Biological interaction networks based on non-parametric estimation

Kalliopi D. Kalantzaki*, Ekaterini S. Bei, Minos Garofalakis and Michalis Zervakis

Department of Electronic and Computer Engineering,
Technical University of Crete,
University Campus,
73100 Chania, Crete, Greece
Email: kkalantzaki@isc.tuc.gr
Email: abei@isc.tuc.gr
Email: minos@acm.org
E-mail: michalis@display.tuc.gr
*Corresponding author

Abstract: Biological networks are often described as probabilistic graphs in the context of gene and protein sequence analysis in molecular biology. Microarrays and proteomics technologies facilitate the monitoring of expression levels over thousands of biological units over time. Several experimental efforts have appeared aiming to unveiling pairwise interactions, with many graphical models being introduced in order to discover associations from expression-data analysis. However, the small size of samples compared to the number of observed genes/proteins makes the inference of the network structure quite challenging. In this study, we generate gene–protein networks from sparse experimental temporal data using two methods, partial correlations and Kernel Density Estimation (KDE), in an attempt to capture genetic interactions. Applying KDE method we model the genetic associations as Gaussians approximations, while through the dynamic Gaussian analysis we aim to identify relationships between genes and proteins at different time stages. The statistical results demonstrate valid biological interactions and indicate potential new indirect relations that deserve further biological examination for validation.

Keywords: GGM; Gaussian graphical model; KDE; kernel density estimation; sparse temporal expansion; network construction; *Arabidopsis thaliana*.

Reference to this paper should be made as follows: Kalantzaki, K.D., Bei, E.S., Garofalakis, M. and Zervakis, M. (2013) 'Biological interaction networks based on non-parametric estimation', *Int. J. Biomedical Engineering and Technology*, Vol. 13, No. 4, pp.383–409.

Biographical notes: Kalliopi D. Kalantzaki received the Diploma degree in Electrical and Computer Engineering on May 2010 and the MSc degree on July 2013 from Technical University of Crete (TUC). Her Master Diploma thesis is entitled 'Graphical Models in Genomic Networks'. She is currently a PhD student at the Technical University of Crete and also a research fellow at the Digital Image and Signal Processing Laboratory. Her research interests include bio data mining and graphical models. She participates at the OASYS and YPERThEN projects while her work has been presented in conference proceedings and published in scientific journals.

Ekaterini S. Bei received the Diploma degree in Biochemistry from the Humboldt Universität zu Berlin, Germany, and the PhD degree from the University of Athens (NKUA), Medical School, Laboratory of Biological Chemistry, in 2008. Her PhD thesis is entitled 'Glucocorticoid receptor and its cross-talk with other cellular signaling factors in depression'. She is currently a Research Fellow in the Department of Electronic and Computer Engineering of Technical University of Crete. Her research interests include biomedical data mining and decision support systems in healthcare. She participates at two NSRF 2007-13 projects and an INTERREG project. Her work has been presented in conference proceedings and published in scientific journals.

Minos Garofalakis received the Diploma degree in Computer Engineering and Informatics (School of Engineering Valedictorian) from the University of Patras, Greece, and the MSc and PhD degrees in Computer Science from the University of Wisconsin-Madison in 1994 and 1998, respectively. He is a Professor at the Department of Electronic & Computer Engineering of the Technical University of Crete. His research interests include database systems, centralised/distributed data streams, data synopses, approximate query processing, uncertain databases, big-data analytics, and data mining. He has published over 110 scientific papers and his work has received over 7500 citations. He is an ACM Distinguished Scientist (2011) and a member of the IEEE.

Michalis Zervakis holds a PhD degree from the University of Toronto, Department of Electrical Engineering, since 1990. He joined the Technical University of Crete (TUC) on January 1995, where he is serving as Professor at the Department of Electronic and Computer Engineering. He is the director of the Digital Image and Signal Processing Laboratory at the Technical University of Crete. His research interests include modern aspects of signal processing, multi-channel and multi-band signal processing, wavelet analysis for data/image processing and compression, neural networks and fuzzy logic, imaging systems and integrated automation systems. Developments also include DSP-based real-time implementation. He has published over 100 scientific papers in these areas. He is a member of the IEEE.

This paper is a revised and expanded version of a paper entitled 'Biological interaction networks based on non-parametric estimation' presented at the '5th International Conference on Advances in Medical Signal and Information Processing (MEDSIP'2012)', 5–6 July, Liverpool, UK.

1 Introduction

In recent years, the description of genome sequences has resulted in large amounts of gene and protein expression data. The simultaneous examination of thousands genomic units gave a new perspective in the field of bioinformatics as it made possible the study of biological networks. Common approaches to systems biology are based on mathematical representation of biological processes aiming at a deeper understanding of biochemical interactions between genes and genes products.

The latest high-throughput microarray technologies allow the simultaneous measurements of expression levels. These technologies have given insight into microbiology since its invention (Schena et al., 1995) with large amount of data being generated. The

extended study of these data sets has provided a new perspective in gene–gene network association studies with the network construction from experimental data being a promising approach in modelling functional processes.

While a variety of computational methods have been considered for constructing gene–protein networks from observational expression data, such as linear models (Deng et al., 2005), Boolean network models (Huang, 1999; Shmulevich et al., 2002), Bayesian networks (Friedman et al., 1999; Friedman et al. 2000; Imoto et al., 2003; Kim et al., 2003), Gaussian networks (Koller and Friedman, 2009; Schäfer and Strimmer, 2005a), which aim to provide suitable mathematical models for describing stochastic network-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 behaviour between different time slices (Cho et al., 2008; Kim et al., 2003; Zou and Conzen, 2005).

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: first, the lack of experimental samples (genes/proteins) when the number of the features under examination has greatly increased. More precisely, in a typical microarray data set 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 (Schäfer and Strimmer, 2005b; Wong et al., 2003). Second, the information contained in expression data is limited by tissue quality, the experimental design, noise, and measurement errors. These factors negatively affect the estimation of causal relationships in network structure and the derivations of dependencies enclosed between neighboured genes/proteins (Wong et al., 2003).

A common graphical representation scheme is the Gaussian model firstly introduced by Waddell and Kishino (2000). However, there is a critical detail in applying Gaussian modelling. If the number of samples is far smaller than the number of features, then this framework is inefficient. The covariance matrix, embodying the interactions between genes/proteins, is often not positive definite, which rendering the computation of the partial correlation matrix.

Given these challenges, it becomes obvious that graphical models need additional tools to overcome such obstacles. In this paper, we propose a new methodology for modelling dynamic Gaussian Graphical Models (GGM) from sparse data. More specifically, we focus on the problem of completing the information loss in time varying Gaussian networks through the non-parametric framework of Kernel Density Estimation (KDE; Hansen, 2004). Our approach exploits the idea that Gaussian densities describe sufficiently biological interactions and that neighbouring gene/proteins can be described by conditional probabilities as approximations of Gaussians with non-linear parameters. In addition, due to the fact that GGMs are widely known as non-directed graphs, we introduce directions based on Bayesian Information Criterion (BIC). This makes interactions within the graph conceptually more representative to biological processes.

The paper is organised as follows. In Section 2, we provide a review of kernel-based density estimation and summarise approaches in network construction from experimental data. In this section, we also incorporate methodologies for revealing direct associations between genes/proteins. We continue in Section 3 by introducing our approach in representing non-linear dependencies between genes/proteins using a dynamic Gaussian model. In Section 4, we present results in applying the proposed modelling scheme and discuss our findings. The final section presents the conclusion and future work.

2 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 GGM. The second approach examines dependency between nodes using a non-parametric approximation of the missing experimental data through KDE. After this design step, we examine the assignment of directions to the edges of the produced networks using BIC.

2.1 Gaussian graphical model

GGMs (Dobra et al., 2004; Wu et al., 2003) 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 this 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 (PC) $\Pi = (\pi_{ik})$. 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}, with elements:

$$\pi_{ik} = -\frac{S_{ik}^{-1}}{\sqrt{S_{ii}^{-1}S_{kk}^{-1}}} \tag{1}$$

Given the experimental data, the covariance matrix is computed and then it is inverted. From equation (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 data set 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 (Wu et al., 2003), an approximation of the standard matrix inverse, based on the Singular Value Decomposition (SVD).

2.2 Kernel density estimation

KDE (Cai, 2001; Hansen, 2004) is a non-parametric framework that can predict the Probability Density Function (PDF) of a random variable. Given a limited genomic independent identically distributed data set $X = (x_1,\ldots,x_n)$, KDE allows to simulate the PDF of X as follows:

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

$K(.)$ is a symmetric positive definite Gaussian function $K(u) = \frac{1}{2\pi} e^{\frac{-1}{2}u^2}$, n is data set's size of the gene/protein X and $h > 0$ is a smoothing parameter, the bandwidth that controls the extent of the kernel (Wang and Mirota, 2010).

Under the assumption that gene and gene-products share similarities in data sets, the problem of network construction is reduced to examination of independence between nodes X_i and X_k through the cross correlation test:

$$f_h(X_i, X_k) = f_h(X_i) * f_h(X_k) \tag{3}$$

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. This means X_i and X_k share common information characteristics that imply interaction.

2.3 Edge orientation

Up to this point, we have reviewed two approaches in revealing the network structure, thus providing an intuition on whether two nodes interact. Nevertheless, they do not imply anything about 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 (Chaibub Neto et al., 2008; Chen et al., 2006).

$$f(X_k \mid X_i) f(X_i) = f(X_i, X_k) = f(X_k) f(X_i \mid X_k) \tag{4}$$

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.

$$BIC = -2 \log \hat{L} + K \log N \tag{5}$$

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

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

In more complex networks' edges are oriented by splitting the graph structure into smaller sub-networks. 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 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.

2.4 Linear Gaussian graphical model

Linear Gaussian Graphical Model (LGGM) (Werhli et al., 2006) 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 $X_i \sim N\left(\sum_k w_{ik} x_k, \sigma\right)$. Here 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 modelling 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.

2.5 Dynamic Gaussian model

Dynamic Gaussian Networks (DGN) (Kim et al., 2003; Murphy and Mian, 1999) 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 summarised as $p \times n$ $X = \left(X_1, \ldots, X_p\right)^T$ whose i-th row vector $X_i = \left(x_{i1}, \ldots, x_{in}\right)^T$ 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:

$$f\left(X_{11}, ..X_{pn}\right) = f\left(X_1\right) f\left(X_2 \mid X_1\right) .. f\left(X_p \mid X_{p-1}\right) \tag{6}$$

$$f\left(X_i \mid X_{i-1}\right) = f\left(X_{i_1} \mid P_{a(i-1),1}\right) * .. * f\left(X_{i_n} \mid P_{a(i-1),n}\right) \tag{7}$$

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

Thus, in DGNs transition between different time slices is modelled as a product of conditional probabilities where the parents of node X_{i-1} are bequeathed to X_i.

3 Proposed method

Exploiting the above tools, two networks are generated each following a different approach in revealing genetic associations (namely, PC and KDE). In this section, we augment these networks with a novel framework for estimating dependencies

between genes/proteins by enforcing a non-linear structure in modelling the parameters of their conditional probability distributions. More specifically, we represent conditional probabilities as Gaussian distributions through KDE.

3.1 Conditional probability 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. Consequently, interactions with parental nodes are modelled by the conditional distribution of each gene. We use KDE as a non-parametric framework in order to capture the dependencies from parental nodes in the 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 *i*-th gene/protein. Let P_{a_i} be the parents of gene/protein X_i then direct dependencies are encoded according to Bayes' theorem as:

$$f\left(X_i \mid P_{a_i}\right) = \frac{f\left(X_i, P_{a_i}\right)}{f\left(P_{a_i}\right)} \tag{8}$$

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

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

Replacing equation (8) with equation (9) we obtain

$$f\left(X_i \mid P_{a_i}\right) = \frac{\sum_{j=1}^{p} K_{h_1}\left(x - x_{ij}\right) K_{h_2}\left(p_{a_i} - p_{a_{ij}}\right)}{\sum_{j=1}^{p} K_{h_2}\left(p_{a_i} - p_{a_{ij}}\right)} \tag{10}$$

where $K(.)$ is a Gaussian kernel function described as equation (2), p is data set's size and $h_1 = c_1 n^{-1/6}$, $h_2 = c_2 n^{-1/6}$ for $c_1, c_2 > 0$ are the smoothing parameters selected as optimal approximations of Gaussians basis functions (Davis, 1998; Yu et al., 2010).

Equation (10) implies that the conditional density estimate is an asymptotic approximation of Gaussian (Fan et al., 1996; Hansen, 2004) $N\left(\theta_1, \sigma_1^2\right)$ with $R(K) = \int K(u)^2\, du$ and parameters as follows:

$$\theta_1 = \frac{\sigma_\kappa^2}{2\sqrt{c_1 c_2}}\left(c_1^2 f^{(2)}\left(X_i \mid P_{a_i}\right) + c_2^2 f_{(2)}\left(X_i \mid P_{a_i}\right) + 2c_2^2 f_{(1)}\left(X_i \mid P_{a_i}\right) f^{(1)}\left(X_i \mid P_{a_i}\right)\right. \tag{11}$$

$$\sigma_1^2 = \frac{R(K)^2\, f\left(X_i \mid P_{a_i}\right)}{c_1 c_2 f\left(P_{a_i}\right)} \tag{12}$$

Hence, equations (11) and (12) encode a Gaussian model that captures non-linear dependencies of network parameters. If a gene/protein has no parents, the mean and variance are 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 are no longer an obstacle due to estimation from the remaining samples.

4 Results and discussion

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

4.1 Network construction and direct relations

The data samples we used for testing concern the developing *Arabidopsis thaliana* seeds (Hajduch et al., 2010; Wille et al., 2004) harvested at 5, 7, 9, 11, and 13 days after flowering using Affymetrix ATH1 chips. We isolated the carbohydrate metabolism pathway including 7 'significant' and 6 'unrelated' genes and studies the network associated with this pathway. Genes that encode invertases (At1g35580, At5g22510) or sucrose synthases (At3g43190, At4g02280, At5g20830, At5g37180, and At5g49190), both being important enzymes in the metabolism of sucrose, were designated as 'significant' genes (Koch, 2004). In order to test our proposed algorithm, we included more than one sucrose synthase genes as internal controls. As 'unrelated' genes we designated six genes that are involved in other biological processes (intracellular traffic, energy, protein destination and storage, disease/defence) in carbohydrate metabolism (Hajduch et al., 2010; Lamesch et al., 2012). These 'unrelated' genes are either expressed in seeds (At1g54050 and At3g17520) or not expressed in seeds (At1g13140, At2g39470, At4g14630, At4g15010) and are identified as biomarkers for specific organs (flowers, leaves, roots, siliques) in *Arabidopsis*. Overall, we studied 113 genes and 27 gene–protein pairs, for all stages of growth. Our goal was to verify known gene–protein interactions, direct associations between genes as well as to highlight how the pathway is affected by significant factors.

Table 1 presents the number of verified gene–protein pairs. The first column describes different thresholds on partial correlation set on PC for (1), while the second column provides the thresholds of absolute difference of (3) for KDE. The third and fourth columns summarise 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 1 Network structure for various thresholds with PC and KDE algorithms

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

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

Table 2 shows the verified interactions between genes as well as interactions of proteins. We compared the performance of the two approaches taking into account the existent information on gene–gene and protein–protein interactions from two related databases, namely ATTED-II, the Arabidopsis gene co-expression database (Obayashi et al., 2009) and AtPIN, *A. thaliana* Protein Interaction Network (Brandão et al., 2009). The former provides 3321 genes (interacting directly or indirectly), while the latter provides 1092 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 (Liu et al., 2009). The high number of protein interactions may be relevant to a number of physically interacting proteins, but also to a number of interacting proteins that are not physically (directly) connected.

Table 2 Gene–gene and protein–protein interactions for various thresholds

Threshold		*Verified Gene Interactions*		*Verified Protein Interactions*	
PC	*KDE*	*PC*	*KDE*	*PC*	*KDE*
≥0.1	≤0.1	58/62	0/62	240/729	46/729
≥0.2	≤0.2	52/62	3/62	212/729	76/729
≥0.3	≤0.3	48/62	6/62	182/729	108/729
≥0.4	≤0.4	44/62	19/62	158/729	148/729
≥0.5	≤0.5	39/62	34/62	130/729	184/729

Tables 1 and 2 provide a notion of the identified number of verified interactions. Comparing the performance of two methodologies, KDE appears to behave better in capturing the above biological associations. More precisely, KDE identifies up to 81% of known gene/protein interactions, up to 96% known gene–gene interactions and up to 36% existent protein–protein interactions. These percentages for PC are 70%, 93% and 33%, respectively. Finally, to assess the network reconstruction ability, we counted true positives TP (correctly identified true edges), false positives FP (spurious edges), true negatives TN (correctly identified zero-edges) and false negatives FN (not recognised true edges) edges. In order to assess the number of TP, FP edges, we employed the q-value approach (Storey and Tibshirani, 2003) that establishes the statistically significant edges to be marked as false positives. Applying the two-sample *t*-test (Huber et al., 2002) on the extracted network edges, we found the amount of TP, FP, TN and FN edges of the predicted network structure. Figure 1 summarises the true positive rate for both algorithms, meaning framework's ability to detect existent interactions.

In order to find the optimal threshold for each algorithm, the size of the graph has to be taken into consideration. This is necessitated by the fact that as graph becomes denser, more interactions are generated. Thus, the probability of capturing pre-existent associations increases. We use the *F*-score metric, $F=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$ (Davis and Goadrich, 2006). The beta parameter is set to 2, as to represent the metrics of recall and precision evenly weighted. These results reveal appropriate threshold $th = 0.3$ for KDE and $th = 0.5$ for PC. For the selected thresholds we present in Figure 2 the Receiver Operating Characteristic (ROC) curves for both algorithms (Davis and Goadrich, 2006). KDE outperforms PC with the Areas under the Curve (AUC) reaching 84% and 77%, respectively.

Figure 1 True positive rate for the verified gene or/and protein interactions for KDE and PC algorithms

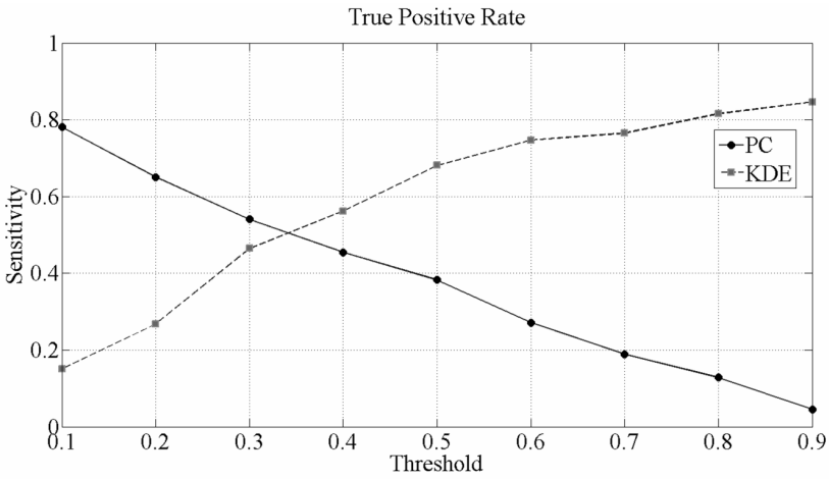
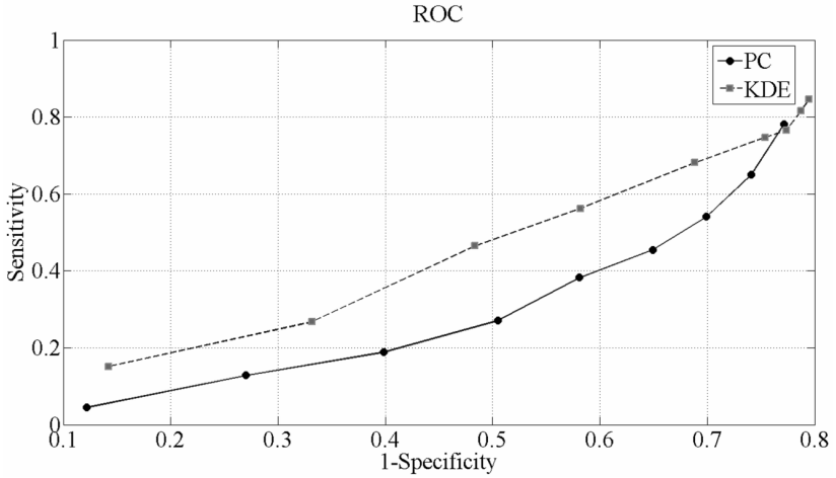


Figure 2 Receiver Operating Characteristic (ROC) of KDE and PC algorithms



In studies associated with the biological relevance of molecule interactions, one intriguing issue is the consideration of false positive associations in the definition of the 'ground truth' network (Brandão et al., 2009; Obayashi et al., 2009; Wang et al., 2012). Traditionally, true positive or negative associations involve only biologically well-established interactions. Nevertheless, such a consideration can only assess a very small number of interactions, leaving the vast majority of unidentified interactions as potentially false positive generators. Under this consideration, the false positive number is typically large, since we can only consider as ground-truth positives the direct interactions that have been biologically confirmed (Brandão et al., 2009; Hajduch et al., 2010; Obayashi et al., 2009). In practice, however, the majority of molecules in the neighbour of a gene or protein participate in similar biological processes and, as such, the entire neighbourhood may trigger many more direct interactions, which have not been experimentally established yet. Thus, an alternative consideration in the definition of ground-truth positives would include indirect associations stemming from all connections in the neighbourhood of established ones, as a valid assumption that also affects the determination of relevant false positive interactions. The validity of false positives can then be supported by statistical measures, such as q-values. Based on this rationale, we adopted the use of q-values in our study, enabling the use of many experimentally verified indirect edges as TP associations and drastically limiting the number of false positive interactions.

Addressing these aspects in more detail, Figure 3 presents a snapshot of the constructed KDE network where the At3g55650 (a) and At3g03250 (b) genes are directly connected to their neighbours. The central gene is associated with biologically verified genes (indicated by green squares), whereas triangle and circular elements indicate false positive associations including possible unidentified interactions. For gene At3g55650, KDE has successfully captured all known biological interactions. Alternatively, apart from all known associations of gene At3g03250, KDE derives many data-driven interactions, which traditionally would be considered as false positives (brown circles). With our approach, these interactions deserve closer attention, since some of them could indeed be encountered as true positives as their molecules participate in the same

processes as their neighbouring genes (green squares). The use of the statistical q-value measure supports the validity of this consideration, as it does not reject any molecule as false positives. The proposed consideration (with its associated assumptions) is further supported by recent studies, since some interacting molecules may be the result of indirect connections (not physically interacting proteins) with At3g03250 (Kim et al., 2006). Similarly, At1g13440, At3g12780, At2g36580 and the At3g03250 in Figure 3 are all identified as 14-3-3 client and binding proteins (Swatek et al., 2011). Recent studies report that the 14-3-3 proteins interact dynamically with proteins engaged in plant nitrogen and carbon metabolism, and appear to possess a modulatory role in Arabidopsis seed development. They also suggest an important role for 14-3-3 proteins in the homeostatic control of crucial glycolytic intermediates, such as the phosphoenolpyruvate (PEP) (Swatek et al., 2011). Thus, it can be argued that the central molecule At3g03250 might indeed be interacting with At1g13440, At3g12780 and At2g36580 through 14-3-3 proteins that are involved in carbohydrate metabolic process, including glycolysis of developing *A. thaliana* seed. Furthermore, due to the verification of the interaction between the central gene At3g03250 and the At1g09780 gene that encodes a phosphoglycerate mutase isozyme, we expected that the observed interactions with two other isozymes of phosphoglycerate mutase (At3g08590, At4g09520) could also be considered as true positives, as they all catalyse the same reversible reaction (3-phospho-D-glycerate to 2-phospho-D-glycerate). In addition to indirect relations within the pathway, there exist indirect interactions across related pathways, which affect the carbohydrate metabolism and need to be taken into similar consideration. In this form, the statistically derived interactions could substantiate valid assumptions for biological consideration.

Figure 3 Snapshot of direct connected genes/proteins in the KDE (0.3) network for selected (central) genes. Square green molecules indicate true positive relations; brown triangles reflect false positives and brown ellipses denote controversial false positives. Grey and orange edges show genetic and proteomic associations, respectively

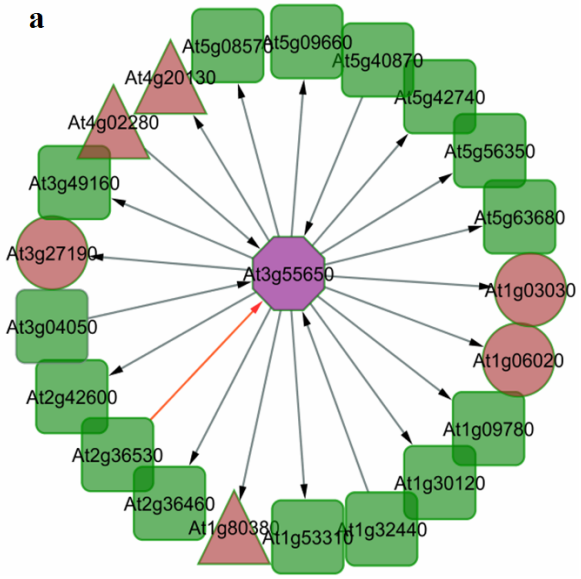
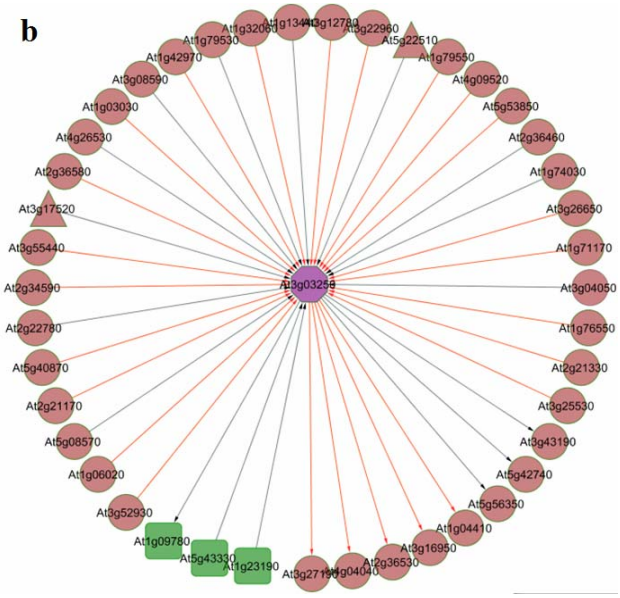


Figure 3 Snapshot of direct connected genes/proteins in the KDE (0.3) network for selected (central) genes. Square green molecules indicate true positive relations; brown triangles reflect false positives and brown ellipses denote controversial false positives. Grey and orange edges show genetic and proteomic associations, respectively (continued)



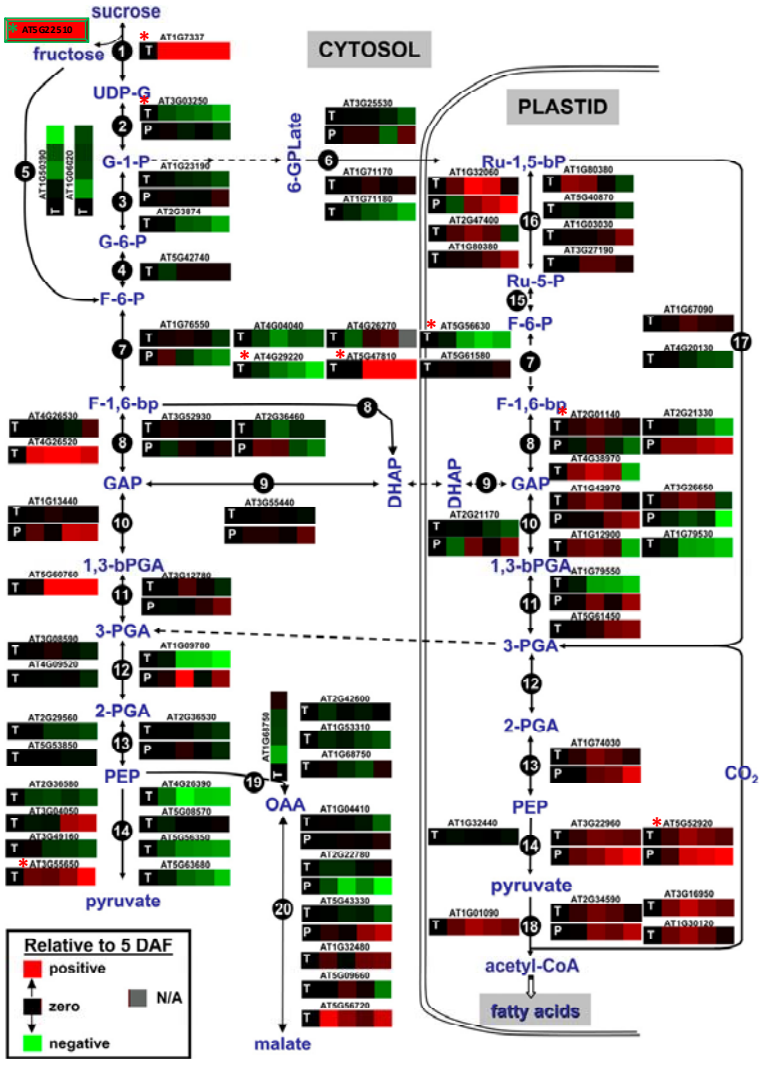
4.2 Direct and indirect implications of activation

For studying the algorithmic implications on network associations, we first examine the direct and indirect genetic implications for different time stages. For this purpose, we included in the studied carbohydrate pathway a set of carefully selected genes (7 'significant', 6 'unrelated' genes); a total number of 113 genes is examined. From this the total set we select a sub-group of genes whose expression is a priori known to affect the rest of genes/proteins involved in the pathway. Then, based on the experimental values of those few genes, we predict other genes/proteins amendable to be overexpressed (activated) or underexpressed (inhibited). We further examine if those predicted associations are verified according to Hajduch et al.'s (2010).

We now proceed with the verification of expression profiles for the studied genes and proteins for the five stages of growth in association with the related study of Hajduch et al. (2010) presented in the pathway in Figure 4. In fact, we consider alterations/associations of expression caused by the fixed initial expression of a sub-group of genes, a priori known to affect the rest. For this purpose, we isolated the genes At2g01140, At3g03250, At5g52920, At1g73370, At5g47810, At5g56630, At3g55650, At4g29220 and At5g22510, which show different expressions during the five stages of growth. Furthermore, they hold an important role in carbohydrate metabolism and their study is expected to reveal an impact on the genes/proteins involved in the pathway. To determine the impact of expression levels, we pose inference queries conditioned on the observation of each of the above genes. For instance, the probability of gene At3g26650 to be inhibited when At2g01140 gene is activated that is summarised as the conditional probability of the first given the expression level of the latter. In order to model the

activation and inhibition, we set a threshold as a mean of the experimental values for each gene. We mention here that in our consideration, a gene can either be activated or inhibited. Thus, a gene/protein is considered inhibited if its expression is lower than the experimental mean of this gene/protein in the data set; otherwise, it is labelled as activated. In order to also account for uncertainties in the close region of the threshold, the computation of the conditional probability takes soft bounds on activation/inhibition, considering the mean plus/minus one standard deviation, respectively.

Figure 4 Schematic view of carbohydrate metabolism during seed filling of Arabidopsis. Expression (heat) maps of individual protein (P) and transcript (T) expression based on proteomics and microarray experiments as relative value to 5 DAF are shown. Protein/transcript pairs are under one ATG number. Asterisks in red and bold depict the nine observed genes. At5g22510 (top) is external to the pathway and considered as 'significant' gene



Source: Hajduch et al. (2010)

Table 3 summarises the predicted expression levels from our proposed method on the KDE network. The first column shows the targeted genes for inference. The rest columns show the predicted expression profiles for the other involved genes/proteins in the pathway for all stages of growth. The presented expression profiles were selected as the most significant with the highest probability to occur. The 'unrelated' gene At5g22510 appears in many cases to have opposite expression compared to At1g73370. Both genes encode enzymes that catalyse the sucrose cleavage in plants but the final products of their enzymatic pathways are different in important aspects (Koch, 2004). Additionally, Table 3 presents the outcome of the predicted associations when At1g73370 and At3g03250 are simultaneously observed as activated and inhibited, respectively. There are many reasons for posing this query. We want to examine the robustness of the proposed method for a complex of important genes. Furthermore, queries with opposite genetic behaviours correspond to a more realistic interpretation. Finally, we observe that genes At1g73370 and At3g03250 have particular biological significance in *A. thaliana* carbohydrate metabolism. More specifically, At1g73370 encodes a sucrose synthase, which is implicated in sucrose metabolism and is vital for homeostatic regulation between metabolic pathways and sucrose signals. In addition, genes that encode sucrose synthase enzymes are responsive to the action of their own enzyme products (Koch, 2004). The At3g03250 gene encodes the UDP-glucose pyrophosphorylase, a key enzyme for carbohydrate metabolism that is essential in Arabidopsis (Meng et al., 2009). It is reported that the At3g03250 gene is co-regulated with genes implicated in carbohydrate metabolism, late embryogenesis and seed loading (Meng et al., 2009). Despite the sparse available data, our proposed method detects correctly the expression variability of many genes and/or proteins which is presented in Table 3 (e.g. At1g13440, p512; At2g21170, p2392).

Our analysis of sparse experimental data in Table 3 allows the generation of gene–protein networks and illustrates three key points focusing on the outcome interactions of the 'significant' genes associated with the KDE method. First, we observe that the target genes from the 1st column interact with genes from other columns, most of which are involved in carbohydrate metabolism. These gene-pairs are indirectly interconnected according to ATTED-II (Obayashi et al., 2009). Second, we highlight new gene–protein interactions between the 'significant' genes and proteins (3rd, 5th, 7th, 9th, 11th columns). We highlight two indicative examples: (a) 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 signalling (Lu et al., 2012) and (b) plastidial glyceraldehyde 3-phosphate dehydrogenase, a subunit (GAPA) that participates in the reductive carbon cycle and also is involved in response to sucrose stimulus (Muñoz-Bertomeu et al., 2009). Third, we reveal potentially useful, new gene–gene (direct or indirect) interactions between the target genes and the genes showed in other columns, including interactions with the seemingly 'unrelated' genes. Interestingly, the 'unrelated' gene At3g17520 has inference significance and is a member of the group 4 late embryogenesis abundant (LEA) protein genes (Lamesch et al., 2012). The presence of their encoded LEA proteins is related to the adaptive response of higher plants caused by adverse conditions to maintain normal metabolism (Hong-Bo et al., 2005). The observed gene–gene and gene–protein interactions between the various 'significant' genes with LEA gene or GAPA and FBA protein should be experimentally analysed in order to find their possible associations or cross-talks between carbohydrate metabolism and other pathways during seed development in *A. thaliana*.

Table 3 Predicted interactions from inference. Interactions of observed with other genes in the KDE network at different time points. The observed genes are selected based on their inference significance

Observed Genes	*Predicted interactions/Day-KDE (0.3)*									
	1st Day		*2nd Day*		*3rd Day*		*4th day*		*5th day*	
At1g73370 (activated)	At2g21330↑	p2225↓(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)
	At3g52930↑	-	At3g52930↑	-	At3g52930↑	p2173↓ (At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)
	At3g26650↑	-	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)	At3g26650↑	p532↑(At3g26650)
	-	p496↑ (At2g36460)	At2g36460↑	p496↑(At2g36460)	-	p496↑(At2g36460)	At2g36460↓	p496↑(At2g36460)	-	p496↑(At2g36460)
	At2g21170↑	p2392↓(At2g21170)	At2g21170↑	-	At2g21170↑	p2392↓(At2g21170)	At2g21170↑	-	At2g21170↑	p2392↓(At2g21170)
	At2g01140↓	-	At2g01140↓	-	At2g01140↓	-	At2g01140↓	p2218↓(At2g01140)	-	-
	At3g17520a↑	p2322↓(At3g17520)	-	p2322↓(At3g17520)	At3g17520a↑	p2322↓(At3g17520)	-	p2322↓(At3g17520)	At3g17520a↑	p2322↓(At3g17520)
	-	p1877↓(At3g22960)	-	p1877↓ (At3g22960)	-	p1877↓ (At3g22960)	At3g22960↓	p1877↓(At3g22960)	At3g22960↓	p1877↓(At3g22960)
	-	-	-	p512↓(At1g13440)	At1g13440↓	-	-	p512↓(At1g13440)	-	-
	-	-	-	p175↓(At1g76550)	-	-	-	p175↓(At1g76550)	-	-
	At3g25530↑	-	-	-	At3g25530↓	-	At3g25530↓	-	-	-
	At1g42970↑	-	-	-	At1g42970↑	-	-	-	At1g42970↑	-
	At3g12780↑	-	-	-	At3g12780↑	-	At3g12780↑	-	At3g12780↑	-
	At5g47810↑	-	At5g47810↑	-	At5g47810↑	-	At5g47810↑	-	-	-
	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-	At1g09780↑	-
	At5g60760↑	-	-	-	At5g60760↑	-	At5g60760↑	-	At5g60760↑	-
	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-
	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-	At4g02280↑	-
	At1g50390↑	-	-	-	At1g50390↓	-	At1g50390↓	-	At1g50390↓	-
	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-	At1g32060↑	-
	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-	At5g49190↓	-

Table 3 Predicted interactions from inference. Interactions of observed with other genes in the KDE network at different time points. The observed genes are selected based on their inference significance (continued)

Observed Genes	*Predicted interactions/Day-KDE (0.3)*									
	1st Day		*2nd Day*		*3rd Day*		*4th day*		*5th day*	
At5g22510 (activated)	At2g21330↓	p2225↓(At2g21330)	At2g21330↑	p2225↓(At2g21330)	-	-	At2g21330↑	p2225↑(At2g21330)	At2g21330↑	p2225↑(At2g21330)
	At3g52930↑	-	At3g52930↑	-	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)	At3g52930↑	p2173↑(At3g52930)
	At3g26650↓	-	At3g26650↓	-	At3g26650↓	p532↓ (At3g26650)	-	-	-	-
	-	p496↓ (At2g36460)	At2g36460↑	p496↓ (At2g36460)	-	p496↑(At2g36460)	At2g36460↑	p496↑(At2g36460)	-	p496↑(At2g36460)
	At2g21170↑	-	At2g21170↓	-	At2g21170↓	-	At2g21170↓	-	-	-
	At2g01140↓	-	At2g01140↓	-	At2g01140↓	p2218↓ (At2g01140)	At2g01140↓	-	-	p2218↓(At2g01140)
	At3g17520a↑	-	-	-	At3g17520a↑	-	-	-	At3g17520a↑	-
	-	p512↑(At1g13440)	-	-	At1g13440↓	p512↓(At1g13440)	-	-	At1g13440↑	p512↓ (At1g13440)
	At1g42970↑	-	-	-	At1g42970↑	-	-	-	At1g42970↑	-
	At3g12780↑	-	-	-	At3g12780↑	-	At3g12780↑	p2124↓ (At3g12780)	At3g12780↑	p2124↓ (At3g12780)
	At5g60760↑	-	-	-	-	-	-	-	-	-
	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-	At1g79550↑	-
	At4g02280↑	-	At4g02280↑	-	At1g79550↑	-	At4g02280↑	-	At4g02280↑	-
	-	-	-	-	-	-	At1g32060↓	-	At1g32060↓	-
	-	-	-	p2035↓(At1g74030)	-	-	-	p2035↓(At1g74030)	-	p2035↓(At1g74030)
	At4g38970↓	-	-	-	-	-	At4g38970↓	-	At4g38970↓	-
	At1g73370↑	-	At1g73370↓	-	-	-	At1g73370↑	-	At1g73370↑	-

Table 3 Predicted interactions from inference. Interactions of observed with other genes in the KDE network at different time points. The observed genes are selected based on their inference significance (continued)

Observed Genes	*Predicted interactions/Day-KDE (0.3)*									
	1st Day		*2nd Day*		*3rd Day*		*4th day*		*5th day*	
	At2g21330↓	p2225↑(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)	At2g21330↓	p2225↓(At2g21330)
	At3g52930↓	-	At3g52930↓	-	At3g52930↓	p2173↑(At3g52930)	At3g52930↓	-	At3g52930↓	p2173↓(At3g52930)
	At3g26650↓	-	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)	At3g26650↓	p532↓(At3g26650)
	-	p496↓(At2g36460)	At2g36460↓	p496↓(At2g36460)	-	p496↓(At2g36460)	At2g36460↑	p496↓(At2g36460)	-	p496↓(At2g36460)
	At2g21170↓	p2392↑(At2g21170)	At2g21170↓	-	At2g21170↓	p2392↑(At2g21170)	At2g21170↓	-	At2g21170↓	p2392↑(At2g21170)
	At2g01140↑	-	At2g01140↑	-	At2g01140↑	-	-	p2218↑(At2g01140)		
At1g73370 (activated)	At3g17520a↓	p23222↑(At3g17520)	-	p23222↑(At3g17520)	At3g17520a↓	p23222↑(At3g17520)	-	p23222↑(At3g17520)	At3g17520a↓	p23222↑(At3g17520)
	-	-	-	p512↑(At1g13440)	At1g13440↑	-	-	p512↑(At1g13440)	At1g13440↓	p512↓ (At1g13440)
At3g03250 (inhibited)	At1g42970↓	-	-	-	At1g42970↓	-	-	-	At1g42970↓	-
	At3g12780↓	-	-	-	At3g12780↓	-	At3g12780↓	-	At3g12780↓	-
	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-	At1g79550↓	-
	-	-	At1g76550↓	p175↑(At1g76550)	At1g76550↓	-	At1g76550↓	p175↑(At1g76550)	At1g76550↓	-
	At1g50390↓	-	At1g50390↓	-	At1g50390↓	-	-	-	-	-
	At1g09780↓	-	At1g09780↓	-	At1g09780↓	-	At1g09780↓	-	At1g09780↓	-
	-	-	-	-	At5g63680↓	-	At5g63680↓	-	At5g63680↓	-
	At5g49190↑	-	At5g49190↑	-	At5g49190↑	-	At5g49190↑	-	At5g49190↑	-

Note: Abbreviations: ↑ indicates activation of gene/protein; ↓indicates inhibition of gene/protein; a indicates 'unrelated' gene.

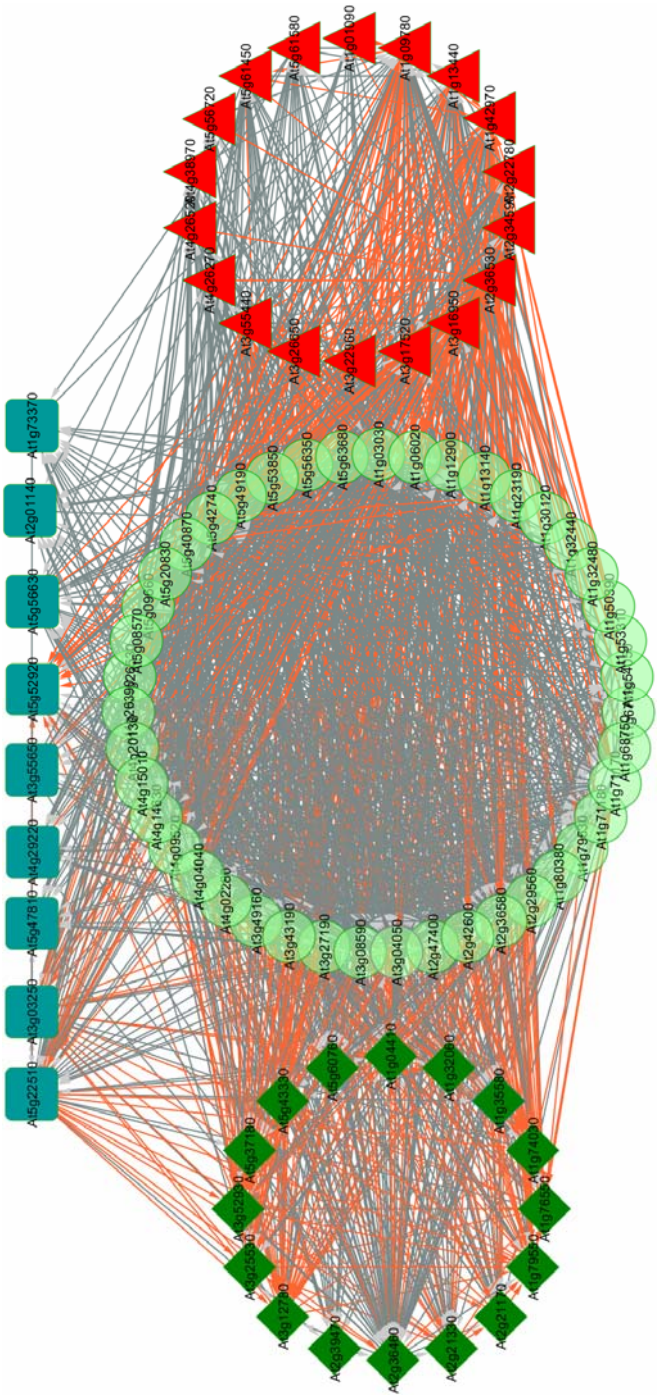
In a related attempt of Hajduch et al. (2010) to examine the behaviour of genes and corresponding proteins, the expressions from 2nd to 5th stages of seed development is compared to the corresponding 1st stage (Figure 4). In the associated colour map, red regions imply concentration increase compared to 1st stage, green regions indicate decrease, while black regions reflect no change in concentration. These activation/ inhibition results are also checked through Figure 4 and the verified associations are organised in terms of correctly predicted activations (red triangles), inhibitions (green diamonds) and no significant activations (light-green circles). The resulting network from our formulation is shown in Figure 5 illustrating these gene associations caused by the selected nine genes and which are consistently observed for all five stages of development. Regarding the expression of pairs in Figure 4 (26 pairs), the study of Hajduch et al. (2010) based on linear regression, reveals disagreement with the heat-map in some gene–protein pairs with opposite expression profiles. This effect is also derived from our approach for the gene–protein pairs At2g21330, At3g52930, At3g26650, At2g36460, At1g13440 and At1g76550, as presented in Table 3. For the remaining gene–protein pairs, the predicted expressions from our model attain low probabilities, with one exception of the At2g21170 pair that expresses discordance of expressions in time and is attributed to post transcriptional regulation.

In the remaining sections we compare the performance of our proposed method with LGGM applied on the KDE and PC networks. The genes At2g01140, At3g03250, At5g52920, At1g73370, At5g47810, At5g56630, At3g55650, At4g29220 and At5g22510 are simultaneous observed due to their importance in the involved biological processes into the pathway. Figure 5 presents the predicted expressions of our proposed method on KDE network. The predicted outcomes indicate high probabilities and same expression profiles for all stages of development.

Figure 6a and 6b shows how the heat-maps of the predicted expression profiles for all five days applying our proposed method on the networks produced by KDE and PC. Figure 6c and 6d presents the results after applying the LGGM approach on the respective networks. Regions on the heat-maps marked as red represent the predicted activation, while green regions illustrate the predicted inhibition. The intensities for both cases reflect signed probabilities with the inhibited predictions set as negative values. All predicted outcomes were chosen with probabilities higher than 0.4, while black regions in the heat-maps indicate cases with probabilities smaller than the above threshold. All cases represent predicted results conditioned on the observation of the nine above mentioned genes. Clearly, the proposed method enables both networks (constructed by KDE and PC) to achieve higher numbers of predicted interactions compared to the LGGM approach. Moreover, while our model captures inhibited in addition to activated expressions, the LGGM approach fails in identifying expressions with high probabilities for both types of networks. This illustrates another aspect of the proposed method as it predicts expression of genes for both activation and inhibition with high probabilities.

In order to validate the above observations, we compare the four derived heat-maps with the results of Hajduch et al. (2010). More specifically, we compute the precision of our outcome in relation to the results presented in Figure 4 and set each predicted expression as true (false) positive if it agrees (disagrees) with the corresponding prediction of Figure 4. However, the true positive relations are difficult to define in the entire data set. In our consideration, they involve only the direct connections in Figure 4 to the observed genes/proteins, in addition to their neighbouring molecules. The main rationale for this assumption is that adjacent molecules are expected to engage in similar biological processes and interact with the observed ones.

Figure 5 Predicted expressions for the nine observed genes (blue squares) for all stages of development. Inhibited molecules are marked as green diamonds and activated as red triangles. Genetic interactions are presented as grey and proteomic as orange



Note: Created by Cytoscape Platform.

Figure 6 Heat maps of the predicted expressions for all stages of seed development. White regions imply inhibition, grey regions imply activation, and black regions reflect none predicted expression. The horizontal axes represent the five stages of seed development. (a) Proposed method on KDE network; (b) LGGM method on KDE network; (c) Proposed method on PC network; (d) LGGM method on PC network

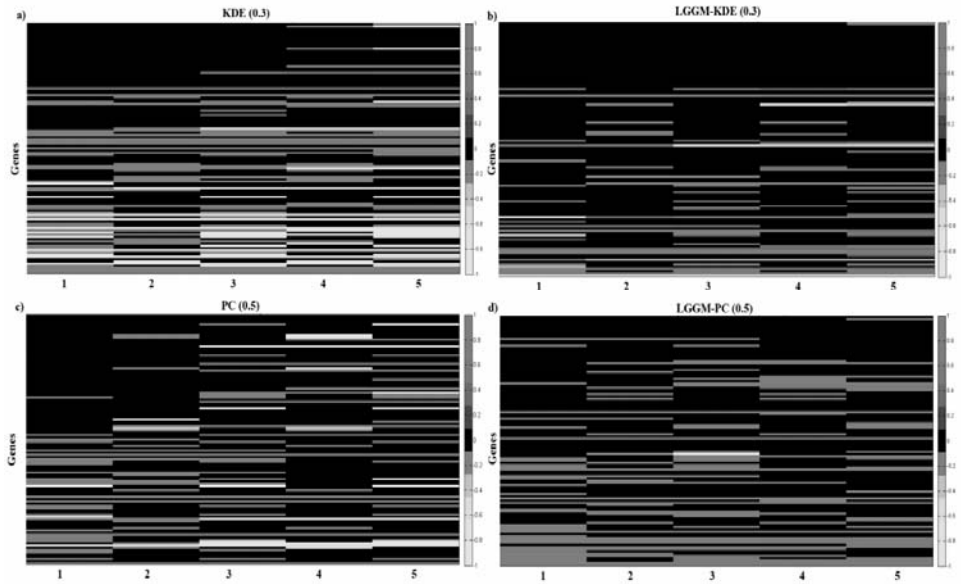


Figure 7 Precision-recall curves for different levels of probability according to the het-maps of Figure 6. The rightmost point in each figure reflects probability over 0.4; while the leftmost point is for probability 0.9, evenly spaced; (a) Proposed method on KDE network; (b) LGGM method on KDE network; (c) Proposed method on PC network; (d) LGGM method on PC network

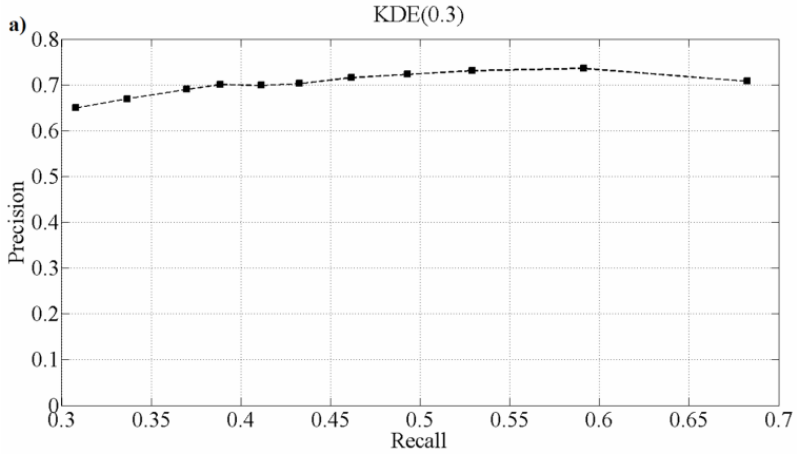


Figure 7 Precision-recall curves for different levels of probability according to the het-maps of Figure 6. The rightmost point in each figure reflects probability over 0.4; while the leftmost point is for probability 0.9, evenly spaced; (a) Proposed method on KDE network; (b) LGGM method on KDE network; (c) Proposed method on PC network; (d) LGGM method on PC network (continued)

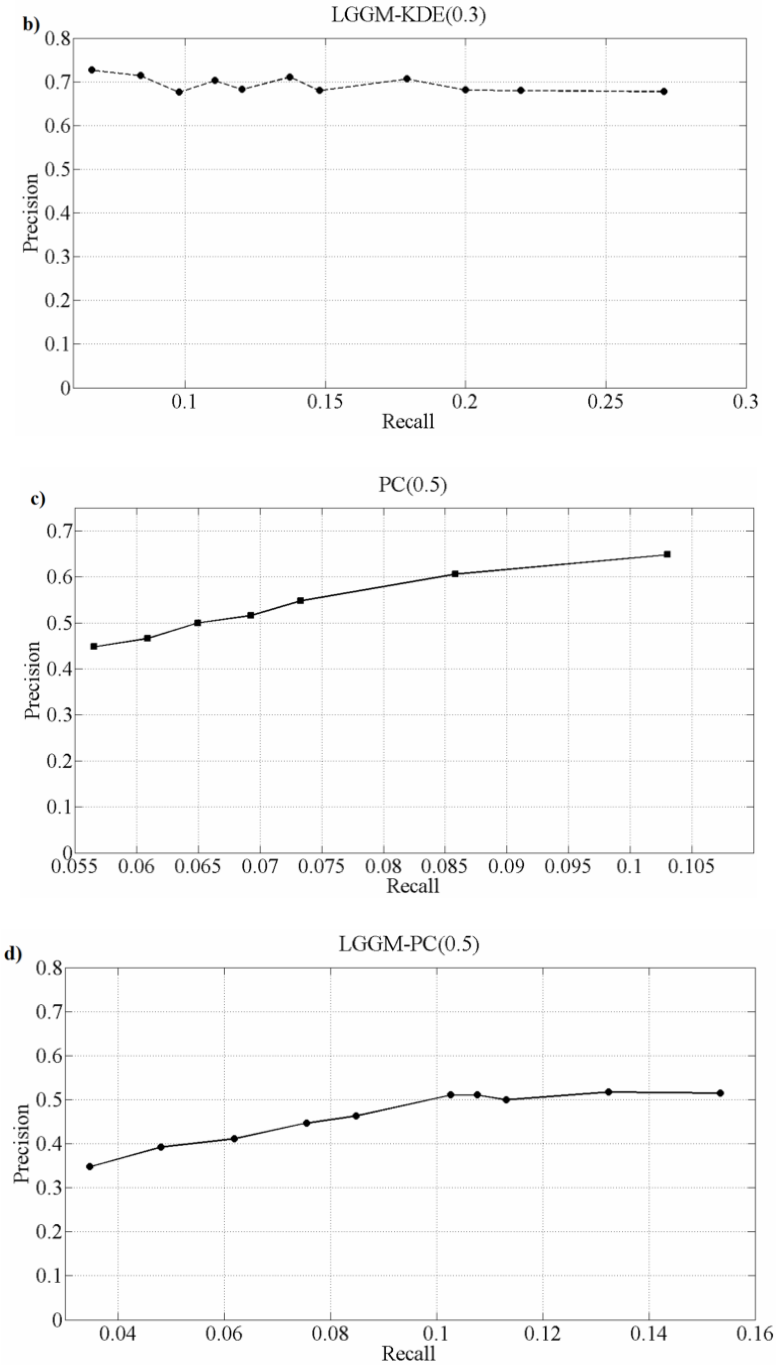


Figure 7 describes the precision-recall curves on KDE and PC networks for different levels of probability, higher than 0.4. Our goal is to successfully predict the activation/ inhibition (246 red/green boxes of Figure 4) of seed development. Moving from right to left, the marked points indicate probabilities from 0.4 to 0.9, evenly spaced. Thus, the rightmost point indicates probability higher than 0.45, while the leftmost point reflects probability higher than 0.9 for the targeted predictions. The proposed method applied on the KDE network (Figure 7a) reaches high levels of precision and recall (74% and 70%, respectively). It identifies successfully up to 65% of true positives with probability higher than 0.9; as the subset of the predicted interactions is augmented (with lower thresholds on probability), the precision also rises, reaching its highest score at probability 74%. This is an indication that the proposed framework is able to discriminate the expression profiles giving high scores to true positives. In addition, it reaches high levels of recall for probabilities higher than 0.45, indicating that the model is able to make predictions for all different stages. As the level of probabilities increases, the recall is reduced since the model becomes stricter and the amount of predicted outcomes is smaller than the ideal goal of 246 true positives. In contrast to our framework, the other studied approaches can show similar precision scores but they fail in reaching high recall; in fact they fail to make predictions for the given data set, even for smaller levels of probability. Overall, our method outperforms others in revealing predicted temporal expressions in terms of recall and precision, in addition to discriminating activated and inhibited genes/proteins. Interestingly, while the LGGM-KDE (Figure 7b) approach has low recall, it has higher precision levels compared to LGGM-PC (Figure 7c) and PC (Figure 7d). This suggests that the network structure based on KDE has successfully captured many biological interactions verifying the superiority of KDE over PC. Moreover, our proposed method allows generating gene–protein networks that verify existing interactions of genes/proteins and also reveal new gene–gene (direct or indirect) and gene–protein interactions within the examined carbohydrate metabolic pathway. In the same way, it illustrates interactions with other 'unrelated' molecules (e.g. LEA protein genes and their encoded LEA proteins), which may indicate a possible cross-talk between carbohydrate metabolism and other signalling pathways during seed development in *A. thaliana*. Finally, capturing temporal expressions for all different stages, we achieve to recover co-expressed molecules, which in turn might reveal possible functionally related genes and give insights about the genetic regulatory systems of the elaborated pathways.

5 Conclusion

In this paper, we develop a framework for network creation towards examining gene and/or protein associations at different stages of organism development. The associations identified by KDE have significant overlap with verified associations between the participating genes/proteins, as the majority of the genes/proteins are located close to the processes of the carbohydrate metabolism pathway. On the contrary, the PC approach appears to capture less of those associations. Thus, ROC and precision-recall curves indicate that KDE performs better on network construction. This also supports the claim that KDE performs better in modelling the genetic associations with sparse experimental data compared to other related algorithms. Considering the modelling of conditional dependencies, both heat-maps and precision curves indicate that genetic associations

enclose more complex dependencies, whereas linear Gaussian approaches lack the ability to model/predict such relations. Ongoing research is under investigation in an attempt to reveal the potential benefits of this methodology on human cancer, so as to highlight important gene profiles of the participant genes in dysregulated pathways in cancer diseases (oral, breast) (Kalantzaki et al., 2013).

The most important contribution of this study is the provision of a different perspective in revealing the identity of genetic interactions. Network construction is complex problem, which has been studied with simplifications at the different layers of genetic information. Unrealistic assumptions often cause the generation of poor results on precision, especially in such complex organisms. The direct interactions are to a large extent unknown, especially if we take into account all the possible pathways that affect groups of genes. In addition, the available knowledge of direct interactions is established under specific conditions, which also seem to change when abnormalities happen. These issues imply the need to re-examine the generation mechanism for expression profiles in relation to underlying genetic factors and their direct or indirect relevance in specific pathways. In this direction, our approach enables the verification of relations in the expression profiles from the underlying interactions, and can be used as a first step in studying whether indirect effects of important genetic molecules verify to a good extent the expression profiles of genes involved in the pathway as well as in uncovering regulatory systems of these genes.

Acknowledgements

We sincerely thank Dr. Jay J. Thelen (Department of Biochemistry, University of Missouri, Columbia) for kindly providing us with the experimental data used in this study. This research supported by 'OASYS' project funded by the NSRF 2007-13 of the Greek Ministry of Development and by 'YPERThEN' project, which is funded by the EU and funds from Greece and Cyprus.

References

Brandão, M.M., Dantas, L.L. and Silva-Filho, M.C. (2009) 'AtPIN: *Arabidopsis thaliana* protein interaction network', *BMC Bioinformatics*, Vol. 10, No. 454, doi: 10.1186/1471-2105-10-454.

Cai, Z. (2001) 'Weighted Nadaraya–Watson regression estimation', *Statistics & Probability Letters*, Vol. 51, No. 3, pp.307–318.

Chaibub Neto, E., Ferrara, C.T., Attie, A.D. and Yandell, B.S. (2008) 'Inferring causal phenotype networks from segregating populations', *Genetics*, Vol. 179, No. 2, pp.1089–1100.

Chen, X-W., Anantha, G. and Wang, X. (2006) 'An effective structure learning method for constructing gene networks', *Bioinformatics (Oxford, England)*, Vol. 22, No. 11, pp.1367–1374.

Cho, H.C., Fadali, M.S. and Lee, K.S. (2008) 'Online probability density estimation of nonstationary random signal using dynamic Bayesian networks', *International Journal of Control, Automation, and Systems*, Vol. 6, No. 1, pp.109–118.

Davis, D.T. (1998) 'Expanding Gaussian kernels for multivariate conditional density estimation', *IEEE Transactions on Signal Processing*, Vol. 46, No. 1, pp.269–275.

Davis, J. and Goadrich, M. (2006) 'The relationship between precision-recall and ROC curves', *Proceedings of the 23rd International Conference on Machine Learning – ICML'06*, ACM Press, New York, New York, USA, pp.233–240.

Deng, X., Geng, H. and Ali, H. (2005) 'EXAMINE: a computational approach to reconstructing gene regulatory networks', *Bio Systems*, Vol. 81, No. 2, pp.125–136.

Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G. and West, M. (2004) 'Sparse graphical models for exploring gene expression data', *Journal of Multivariate Analysis*, Vol. 90, No. 1, pp.196–212.

Fan, J., Yao, Q. and Tong, H. (1996) 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems', *Biometrika*, Vol. 83, No. 1, pp.189–206.

Friedman, N., Iftach, N. and Peér, D. (1999) 'Learning Bayesian network structure from massive datasets: the 'sparse candidate' algorithm', *UAI'99 Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp.206–215.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) 'Using Bayesian networks to analyze expression data', *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, Vol. 7, Nos. 3/4, pp.601–620.

Hajduch, M., Hearne, L.B., Miernyk, J., Casteel, J.E., Joshi, T., Agrawal, G.K., Song, Z., Zhou, M., Xu, D. and Thelen, J.J. (2010) 'Systems analysis of seed filling in Arabidopsis: using general linear modeling to assess concordance of transcript and protein expression', *Plant Physiology*, Vol. 152, No. 4, pp.2078–2087.

Hansen, B.E. (2004) 'Nonparametric conditional density estimation', Preliminary manuscript.

Hong-Bo, S., Zong-Suo, L. and Ming-An, S. (2005) 'LEA proteins in higher plants: structure, function, gene expression and regulation', *Colloids and Surfaces: B, Biointerfaces*, Vol. 45, Nos. 3/4, pp.131–135.

Huang, S. (1999) 'Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery', *Journal of Molecular Medicine*, Vol. 77, No. 6, pp.469–480.

Huber, W., Heydebreck von, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) 'Variance stabilization applied to microarray data calibration and to the quantification of differential expression', *Bioinformatics*, Vol. 18, No. 1, pp.S96–S104.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2003) 'Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks', *Proceedings/IEEE Computer Society Bioinformatics Conference*, Vol. 2, pp.104–113.

Kalantzaki, K., Bei, E., Exarchos, K.P., Zervakis, M., Garofalakis, M. and Fotiadis, D.I. (2013) 'Nonparametric network design and analysis of disease genes in oral cancer progression', *IEEE Journal of Biomedical and Health Informatics*, Vol. PP, No. 99, pp.2168–2194.

Kim, S.Y., Imoto, S. and Miyano, S. (2003) 'Inferring gene networks from time series microarray data using dynamic Bayesian networks', *Briefings in Bioinformatics*, Vol. 4 No. 3, pp.228–235.

Kim, P.M., Lu, L.J., Xia, Y. and Gerstein, M.B. (2006) 'Relating three-dimensional structures to protein networks provides evolutionary insights', *Science*, Vol. 314 No. 5807, pp.1938–41.

Koch, K. (2004) 'Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development', *Current Opinion in Plant Biology*, Vol. 7 No. 3, pp.235–446.

Koller, D. and Friedman, N. (2009) *Probabilistic Graphical Models Principles and Techniques*, in Koller, D. and Friedman, N. (Eds), The MIT Press, Cambridge MA.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) 'The arabidopsis information resource (TAIR): improved gene annotation and new tools', *Nucleic Acids Research*, Vol. 40, pp.D1202–D1210.

Liu, G., Wong, L. and Chua, H.N. (2009) 'Complex discovery from weighted PPI networks', *Bioinformatics*, Vol. 25 No. 15, pp.1891–1897.

Lu, W., Tang, X., Huo, Y., Xu, R., Qi, S., Huang, J., Zheng, C. and Wu, C.A. (2012) 'Identification and characterization of fructose 1,6-bisphosphate aldolase genes in Arabidopsis reveal a gene family with diverse responses to abiotic stresses', *Gene*, Vol. 503, No. 1, pp.65–74.

Meng, M., Geisler, M., Johansson, H., Harholt, J., Scheller, H. V, Mellerowicz, E.J. and Kleczkowski, L. (2009) 'UDP-glucose pyrophosphorylase is not rate limiting, but is essential in Arabidopsis', *Plant & Cell Physiology*, Vol. 50, No. 5, pp.998–1011.

Muñoz-Bertomeu, J., Cascales-Miñana, B., Mulet, J.M., Baroja-Fernández, E., Pozueta-Romero, J., Kuhn, J.M., Segura, J. and Ros, R. (2009) 'Plastidial glyceraldehyde-3-phosphate dehydrogenase deficiency leads to altered root development and affects the sugar and amino acid balance in Arabidopsis', *Plant Physiology*, Vol. 151, No. 2, pp.541–558.

Murphy, K. and Mian, S. (1999) *Modelling Gene Expression Data using Dynamic Bayesian Networks*. Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) 'ATTED-II provides coexpressed gene networks for Arabidopsis', *Nucleic Acids Research*, Vol. 37, pp.D987–D991.

Schäfer, J. and Strimmer, K. (2005a) 'An empirical Bayes approach to inferring large-scale gene association networks', *Bioinformatics*, Vol. 21, No. 6, pp.754–764.

Schäfer, J. and Strimmer, K. (2005b) 'An empirical Bayes approach to inferring large-scale gene association networks', *Bioinformatics*, Vol. 21, No. 6, pp.754–764.

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science*, Vol. 270, No. 5235, pp.467–470.

Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (2002) 'Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks', *Bioinformatics*, Vol. 18, No. 2, pp.261–274.

Storey, J.D. and Tibshirani, R. (2003) 'Statistical significance for genome-wide experiments', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 100, pp.9440–9445.

Swatek, K.N., Graham, K., Agrawal, G.K. and Thelen, J.J. (2011) 'The 14-3-3 isoforms chi and epsilon differentially bind client proteins from developing arabidopsis seed', *Journal of Proteome Research*, Vol. 10, No. 9, pp.4076–4087.

Waddell, P.J. and Kishino, H. (2000) 'Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data', *Genome Informatics Workshop on Genome Informatics*, Vol. 11, pp.129–140.

Wang, C., Marshall, A., Zhang, D. and Wilson, Z. (2012) 'NAP: an integrated knowledge base for Arabidopsis protein interaction network analysis', *Plant Physiology*, Vol. 158, No. 4, pp.1523–1533.

Wang, H. and Mirota, D.G.D.H. (2010) 'A generalized kernel consensus-based robust estimator', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 1, pp.178–184.

Werhli, A.V, Grzegorczyk, M. and Husmeier, D. (2006) 'Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks', *Bioinformatics*, Vol. 22, No. 20, pp.2523–2531.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. and Buhlmann, P. (2004) 'Sparse graphical Gaussian modeling of the isoprenoid gene network', *Arabidopsis thaliana*', *Genome Biology*, Vol. 5, No. 11, R92p.

Wong, B.F., Carter, C.K. and Kohn, R. (2003) 'Efficient estimation of covariance selection models', *Biometrika*, Vol. 90, No. 4, pp.809–830.

Wu, X., Ye, Y. and Subramanian, R.K. (2003) 'interactive analysis of gene interactions using graphical Gaussian model', *3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pp.63–69.

Yu, K., Ally, A.K. and Hand, D.J. (2010) 'Kernel quantile-based estimation of expected shortfall', Vol. 12, No. 4, pp.15–32.

Zou, M. and Conzen, S.D. (2005) 'A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data', *Bioinformatics*, Vol. 21, No. 1, pp.71–79.