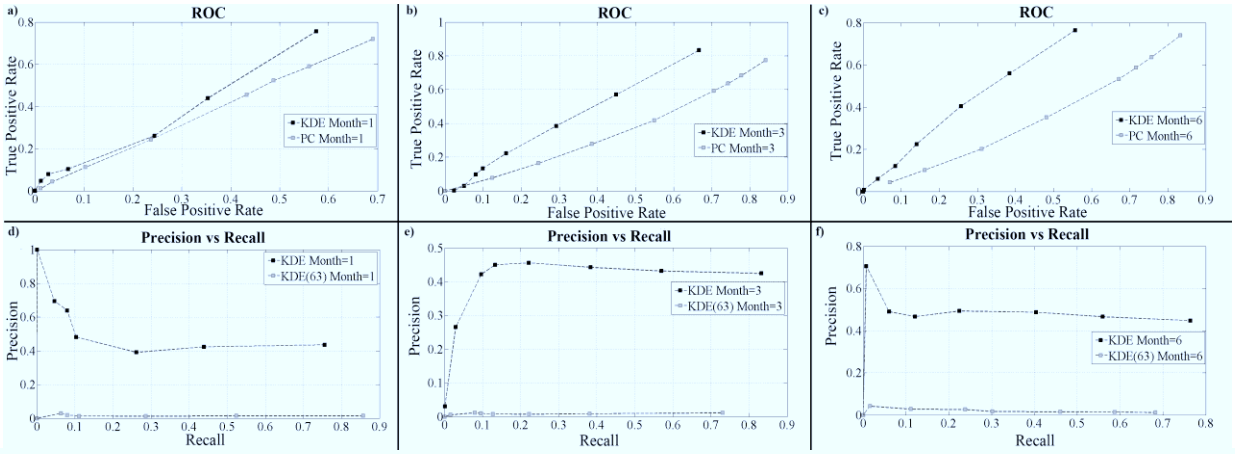

Figure 1. ROC comparison between KDE and PC for the blood samples **(a-c)**. For the ninth month, PC cannot derive a reliable network structure. Apparently KDE covers larger AUC for all cases; Precision vs Recall comparison between KDE and PC for all cases on blood samples **(d-f)**. KDE(63) and PC(63) 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. *(Fig. 1)*

Based on our findings, we suggest that MET is characterized by topological alterations based on degree (number of associations with molecules) in each time slice accompanied with functional alterations (different interactions with genes and/or gene products in each time slice) during oral cancer progression, which are in accordance with its biological role as proto-oncogene [11]. As many of these genes are basic components of multistep tumorigenesis, and since the role of MET in cancer development and progression has long been demonstrated, MET might be viewed as a key regulator of oral carcinogenesis. Furthermore, the *in vivo* MET molecular network might be an important determinant for the screening of patients at the time of diagnosis, during oral cancer progression and for effective therapy. Towards this direction we can define a unique motif for each time-stage matching the clinical characteristics of a specific patient group. The proper analysis of motifs in tumor progression would enable the more accurate categorization of disease stage and indicate the proper therapy, targeting for example a specific signaling pathway. Based on the current evaluation metrics of our framework, the predictive accuracy of such screening can be considered sufficient.

IV. Conclusion

Clearly, the KDE approach models quite well the verified direct and indirect associations between the participating genes. On the contrary, the PC approach appears to capture less of those associations. Thus, our results indicate that KDE performs better on the network construction. With this analysis 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. Our network framework reveals important information about the functional alterations of gene-gene and/or gene-gene product interactions, which take place in particular time stages of a complex disease, such as oral cancer.

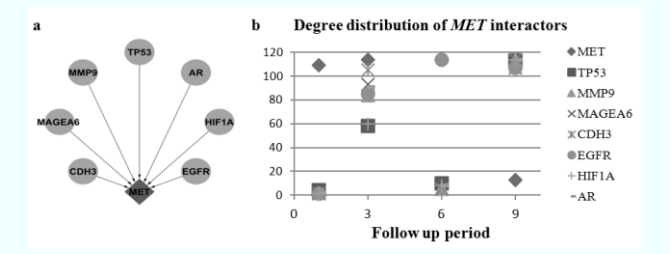

Figure 2. **(a)** Network intersection of all time slices on blood samples. MET places the central role. **(b)** Degree distribution of MET interactors.

References

[1] J. Campo-Trapero, J. Cano-Sánchez, B. Palacios-Sánchez, J. J. Sánchez-Gutierrez, M. a González-Moles, and A. Bascones-Martínez, "Update on molecular pathology in oral cancer and precancer.," Anticancer research, vol. 28, no. 2B, pp. 1197–1206, 2008.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.