Advances in Complex Systems

EDGE WEIGHTING OF GENE EXPRESSION GRAPHS

GRAINNE KERR, DIMITRI PERRIN*, HEATHER J. RUSKIN AND MARTIN CRANE
Centre for Scientific Computing & Complex Systems Modelling
Dublin City University
Dublin, Ireland
{*gkerr, dperrin, hruskin, mcrane*}*@computing.dcu.ie*

Received November 2008
Revised July 2009

In recent years, considerable research efforts have been directed to microarray technologies and their role in providing simultaneous information on expression profiles for thousands of genes. These data, when subjected to clustering and classification procedures, can assist in identifying patterns and providing insight on biological processes. To understand the properties of complex gene expression datasets, graphical representations can be used. Intuitively, the data can be represented in terms of a bipartite graph, with weighted edges corresponding to gene-sample node couples in the dataset. Biologically meaningful subgraphs can be sought, but performance can be influenced *both* by the search algorithm, and, by the graph weighting scheme and both merit rigorous investigation. In this paper we focus on edge-weighting schemes for bipartite graphical representation of gene expression. Two novel methods are presented: the first is based on empirical evidence; the second on a geometric distribution. The schemes are compared for several real datasets, assessing efficiency of performance based on four essential properties: *robustness to noise and missing values*, *discrimination*, *parameter influence on scheme efficiency* and *reusability*. Recommendations and limitations are briefly discussed.

Keywords: Edge Weighting; Weighted Graphs; Gene Expression; Bi-clustering.

1. Introduction

Clustering techniques are fundamental for the exploration of gene expression data, [1–4]. Gene profiles are grouped into K subsets (clusters), with K not necessarily known *a priori*, such that elements in the same subset are associated with one another. The fundamental biological premise underlying these approaches is that genes, which display similar expression patterns, are *co-regulated* and may share a common function or contribute to a common pathway. Identification of patterns is made more difficult by the fact that association (through similarity measures or adherence to some co-regulation model) among a *subset of genes* may be determined by a *subset of samples* giving rise to *biclusters*, [5]. Moreover, genes may belong

*Corresponding author

to a number of biclusters, with varying degrees of membership, [6]. This, and the fact that gene expression profiles often originate from very noisy experimental measurements, makes computational solutions to the clustering problem difficult and patterns difficult to interpret.

Graph theoretical modelling is proving a useful tool in the analysis of large complex biological datasets. For instance, protein-protein interactions can be modelled by *undirected* graphs, [7], where nodes represent proteins and an edge connects two nodes if the proteins *physically combine*. Transcriptional factor binding sites can be identified through the use of *undirected weighted* graphs, where weights of edges capture the similarity between aligned nucleotides in an input set of promoters, [8]. Further, metabolic networks can be represented as *bipartite* graphs. In this case an edge connects a reaction node to a compound node, representing either substrate or product relationships, [9].

Gene expression can also be modelled using graphical techniques, where the nodes represent genes and edges represent similarities in gene expression. In the pioneering work of Sharan et al. [10] and Hartuv et al. [11], a probabilistic model for edge weighting in gene expression was introduced. Here, the probabilistic weight of an edge is derived from the similarity measure between the two gene vectors, and reflects the probability that these two genes are 'mates' (i.e. belong in the same cluster). Building on this work, Tanay et al. [4, 12], used a *weighted bipartite* graph to model gene expression, where two node types are used to represent genes and experimental properties and in this scheme, the weight of an edge, e_{ij} was designed to incorporate the *probabilities* of a gene i having experimental property j. These are powerful techniques, in that the sum of edge weights in a cluster denote its statistical significance. Alternatively, weighting schemes for one-mode gene expression graphs have been put forward by Carlon et al. [13] and Zhange et al. [14]. In this work, edge weights were created using use an adjacency function. This function is based on a similarity measure and assumes that connections between nodes approximate a scale-free topology. Analysis of complex weighted networks was also considered in [15], (not specifically gene-expression data) although the authors did not investigate the weighting scheme itself.

Graphical clustering techniques involve identification of similar, (e.g densely connected), subgraphs, where usual limitations (such as NP-completeness) apply. A variety of evolutionary and heuristic algorithms can be used to find subgraphs and include Genetic algorithms, [16], Simulated Annealing, [17], the MinCut algorithm, [18], CLICK, [10] and Samba, [4]. These algorithms all rely on an objective function to evaluate the sub-graphs found. Fundamentally, this function relies on the concept of an edge weight and its derivation.

Substantial efforts have been devoted to the development of weighted graph-based clustering methods for biological tasks, (see references above). However, it is important to recognise *the essential role a weighting scheme itself plays on cluster determination*, and thus to investigate the intrinsic nature of these schemes in iso-

lation. To this end, we introduce and compare two new weighting schemes: (i) a distribution-based method, (where edge weights are estimated from a pre-defined probability distribution); (ii) an empirical-based method, (where edge weights are determined by the experimental data). We also discuss performance measures for these (and the Tanay) edge weighting schemes, and present a comparative validation based on application to several real datasets.

2. Approach and Evaluation Framework

2.1. *Gene Expression as a Bipartite Graph*

Gene expression data is typically presented as a matrix, X, where rows ($i = 1 \dots n$) correspond to gene vectors, $g_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, which record expression values for gene i across p experiments. A *bipartite* graph is a triplet $G = (\top, \bot, E)$, where $\top$ is the set of *top* nodes, $\bot$ the set of *bottom* nodes, and $E \subseteq \top \times \bot$ is the set of edges, which links top and bottom nodes. $\top \cap \bot = \emptyset$, is a defining property of bipartite graphs. We define a *weighted* bipartite graph to model gene expression under a number of experiments as, $G = (\top, \bot, E, W)$, where $\top$ corresponds to the set of genes, $|\top| = n$, $\bot$ to the set of samples, $|\bot| = p$, E is the set of edges between genes and samples, and $W = (w_{ij})$ where $w_{ij} \in \mathbb{R}$ denotes the weight of the edge e_{ij} between 'gene' node i and 'sample' node j, (a sample refers to a microarray experiment carried out on mRNA extracted from biological samples; each microarray experiment is a column in the gene expression dataset). If every gene sample couple is connected by an edge, (i.e. $|E| = p \times n$), then the gene expression network is said to be *fully connected*, otherwise it is *partially connected*.

As mentioned previously, essential properties of the dataset can be transformed to bipartite graph properties by appropriate representation. However, care must be taken not to lose important information in the transformation process. This allows us to study the gene-sample inter-activity using the powerful tools and notions provided for classical networks. Here, we are concerned with what constitutes an (important) edge in the context of gene expression and how this edge is weighted.

2.2. *Definition of Assessment Properties*

It is important, both for the evaluation of any clustering technique and for its *reusability*, that weighting schemes are validated independently of the subsequent network analysis. A clustering technique should not be considered as a "black box", as each technique will find optimal clusters based on internal search criteria and optimisation functions, both of which may differ considerably between approaches. Assessment of a clustering technique's core features is crucial to improve overall performance, and should be chosen to reflect the properties of the dataset to be analysed. For a given graph-clustering technique, reliable results reflect a well-designed weighting scheme, together with a reasonably robust and efficient search algorithm, and both aspects are susceptible to refinement. Consequently, we pro-

pose a graphical edge weighting assessment procedure for gene-sample networks, based upon four properties, as detailed below.

2.2.1. *Discrimination:*

Ability of the method to "rate" highly those gene-sample couples which contribute to a cluster. The range and distribution of edge weights establish how well a given scheme distinguishes between relevant and irrelevant gene-sample couples.

2.2.2. *Reusability:*

Independence of the proposed scheme and the subsequent clustering technique. This considers which specific properties of the proposed weighting scheme must be accommodated by the subsequent clustering technique.

2.2.3. *Robustness:*

Ability of a given weighting scheme to deal with noise and missing values. This involves investigation of distortion of weights caused by different levels of noise and missing values.

2.2.4. *Parameter Influence:*

Any weighting scheme ideally requires minimal specification of input parameters. We thus examined input parameter influence on discrimination and robustness, as well as on the distribution of weights themselves.

2.3. *Datasets*

The three weighting schemes were tested on *three* datasets. (i) The *Yeast Cell Cycle Data*, provided by [19], contains time-course expression profiles for more than 6000 genes, with 17 time points for each gene (taken at 10-min intervals) and covers nearly two yeast cell cycles (160 min). The raw gene expression profiles were downloaded from `http://genomics.stanford.edu`. (ii) A *Lymphoma Dataset* (downloaded from supplementary web by [1], `http://llmpp.nih.gov/lymphoma/index.shtml`), relates to an experiment to characterize gene expression in Diffuse Large B-cell lymphoma (DLBCL). The complete dataset contains expression levels for 4,026 genes and 96 samples. Finally, (iii) Gefitinib Treated *Kasumi Cell Line dataset*. Here Kasumi cells were treated with gefitinib or dimethyl sulfoxide (DMSO) control in duplicate, for 6 hours and, in triplicate for 24 hours. This results in a dataset of 22283 genes and 10 samples [20], and is available from the MIT Broad Institute website, (`http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi`).

3. Edge Weighting Schemes

3.1. *Tanay Scheme*

This method is incorporated into the SAMBA clustering algorithm [4] and is available with the EXPANDER software suite [21], (which also offers a number of other clustering options and is available from: `http://www.cs.tau.ac.il/~rshamir/expander/expander.html`). As this software does not need to provide output on the weight of each gene-sample couple in the dataset, completion of our analysis required us to recode this scheme to allow access to the weight information during the algorithmic process. The information for this scheme was taken from [4] and [22].

Note that the bipartite representation of the data in the SAMBA algorithm, which uses this weighting scheme, has node sets $\{genes\}$ and $\{properties\}$. The *property* set includes[a] nodes for properties of strong, moderate and weak induction/repression for each particular sample j. Thus $|\{properties\}| = 6 \times p$, where $p =$ number of samples. A gene i is defined to be: weakly induced in sample vector j if the expression value x_{ij} is ranked $\geq x_{\text{weak Induced}}$ in j; moderately induced if it is ranked $\geq x_{\text{moderate Induced}}$ in j; and strongly induced if it is ranked $\geq x_{\text{strong Induced}}$ in j. Similarly it is: weakly repressed if it is ranked $\leq y_{\text{weak repressed}}$ in j; moderately repressed if it is ranked $\leq y_{\text{moderate repressed}}$ in j; and strongly repressed if it is ranked $\leq y_{\text{strong repressed}}$ in j. These thresholds are arbitrarily chosen and are fixed for all samples. For this analysis we used thresholds of $0.97, 0.90, 0.87, 0.03, 0.10, 0.13$, respectively for each of the values above.

Briefly, with this method let $\phi(i, j)$ be the probability that gene i has property j (i.e. weakly/moderately/strongly induced/repressed in sample j. These 'categories' are not mutually exclusive, i.e. if a gene i is weakly induced with a high probability, it may also be strongly induced, albeit with a smaller probability.), see [4, 22] for more details on this calculation. The majority of gene sample couples will have $\phi = 0$. This probability is assigned as the weight of a given edge, e_{ij}, scaled with the log-likelihood of obtaining that edge by chance. Thus each gene-sample couple in the dataset has weight Eq.1:

$$w_{ij} = \phi(i, j) \times log\frac{P_c}{P_{i,j}} + (1 - \phi(i, j)) \times log\frac{1 - P_c}{1 - P_{i,j}} \tag{1}$$

The probability $P_{i,j}$ is the fraction of random bipartite graphs, with degree sequence identical to G, that contain e_{ij} (and can be estimated using Monte-Carlo methods). P_c is based on the assumption that an interesting edge occurs with a constant probability $> max_{(i,j)\in I\times J} P_{i,j}$. For this work a P_c value of 0.9 was used [4].

[a]This algorithm was designed to be applied to a compendium of data-sources, reflecting a set of "properties", not just gene expression data. For this analysis we are restricting it to gene expression data

3.2. *Empirical-based Weighting*

The empirical-based weighting scheme is data-driven in the sense that it is determined from direct analysis of the dataset, i.e. based on observations obtained. Fundamental to the method proposed here is that genes which, for a given sample, have either high or low expression (equivalent to induction or repression) are more likely to contribute to a function, or have a functional response, than for those for which expression values remain unaffected. Affected genes can thus be extracted for further analysis.

A high/low expression value for gene, i under sample j, is determined, relative to other expression values in gene vector i, i.e. across rather than within samples. The motivation for this is that, with microarray technology, direct comparison of expression measures within arrays is problematic, because fluorescent intensities are not the same across genes. The measured intensities are roughly proportional to mRNA abundance but the proportionality factor is different for each gene. Specifically, this means that between-sample, within-gene comparisons are appropriate, but within-sample, between-gene comparisons are not straightforward [If, for example, in sample j_1, genes a and b, have measured expression of 100 and 200 respectively. These observed data do not reflect the real relative abundance of mRNA for these two genes - there could be more mRNA for gene a. If in a second sample, j_2, gene a has an expression measure of 200, we could conclude that the abundance of mRNA for a in sample j_2 is likely to be higher than that observed in sample j_1, [23].]

Under this weighting scheme, an edge e_{ij} exists when the i^{th} gene shows "significant" induction or repression, relative to its mean level of expression, for sample j. It makes use of Chebyshev's inequality [24], as no distributional form of expression values for each gene is assumed. Chebyshev's inequality, for any real number $\kappa > 0$, can be written:

$$Pr(|X - \mu| \geq \kappa\sigma) \leq \frac{1}{\kappa^2} \tag{2}$$

with random variable X, μ the expected value of X and σ^2 the variance.

This scheme uses a two step process (Figure 1). The first step involves identification of those expression values $X = x_{ij}$, $(i = 1 \ldots n,\ j = 1 \ldots p)$, of interest which, for a given sample j, are $\geq \kappa\sigma$ from the mean expression of gene vector i. From Eq. 2, for example, the associated probability of an expression value $\geq 3.16\sigma$ from the mean of gene i is less than 0.10. Expression values $\geq 3.16\sigma$ from the mean would indicate a strong response of gene i to sample j. Note: this method assumes that for the majority of samples, the expression of gene i will not be affected. If, on the other hand, the aim of the experiment is to elicit an gene expression response across the majority of samples, a baseline sample point could be used as an assumed mean (e.g. use sample-point in the absence of stimulus).

Clearly, categories can be established to highlight those expression values which

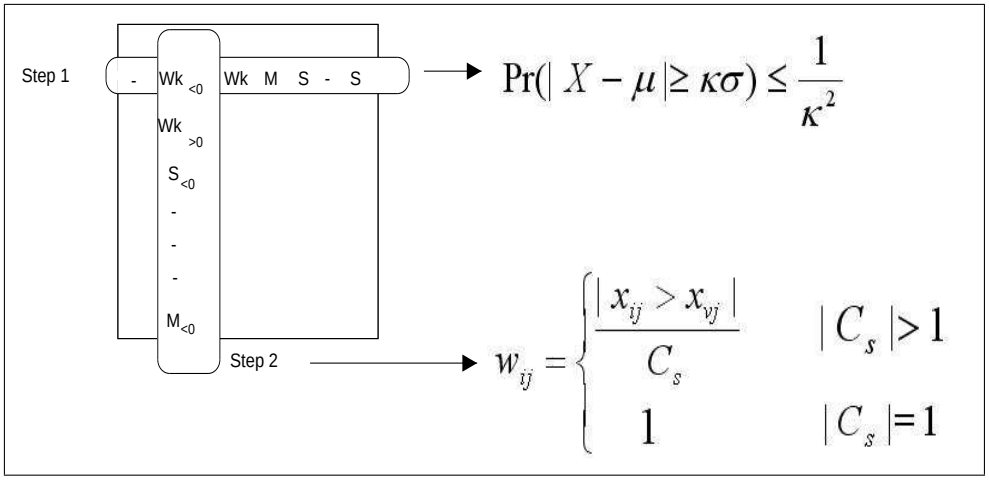

Fig. 1. Two-step process of empirical scheme. Step 1: A univariate analysis of each gene vector is carried out to determine strength of response. Step 2: A univariate analysis of each sample vector is performed and used to order gene response.

indicate a *weak response*, *moderate response* and *strong response*, where κ indicates the threshold between categories. For example, expression values which are $\geq 2.58\sigma$, $\geq 3.16\sigma$ and $\geq 4.47\sigma$ from μ fall into *non-overlapping* categories of weak, moderate and strong respectively, (i.e. a gene-sample couple will not be categorised as moderately *and* strongly responsive), corresponding to probabilities = 0.15, 0.10 and 0.05. In practice, these probabilities (reflecting the thresholds between categories) are chosen by a careful threshold estimation procedure (see Section 3.2.1), and may change based on the dataset under consideration. Obviously the number of categories can be extended for fine-grained response, while thresholds between categories can also be adjusted. For this analysis, the categories weak, moderate and strong, as defined above and in [4], were used. Thresholds chosen are discussed further in Section 3.2.1.

In the second step, a univariate analysis of expression values within each sample vector is carried out and used to order gene response. For each sample vector, j, expression values x_{ij} which indicate strong response of the i^{th} gene under j, (as determined in step one) are selected. Similarly, genes which show moderate and weak response under j can be identified. For each of the three "strength of response" categories, a gene may be repressed or induced; ($X-\mu < 0$, or $X-\mu > 0$ respectively), giving six sub-categories in total. For each sub-category, C_s, $s = 1 \ldots 6$ and for each sample variable, $j = 1 \ldots p$, the empirical probability of $x_{ij} \in C_s$ is calculated as $|x_{ij} \geq x_{vj}|/|C_s|$, $x_{vj} \in C_s$, $i \neq v$, (probability = 1 if $|C_s| = 1$), and hence the edge weight, in the bipartite gene expression graph is obtained.

The weight is thus a direct reflection of obtaining a given expression level *in an induced/repressed response category for a particular sample*. So, if many genes react

strongly in the sample, the weight is smaller, while if only a few react strongly, the weight will be larger. Note that, with this weighting scheme, a given sample (experiment) may also have no reacting genes.

The weighted graph can be broken down into an independent subgraph for each sub-category. Gene-sample edges in these different response groups may well have similar weight "values" but are distinguished, in terms of absolute levels of expression, by the category into which they fall.The reason for this approach is that if the dataset was not categorised, a weak response and a strong response gene-sample couple would have very different weights, with the consequence that the strong response couple would dominate the analysis and obscure more subtle patterns. Hence, the analysis does not depend directly on levels of expression but rather on strength of response. Within a category, the weight can nevertheless be interpreted directly in terms of the relative probability of gene-sample response. Thus the *higher the weight*, the more confidence that a relationship exists between gene and sample *in that category*.

3.2.1. *Threshold Estimation*

This scheme requires specification of number of categories and threshold values for each. To decide on thresholds between categories, graphs from real datasets, $G = (\top, \bot, E)$ are compared to graphs from random datasets, $G_{Rand} = (\top_{Rand}, \bot_{Rand}, E_{Rand})$, for a range of thresholds, (Fig. 2). A random dataset of the same dimension as the input dataset was created, where for each row (gene) i, random numbers were selected from a Normal distribution of mean, $\mu_{i\ Rand} = \mu_{i\ Real}$, and standard deviation, $\sigma_{i\ Rand} = \sigma_{i\ Real}$ (since for a gene which does not respond to any sample, expression would be relatively constant with no deviation from Normal). For each threshold choice in this analysis, 100 random datasets were created to estimate cut-offs, with comparisons based on averaging over these.

A null model assumes each edge in the graph to be created with a probability $= (|E_{Rand}|/|\top_{Rand}|.|\bot_{Rand}|)$, while an alternative assumes an edge was created with probability $= (|E|/|\top|.|\bot|)$. The level at which the log of the ratio of these two probabilities is maximised is taken to be optimal in terms of real effect observed, (the log ratio is used as it is easily manipulated). Clearly, one criterion for definition of maximum threshold is that at least one gene-sample couple must be identified in the real and random graph. Thresholds are then tested in probability increments of 0.02, to determine percentage inclusion of expression values. Once the "strong response" threshold is found, moderate and weak response thresholds can be established similarly.

3.3. *Distribution-based Weighting*

The motivation for the distribution-based scheme, proposed here, is that it is difficult, if not impossible, to give an absolute characterisation of an important gene-sample couple. An absolute expression level, on its own, means very little, as noted

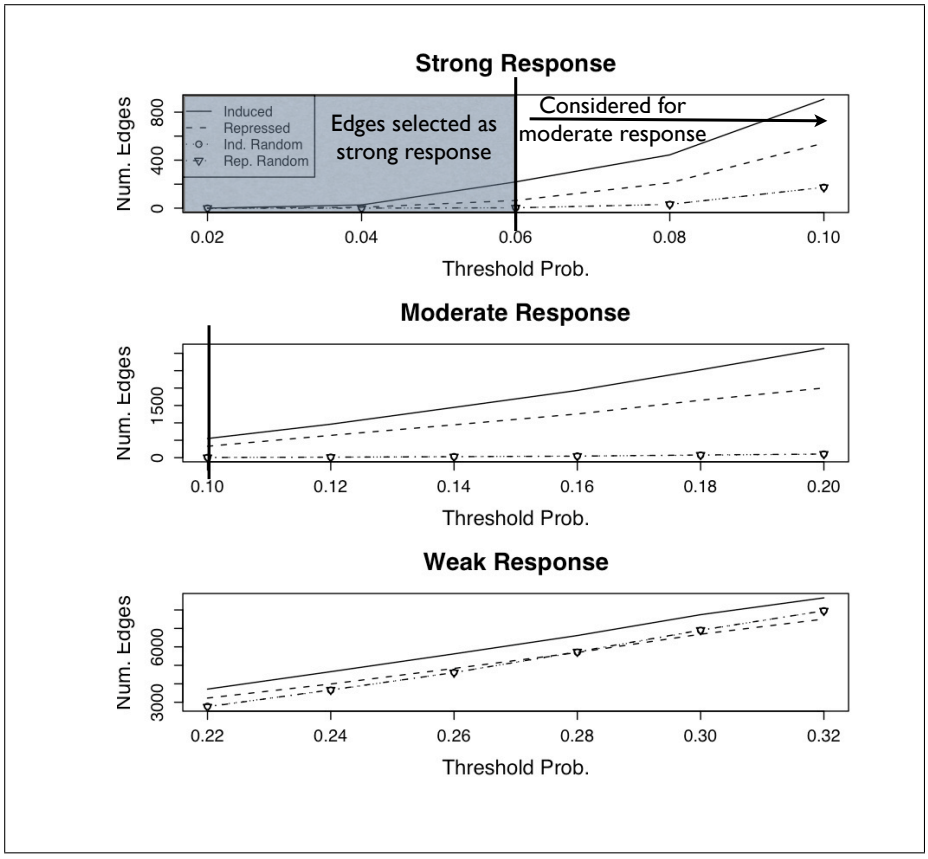

Fig. 2. Threshold Analysis of Lymphoma data. X-axis indicates the probability threshold $\frac{1}{k^2}$, the y-axis indicates the number of edges selected.

in Section 3.2. On the other hand, the study of the expression level of a given couple, relative to that for other couples can provide interesting insight on functionality. An 'interesting couple' is one where we can highlight a significant effect for a given gene responding to a sample.

This implies an expression level, differing notably from the 'average' expression observed for this gene across all samples, (and used by empirical-based approach). This approach is adaptable, given that if the goal of the experiment is to identify samples that elicit uniformly strong or weak responses from profiled genes across all sample points, (as would be the case when a gene's expression quickly reaches a plateau following a perturbation and remains active for the subsequent duration of the experiments) a baseline sample point could be used to quantify an 'interesting gene-sample couple rather than the average of the gene vector. An 'interesting' couple also implies some deviation from the effect this sample has on the whole set

of genes: if a certain sample leads to over-expression of virtually all the genes, it might be of limited interest to consider its effect on one gene, as this would not be in any way unique.

The dataset is considered as a matrix, X, containing only positive values[b], the expression levels. Each gene vector is scaled to have a mean of 1, by dividing the expression level, x_{ij}, by the row mean for gene i. Thus, genes $i \in I$ with expression value $x_{ij} \leq 1$ are considered repressed and $x_{ij} > 1$ are considered induced in sample j. For a given gene, we therefore obtain a series of positive values, of average 1, with values x, s.t. $0 \leq x < 1$ when the gene is under-expressed in a certain sample, and s.t. $x > 1$ for a gene over-expressed in the sample.

For each sample, in order to differentiate between genes that show specific behaviour and those that react similarly to other genes, a geometric series, $(a, aR, aR^2, aR^3, \ldots)$ is useful to create tightly defined categories for values close to 1 (= mean expression level for a gene across all samples, as above), and broader categories as expression deviates further from this value. Such series are used because of the skewness of the data: many genes show very little response to a given sample, having an expression value equal to 1. (Note, this property is not unique to a geometric series; however a geometric series is sufficient for the purpose of this analysis and can be manipulated easily.) Two series are used: one for values greater than 1, and the other for values smaller than 1. A ratio R for each geometric series is calculated using Eq. 3, where N_c is the number of *categories* needed, and *Min* and *Max* are, respectively, the smallest and largest value of the partial set considered for the series. For the $[1; \infty)$ part of the dataset, $Min = 1$, and the category boundaries are, therefore, 1, R, R^2, etc., with $R \geq 1$. For the $[0; 1]$ part of the dataset, $Max = 1$, and the category boundaries are, therefore, 1, $\frac{1}{R}$, $\frac{1}{R^2}$, etc., with $R \geq 1$. For this part of the dataset, there can be a problem if $Min = 0$, since R cannot be calculated in such cases. This is typically due to missing values. One solution here is to use the average value of R, observed for samples where $Min \neq 0$, ensuring that the distribution obtained is consistent with the data. Alternatively, the smallest value larger than zero could be used to calculate R.

$$R = \left(\frac{Max}{Min}\right)^{1/N_c} \tag{3}$$

Once the categories have been created and populated by expression values of the dataset, weights are assigned to gene-sample couples, depending on the size of the category to which each couple belongs. As most optimization techniques traditionally *minimize* a given objective function, we want to have negative weights for "interesting" couples, and positive otherwise. To obtain size-dependent weights,

[b]Some microarray datasets are only made available after they are transformed into log space, (a result of the normalisation process), thus leading to some negative values for low expression levels. For such cases, the first step is to transform the data back from log-space, to deal exclusively with positive values.

the average population of all categories, $\bar{S}$, is subtracted from the population of that category to which the couple belongs, S_C, (Eq. 4): hence gene-sample couples in "small" categories are negatively weighted, while larger ones have positive values. To avoid extreme weight values for datasets with very large number of genes, n, weights are normalised by dividing by n.

$$w_{ij} = \frac{S_C - \bar{S}}{n} \tag{4}$$

These weights should be interpreted as follows: the lower the weight w_{ij} of edge e_{ij}, the more significantly the expression level of gene i deviates from what is observed for the majority of genes for sample j.

4. Scheme evaluation

In this Section, we use the 4-point framework introduced in Section 2.2 to analyse our two novel weighting schemes, and to compare them with the scheme introduced by Tanay *et al.* [4].

4.1. *Reusability*

The empirical weighting scheme results in a partially connected graph for each subcategory, since genes which do not show a significant change for a given sample do not generate an edge (i.e. a partially connected graph consists of an edge *and* non-edge set).

As genes which do show a significant change in expression generate an edge with a positive weight, search criteria which maximize an objective function would be biased towards having as many edges in a cluster as possible, (non-edges do not have an associated negative weight to offset the positive weights of edges).

Subsequent clustering techniques do need to allow for this non-edge set, by optimisation of the objective function excluding/weighting nodes not connected by an edge. This scheme requires a dedicated algorithm for subsequent biclustering, which maintains the lists of gene-sample couples in each category, (i.e. strongly induced, moderately induced, etc.).

The distribution-based scheme is independent of subsequent clustering approaches, since it was designed without an explicit approach in mind. It is versatile and can be used with most clustering techniques and with other types of large datasets, biological or otherwise.

Likewise, the Tanay scheme is independent of the subsequent clustering technique, as it results in positive edges for interesting gene-sample couples and negative edges for non-interesting gene-sample couples. Indeed, the scheme was specifically designed for an additive scoring system, where the sum of the edge weights in a subgraph corresponds to its statistical significance, Eq. 1, (see [22] for more details). This is a desirable property which greatly aids the complicated step of cluster validation. This scheme has also been applied to a compendium of information (and

not just to gene expression data) including data from transcription factor location studies, sythenetic lethality studies and protein protein interactions, [12]. Each experimental measurement was transformed into a probability of having that specific experimental property occurring (recall with the Tanay model, the bipartite graph contains a gene node set and a set of experimental 'properties). Although the integration of hetergenous datasources is non-trivial, the Tanay scheme is a powerful one, which exploits additional information that is not observable at the transcription level alone.

4.2. *Parameter influence*

The distribution and empirical weighting schemes under consideration are controlled by a single parameter and are, therefore, easily configurable while offering some flexibility.

The distribution-based scheme is controlled by the *number of categories*, N_c. The influence of this parameter is assessed through examination of the robustness and discrimination achieved, as detailed below.

With respect to the empirical scheme the parameter that influences results is κ, (Eq. 2), which determines thresholds between categories. Table 1 illustrates the results of the threshold analysis for the *Lymphoma* dataset. The maximum threshold for which any gene-sample couple was identified in the real dataset was $\kappa = 7.07$ (probability ≤ 0.02, (Eq. 2)). Thresholds of $\kappa = 5, 4.08, 3.58$ and 3.162, (i.e. probabilities $(\frac{1}{\kappa^2}) \leq 0.04, 0.06, 0.08$ and 0.10 respectively) were then tested. The log ratio of the probabilities is maximised at $\kappa = 4.08$, and this was taken to be the strong response threshold. Gene-sample couples which were identified as strongly responding where then removed from the analysis. Next, values of $\kappa = 3.58, 3.162.88, 2.67, 2.5$ and 2.35 where tested for the moderate response threshold. The log ratio of the probability between real and random was maximised at $\kappa = 3.16$, and this was set as the moderate response threshold. Gene-sample couples which were identified as moderately responding where then removed from the analysis. Finally, values of $\kappa = 2.88$, 2.67, 2.5, 2.35, 2.23, 2.13, 2.04, 1.96 and 1.88 where tested for the weak response threshold. The log ratio of the probability, between real and random, was maximised at $\kappa = 2.23$, and this was set as the weak response threshold. The rationale for choosing these threshold cut-offs is that the distinction between real and random graphs is maximised at these points, while assigning a probability to each gene-sample couple in a category (e.g. the probability of a gene-sample couple occurring by chance in the moderately responsive category is < 0.08).

Table 2 provides results of threshold analysis for all three test datasets..

The main parameters affecting the Tanay weighting scheme are, again, thresholds between categories, and P_c, (Eq. 1). This scheme adopts a 'hard' thresholding approach, whereby thresholds between categories are arbitrarily chosen, based on normalized ranked values within each sample and are not data dependent. For exam-

Strong					
κ	7.07	5	4.08	3.58	3.16
P	0.02	0.04	0.06	0.08	0.10
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	*undef*	*undef*	3.98	2.45	1.5
Moderate					
κ	3.162	2.88	2.67	2.5	2.35
P	0.10	0.12	0.14	0.16	0.18
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	4.35	3.9	3.54	3.31	3.20
Weak					
κ	2.23	2.13	2.04	1.96	1.88
P	0.20	0.22	0.24	0.26	0.28
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	0.063	0.035	0.019	0.002	-0.01

Table 1. Threshold Analysis , Lymphoma data. κ= the number of standard deviations from mean (Eq. 2). P = $\frac{1}{\kappa^2}$, is the probability that values are $\kappa\sigma$ from mean. The maximum log-ratio is taken as the threshold between categories. The *undef* entries indicate where the log-ratio could not be calculated due to division by zero, i.e. no edges were identified for the random graph.

	Strong	Moderate	Weak
Lymphoma	4.08	3.162	2.23
Yeast Cell Cycle	3.16	2.88	2.23
Kasumi	2.77	2.58	2.23

Table 2. κ thresholds identified for each of the tested datasets.

ple, the top and bottom t% of normalized ranked values are selected as moderately expressed, where t is an arbitrarily chosen input parameter. This 'hard thresholding' has consequences for the deterioration of the scheme when noise and missing values are added to the data. As the t is lowered, a higher percentage of edges will be identified, even if none exist (i.e. as not data dependent, it will always select the top t%). For an analysis of parameter P_c see [4].

4.3. *Robustness*

Noise and missing values were included in the dataset in the tests reported here in order to mimic measurement error of different amounts. Noise was randomly "added/subtracted" to each value in the dataset as a percentage (up to 10%) of the original value. To replace data with missing values, up to 10% of expression values from the original dataset were randomly selected and removed[c] Commonly,

[c]In the absence of more specific knowledge of the distribution of missing values in datasets tested, it was assumed distribution was close to random and hence that random selection of missing values was a reasonable solution for the testing under consideration.

missing values in the experimental gene expression matrix are replaced by zeroes or by an average expression level of the gene, ("row average"). Such methods, however, do not take into account the correlation structure of the data, and more sophisticated options include methods of K-Nearest Neighbour (KNN) and Support Vector Decomposition type, [25]. Missing value estimation methods have generated a considerable literature in their own right, so to test our weighting schemes the common practice of replacing missing values by the row mean was adopted. The missing value estimation procedure chosen will, of course, affect the final form of the data. For instance, replacement by the row mean will reduce the standard deviation of each gene vector under consideration.

For this analysis, we defined *"Average Absolute Variation"* as the average difference in edge weights, compared to 0% noise/missing values and *"Stable weights"* to be those for which the variation is less than the level of noise/missing values added.

The influence of noise and missing values are summarised respectively in Table 3 for the *Lymphoma* dataset. Results for other datasets, (not displayed here), are consistent with these.

4.3.1. *Empirical Weighting Scheme*

The absolute variation in weights is extremely low for the empirical scheme since the technique examines extreme values, i.e. values which appear in the tail of the distributions of each gene variable. In addition, weights are not based directly on a given expression value, but on that expression value *relative* to other values in the category for a particular sample (Step 2 of scheme, see Fig. 1). The category is in turn defined *relative* to expected value of the gene variable (Step 1 of scheme, see Fig. 1). As "missing" values are replaced by the row mean, this does not greatly affect extreme values. Equally, even noise added at 10% level of the original values does not affect *relative* values, thus, perturbations in the data have small effect on weights assigned.

Similar to results shown in Table 3, for the *Kasumi* dataset, average absolute variation in edge weights is $\sim 0.26\%$ for an added noise level of 10% (data not shown), while denoting 10% of the dataset as missing values, gives average absolute variation in values $\sim 0.3\%$, with stable weights accounting for around 99.5%. For the *Yeast Cell Cycle* data, the corresponding values for 10% noise added were: $\sim 0.22\%$ (absolute variation) and $\sim 99.65\%$ (stable weights); and for missing values at 10% were: $\sim 0.15\%$ (absolute variation) and $\sim 99.69\%$ (stable weights).

4.3.2. *Distribution-Based Weighting Scheme*

The distribution-based weighting scheme is more generic, as it does not rely as much on the underpinning biological information: as such, it is less robust to noise and missing values, which alter the distribution of expression levels.

Influence on the weights used is "reasonable" for low noise perturbation. Specif-

% Noise level	1.5	2.5	5	10
Distribution Based				
% Average Absolute variation	3.10	4.90	9	15.80
% "stable" weights	83	84.5	89.1	88.6
Empirical Based				
% Average Absolute variation	0.06	0.02	0.03	0.05
% "stable" weights	99.66	99.68	99.68	99.65
Tanay Scheme				
% Average Absolute variation	2×10^{-3}	2×10^{-3}	2×10^{-3}	3×10^{-3}
% "stable" weights	99.98	99.99	99.99	99.99
%missing values	1.5	2.5	5	10
Distribution Scheme				
% Average absolute variation	17.70	27.30	48.86	92.00
% of "stable" weights	37.40	33.40	32.50	30.20
Empirical Scheme				
% Average absolute variation	0.04	0.04	0.09	0.16
% of "stable" weights	99.70	99.72	99.64	99.62
Tanay Scheme				
% Average absolute variation	2×10^{-3}	3×10^{-3}	3×10^{-3}	6×10^{-3}
% of "stable" weights	99.98	99.99	99.99	99.99

Table 3. Influence of noise level and missing values on weights assigned (for 20 categories for distribution based scheme). Stable weights refer to the percentage of weights for which the level of variation is less than the level of perturbation added.

ically, the percentage of stable weights is a helpful indicator, given the nature of the scheme: when a gene-sample couple falls into a new category due to added noise, this changes the weights for all couples in the new category as well as all those in the old one. The % stable weights value may, therefore, give more insight into scheme robustness, rather than using only the average absolute variation of the weights, (Table 3). Clearly, as the number of categories increases, each category interval decreases, and gene-sample couples are more likely to change categories, leading to a smaller proportion of stable weights, Figure 3. Stable weights are re-defined based on the level of deviation smaller than the added perturbation. For very low perturbations, this rule becomes very difficult to satisfy, and creates the initial dip, which can therefore be discarded. The proportion of stable weights can then be taken as almost constant, (for a fixed number of categories).

With respect to missing values, the scheme is far less robust than the empirical and Tanay schemes with respect to noise perturbations. However, it is important to note that *the sign* of the remaining weights is not lost: a positive weight does not become negative, (except for cases when more than half the values for a given gene are missing; in those cases it would almost certainly be excluded from the

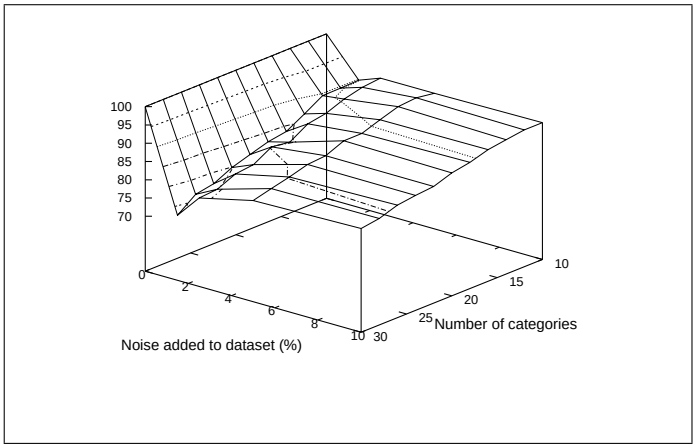

Fig. 3. Influence of noise - Proportion of stable weights (%)

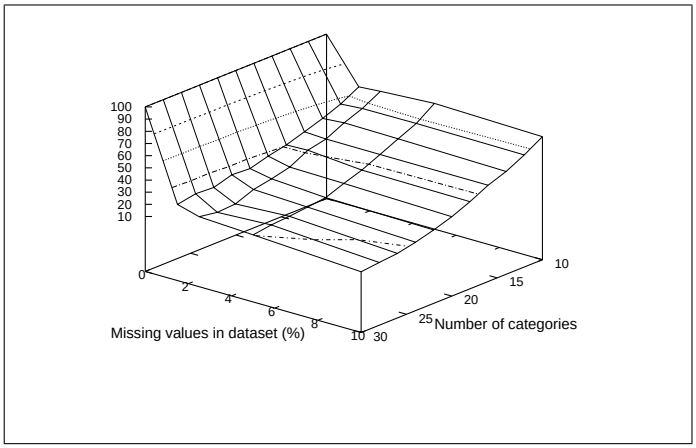

Fig. 4. Influence of missing values - Proportion of stable weights (%)

dataset before the scheme is applied). What is partially lost is the degree of over-expression, (or under-expression), rather than the knowledge that this change of expression occurs. A representation of missing value influence, depending on the number of categories, is given in Figure 4.

As noted previously, the results displayed correspond to untreated data, i.e. where the mean replaces missing values, thus creating larger perturbations in the weights. Future work might reasonably include tests of the effect of the various correction techniques on scheme robustness.

4.3.3. *Tanay Scheme*

Perturbations in the data have very little effect on weights derived with the Tanay scheme, (similar results for all tested datasets, data not shown for Yeast and Kasumi datasets). Slightly surprised by this result, missing values were tested up to a level of 80%, but effects remained minimal (0.01%, average variations and 99.99% stable weights). It may be the case that this scheme identifies 'interesting' gene-sample couples, even if none exist, due to the 'hard' threshold nature of the scheme (Section 4.2). If 10% noise added to the dataset, thresholds still depend on ranking and not mean level, thus approximately the same gene-sample couples are selected as interesting as relative ranked position is not changed (i.e. if 10% random noise is added to every value in the data, ranked positions remain relatively unaffected). Missing values have little effect, as these are replaced by the mean, therefore more extreme values are replaced with central values, but the same thresholds are used.

4.4. *Discrimination*

For this analysis, 'random graph' refers to graphs created from random datasets, as described in Section 3.2.1.

4.4.1. *Empirical Weighting Scheme*

From the threshold analysis, described above, maximum discrimination between empirical and random graphs is achieved. As expected, the largest number of gene-sample couples falls in the weak response category. Discriminating between gene responses depends on the category thresholds used. From Table 4, for strong and moderate response, the probability of an edge existing between a gene and sample node in the real graph is greater than that for the random graph, *indicating that significant structure is present*. For a weak response, the ratio of probabilities is smaller and it is less convincing that real differences exist. Nevertheless, an examination of the average degree of sample nodes in the real graphs indicates that the average number of genes responding is higher than expected. For example, for the weak repression sub-category, a sample node is, on average, connected to $\sim 3\%$ of gene nodes compared to $\sim 1.5\%$ in the random graph, $(d_\perp/d_\top)$. The average degree of a sample node in the real graph is also higher than expected, $(> m/n_\perp)$. This suggests that, although the ratio of probabilities in weak response is not high, some pattern structure is present and the method is capable of identifying indicative gene-sample couples.

4.4.2. *Distribution-Based Weighting Scheme*

By construction, there are no "damaging" false-positives or false-negatives, at least in theory. In practice, false-positive or false-negative edges may have weights very close to zero, (either positive or negative depending on the number of categories).

	Repressed			Induced		
	Weak	Moderate.	Strong.	Weak.	Moderate.	Strong.
$n_\top$	2572	456	111	2829	719	304
$n_\perp$	63	26	28	81	53	54
m	4324	506	114	4939	822	316
$d_\top$	1.69	1.11	1.03	1.74	1.148	1.04
$d_\perp$	72.29	21.66	4.75	64.09	17.75	6.08
δ	0.0112	0.002	0.0003	0.0127	0.002	0.0008
$n_{\top rand}$	2774	8	5	2718	7	3
$n_{\perp rand}$	87	7	5	87	7	3
m_{rand}	4007	9	5	4020	8	4
$d_{\top rand}$	1.477	1	1	1.462	1	1
$d_{\perp rand}$	46.57	1.28	1	45.875	1.14	1
δ_{rand}	1×10^{-2}	2×10^{-5}	1×10^{-5}	1×10^{-2}	2×10^{-5}	1×10^{-5}

Table 4. Categories for Lymphoma data, created by cut off thresholds 0.20 (weak induction/repression), 0.10 (moderate induction/repression), and 0.06 (strong induction/repression). $n_\top$ = the active set of genes (gene nodes with degree ≥ 1), $n_\perp$ = active set of samples, m = number of edges, $d_\top$ = average degree of active set of genes, $d_\perp$ = average degree of active set of samples, δ = bipartite density (i.e the fraction of existing links with respect to possible ones.)

A significant change in expression patterns can not lead to a positive weight, while negative weights are only obtained where there *is* a significant change. Since absolute, (positive vs. negative), discrimination is guaranteed, the focus here is to assess *relative* discrimination, i.e. distribution of weight values. Results, of the influence of the number of categories, on this discrimination are displayed in Table 5.

The size of the categories reflects the biological significance of the gene-sample couples in it. Additionally, size of each category is influenced by the number of category available. Biological significance of the weights is a direct consequence of the weighting process: the lower the weight $w_{i,j}$ of edge $e_{i,j}$, the more significantly the expression level of gene i deviates from what is observed for the majority of genes for sample j (i.e. there will be less gene-sample couples in the category), Eq. 4. The scheme operates a preselection of potentially interesting gene-sample pairs.

A first observation, based on the proportion of negative weights, is that discrimination is good: there are on average, just under 41,019 ($\sim 11\%$) negative weights (out of 386,496), implying that 11% of gene-sample couples in the dataset are of interest. This value varies between 35,542 (30 categories) and 49,242 (10 categories), with a standard deviation of $\sim$ 3908. The scheme selected $8.75\% \pm 0.21\%$ gene-sample couples from a random dataset, (Section 3.2.1), for $N_c = 14$. Discrimination is satisfactory for any number of categories, (N_c), in the range tested, (Z-test, $p < 0.0001$). Weights obtained with fewer categories appear more discriminatory in general, although this is less evident at $N_c = 10$, which indicates that $N_c < 10$ would be ill-advised. With less than 10 categories, there is an artificial increase

	Distribution Weights						
k	≤-6	[-6;-4]	[-4;-2]	[-2;0]	[0;2]	[2;4]	≥ 4
10	14977	10207	8779	15279	13337	12969	310948
12	7708	13206	10917	11324	11453	17783	314105
14	733	14252	14998	13327	14293	17606	311287
16	0	9778	16293	17884	14973	17616	309952
18	0	4723	19030	18740	18141	17629	308233
20	0	0	20095	20116	21275	20942	304068
22	0	0	17435	22792	23217	19002	304050
24	0	0	14050	24584	24329	24004	299529
26	0	0	11334	26286	26337	28134	294405
28	0	0	8325	28679	27582	32757	289153
30	0	0	4751	30791	29200	35930	285824

Table 5. Influence of the number of categories on discrimination for Lymphoma data: distribution of weight values. Values in cells represent the number of gene-sample couples in various ranges

of extreme weights (i.e. very low or high weights), indicating that the biological significance of the weights is lost. Given that $N_c = 12$ to 16 categories also corresponds to improved robustness (Section 4.3), using this range is recommended. This recommendation also applies to other datasets, for which results obtained are similar.

4.4.3. *Tanay Scheme*

From our analysis of this scheme (Table 6), we observed that (a) a smaller number of total positive weights were identified, compared to those selected from corresponding graphs generated from random dataset, (Section 3.2.1), and (b) the number of positive weights in each category is roughly equivalent to random, (with the exception of Kasumi Moderate and Strongly repressed categories). Observations (a) and (b) imply that the degree 'sharing' of gene-sample couples in the real dataset (i.e. gene-sample couples having a positive edge in the weakly induced category, also feature in the strongly induced category - categories are not mutually exclusive). Since 'hard' thresholds between categories were used and arbitrarily chosen, the order of the number of edges in each category is $Strong < Moderate < Weak$. Note also that those gene-sample couples, evaluated as strongly reacting, will have a magnified impact on any clustering procedure on the resulting graph, due to the overlap between categories.

4.5. *Discussion and Conclusion*

From the investigations above, it is clear that interpretations of edge weights in graphical gene expression schemes can differ considerably. Primarily, the empirical

Datasets	Lymphoma	Yeast Cell	Kasumi
% total edges	2.7	3.34	2.13
	(2.9 ± 0.007)	(4.3 ± 0.01)	(3.6 ± 0.02)
Induced			
% strong	1.2	1.1	0.8
	(1.1 ± 0.009)	(1.2 ± 0.02)	(0.8 ± 0.004)
% moderate	3.2	4.1	1.0
	(3.1 ± 0.01)	(3.9 ± 0.03)	(1.0 ± 0.005)
% weak	3.9	5.0	1.1
	(3.9 ± 0.02)	(4.8 ± 0.03)	(1.1 ± 0.007)
Repressed			
% strong	1.2	1.1	1.3
	(1.1 ± 0.009)	(1.2 ± 0.02)	(1.4 ± 0.06)
% moderate	3.2	3.8	3.8
	(3.1 ± 0.01)	(3.9 ± 0.02)	(3.4 ± 0.06)
% weak	4.0	4.7	4.7
	(3.8 ± 0.01)	(4.8 ± 0.03)	(4.1 ± 0.12)

Table 6. Tanay scheme - percentage of total possible edges is taken as $\frac{|E|}{n \times p \times 6}$ whereas percentage of edges in each category is taken with respect to total possible edges in that category i.e. $\frac{|E|}{n \times p}$, (n = number of genes, p = number of samples). Bracketed values represent results from random graphs.

and Tanay weighting scheme result in *positive* weights for "interesting" gene-sample couples, while the distribution technique leads to significant effects reflected in *negative* weights. If the optimisation technique uses a minimisation-based objective function, the empirical and Tanay weights could simply be negated, (similarly for distribution weights if a maximisation function is used). For *both* empirical and distribution schemes presented, edge weights linking effected genes for a particular sample j, are defined *relative to other gene expression values for sample* j. This is an important corollary, as absolute level of gene expression is not directly accounted for, only the fact that it does change, with the significance of this change relative to the majority of genes. Relative evaluation is also an intrinsic feature of the Tanay based scheme, as the initial probability $\phi(i, j)$, (Eq. 1) is based on ranks. However, the selection of pre-determined hard thresholds between ranks has a large effect on robustness and discrimination as it is not data-dependent. Conversely, the issue of data dependent threshold estimation is addressed in the two novel schemes presented, but is non-trivial as numerous thresholds need to be assessed, which is computationally expensive.

The empirical-based scheme is more specific, in the sense that fewer gene-sample couples are identified, compared to the distribution-based and Tanay scheme. For example, for the *Lymphoma* dataset, the empirical based scheme extracts $11,021$

gene-sample couples ($\sim$ 3% at optimal threshold levels). The distribution-based scheme extracts 43,155 couples ($\sim$ 11%) (for 12 categories), although 11,324 of these are close to 0 (and hence do not have a major impact on clustering), this still leaves 31,831 ($\sim$ 8%), "interesting couples". The Tanay scheme extracts $\sim$ 16% for the same dataset.

The empirical-based method deteriorates slowly with perturbations in data, hence for data known to contain many missing values and/or noise it may be a better choice. Weights overall for the Tanay scheme seem little affected by noise and missing values, which indicates that this scheme will assign high weights to gene-sample couples, (even if none are present). If the sample size (i.e. the number of microarray experiments), for each gene is small, the performance of the empirical-based scheme deteriorates with respect to its random graph comparison basis, (difficult to estimate μ and σ of each gene variable) while the thresholds between categories become increasingly difficult to identify. In this situation, the distribution-based scheme would be the better choice.

Although the aim of this work was solely the presentation and investigation of weighting schemes in isolation, comparison of clusters found from graphs, created using the various weighting schemes presented, would inevitably provide a more complete assessment of the weighting schemes. Throughout the assessment presented here, properties which would affect subsequent clustering where indicated. For example, positive edge weights, created by the the empirical and Tanay weighting scheme require a maxisation function, while negative weights, created by the distribution based scheme, require a minimisation function. Also, as noted previously, the empirical scheme results in a set partially-connected graphs, while the distributed and Tanay scheme results in a fully-connected graph. This has consequences for the subsequent clustering procedure used to group the edge weights. Therefore, its clear that the same clustering procedure cannot be universally applied to graphs resulting from all weighting schemes, and an analysis of clusters produced by these schemes would also involve a detailed analysis of the clustering procedure itself, to account for differences in these, and in the weighting scheme. Such an analysis is outside the scope of this paper, although it is noted that this is an important continuation of this work.

Using graphical techniques to extract meaningful information from biological data is an intuitive and popular method. In this paper, we limited our investigation to bipartite graphs as this representation captures essential properties of the gene expression dataset and allows for the extraction of biclusters most suitablet for data in this domain, [6].

An investigation into weighting of gene expression networks is long overdue. In this paper we proposed and compared two weighting schemes applied to gene expression bipartite graphs with a view to extracting meaningful biclusters from the data. We also compared the properties of these novel schemes to the innovative work of Tanay et al. [4, 12, 22]. The importance of assessing edge-weighting schemes

was highlighted and we showed that edge weights must be considered independently from the clustering procedure, since alternative edge weight derivation can lead to different interpretations of the data. This type of assessment framework for weighting is equally crucial in the context of other types of large dataset, biological or other.

Further investigation on extraction of meaningful graphical relationships through choice of edge-weighting schemes should include incorporation of information from related sources, (such as protein interaction and promoter information etc.), to refine weights and improve handling of missing values. Investigation of automatic threshold estimation for category sub-divisions is also indicated. Comparison of clusterings using various weighting procedures is a required next step.

Acknowledgement

D. Perrin would like to acknowledge support from the Irish Research Council for Science, Engineering and Technology (Embark Initiative).

References

[1] A. A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.

[3] P. Tamayo et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96(6):2907–2912, 1999.

[4] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 Suppl 1:136–44, 2002.

[5] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Compt. Biol. Bioinform.*, 1(1):24–45, 2004.

[6] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan. Techniques for clustering gene expression data. *Comp. Biol. Med.*, 38(3):283–293, 2008.

[7] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57, 2004.

[8] T. E. Reddy, C. DeLisi, and B. E. Shakhnovich. Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS Comput. Biol.*, 3(5):90, 2007.

[9] R. Bourqui, L. Cottret, V. Lacroix, D. Auber, P. Mary, M. F. Sagot, and F. Jourdan. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Sys. Biol.*, 1:29, 2007.

[10] R. Sharan and R. Shamir. CLICK: a clustering algorithm with applications to gene expression analysis. *ISMB '00*, 8, 2000.

[11] E. Hartuv, A. O. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering cdna fingerprints. *Genomics*, 66(3):249 – 256, 2000.

[12] A. Tanay, I. Steinfeld, M. Kupiec, and R. Shamir. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol. Sys. Biol.*, 1:2005.0002, 2005.

[13] M. R. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7:40, 2006.

[14] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4:Article17, 2005.

[15] J. Saramaki, M. Kivela, J. P. Onnela, K. Kaski, and J. Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 75:027105–027109, 2007.

[16] V. Di Gesu, R. Giancarlo, G. Lo Bosco, A. Raimondi, and D. Scaturro. Genclust: a genetic algorithm for clustering gene expression data. *BMC Bioinformatics*, 6:289, 2005.

[17] K. Bryan, P. Cunningham, and N. Bolshakova. Application of simulated annealing to the biclustering of gene expression data. *T-ITB*, 10(3):519–525, 2006.

[18] C. H. Q. Ding, X. He, H. Zha, M. Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *ICDM '01*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.

[19] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2(1):65–73, 1998.

[20] K. Stegmaier et al. Gefitinib induces myeloid differentiation of acute myeloid leukemia. *Blood*, 106(8):2841–2848, 2005.

[21] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon. Expander – an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6:232, 2005.

[22] A. Tanay. *Computational analysis of transcriptional programs: function and evolution. Ph.D. Thesis.* PhD thesis, The Blavatnik School of Computer Science, Tel Aviv University, 2005.

[23] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*, pages 183–4. Springer-Verlag New York, Inc, Secaucus, NJ, USA, 2005.

[24] P. L. Chebyshev. Des Valeurs Moyennes,. *Journal de Mathématiques Pures et Appliquées*, 12(2):177–8, 1867.

[25] O. Troyanskaya et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.