Journal of
Health & Medical Informatics

Xiang et al., J Health Med Informat 2011, S2
DOI: 10.4172/2157-7420.S2-001

Research Article Open Access

Divergence Weighted Independence Graphs for the Exploratory Analysis of Biological Expression Data

Yang Xiang[1], Marja Talikka[1], Vincenzo Belcastro[1], Peter Sperisen[1], Manuel C. Peitsch[1], Julia Hoeng[1]* and Joe Whittaker[2]*

[1]Department of Biological System Research, Philip Morris International R&D, Neuchatel, Switzerland
[2]Department of Mathematics and Statistics, Lancaster University, UK

Abstract

Motivation: Understanding biological processes requires tools for the exploratory analysis of multivariate data generated from in vitro and in vivo experiments. Part of such analyses is to visualise the interrelationships between observed variables.

Results: We build on recent work using partial correlation, graphical Gaussian models, and stability selection to add divergence weighted independence graphs (DWIGs) to this toolbox. We measure all quantities in information units (bits and millibits), to give a common quantification of the strength of associations between variables and of the information explained by a fitted graphical model. The marginal mutual information (MI) and conditional MI between variables directly account for components of the information explained. The conditional MIs are displayed as edge weights in the independence graph of the variables, making the complete graph informative as to the unique association between those variables. The summary table of the information decomposition 'total = explained + residual' provides a simple comparison of graphical models suggested by different search routines, including stabilised versions. We demonstrate the relevance of the conditional MI statistics to the graphical model of the data by analysing simulated data from the insulin pathway with a known ground truth. Here the method of thresholding these statistics to suggest a network performs at least as well as several other network searching algorithms. In searching a biological data set for novel insight, we contrast the DWIGs from the fitted maximum weight spanning tree and from the fitted model of a stabilised ARACNE network. DWIG is a powerful tool for the display of properties of the fitted model or of the empirical data directly.

Keywords: Bronchoalveolar lavage fluid (BALF); Divergence weighted independence graphs (DWIG); Conditional mutual information (CMI); Graphical gaussian model (GGM)

Abbrevitations: BALF: Bronchoalveolar lavage fluid; DWIG: Divergence weighted independence graphs; CMI: Conditional mutual information; GGM: Graphical gaussian model

Introduction

Graphical models have a long history and one with strong connections to biology, especially the work of Sewall Wright [1]. The explicit relationship between the ideas of conditional independence and graph theory was made in Darroch et al. [2] and the texts of Lauritzen and Whittaker [3,4] and, more recently, Koller and Friedman [5] outline this theory.

Bioinformatics has given impetus to the study of graphical models in the past decade with its interest in large networks generated by molecular interactions and gene transcription experiments; for instance one can cite the work of Butte, de la Fuente, Dobra , Ma, Magwene, and Toh [6-11]. Additionally to an enhanced motivation, the interest from bioinformaticians has brought new methodology with a focus on mutual information (MI) [12], and the application of MI to graphical model search [13,14].

In mainstream statistics the past decade has seen novel developments in variable selection and regularisation and its application to graphical model search and estimation, for instance [15-17]. An exposition of the field is given in [18], where much research emanated from the seminal paper of Tibshiranie [19]. Recently stability selection [20] helps to improve the generalisability of found models to different data sets. Several model search routines within the space of Gaussian graphical models are now implemented in R packages [21], including the Chow Liu tree [22,23], the ARACNE [13,24] and Glasso [15] routines.

There are many biological experiments which lead to the analysis of tens rather than thousands of variables. Some arise by taking subsets of larger arrays and while of relatively lower dimension the data is still inherently multivariate. Here we propose divergence weighted independence graphs (DWIGs) be added to the tool kit of the statistical analyst. We outline here the steps taken in using this methodology.

Preliminary to our procedure is standard variable screening involving the selection of interesting subsets, adjustments for treatment effects, transformations to normality and imputation for missing values. This uses traditional statistical tools. A first pass of the data is to display the DWIG calculated from the sample variance matrix. This may be for all variables, or for a subset of variables. The weights are conditional MIs for each pair of variables adjusted for all the other variables in the

***Corresponding authors:** Julia Hoeng, Department of Biological System Research, Philip Morris International R&D, Neuchatel, Switzerland, Tel: +41 (58) 242 2214; Fax: +41 (58) 242 2811; E-mail: Julia.Hoeng@pmi.com

Joe Whittaker, Department of Mathematics and Statistics, Lancaster University, UK, E-mail: joe.whittaker@lancaster.ac.uk

Received November 19, 2011; **Accepted** December 12, 2011; **Published** December 16, 2011

Citation: Xiang Y, Talikka M, Belcastro V, Sperisen P, Peitsch MC, et al. (2011) Divergence Weighted Independence Graphs for the Exploratory Analysis of Biological Expression Data. J Health Med Informat S2. doi:10.4172/2157-7420.S2-001

set. The calculation is made using formulae for partial correlations of normally distributed variables, described in the Supplement. The plot is implemented using the dot routine of [25]. Because of the weighting the complete graph is informative about the relative strengths of interactions between these variables. When there are too many variables to display easily, a subgroup of more informative variables may be chosen.

A concern in applying DWIGs to large numbers of variables is that the conditional MIs are poorly estimated because the set of conditioning variables is also large. A remedy is to impose constraints so that the near zero conditional MIs become exactly zero and so better estimate the larger ones. This can be achieved by thresholding the conditional MIs. However it is better to employ graphical model search to find an 'optimal' graph and to compute the conditional MIs using the variance matrix from the fitted model. The DWIG can be displayed for any Gaussian graphical model search routine, but we suggest implementing stability selection to improve the generalisability of the fitted graph.

Assessing how well one model fits, or comparing its fit to another, is an important part of the analysis. The total information to explain is a measure of how much dependence exists in the data. With a fitted variance matrix, the total information satisfies the identity that 'total = explained + residual', and the edge weights in the DWIG constitute part of this explained information. The summary table of this information decomposition provides a simple comparison of graphical models suggested by different search routines, including stabilised versions of such.

Output from this analysis are one or more DWIGs calculated from the sample and from fitted variance matrices. The graph is easily comprehensible because weak associations are almost invisible making strong interactions show clear patterns, such as clustering and chains, and thereby highlighting markers of potential biological interest. The MIs displayed that quantifies these relationships is conditional, adjusted for common association with other neighbouring variables, so improving the possibility of identifying unique interactions and causal relations.

In keeping with an exploratory analysis this procedure may well go through several iterations, starting with a relatively small subset and then choosing different but overlapping subsets of variables for analysis and for comparison.

The novel highlights of the proposed method of constructing such graphs are firstly, choosing conditional MI measures that directly relate to the Markov properties of the graph, and secondly, that the construction can be applied both to observed and to fitted variance covariance matrices.

The principal features of the methodology are described above, and their mathematical basis is outlined in the Supplement. In the Results section following we discuss two substantive examples of this technique. First, using simulated data from the well known insulin pathway, we show that thresholding empirical conditional MI weights performs better than, or at least as well as, other network searching algorithms such as CLR, Banjo, MRNET, and ARACNE. Secondly in the search for novel biological insight from a data set of biological interest, we compute the DWIGs of the most informative subset of variables, of the best tree representation of the data, and give a comparison of the DWIGs from the treated and control groups obtained by fitting models elucidated from a stabilised network search. Finally we end with a brief summary in the Discussion section.

Methods

Divergence weighted independence graph

The main definitions and formulae for the construction of divergence weighted independence graphs are given in the Supplement, together with references. Here we briefly summarise how the methodology relates to the concepts of statistical independence and information. There are two basic notions: the conditional independence graph that underpins the theory of graphical models; and the mutual information (MI) measure from classical information theory. In the analysis of a units-by-variables data matrix it is the variables that are identified with the nodes of the graph. The edge weights are conditional MIs and reflect the strength of association between the two variables, conditioned on all other variables in the graph. These weights are measured in millibits derived from entropy. When a weight is exactly zero it indicates the conditional independence of the two variables.

Making the assumption that the joint distribution of the samples is multivariate normal leads to (well-known) formulae for computing these weights from the empirical variance covariance matrix. Additionally this gives formulae for determining which the most informative variables are. Fitting a graphical Gaussian model (GGM) to data clarifies the relationships between the variables, and the generalised likelihood ratio test (deviance) permits an assessment of how well the model mimics the data. Computing the edge weights from the fitted variance matrix gives a weighted independence graph appropriate to the fitted model. Calculation of the deviance from fitting a GGM leads to a decomposition in explained information with components that relate to the edge weights.

Stability selection using subsampling

The parameters of a graphical Gaussian model are partial correlation coefficients, and their sample estimates are often poor when the sample size is not big. Graphical model search is difficult because of this together with the fact that the number is large and the parameters are inter-related. Meinshausen and Buhlmann [20] propose a new procedure using subsampling for structure estimation and variable selection in high dimensions. It gives finite sample control for the error rate of some false discoveries and can be used to tune regularisation parameters. We use it here, in combination with the ARACNE search engine, to provide stability of edge selection in face of perturbations of the data.

The subsampling procedure estimates the probability that a given edge is selected when running the search procedure on a randomly chosen subsample size [n/2] from the data. The stable set of edges is the set for which this probability exceeds a threshold, which we set to 0.50. This threshold is motivated by analogy to the median probability model of Barbieri and Berger [26] in Bayesian model selection, where variables are selected with marginal posterior probability of at least 1/2 as opposed to those variables with the highest joint posterior probability.

A/J mouse

Female A/J mice, about 6 months of age, were exposed to diluted mainstream smoke from the Reference Cigarette 2R4F for 2, 3, and 4

h per day, 5 days per week during 3 and 5 months. The mean smoke concentration throughout the study was 735 μg total particulate matter/l. All mice were exsanguinated under deep pentobarbital anesthesia. The tracheae was cannulated and the lungs lavaged with 1 ml of Ca^{2+}- and Mg^{2+}- free phosphate-buffered saline (PBS). After centrifugation the supernatants from the 1st lavage cycle were aliquoted in adequate volumes, stored below -60°C and sent to RBM (Rules Based Medicine, http://www.rulesbasedmedicine.com/) for MAP (multi-analyte profile) analysis.

Results

We describe the results of using divergence weighted independence graphs in two examples. The first is a simulated data set from a known pathway model, chosen to compare our empirical findings with the generating model. The second is a case study into experimental data in which the DWIG is employed to recover known biological interactions and to discover novel ones.

Insulin pathway

Data was simulated from the mathematical model of Sedaghat et al. [27] without feedback that includes postreceptor events involved in insulin signaling (Figure 1). When insulin binds to its receptor located at the cell surface, the receptor (INSR) undergoes autophosphorylation. The resulting increase in receptor tyrosine kinase activity leads to phosphorylation of insulin receptor substrate 1 (IRS1) and subsequent activation of phosphatidylinositol 3-kinase (PI3K). Further activation of downstream kinases stimulates the translocation of the glucose transporters and glucose uptake by the tissue.

We began simulation of this model by generating time courses for all state variables in response to a maximally stimulating step input of 10^{-7} M insulin that was turned off after 15 minutes. There are 21 variables described in this mathematical model [27]. The kinetic parameters and initial protein concentrations or percentages of every virtual animal were generated by adding a random noise to the corresponding population values given in this paper, according to

$$\theta' = \theta \times (1+\varepsilon), \quad with \quad \varepsilon \sim N(0,\sigma^2) \qquad (1)$$

where θ is any specific kinetic parameter or initial protein concentration/percentage, and the standard deviation $\sigma \in [0.05, 0.20]$. The simulation, based on a system of differential equations, was performed for each virtual animal for time $t \in [0,60)]$ minutes. Different values of $\sigma \in [0.05, 0.20]$ were utilised and the results of different values of σ are quite similar. The results with $\sigma = 0.15$ are shown below while the results of other values of σ are shown in the supplementary material.

Thirty two virtual animals were sacrificed at each time point, t =5, 10, 20, 30, 40 mins, and the simulated protein concentrations and percentages were recorded. A measurement error was added to the simulated protein concentration or percentage using the same formula as (1). In total 160 virtual animals were sacrificed. This simulated data mimic not only the within-individual stochastic variation, but also between individual variability.

The data required some preliminary data analysis: some variables have been removed because of a preponderance of zero values; the ones retained had a few outliers including zeros which required imputation; all variables required a log transform to improve the linearity of the relationships and the plausibility of a normal distribution assumption. From this we selected several subsets of variables for further analysis; one, termed 'nonzero', presented here to illustrate our methods, consists of x_5, x_6, ..., x_{12} with definition in Table 1, though below we display the x labels for convenience.

A flavour of the data is given by the pairs plot (Figure 2) which shows substantial dependence.

We are looking for a description of the key interactions and associations between the variables, and a measurement of the total amount of dependence there is to explain.

Insulin: contrasting DWIGs based on marginal and conditional MI: The total information to explain is the divergence between complete dependence and complete independence given at expression (10) in the Supplementary section. For the insulin data this total dependence amounts to 10315 millibits of information for these 8 variables.

The DWIGs computed from the marginal and the conditional

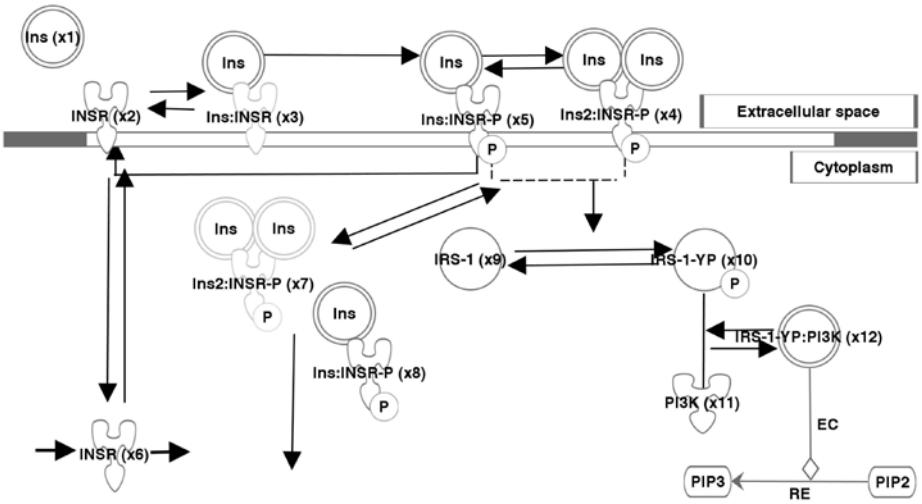

Figure 1: The schematic plot of the insulin pathway without feedback [27]. The black arrows denote the direction of signal transduction in the insulin pathway. Only the variables used in the DWIG analysis were plotted. This plot was generated through the use of IPA (Ingenuity Systems, www.ingenuity.com).

Label	State variable
x1	Ins, insulin input
x2	INSR, concentration of unbound surface insulin receptors
x3	Ins:INSR, concentration of unphosphorylated once-bound surface receptors
x4	Ins2:INSR-P, concentration of phosphorylated twice-bound surface receptors
x5	surface Ins:INSR-P, i.e. concentration of phosphorylated once-bound surface receptors
x6	Intracellular INSR, i.e. concentration of unbound unphosphorylated intracellular receptors
x7	Intracellular Ins2:INSR-P, i.e. concentration of phosphorylated twice-bound intracellular receptors
x8	Intracellular Ins:INSR-P, i.e. concentration of phosphorylated once-bound intracellular receptors
x9	IRS-1, i.e. concentration of unphosphorylated insulin receptor substrate-1
x10	IRS-1-YP, i.e. concentration of tyrosine-phosphorylated IRS1
x11	Unactivated PI3K, i.e. concentration of unactivated PI 3-kinase
x12	IRS-1-YP:PI3K (activated), i.e. concentration of tyrosine-phosphorylated IRS1/activated PI3-kinase complex

Table 1: Definition of variables in the insulin model.

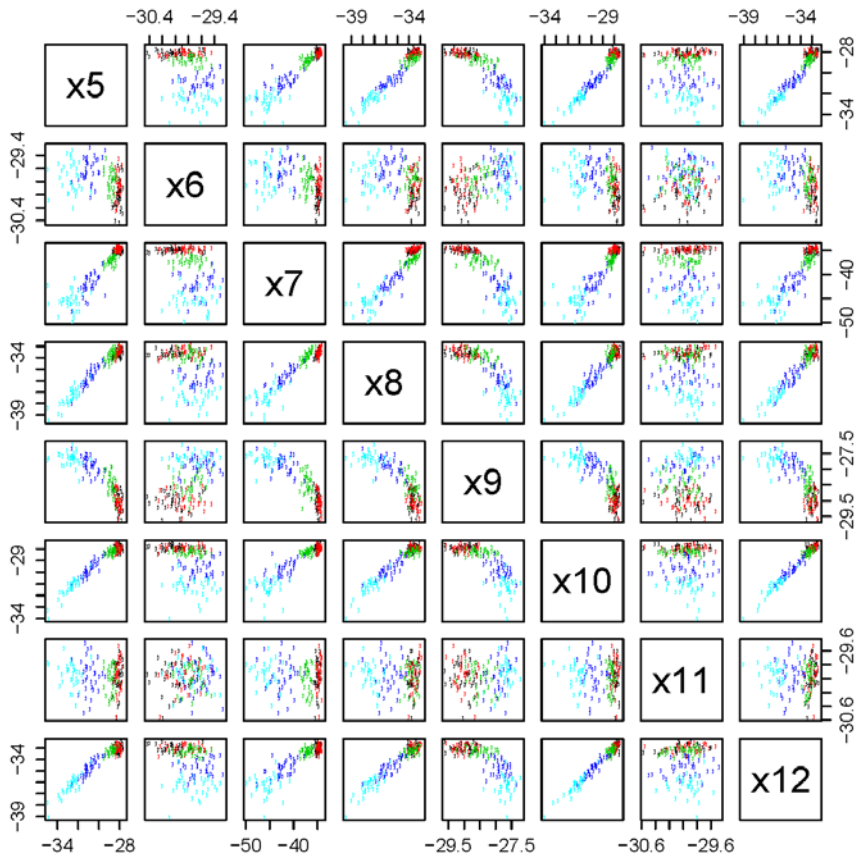

Figure 2: Pairs plots for nonzero subset. The units at the 5 sacrifice times are colour coded. Black, red, green, blue, and cyan, correspond to the 5, 10, 20, 30, 40 time groups respectively.

MI break this quantity down into those parts attributable to pairwise relationships; these are displayed in Figure 3.

The marginal MIs displayed in Figure 3A are calculated from the sample correlations using expression (6) of the Supplement. To give an idea of the transformation between correlation and divergence note that a correlation of 0.2 corresponds to 30.1mbits while one of 0.8 corresponds to 754.7 mbits. Figure 3A is dominated by the thick black lines, confirming the substantial correlations between most pairs of variables already evident in the pairs plot of Figure 2. Surprisingly x_6 and x_{11} appear disconnected from the other variables.

The contrast between the marginal and conditional DWIGs is striking and the conditional measures displayed in Figure 3B using expression (5) of the Supplement tell a very different story. The levels of the conditional MIs are lower than the corresponding marginals and

some edges have almost disappeared. The strong relationship $x_{10} - x_{12}$ persists. The relationships $x_5 - x_8$, $x_5 - x_{10}$, and $x_5 - x_{12}$ are substantially moderated by conditioning, and that between (x_{10}, x_{12}) and x_8 has almost disappeared. One can conclude that $(x_7, x_8) \perp (x_{10}, x_{12}) \mid (x_5, x_9)$. A relatively strong relationship is now seen between $x_5 - x_9$.

The relationships $x_{11} - x_{12}$ and $x_{10} - x_{11}$ now appear sizeable, which is due to using a different setting for the max displayed in the marginal and conditional DWIGs, and realising that conditioning has not particularly decreased the magnitude of these dependences. The dependence between x_6 and the other variables is roughly of the same magnitude as in the marginal DWIG.

The strongest edge displayed in the DWIG algorithm is between x_{10} and x_{12}. The link between the concentration of IRS-1-YP and the concentration of IRS-1-YP:PI3K complex perfectly illustrates an equilibrium that is expected from the biological understanding of the system; once the IRS-1 receptor is tyrosine phosphorylated (IRS-1-YP), it forms a complex with PI3K [28].

Insulin: a fitted GGM: The conditional MIs displayed in Figure 3B are intimately related to the independence graph of the variables. When any one of these is 0 the corresponding conditional independence statement holds. When a graph is defined by the set of those conditional MIs that are 0 it has the Markov separation properties of an independence graph [3,4]. Consequently the relative invisibility of the displayed DWIG does suggest a specific graph, which can be formally defined by thresholding the MI. Assuming that the data is well represented by a joint multivariate normal distribution the graphical Gaussian model (GGM) of the thresholded graph may be evaluated. The fitting procedure has been briefly described in the Methods section.

A comparison of these two DWIGs is displayed in Figure 4.

The numerical values for the information explained from the identity 'total=residual+explained' at expression (13) of the Supplement are 10315 = 299 + 10016 mbits, computed from the sample and fitted variance matrices, show that the vast majority of dependence is accounted for by this model with 12 edges.

For this data the thresholded conditional DWIG is very close to the DWIG computed from the fitted variance matrix, seen in the two panels of Figure 4. This is not always the case and (i) estimating the conditional information empirically for a large conditioning set can be unreliable, and (ii) model search requires a better procedure than thresholding in order to find a well fitting model.

Insulin: comparison with other four methods: Displaying a DWIG is mainly proposed as an exploratory tool for descriptive multivariate analysis, however it does of course suggest some graphical models of interest, and here we make a limited comparison with some widely used reverse engineering methods, ARACNE [13], Banjo [29], CLR [30], MRNET [31]. The performance of these methods is based on our simulation data where the ground truth is known.

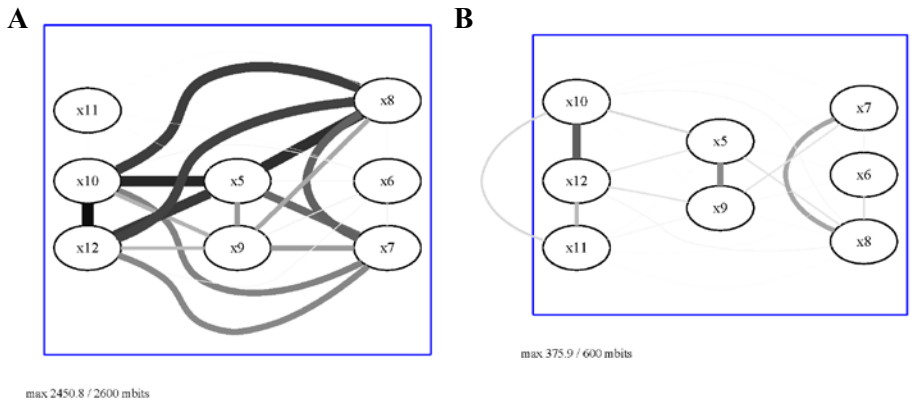

Figure 3: Contrasting DWIGs using marginal and conditional MI for the insulin data. (A) Marginal DWIG, with maximum divergence set to 2600mbits. (B) conditional DWIG with maximum set to 600mbits. The edge width and gray scale are linearly related to the divergence, the millibit measure of edge strength. The maximum width and tone are user-defined. The unit used in the display is millibits. There are 28 edges in both graphs corresponding to all pairs of the 8 variables, but their visibility depends on their strength relative to the set maximum. The actual conditional MIs measured in millibits used to calculate the weights for B are reported in adjacency Table of DWIG in the supplementary data.

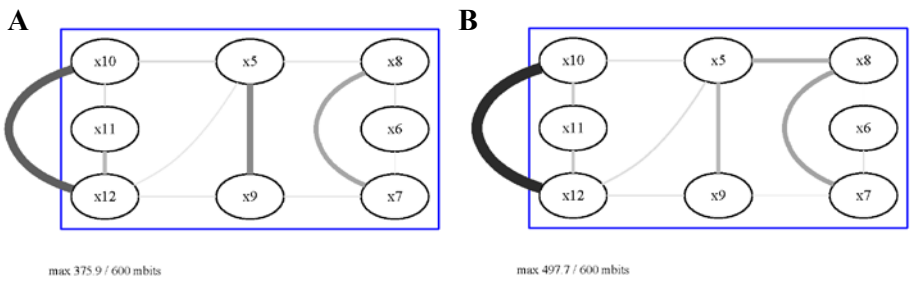

Figure 4: The thresholded sample DWIG and the DWIG of the fitted model for the insulin data. (A) The thresholded sample DWIG. (B) The conditional DWIG of the fitted model defined by thresholding. Both graphs display conditional MIs. The graphs are very similar, with just the weight on the edges relating $x_{10} - x_{12}$, $x_5 - x_9$ and $x_5 - x_{12}$ differing.

The weighted adjacency matrices given by DWIG, ARACNE, Banjo, CLR, and MRNET, are in the supplementary material. The ROC diagram was built from these matrices and is displayed in Figure 5. The area under curve (AUC) from DWIG, ARACNE, Banjo, CLR, and MRNET, are 0.9, 0.64, 0.71, 0.76, and 0.67 respectively, with the thresholded DWIG outperforming the other methods. However, the most attractive feature of DWIG for us is to display properties of the fitted model or of the empirical data directly rather than selecting the best model.

A/J mouse study

The methodology was also tested in the context of smoke-induced perturbations in A/J mice; cigarette smoke is a known cause for emphysematic changes in rodent lung [32]. A/J mice exposed to cigarette smoke for different durations daily, 5 days per week, showed a pronounced lung inflammation at the endpoints after 3 and 5 months of exposure. Successful measurements for 25 analytes in the bronchoalveolar lavage fluid (BALF) were performed for MAP. BALF samples contain cells and proteins which are involved in many processes that are critical for host defense in both mice and humans [33]. These 25 cytokines/chemokines measured successfully in BALF are shown in Table 2. While this is experimental data, our data mining analysis of the outcome variables is conducted in the belief that their interactions would flag some of the underlying biological processes.

There were 114 mice used in the study, balanced between the four treatment levels and the two time points, as shown in Table 3. We could not analyse protein covariation in separate dose levels and exposure lengths because of the small size of the groups and so took residuals from fitting a linear additive model (on the log scale) adjusting for treatment and endpoint. This provided a combined data set of dimension (114, 25) for initial analysis using graphical models and DWIGs. Finally for a comparison of the exposed mice (treatment levels 1, 2, 3) with the controls we used the corresponding data sets of dimension (83, 25) and (31, 25).

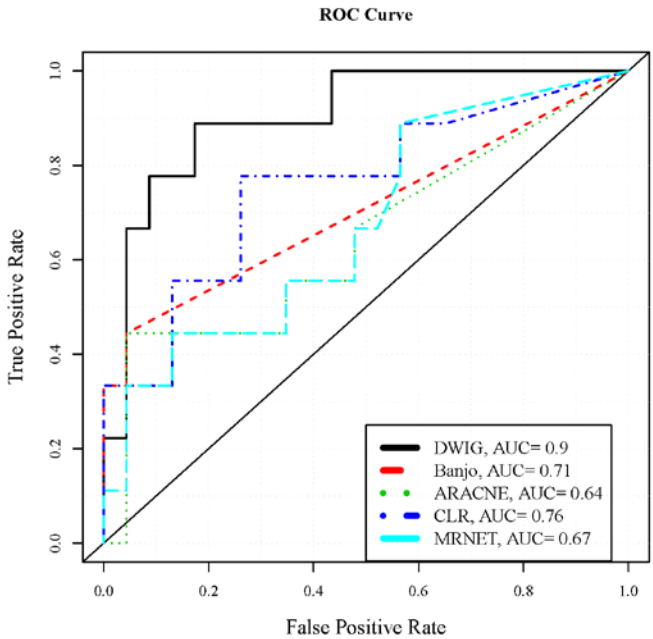

Figure 5: ROC curves of DWIG, ARACNE, Banjo, CLR, and MRNET. The ROC curves were built according to the adjacency matrix constructed from the simulated insulin dataset with standard deviation σ =0.15.

Abbreviation used	Protein name
CD40	CD40
CD40L	CD40 ligand
Eotaxin	Eotaxin
FGF9	Fibroblast growth factor 9
GCP2	Granulocyte chemotactic peptide-2
Haptogl	Haptoglobin
IgA	Immunoglobulin A
IL10	Interleukin-10
IL1b	Interleukin-1 beta
Insulin	Insulin
IP10	Inducible Protein-10
Leptin	Leptin
LIF	Leukemia inhibitory factor
MCP1	Monocyte chemotactic protein-1
MCP3	Monocyte chemotactic protein-3
MCSF	Macrophage colony stimulating factor
MDC	Macrophage-derived chemoattractant
MIP1b	Macrophage inflammatory protein -1 beta
MIP1g	Macrophage inflammatory protein -1 gamma
MIP2	Macrophage inflammatory protein -2
MMP9	Matrix metalloproteinase-9
Myogl	Myoglobin
TF	Tissue factor
VCAM1	Vascular cell adhesion molecule-1
VEGF	Vascular endothelial growth factor

Table 2: Symbols and names for the 25 cytokines/chemokines analyses in the BALF of A/J mice dataset.

Number of animals	3 months	5 months
0 ug TPM/(l*day)	15	16
1500 ug TPM/(l*day)	14	12
2250 ug TPM/(l*day)	13	15
3000 ug TPM/(l*day)	14	15

Table 3: Experimental design for the A/J mice dataset. TPM: Total particulate Matter.

A/J mouse: most informative variables: While 25 variables is not a large number by current standards there are still some 300 pairwise interactions to inspect. The 9 most informative variables were identified by using expression (7) in the Supplement. We chose to look at the DWIG for the subset of the 9 most informative variables from the combined dataset as a starting point, see Figure 6.

This DWIG is thresholded, so that all edges with MIs less than 57.7mbits (the 80% quantile) are excluded, to clarify the picture. It is seen that the strongest interaction, after conditioning on the other 23 variables, is between MCP1 and MCP3. The magnitude of this MI is about 700mbits, which corresponds to a (partial) correlation of 0.8, and is much the highest; the others are in the range 57-150mbits. Interestingly the graph shows that it is MCP3 and not MCP1 that is associated with the other variables.

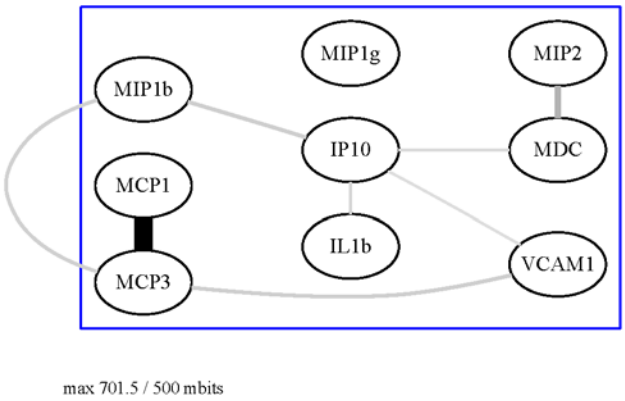

Figure 6: DWIG of the 9 most informative variables from the combined data-set, adjusted for the other 16 variables.

with the best fitting graphical model. The residual deviance and degrees of freedom (674.6, 264) are not consistent with chance, suggesting this is not the 'true' graph. However it does explain more of the total information against independence, and in that sense provides the best parsimonious approximation of the data. We point out that edge set suggested by each of these routines differs.

A/J mouse: DWIG from the fitted stabilised network: The graph which we wish to interpret biologically is the DWIG for the treated group. Our concern with the small sample sizes prompted us to employ stabilised versions of the search routine, where the data is perturbed by a random selection of half the units. The tuning parameters were the same for each random selection, and the edges selected into the graph are the ones discovered in more than 50% of the repetitions. The resulting stabilised graph was then fitted to the combined data to produce the corresponding row of Table 4 and the appropriate conditional MI weights for the DWIG. The fit of the stabilised graph

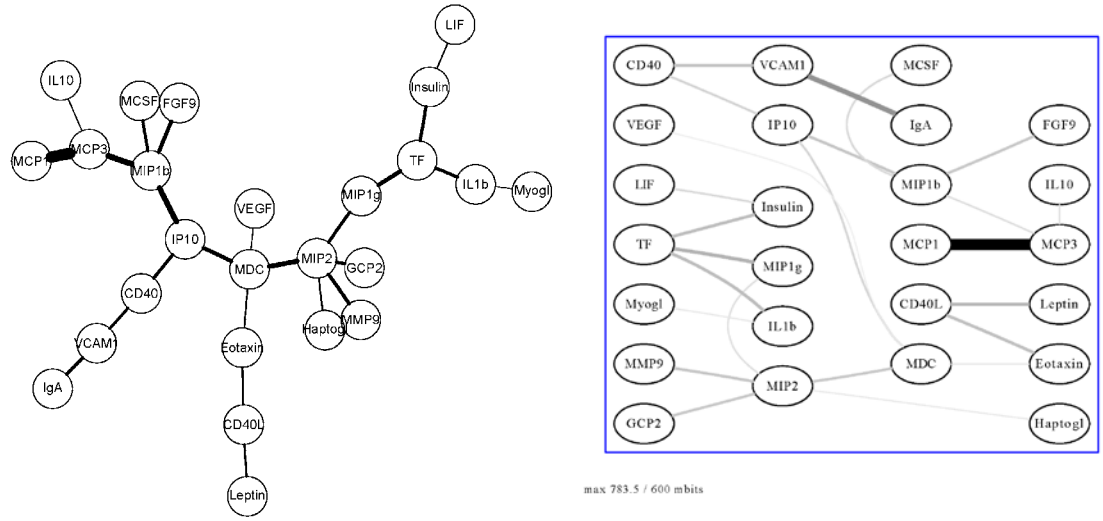

Figure 7: Two layouts of the DWIG of the CL-tree fitted to the combined dataset.

A/J mouse: DWIG from the CL-tree: The divergence weighted CL-tree gave an immediate overview of all variables together. Two displays of the graph of the tree are given in Figure 7: the left hand panel makes the tree structure prominent; the right hand panel has variables located in the same vertical ranks as the graph of the chosen model displayed later in Figure 8.

The quality of the fit of the tree and the amount of explained information is contained in Table 4.

A/J mouse: search: The tree constraint visibly includes some rather weak edges, and perhaps, invisibly, excludes some stronger edges erroneously. To get a better representation of the data we ran several search routines including CL-tree [22,23], ARACNE [13], MRnet [31], Glasso [15], PCalg [16] to search over the space of undirected graphical models. We set a common threshold of detecting 35 edges in the output adjacency matrix. Table 4 summarises the results of this search by the number of edges, the decomposition of the total information into explained and unexplained information, and the residual deviance and degrees of freedom.

From this Table it is seen that the ARACNE search routine came up

Procedure	edges	expl.(mbits)	unexpl.(mbits)	deviance	df
Indep	0	0	14456	2231	300
Tree	24	8720.4	5735.3	885.1	276
ARACNE	36	10084.5	4371.2	674.6	264
ARACNE Stab.	24	8724.2	5731.5	884.6	276
Mrnet	35	9084	5372	829	265
Glasso	35	8654.3	5801.4	895.3	265
PCalg	35	9792.1	4663.6	719.8	265

Table 4: A/J mouse: information decomposition.

is not as good as the original ARACNE search, but has substantially fewer edges and is protected against capitalisation from fortuitous data configurations. This exercise was repeated separately for the treated and control groups, and because of smaller sample size, two search routines, ARACNE and Glasso, were used in tandem by including only edges detected by both. The DWIGs for the two groups based on fitting the selected model are given in Figure 8.

A/J mouse: network biology: As expected, while some of the edges were in common between the control and smoke treated DWIG

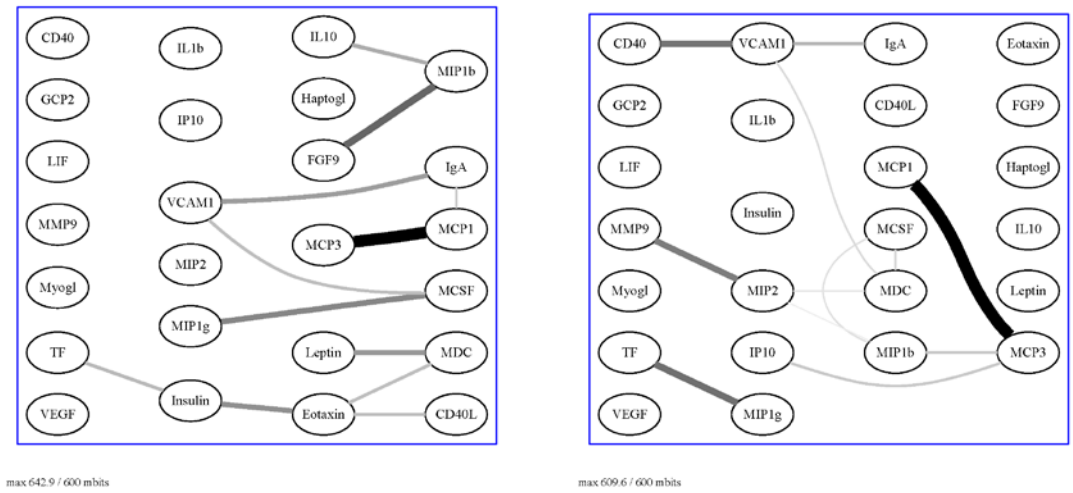

Figure 8: DWIGs of the fitted graphical models obtained by using the common edges from the stabilised versions of an ARACNE and a Glasso search. Left panel:control group; Right panel: treatment group.

profiles (VCAM1-IgA and MCP3-MCP1), several edges disappeared and new edges were formed following the smoke exposure. To the best of our knowledge, these edges signify putative novel interactions occurring in the lungs of smoke-exposed animals that develop COPD-like pathology. Some of the molecules have been previously shown causally linked in the lung context; there are reports demonstrating relationship between MMP9 and MIP2 [34-36] as well as CD40 and VCAM1 [37,38], two important edges in the DWIG network. Even though these relationships were found in diverse experimental setups and in the context of lung challenges unrelated to smoking, they may denote common mechanisms involved in lung host defense and signaling. While the biological relevance of the DWIG network needs to be experimentally verified, the algorithm presented in this paper provides a tool to discover potentially interesting relationships between biological variables.

Discussion

One objective for the multivariate exploratory analysis of observed random variables is to summarise their interactions in the original coordinate system. Using the mathematical framework of conditional independence, mutual information and graphical modelling, we suggest that the DWIG, which is a tool to make informative visual displays for chosen subsets of variables, goes some way to meet this challenge.

This paper makes several contributions. Firstly, to measure interaction using conditional MI, because conditioning allows scientific statements of unique attribution to be made from observational studies and use a common currency of millibits to allow comparison across studies. Secondly, to use graphical Gaussian modelling to reduce the size of the effective conditioning set to better estimate the conditional MIs; this assumption also provides an assessment of fit and quantifies the information explained by dependence in the information decomposition. Thirdly, to employ stability selection as a tool in graphical model search.

Currently the DWIG is used for biological expression data with fewer than 100 variables, e.g. protein expression data set, as estimation of the covariance matrix is problematic when the number of variables exceeds the number of observations. However, biological array experiments that generate many observed variables but with few observations on each, e.g. gene expression data set, could be analysed by DWIG with restricting the dimension of the conditioning set a priori using the qp-graph methodology.

Among several ways to extend the applicability of DWIGs we mention one here. The underlying rationale for a DWIG is predicated on the assumption of Gaussian data where associations can be measured by the ordinary Pearson correlation coefficient. Real data is more complex and effort is needed to ensure that Gaussianity approximately holds in practise: for instance, non-linearities in relationships may be reduced by transformation, and missing data on some variables may need imputation. Different continuous distributions require bespoke treatment, sometimes even to the exclusion of certain variables. Incorporating robust or shrunken estimates of the covariance [39,40] would make the methodology less sensitive to non-Gaussian data.

Conclusions

A novel display, the DWIG for exploratory analysis of biological expression data, is established. Using simulated data from the well known insulin pathway, we show that the thresholded sample DWIG (thresholded empirical conditional MI weights) performs better than, or at least as well as, other network searching algorithms such as CLR, Banjo, MRNET, and ARACNE. DWIG was applied to a MAP data set obtained from BALF samples of female A/J mice exposed to cigarette mainstream smoke for 3 and 5 months. An association network was built. The DWIG is a powerful tool for the display of a biologically meaningful network estimated from biological expression data set with small or medium number of variables.

Supplementary Data

The R-code for producing DWIGs is currently available at http://www.maths.lancs.ac.uk/%7Ewhittake.

Acknowledgements

This work would not be possible without the open source software R project [21], its extensive package repository CRAN and the interface to the open source utility Graphviz (25) invaluable for graph layout.

References

1. Wright S (1923) The Theory of Path Coefficients a Reply to Niles's Criticism. Genetics 8: 239-255.
2. Darroch JN, Lauritzen SL, Speed TP (1980) Markov fields and log-linear interaction models for contingency tables. Ann Statist 8: 522-539.
3. Lauritzen SL (1996) Graphical Models.
4. Whittaker J (1990) Graphical Models in Applied Multivariate Statistics.
5. Koller D, Friedman N (2009) Probabilistic Graphical Models: Principles and Techniques The MIT Press, Cambridge, Massachusetts.
6. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 97: 12182-12186.
7. de la Fuente A, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20: 3565-3574.
8. Dobra A, Hans C, Nevins JR, Yao G, West M (2004) Sparse graphical models for exploring gene expression data. J Multivar Anal 90: 196-212.
9. Ma S, Gong Q, Bohnert HJ (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Res 17: 1614-1625.
10. Magwene PM, Kim J (2004) Estimating genomic coexpression networks using first-order conditional independence. Genome Biol 5: R100.
11. Toh H, Horimoto K (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 18: 287-297.
12. Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18 Suppl 2: S231-S240.
13. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMCBioinformatics 7 Suppl 1: S7.
14. Reverter A, Chan EK (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics 24: 2491-2497.
15. Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432-441.
16. Kalisch M, Buhlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J Mach Learn Res 8: 613-636.
17. Kramer N, Schafer J, Boulesteix AL (2009) Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics 10: 384.
18. Hastie T, Tibshirani R, Friedman J (2004) Elements of Statistical Learning Springer, New York.
19. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Statist Society B 58: 267-288.
20. Meinshausen N, Buhlmann P (2010) Stability selection. J Royal Statist Society B 72: 417-473.
21. Team RDC (2011) R: A Language and Environment for Statistical Computing; Vienna, Austria.
22. Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information. Theory 14: 462-467.
23. Edwards D, de Abreu GC, Labouriau R (2010) Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinformatics 11: 18.
24. Meyer PE, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9: 461.
25. Ellson J, Gansner ER, Koutsofios E, North S, Woodhull G (2004) Graphviz and dynagraph - static and dynamic graph drawing tools. Springer-Verlag, Berlin: 127-148.
26. Barbieri M, Berger J (2004) Optimal predictive model selection. The Annals of Statistics 32: 870-897.
27. Sedaghat AR, Sherman A, Quon MJ (2002) A mathematical model of metabolic insulin signaling pathways. Am J Physiol Endocrinol Metab 283: E1084-E1101.
28. Backer JM, Myers MG, Jr., Shoelson SE, Chin DJ, Sun XJ, et al. (1992) Phosphatidylinositol 3'-kinase is activated by association with IRS-1 during insulin stimulation. EMBO J 11: 3469-3479.
29. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20: 3594-3603.
30. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8.
31. Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol: 79879.
32. Braber S, Henricks PA, Nijkamp FP, Kraneveld AD, Folkerts G (2010) Inflammatory changes in the airways of mice caused by cigarette smoke exposure are only partially reversed after smoking cessation. Respir Res 11: 99.
33. Gharib SA, Nguyen E, Altemeier WA, Shaffer SA, Doneanu CE, et al. (2010) Of mice and men: comparative proteomics of bronchoalveolar fluid. Eur Respir J 35: 1388-1395.
34. Albaiceta GM, Gutierrez-Fernandez A, Parra D, Astudillo A, Garcia-Prieto E, et al. (2008) Lack of matrix metalloproteinase-9 worsens ventilator-induced lung injury. Am J Physiol Lung Cell Mol Physiol 294: L535-L543.
35. Lanone S, Zheng T, Zhu Z, Liu W, Lee CG, et al. (2002) Overlapping and enzyme-specific contributions of matrix metalloproteinases-9 and -12 in IL-13-induced inflammation and remodeling. J Clin Invest 110: 463-474.
36. Yoon HK, Cho HY, Kleeberger SR (2007) Protective role of matrix metalloproteinase-9 in ozone-induced airway inflammation. Environ Health Perspect 115: 1557-1563.
37. Lei XF, Ohkawara Y, Stampfli MR, Mastruzzo C, Marr RA, et al. (1998) Disruption of antigen-induced inflammatory responses in CD40 ligand knockout mice. J Clin Invest 101: 1342-1353.
38. Propst SM, Denson R, Rothstein E, Estell K, Schwiebert LM (2000) Proinflammatory and Th2-derived cytokines modulate CD40-mediated expression of inflammatory mediators in airway epithelia: implications for the role of epithelial CD40 in airway inflammation. J Immunol 165: 2214-2221.
39. Schafer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754-764.
40. Yuan M (2007) Model selection and estimation in the Gaussian graphical model. Biometrika 94: 19-35.

This article was originally published in a special issue, **Bioinformatics** handled by Editor(s). Dr. Yixuan Wang, Albany State University, USA