Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Biological Interaction Networks Based on Sparse Temporal Expansion of Graphical Models

K. D. Kalantzaki, E. S. Bei, M. Garofalakis, M. Zervakis
Department of Electronic and Computer Engineering, Technical University of Crete,
Chania, Greece,
kkalantzaki@isc.tuc.gr, abei@isc.tuc.gr, minos@acm.org, michalis@display.tuc.gr

***Abstract*—: Biological networks are often described as probabilistic graphs in the context of gene and protein sequence analysis in molecular biology. Microarrays and proteomics technology allow the monitoring of expression levels over thousands of biological units over time. In experimental efforts we are interested in 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 data using two methods, partial correlations and Kernel Density Estimation, in order to capture genetic interactions. Dynamic Gaussian analysis is used to match special characteristics to genes and proteins at different time stages utilizing the KDE method for expressing Gaussian associations with non-linear parameters.**

Keywords- Gaussian Graphical Model, Kernel Estimation, Sparse Temporal Expansion, Network construction, Arabidopsis thaliana.

I. INTRODUCTION

In recent years the description of genome sequences has given a large amount 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. Several methodologies have been proposed for constructing gene-protein networks based on expression data, such as Bayesian networks [1, 2], Gaussian networks [3] that aim to provide suitable mathematical models for describing stochastic net-like associations and dependence structures in complex high-dimensional data. In addition, dynamic graphical approaches have been introduced that model time dependencies and reveal an interactive behavior between different time slices [2, 4].

Unfortunately, although graphical models are promising for interaction analysis their main drawback is their limited performance when the experimental data are insufficient. This problem has two aspects. Firstly, the lack of features 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 [3, 5].

Secondly, the information contained in expression data is limited by their quality, the experimental design, noise, and measurement errors. These lead to loss of information making a hard task the estimation both of causal relationships in network structure and for the dependencies enclosed between neighbored genes/proteins [5].

A common graphical representation is the Gaussian model firstly introduced by Kishino and Waddell [6]. 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 works poorly. In that case, the covariance matrix, which encloses the interactions between genes/proteins is not positive definite, thus makes impossible the computation of the partial correlation matrix.

Given these challenges, great steps have been undertaken to overcome these obstacles. In this paper, we propose a new methodology in modeling dynamic Gaussian graphical models from sparse data. More specifically, our goal is summarized in filling the information loss in time varying Gaussian networks through the non-parametric framework of Kernel density estimation [7]. Our approach lies under 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. Also, 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.

Our presentation is organized as follows. In the following section we provide a review of kernel based density estimation and approaches in network construction from experimental data. We continue introducing our approach in representing nonlinear relations between genes/proteins using a dynamic Gaussian model. In last section we present our results in applying our algorithms. Finally, conclusion and future work are given.

II. BACKGROUND INFORMATION

We explore two approaches for estimating the structure of a gene-protein network. We generate two different networks reflecting the different approaches in expressing generic interactions between genes and proteins. The first approach focuses on estimating the inverse partial correlation matrix through a statistical probabilistic approach of Gaussian Graphical Model (GGM). The second examines dependency

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

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

between nodes using a non-parametric approximation of the missing experimental data through Kernel density estimation (KDE). After that step, we assign directions to the edges of the produced networks using Bayesian Information Criterion (BIC).

A. Gaussian Graphical Model

Gaussian Graphical Models [6] are undirected probabilistic graphical frameworks also known as covariance selection models. In a GGM network, the identification of conditional independence between nodes is based on the assumption that nodes follow a Gaussian distribution. In such case, interactions between two variables are reduced in 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. A good notion of the strength for these interactions is the partial correlation matrix $\Pi=(\pi_{ik})$. Its coefficients describe the correlation between nodes i and k conditioned on all remaining nodes of the network. In the GGMs this property is reflected in the inverse covariance matrix S, S^{-1}.

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

Given the experimental data, the covariance matrix is computed and then inverted. From (1) the partial correlations, π_{ik} can be found. 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 of data is larger than the number of genes/proteins. Otherwise, the inversion of S is unstable. To overcome this obstacle we invert S through Moore-Penrose pseudo inverse [3, 6], an approximation of the standard matrix inverse, based on the singular value decomposition (SVD).

B. Kernel Density Estimation

Kernel density estimation [7] is a non-parametric framework that can predict the probability density function (pdf) of a random variable. Given a limited genomic i.i.d dataset $X=(x_1,..x_n)$, KDE allows to simulate the pdf of X as follows:

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

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/protein X and h>0 is a smoothing parameter, the bandwidth.

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 cross correlation test:

$$f(X_i,X_k)=f(X_i)*f(X_k) \quad (3)$$

The smaller the absolute difference between two members of the equation, 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.

C. Edge Orientation

Up to this point we have provided two ways in revealing the network structure. These approaches give an intuition whether two nodes interact. But they do not imply anything about causality, denoting 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, between two nodes there are two models, i.e. model M_1 where node X_i is the parent of node X_k, or the opposite, model M_2.

Model selection procedures cannot distinguish the above described models because their distribution 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 [8].

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

Therefore, the distinction between models M_1 and M_2 is made by inferring direction of causation between nodes using a scoring function, the BIC criterion (Bayesian Information Criterion).

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

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

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

D. Linear Gaussian Graphical Model

Linear Gaussian Graphical Model (LGGM) [9] is a classical approach in GGMs that models dependencies between nodes as linear combination of means. Each node X_i is distributed 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.

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. In the following section we introduce another approach where non-linear characteristics are given to the parameters of distribution.

E. Dynamic Gaussian Model

Dynamic Gaussian Networks (DGN) [2, 10] can be viewed 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.

For p microarrays sets and expression levels of n genes/proteins, the data matrix can be summarized as $p \times n$ $X=(X_1,\ldots,X_p)^T$ 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:

$$f(X_{11},..X_{pn})=f(X_1)f(X_2|X_1)..f(X_p|X_{p-1}) \quad (6)$$

$$f(X_i|X_{i-1})=f\left(X_{i_1}\middle|\mathbf{P}_{\mathbf{a(i-1),1}}\right)*..*f\left(X_{i_n}\middle|\mathbf{P}_{\mathbf{a(i-1),n}}\right) \quad (7)$$

where $\mathbf{P}_{\mathbf{a(i-1),\,j}}$ 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.

III. Proposed Method

From the above tools, two networks are generated each following a different approach in revealing genetic associations (GGM and KDE). In this section, we augment these networks with a framework that enforces a non-linear view in modeling the parameters of conditional probability distribution for estimating dependencies between genes/proteins. We represent conditional probabilities as Gaussian distributions through Kernel density estimation.

A. Conditional Propability Distribution

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. Thus interactions with parental nodes are modeled by the conditional distribution of each gene. We use KDE as a non-parametric framework in order to capture the dependencies from parental nodes that underlie on experimental data.

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 $\mathbf{P}_{\mathbf{a_i}}$ be the parents of gene/protein X_i then direct dependencies are encoded as:

$$f(X_i|\mathbf{P}_{\mathbf{a_i}})=\frac{f(X_i,\mathbf{P}_{\mathbf{a_i}})}{f(\mathbf{P}_{\mathbf{a_i}})} \quad (8)$$

In order to model these relations with a coherent mathematical framework based on genomic expressions, we find the joint distributions of (8) with Standard Gaussian Kernel (SGK) as follows [7]:

$$\hat{f}_h(x,y)=\frac{1}{n}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h_2}\right)K\left(\frac{y-y_i}{h_1}\right) \quad (9)$$

From (8), (9)

$$f(X_i|\mathbf{P}_{\mathbf{a_i}})=\frac{\sum_{j=1}^{p}K_{h_1}(x-x_{ij})\,K_{h_2}(p_{a_i}-\mathbf{p}_{\mathbf{a_{ij}}})}{\sum_{j=1}^{p}K_{h_2}(p_{a_i}-\mathbf{p}_{\mathbf{a_{ij}}})} \quad (10)$$

where K(.) is a Gaussian kernel function described as (2), p is dataset's size and $h_1=c_1n^{-1/6}$, $h_2=c_2n^{-1/6}$ for $c_1,c_2>0$ are the smoothing parameters selected as optimal approximations of Gaussians basis functions [11].

Equation (10) implies that the conditional density estimate is an asymptotic approximation of Gaussian [11] $N(\theta_1,\sigma_1^2)$ with $R(K)=\int K(u)^2du$ and parameters as follows:

$$\theta_1=\frac{\sigma_K^2}{2\sqrt{c_1c_2}}(c_1^2f^{(2)}(\mathbf{P}_{\mathbf{a_i}}|X_i)+c_2^2f^{(2)}(\mathbf{P}_{\mathbf{a_i}}|X_i)+2c_2^2f^{(2)}(\mathbf{P}_{\mathbf{a_i}}|X_i)f^{(1)}(\mathbf{P}_{\mathbf{a_i}}|X_i) \quad (11)$$

$$\sigma_1^2=\frac{R(K)^2\,f(\mathbf{P}_{\mathbf{a_i}}|X_i)}{c_1c_2f(\mathbf{P}_{\mathbf{a_i}})} \quad (12)$$

Hence, (11) and (12) encode a Gaussian model that captures non-linear dependencies of Gaussian's parameters. If a gene/protein has no parents the mean and variance is taken from KDE.

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 missing data is no longer an obstacle due to estimation from the remaining samples.

IV. Results

In order to investigate the statistical properties of the proposed framework we start by revealing the network structure using the GGM and KDE approaches. After that step, and for each generated network, the conditional probabilities are found using our proposed algorithm, as well as the LGGM approach. Finally, through inference we evaluate the influence of certain significant factors comparing our results to LGGM. The same framework is applied for different time slices in order to examine time dependencies.

The data samples we used for testing concern the developing Arabidopsis thaliana [12] seeds, 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 (table I) and

TABLE I. Significant And Unrelated Genes/Proteins

Genes	Proteins	Relevance	Biological Process
At3g43190	-	Related	Primary metabolism
At4g02280	-	Related	Primary metabolism
At5g20830	-	Related	Primary metabolism
At5g37180	-	Related	Primary metabolism
At5g49190	-	Related	Primary metabolism
At5g22510	-	Related	Primary metabolism
At1g35580	-	Related	Primary metabolism
At1g13140	-	Unrelated	Energy
At2g39470	-	Unrelated	Energy
At4g14630	-	Unrelated	Protein destination & storage
At4g15010	-	Unrelated	Intracellular traffic
At1g54050	-	Unrelated	Protein destination & storage
At3g17520	p2322, p2323	Unrelated	Disease/defense

TABLE II. GENE-PROTEIN INTERACTIONS

Threshold		Verified Pairs		New Edges		Oriented Edges	
GGM	***KDE***	***GGM***	***KDE***	***GGM***	***KDE***	***GGM***	***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

studies the network associated with this pathway. Under the term significant genes, we included such genes that encode sucrose synthases (At3g43190, At4g02280, At5g20830, At5g37180, At5g49190) or invertases (At1g35580, At5g22510), both important enzymes in carbohydrate (sucrose) metabolism [13]. We included more than one gene that encodes sucrose synthases, in order to use them as internal controls for our proposed algorithm. Under the term unrelated genes, we included four genes that are identified as biomarkers for specific organs (flowers, leaves, roots, siliques) in Arabidopsis and not expressed in seeds (AT1G13140, AT2G39470, AT4G14630, AT4G15010), as well as 2 genes (At1g54050 and At3g17520) that are expressed in seeds but they are not involved in carbohydrate metabolism [12, 14]. Overall, we studied 113 genes and 27 gene-protein pairs, for all stages of growth. Our goal was to verify known gene-protein interactions [11], direct associations between genes as well as to highlight how the pathway is affected by significant factors.

Table II presents the number of verified gene-protein pairs. The first column describes different thresholds on partial correlation set on Moore-Penrose inverse for (1), while the second column provides the thresholds of absolute difference of (3) 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.

Table II shows for the inferred networks with Moore-Penrose pseudo inverse 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

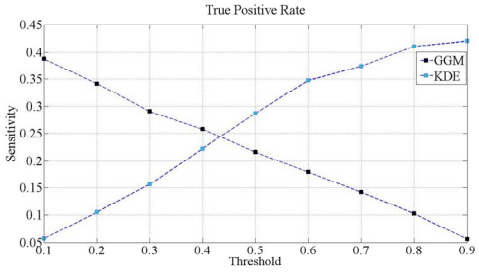

Figure 1. True Positive Rate *(Fig. 1)*

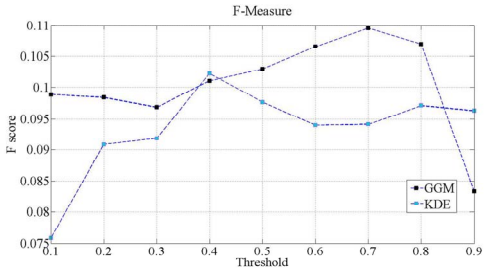

Figure 2. F-measure for KDE and GGM. *(Fig. 2)*

increases. This implies that genes-proteins seem to be less independent. Thus, more interactions are identified and the graph becomes more cohesive.

Table III shows the verified interactions between genes as well as the 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 [15] and AtPIN, *A. thaliana* Protein Interaction Network [16]. 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.

Tables II and III 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. The percentages for GGM 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. Fig. 1 summarizes the true positive rate for both algorithms, meaning framework's ability to detect existent interactions.

TABLE III GENE-GENE AND PROTEIN-PROTEIN ITERACTIONS

Threshold		Verified Gene Interactions		Verified Protein Interactions	
GGM	***KDE***	***GGM***	***KDE***	***GGM***	***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

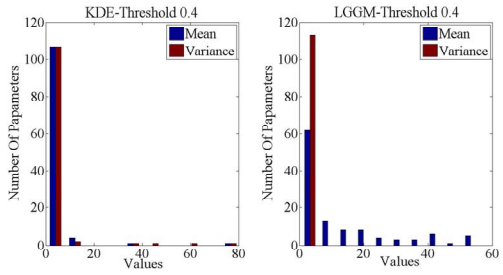

Figure 3. Comparison between KDE and LGGM. *(Fig. 3)*

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. Fig. 2 presents for all thresholds the performance of two methodologies according to F-score metric, $F=\frac{2*precision*recall}{precision\ recall}$. In conclusion, appropriate thresholds for KDE are $0.4 \leq th \leq 0.6$ while for GGM $0.6 \leq th \leq 0.8$.

From a statistical perspective, not all significant edges were found (with low F-score). However, not necessary all false positives edges are spurious edges. In fact, using ANAP (Arabidopsis Network Analysis Pipeline), an interactive Web tool that contains protein interaction data information from 11 public Arabidopsis databases [17], 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.

In the next step, we estimate the conditional probabilities for the produced thresholds. For the generated networks, the mean and variances are compared with the equivalent parameters of the LGGM approach. Fig. 3 and fig. 4 present the histograms of mean and variances for the computed probabilities. For both networks with KDE and GGM, the conditional Gaussian distributions fluctuate close to low means, while the LGGM approach covers a wider range. This is due to the fact that conditional dependencies 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 fig. 5 where is presented the histogram of the expression data.

The last step in evaluating the algorithm was the examination of inference of the effects on genes and proteins for different time stages.

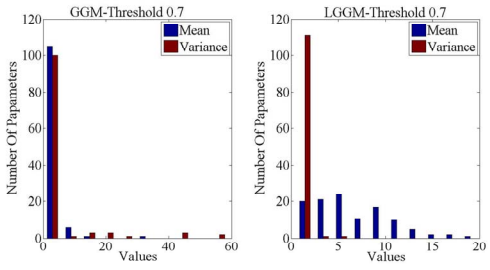

Figure 4. Comparison between GGM and LGGM . *(Fig. 4)*

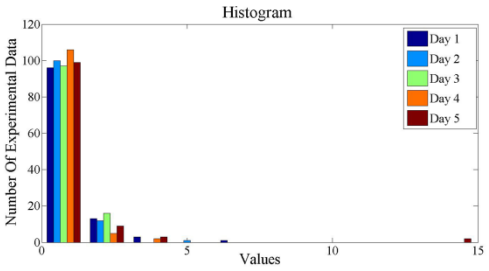

Figure 5. Histogram of the experimental data for all days of growth. *(Fig. 5)*

For this purpose, we included in the studied carbohydrate pathway the 13 genes (7 'significant', 6 'unrelated' genes), as presented in table I. In order to verify algorithms performance we compared our results to the implied interactions of the studied pathway for different time stages.

Table IV summarizes the effects implied by the observations of the relevant genes for this pathway including the above mentioned 13 additional genes. The first column shows the applied algorithms in modeling the conditional probability distributions while the second column provides the implied genes on which the inference is focused. The numbers in parenthesis imply the picked threshold according to fig.2. The third and the fourth column show the implied interactions from inference using KDE, GGM and LGGM in revealing network structure.

Our analysis of sparse experimental data allowed us to generate gene-protein networks and illustrate 3 key results focusing on the outcome interactions of the 'significant' genes of the KDE method (table IV). First, we observe that the implied genes from 2nd column interact with genes from the 3rd column, most of which are involved in carbohydrate metabolism. These gene-pairs are indirectly interconnected, according to ATTED-II [15].

Second, we reveal new gene-gene (directly or indirectly) interactions between the implied genes and the genes showed in the 3rd column, including interactions with the 'unrelated' genes. Interestingly, the 'unrelated' gene At3g17520 has inference significance and is a member of the group 4 late embryogenesis abundant (LEA) protein genes [14]. The presence of their encoded LEA proteins is related to the adaptive response of higher plants caused by adverse conditions to maintain normal metabolism [18].

Third, we highlight new gene-protein interactions between the 'significant' genes and two enzymes (4th column), the 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 [19], and the plastidial glyceraldehyde 3-phosphate dehydrogenase, A subunit (GAPA) that participates in the reductive carbon cycle and also is involved in response to sucrose stimulus [20].

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*.

TABLE IV. SIGNIFICANT INTERACTIONS FROM INFERENCE

	Observed	Produced interactions			
		Gene-Gene			Gene-Protein[b]
KDE (0.4)	***At2g01140***	At2g21170	**At2g21330**[c]	**At1g32060**[c]	p496[b] p532[b]
		At5g60760	At3g12780	At3g26650	
		At2g36460	At1g42970	At1g79550	
		At1g13140	At3g52930	*At3g17520*[a]	
		At1g04410			
	At5g52920	At3g12780	At2g36460	**At2g21330**[c]	p496[b] p532[b]
		At5g60760	At1g09780	At3g26650	
		At1g79550	At2g21170		
	At1g73370	**At2g21330**[c]	At5g60760	At3g12780	p496[b] p532[b]
		At3g26650	At2g36460	At1g42970	
		At2g21170			
	At1g35580	At2g21170	**At2g21330**[c]	At5g60760	p496[b] p532[b]
		At3g12780	At3g26650	At2g36460	
		At1g42970	At1g79550	At2g01140	
		At1g32060[c]			
	At5g22510	At2g21170	At3g12780	At2g36460	p532[b]
		At2g21330	At1g04410	At2g01140	
		At4g14630[d]	*At1g13140*[d]	At1g32060	
		At1g42970	At1g79550		
LGGM (0.4)	***At2g01140***	At5g37180	At4g38970	**At2g21330**[c]	
		At1g35580	**At1g32060**[c]		
	At5g52920	At1g35580	At5g37180	**At2g21330**[c]	
		At1g32060	At4g38970		
	At1g73370	**At2g21330**[c]	At1g09780	At1g35580	
		At5g37180	At4g38970	At1g32060	
	At1g35580	At5g37180	At4g38970	**At2g21330**[c]	
		At1g32060[c]			
	At5g22510	At1g73370	At5g37180		
GGM (0.7)	***At2g01140***	At2g36460	At3g43190	At3g27190	
		At1g13140[c]	**At1g09780**[c]		
	At5g52920	At3g43190	*At3g17520*[a]		
	At1g73370	At3g43190[d]	At1g09780	*At3g17520*[a]	
	At1g35580	At3g43190[d]	At1g09780	***At3g17520***[ac]	
	At5g22510	At3g43190[d]	**At1g09780**[c]	*At3g17520*[a]	
LGGM (0.7)	***At2g01140***	**At1g09780**[c]	At3g52930	At1g35580	
		At1g13140[c]	At1g30120	At2g22780	
		At3g17520[a]			
	At5g52920	At1g09780			
	At1g73370	*At2g39470*	At1g09780		
	At1g35580	***At3g17520***[ac]			
	At5g22510	**At1g09780**[c]			

a. At3g17520 is underlined because of inference significance. b. p496 protein is encoded by At2g36460 and p532 by At3g26650 gene, respectively. c. overlapped genes between KDE(0.4)-LGGM(0.4) and GGM(0.7)-LGGM(0.7) are marked in bold. d. unrelated genes are marked in italic.

V. DISCUSSION AND CONCLUSION

Clearly, KDE models quite well the verified associations between the participating genes/proteins, as the majority of the affected genes/proteins are located close to the processes of the carbohydrate metabolism pathway. On the contrary, GGM seems to capture less of those associations, some of which are also supported by the observed genes. From these observations it is concluded that KDE performs better on the prediction of network construction. Both for KDE and for GGM networks the conditional probabilities are modeled by our non-linear approach. Comparing those networks with the corresponding LGGM, we can identify a major advantage for our framework in revealing indirect biological associations.

REFERENCES

[1] 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.

[2] 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. 228-35, Sep. 2003.

[3] 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.

[4] H. C. Cho, M. S. Fadali, and K. S. Lee, "Online Probability Density Estimation of Nonstationary Random Signal using Dynamic Bayesian Networks," vol. 6, no. 1, 2008.

[5] B. F. Wong, C. K. Carter, and R. Kohn, "Efficient estimation of covariance selection models," Biometrika, vol. 90(4), 809–830, 2003.

[6] U. N. C. Charlotte and K. R. Subramanian, "Interactive Analysis of Gene Interactions Using Graphical Gaussian Model," Gene, pp. 63-69, 2003.

[7] B. E. Hansen, "Nonparametric Conditional Density Estimation," no. November, 2004, www.ssc.wisc.edu/~bhansen.

[8] 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. 1089-100, Jun. 2008.

[9] 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. 2523-31, Oct. 2006.

[10] K. Murphy, and S. Mian, "Modelling Gene Expression Data using Dynamic Bayesian Networks", 1999, http://www4.ncsu.edu/~smsulli2/MA810_Fall2008/SAMSIpapers/murphy99modelling.pdf.

[11] D. T. Davis, "Expanding Gaussian kernels for multivariate conditional density estimation," IEEE Transactions on Signal Processing, vol. 46, no. 1, pp. 269-275, 1998.

[12] M. Hajduch et al., "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. 2078-87, Apr. 2010.

[13] K. Koch, "Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development," Curr. Opin. Plant. Biol, vol. 7(3), pp. 235-246, Jun 2004.

[14] 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, E. Huala, "The Arabidopsis Information Resource (TAIR): improved gene annotation and new Tools," Nucleic. Acids. Res, vol. 40, pp. D1202-1210, Jan 2012.

[15] T. Obayashi, S. Hayashi, M. Saeki, H. Ohta, and K. Kinoshita, "ATTED-II provides coexpressed gene networks for Arabidopsis," Nucleic. Acids. Res, vol. 37(Database issue), pp. D987-991, Jan 2009.

[16] M. M. Brandão, L. L. Dantas, and M. C. Silva-Filho, "AtPIN: Arabidopsis thaliana protein interaction network," BMC. Bioinformatics, vol. 10, pp. 454, Dec 2009.

[17] C. Wang, A. Marshall, D. Zhang, and Z. A. Wilson, "ANAP: an integrated knowledge base for Arabidopsis protein interaction network Analysis," Plant. Physiol, vol. 158(4), pp. 1523-1533, Apr 2012.

[18] S. Hong-Bo, L. Zong-Suo, and S. Ming-An, "LEA proteins in higher plants: structure, function, gene expression and regulation," Colloids. Surf. B. Biointerfaces, vol. 45(3-4), 131-135, Nov 2005.

[19] W. Lu, X. Tang, Y. Huo, R. Xu, S. Qi, J. Huang, C. Zheng, and C. A. 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(1), pp. 65-74, Jul 2012.

[20] 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. Physiol, vol. 151(2), pp. 541-558, Oct 2009.