Published online 19 December 2016

Nucleic Acids Research, 2017, Vol. 45, No. 7 ***e53***
doi: 10.1093/nar/gkw1270

Flexible model-based clustering of mixed binary and continuous data: application to genetic regulation and cancer

Fatin N. Zainul Abidin[1,2] and David R. Westhead[1,*]

[1]School of Molecular and Cellular Biology, University of Leeds, Leeds, West Yorkshire LS2 9JT, UK and [2]Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor D.E., Malaysia

Received October 04, 2016; Revised November 14, 2016; Editorial Decision December 01, 2016; Accepted December 06, 2016

ABSTRACT

Clustering is used widely in 'omics' studies and is often tackled with standard methods, e.g. hierarchical clustering. However, the increasing need for integration of multiple data sets leads to a requirement for clustering methods applicable to mixed data types, where the straightforward application of standard methods is not necessarily the best approach. A particularly common problem involves clustering entities characterized by a mixture of binary data (e.g. presence/absence of mutations, binding, motifs and epigenetic marks) and continuous data (e.g. gene expression, protein abundance, metabolite levels). Here, we present a generic method based on a probabilistic model for clustering this type of data, and illustrate its application to genetic regulation and the clustering of cancer samples. We show that the resulting clusters lead to useful hypotheses: in the case of genetic regulation these concern regulation of groups of genes by specific sets of transcription factors and in the case of cancer samples combinations of gene mutations are related to patterns of gene expression. The clusters have potential mechanistic significance and in the latter case are significantly linked to survival. The method is available as a stand-alone software package (GNU General Public Licence) from http://github.com/BioToolsLeeds/FlexiCoClusteringPackage.git.

INTRODUCTION

Since clustering was first applied to microarray gene expression data for yeast (1), the use of clustered heat maps as an exploratory analysis method has been ubiquitous in the thousands of published studies covering the growing range of high-throughput genome scale data sets in molecular biology research. This is testament to the usefulness of the method, but nevertheless clustering is an art, and despite theoretical methods that quantify the tightness and separation of clusters it can be difficult to decide when a clustering solution is good or bad. Often the clustering solutions chosen in a study are those that yield the most biological insight from the data. Accordingly many different approaches to clustering exist, including standard generically applicable methods like hierarchical clustering and K-means, probabilistic model based approaches and methods specifically tailored to particular data types and problems (e.g. model based clustering for RNA sequencing data (2)). There is no consensus on the best method to apply to a given problem.

Increasingly emphasis in the field is shifting to the integration of multiple genome scale data sets. For example alongside gene expression, studies of cell lines and differentiation (3,4) now often include epigenetic chromatin marks and transcription factor binding, and similarly tumors can be characterized in terms of somatic mutations, copy number aberrations and methylation patterns. In these cases biological insight may still be obtained by clustering, but the entities are characterised by variables of mixed types, so the problem is more complex and less likely to be addressed well by generic methods. A particular challenge is that different data types may contain more or less information, some may be biologically irrelevant, and each has different characteristic levels of random variability (noise). While separate clustering of different data types followed by comparison of resulting clusters can be useful, it can ignore valuable joint information between data types that drives the underlying biology, and this motivates the development of new methodology.

A number of different approaches to clustering entities described by variables of mixed type are now becoming available. For example, Morlini (5) developed an approach to cluster mixed binary and continuous data by treating binary variables as generated from latent continuous variables with a threshold for dichotomy, leading to clustering in a multivariate normal distribution framework. These ideas were extended to ordinal and nominal variables by McParland and Gormley (6). In each case, an expectation maxi-

*To whom correspondence should be addressed. Tel: +44 113 343 3116; Fax: +44 113 343 1935; Email: D.R.Westhead@leeds.ac.uk

mization framework was adopted, requiring manual specification of cluster numbers, and methods were applied to problems with relatively small numbers of variables (<10 continuous and <20 other). Similar ideas have been explored by Browne and Nicholas (7) and by Cai and coworkers in a Bayesian context (8). Alternatively, addressing problems with large numbers of variables of different types and incorporating dimension reduction as an integral component, iCluster (9) and iPF (10) were developed specifically for integrating and clustering mixed genome-scale ('omics') data for disease subtype discovery. In iCluster the link between data types is achieved by assuming a shared underlying latent variable model representing the disease subtypes. A different approach is taken the Bayesian MDI package (11,12), which couples clustering based on each separate data type by linking coefficients specifying the allocation of entities to specific components in a mixture distribution.

Many problems of current interest in our field involve clustering entities described by a mixture of binary and continuous variables: for example, genetic regulation can be described in terms of the presence or absence of transcription factor binding and histone marks in promoter and enhancer elements and this can determine patterns of gene expression (13); equally, tumors can be characterized by the presence or absence of somatic mutations in key signalling genes that may drive downstream changes to oncogenic gene expression patterns. We therefore set out to develop a method to cluster such entities that would be generically applicable to a range of different problems. Our specific goals were a method able handle variable numbers and data set sizes common in the field, and where the optimum number of clusters is unknown and difficult to estimate, making manual experimentation impractical. We sought a method that would give clusters with clear biological interpretability, for instance a pattern of mutation or TF binding that relates to a shared pattern of expression in a cluster of genes, and therefore avoided using dimension reduction as an integral component, assuming that this would be employed at the data preparation stage if necessary to identify the most relevant variables. Satisfying these requirements led to a method complementary to those discussed above that we show to be applicable to several realistic current problems.

We chose a simple model based framework, using a joint probability distribution of binary and continuous variables that is a mixture over an unknown number of clusters. An attractive feature of this probabilistic approach is that it provides a natural treatment of data sets where some variables may be irrelevant, for example passenger mutations in cancer samples or transcription factor binding to DNA without regulatory significance, and where there may be false positives and negatives, for example in chromatin immunoprecipitation data. In outline the method employs a heuristic search for an approximate optimal model followed by refinement using an expectation-maximization procedure. We investigated model selection with simulated data using a range of criteria related to the well-known Akaike-Information-Criterion (AIC) (14), the Bayesian Information Criterion (BIC) (15) and their variants (16,17). The method was applied to two different problems, genetic regulation in yeast based on transcription factor binding and gene expression, and the classification of cancer samples based on somatic mutations and gene expression. This showed the method to be effective in identifying clusters that relate to relevant biomolecular mechanisms, and in the cancer case to survival.

MATERIALS AND METHODS

Model

We consider a set of N entities (data points) i, representing genes, tumor samples etc., each characterised by $r_{ij} \in \{0, 1\}$, $j = 1, \ldots, n_r$ binary variables and $e_{il}, l = 1, \ldots, n_e$ continuous variables. For example the binary variables could indicate the binding or not of a transcription factor in a gene promoter or the presence/absence of a mutation at a particular locus (we allow for n_r such variables), and the continuous variables could be gene expression values in n_e samples, experiments or time points. We assume a probability distribution which is a mixture of N_m components (clusters)

$$p(r_{i1}, \ldots, r_{in_r}, e_{i1}, \ldots, e_{in_e}) = \sum_{m=1}^{N_m} \alpha_m \prod_{j=1}^{n_r} B(r_{ij}; p_{mj}) \prod_{l=1}^{n_e} N(e_{il}; \mu_{ml}, \sigma_{ml})$$

from which data points are assumed to be generated. Here α_m are mixing coefficients $\sum \alpha_m = 1$; B denotes the Bernoulli distribution with parameter p_{mj}, and N is a normal distribution with parameters μ_{ml} and σ_{ml}. In the case of genetic regulation the mixture components represent the well-known concept of a cluster of co-regulated genes, with, for example, Bernoulli parameters p_{mj} representing the probability of binding for particular transcription factors in promoter/enhancer elements, and the μ_{ml}representing a shared average pattern of gene expression, which could be a time or developmental series but is not required to be. In the case of tumour samples, clusters could be related samples where Bernoulli parameters associate mutation probabilities at particular loci with shared patterns of oncogenic gene expression.

Estimating model parameters

Since the number of clusters is unknown and difficult to estimate we adopted an initial heuristic search for an approximately optimal model, followed by refinement of the solution by expectation maximization. The heuristic search employed a Monte-Carlo simulated annealing algorithm (18) to optimize objective functions of the form

$$O(L, k) = -2L + k\lambda(N)$$

where L is the (maximized) log-likelihood from the distribution above, λ is a function of the number of data points N and k is the number of parameters in the model.

We investigated several different functions $\lambda(N)$, including constants ($\lambda = 1.0 - 5.0$) where $\lambda = 2$ corresponds to the standard AIC criterion, $\lambda(N) = \ln N$ (the BIC criterion), the Hannan-Quinn criterion $\lambda(N) = 2 \ln \ln N$ and the consistent AIC (CAIC, $\lambda(N) = 1 + \ln N$). Regarding the choice of objective functions the standard AIC and BIC are the most commonly used, and they arise from fundamentally different theoretical stand points (19). The AIC is

obtained by minimizing the Kullbeck–Liebler distance between the estimated model and an underlying 'true' model, while the BIC maximizes the posterior probability of the model given the data. Both criteria are asymptotic results applicable to large samples, but their large sample behaviour is fundamentally different: the BIC is asymptotically consistent (converges in probability to a single model as $N \rightarrow \infty$) and AIC is not. AIC on the other hand embodies the idea that as the data set grows in size evidence may emerge for a more complex model. We note that in our case we are likely to be some way short of the large sample limit, and that the choice of objective function is likely to be based on empirical considerations. The rationale for investigating different multipliers in the AIC type criteria (with no N dependency in the penalty) was that for small samples the standard AIC is an underestimate of the optimal penalty, but a more complex correction is not suitable numerically for our approach.

Full details of the heuristic search algorithm and equations for expectation maximization are given in Supplementary information, along with details of how the algorithm was parametrised using simulated data. Algorithm parameters in the form of input files for the clustering program for the two biological test cases are also provided in supplementary information.

Test data sets

As an initial test of the methodology and objective functions we examined their ability to find correct solutions for the number of mixture components and the assignment of data points to components in data simulated from the probability distribution above. Simulated data was also used to parametrize the algorithms (see Supplementary information). For the tests described here, we simulated data sets with 100, 200, 500 and 1000 data points, and for each case two data sets comprising clusters of equal sizes of 10 or 20 (e.g. the 500 data point sets were 1. 25 clusters of size 20 and 2. 50 clusters of size 10). All data sets had $n_r = 20$ and $n_e = 20$ and each cluster was specified with a distinct pattern of binary variables p_{mj} and continuous variables μ_{ml} (randomly chosen) and σ_{ml}. One set of simulations modeled the case of tight well separated clusters with low noise, and the second set modeled less well separated clusters with a higher level of noise. In the first set the values of p_{mj} were either 0.9 or 0.1 and all σ_{ml} values were 0.01, in the second set the values of p_{mj} were either 0.9 or 0.4 and all σ_{ml} values were 0.3. Random numbers from appropriate probability distributions were generated using standard functions in the Java programming language.

Data for genetic regulation

To test the methodology in application to genetic regulation we used the well-studied yeast cell cycle, basing our work on gene expression data (18 time-points) from Spellman and co-workers (20) and 103 yeast transcription factors (TFs) from the regulatory map published by Harbison *et al.* (21). In the regulatory map, a TF was assumed to bind if the *P*-value was less than 0.001. Based on our studies with simulated data we considered that our method is suitable for a data set of a few hundred genes and 10–20 regulatory inputs, and this is consistent with estimates from previous studies of the number of genes showing cell cycle related expression and the likely number of TFs involved in cell cycle related regulation (22–24). Accordingly, we began with 525 genes identified by the authors as showing cell cycle related expression. Our preselection of TFs was based on the preselection step for the LeTICE algorithm (25). This is based on the hypothesis that if a TF is active in regulating any of the selected genes, then within the set of genes whose promoters it binds there should be some gene pairs showing highly correlated expression patterns reflecting common regulation, even allowing for the possibility that the TF does not regulate all the genes that it binds. Therefore using the 95th percentile, ρ, of Pearson correlation coefficients over all gene pairs, the proportion of correlations greater than ρ in the gene set that the TF binds is calculated. This is then compared to the proportion of correlations greater than ρ in randomly selected gene sets of the same size, and an empirical p value calculated. If this p value is less than the generous threshold of 0.1 then it is assumed that the TF may regulate some genes and it is retained, otherwise the TF is removed from the set under consideration. In this case 17 TFs were retained for input to the main clustering algorithm, on the assumption that these TFs are the ones likely to be regulating cell cycle genes. Following this the set of 525 genes was further reduced to 328 by eliminating genes not bound by any of the selected TFs.

LeTICE (25) was also used as an alternative method for comparison with our approach. LeTICE is not a generic clustering method but is designed specifically for the problem of genetic regulatory network prediction. It is based on integrating TF binding data with expression pattern data to define a genetic regulatory network, i.e. a set of modules each comprising genes with a common TF binding pattern and a shared pattern of expression. This is achieved by finding the network, B, which maximizes $P(B|L, E)$where L is a matrix of TF binding probabilities and E a matrix of gene expression patterns. As such LeTICE is a method based on a similar premise of integrating TF binding data and expression data to find regulatory relationships, but being based on different underlying methodology it is an ideal comparator, albeit only relevant to the problem of genetic regulation. To provide a direct comparison of algorithms, LeTICE was applied to the dataset described above. Note that LeTICE takes binding p values directly as input and that it has its own TF and gene pre-selection criteria, in this case it selecting 18 TFs and 289 genes. LeTICE was then run with the optimum runtime parameters suggested in the original paper.

As part of this study we also examined the effect of using normalized (where each gene was normalized to zero mean and unit standard deviation) and un-normalised gene expression data. We also compared joint clustering to clustering expression data separately, which can be done by simply omitting binary variables in the input to our program.

In evaluation of our method we considered comparison with the known literature on genetic regulation in the yeast cell cycle, as well as measures of the functional coherence of clusters based on Gene Ontology (GO) using the GOSemSim (26) package in R (with the information content based semantic similarity measure). Since our method can iden-

tify combinatorial regulation (a cluster of genes regulated by more than one TF), and this implies potential interactions between TFs, we also compared these implied interactions with physical and genetic evidence in BioGRID (27). As for selecting a set of relevant regulators related to the well-known cell cycle regulators in the discussion section, KEGG was used to retrieve all genes related to the cell cycle pathway (28).

Data for acute myeloid leukaemia

To test the application of our methodology to data from cancer samples, we applied it to the Acute Myeloid Leukemia (AML) mutation and gene expression data generated by The Cancer Genome ATLAS (TCGA) Research Network (29,30). This is an effective test since classification of AML samples has been the subject of extensive research which can be compared to our results, including the French-American-British (FAB) system which largely relies on cell histopathology, the World Health Organisation classification which includes cytogenetic aberrations, and further work using gene expression alone (31–33), gene mutations (34) and linking gene mutations to expression (35).

Datasets from TCGA were retrieved through using cBioPortal for Cancer Genomics tool (36,37). In the TCGA data, samples were selected based on the availability of mutation and RNA-seq gene expression data. Genes which were mutated in at least two patients were chosen and samples with no mutation were removed, resulting in 170 samples and 154 gene mutations. For gene expression, we chose the 500 genes with highest ranked-based coefficients of variation and standard deviation across these samples (details of samples, mutations and chosen genes are given in Supplementary Table S1). Up and down regulated genes in each cluster were analysed using GenePattern 2.0 (38).

RESULTS

Simulated data

The results of applying the method to simulated data, using 100–1000 data points and 5–100 mixture components are shown in Figure 1 for simulated data with a low level of variability (tight, well separated clusters) and in Supplementary Figure S1 for data with higher variability and less well-defined clusters. Figure 1 shows that for smaller numbers of data points (100–200) the optimization algorithm successfully finds the correct solution and that this is relatively insensitive to the chosen objective function. Only the very low penalty functions ($\lambda = 1.0, 1.5$) generate solutions with lower objective values and more mixture components than the underlying distribution from which the data were generated. With more data points to cluster the optimization procedure finds solutions equal or very similar to the correct solution for $\lambda = 2.0, 2.5$, encompassing the standard AIC and slightly higher penalties, which might be expected on theoretical grounds. However for stronger penalties, including those with N dependency, solutions with higher objective function values and too few clusters are found. We note that this seems to be a failure of the optimization method rather than the objective function, and suggest that it reflects optimization on a surface where the likelihood gives limited 'downhill' information compared to the strong penalty on parameter numbers. Using data simulated with higher variability and less well defined clusters (Supplementary Figure S1) leads to similar conclusions: values of $\lambda = 2.0, 2.5$ yield the best solutions over a range of problem sizes. In this case, some of the higher penalty criteria fail at the level of the objective function (solutions with too few clusters have lower objective values than the correct solutions) rather than optimization method. Overall, these results with simulated data suggest that the method is most successful with AIC type objective functions without N dependency on the penalty term, and that the actual AIC ($\lambda = 2.0$) is an effective choice with simulations suggesting the use of slightly higher penalties for small data sets. Further work with larger simulated data sets (not shown) gave similar results, suggesting as expected that λ values close to 2.0 should be used.

Application to genetic regulation in the yeast cell cycle

We applied our method to the cell cycle data using parameters suggested from the simulation study. Some example clusters are shown in Figure 2 and all clusters are in Supplementary Figure S2, in these cases employing the standard AIC ($\lambda = 2.0$). We also examined the effect of refinement of the clusters using the EM algorithm and results for the marginal densities are shown in Supplementary Figure S3: these results reveal little overlap of the clusters: no genes have significant probability of membership of clusters other than the one assigned by the heuristic search and parameter estimates for the clusters did not change significantly after refinement. Figure 2A illustrates cluster 67, within which genes have a clear cell cycle related expression pattern and linked regulation by three TFs with high probability, corresponding to a clear regulatory hypothesis for this group of genes. On the other hand Figure 2B shows a cluster with no clear TF binding or gene expression pattern, and this case we consider that the cluster contains limited information about the expression and regulation of the corresponding genes. Based on these observations, we chose to analyse our data by first extracting clusters where the expression and binding pattern is clear (average Pearson correlation of expression patterns > 0.5 and at least one TF with binding probability > 0.5). In this case, of the 76 clusters produced (Supplementary Figure S2), 52 met these criteria and we refer to these subsequently as 'clear' clusters.

In Table 1 we show some statistics of different approaches to clustering this data set, where the results in Figure 2 ($\lambda = 2.0$) correspond to the first column. The number of clusters found by each method follows expected patterns: using a stronger penalty term ($\lambda = 2.5$) results in fewer and larger clusters, although this effect is more pronounced that it was in the simulated data examples. Normalizing gene expression also results in fewer clusters. Clustering with expression data only results in fewer larger clusters, indicating that gene expression patterns may be shared when regulation is different and underlining the advantage of a joint clustering approach incorporating regulation. Regarding the results from LeTICE, it should be appreciated that this method assigns genes to 'regulated modules' or to the 'background', a process that corresponds to our approach of focusing on clusters with clear regulation and gene expression. The 14

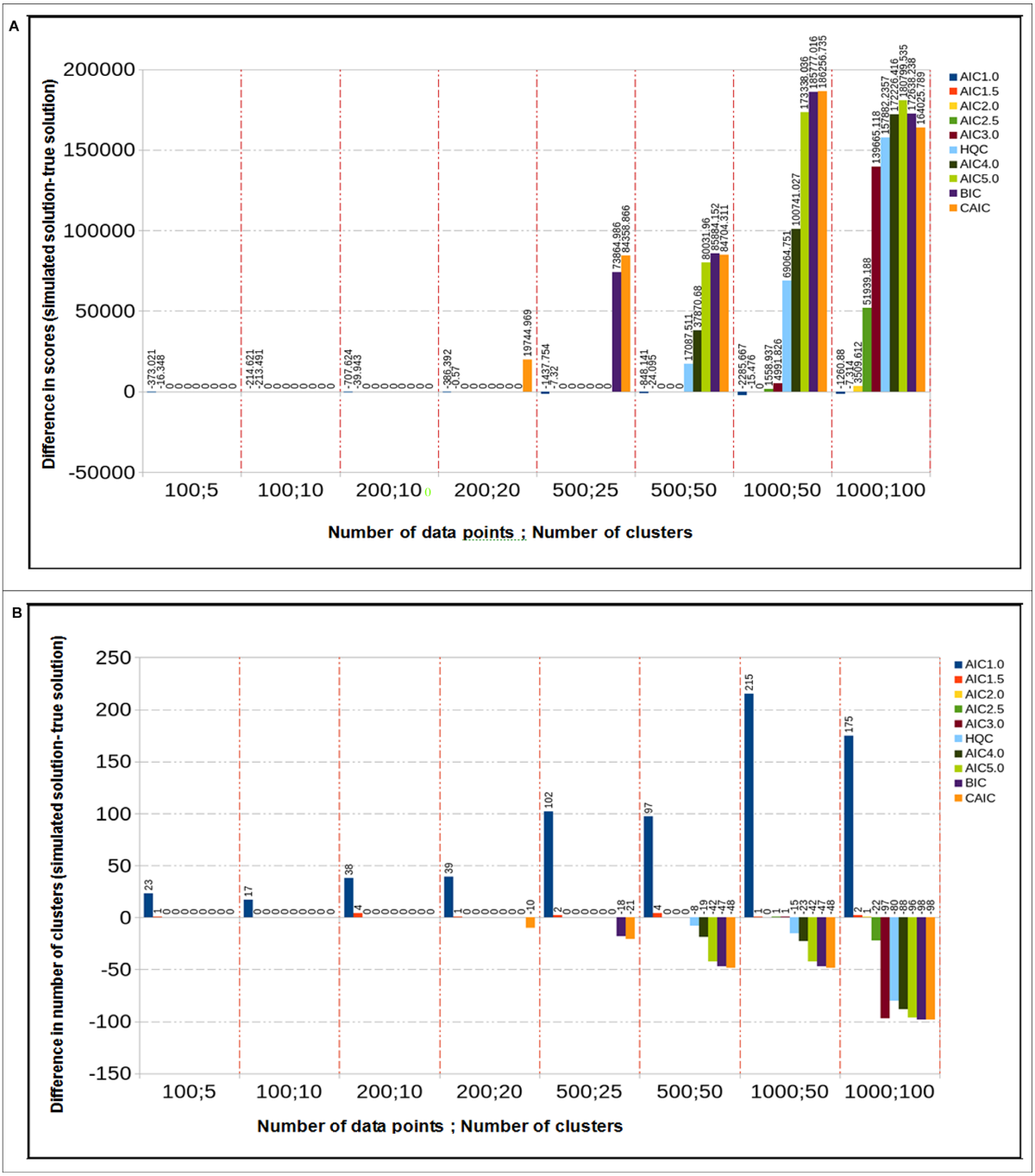

Figure 1. Result from our algorithm using a data set simulated from the probability distribution assumed in the paper for $n_r = 20$ binary variables and $n_e = 20$ continuous variables. In this case, parameters of the simulation correspond to tightly clustered data and relatively little 'noise' (Bernoulli parameters of 0.1 or 0.9 at each regulatory input and expression standard deviations of 0.01). Cases simulated covered 100–1000 data points and 10 or 20 data points per cluster in each case. Panel (**A**) shows the difference in score, and panel (**B**) the difference in the number of clusters, between the solutions found by the algorithm and the known true solutions. Results are shown for several objective functions arranged in order of increasing penalty value λ. Differences of zero in each case indicate that the algorithm found the true solution; negative score differences indicate objective function failures (solutions different to the true solution exist with better scores), and positive score differences indicate search algorithm failure (algorithm stopped at a solution scoring worse than the true solution).

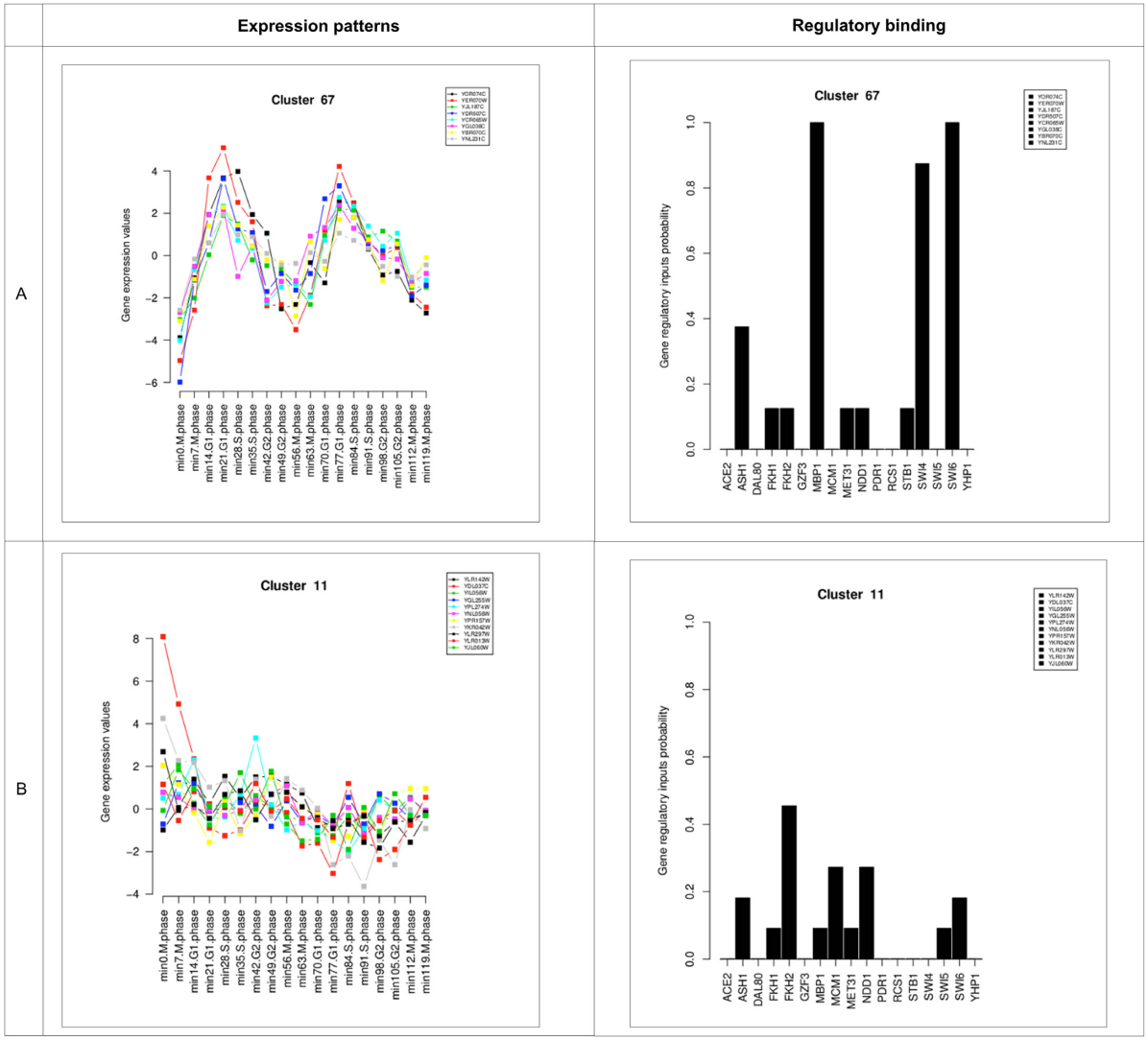

Figure 2. Two examples of clusters showing both expression patterns and regulatory binding patterns generated using the standard AIC objective function and normalized gene expression patterns. Panel A shows expression and regulatory binding patterns plots for genes in cluster 67: here a clear gene expression pattern is associated to a clear regulatory hypothesis involving high probability binding by Mbp1, Swi4 and Swi6. On the contrary, a clear regulatory hypothesis could not be made for cluster 11 in panel B: these genes do not have a very clear cell cycle expression pattern nor do they show a high probability of binding any transcription factor.

clusters produced by LeTICE can be viewed therefore as similar to our results for clear clusters using ($\lambda = 2.5$) and normalized expression data (15 clusters).

We also considered which of the TFs were assigned significant regulatory roles, i.e. which TFs appear with binding probability >0.5 in at least one 'clear' cluster. As expected these were more numerous in methods that produce more and smaller clusters. We observe that all nine of the well-known yeast cell cycle TFs (23), were assigned regulatory roles by our methods using any objective function tested, and with LeTICE this was not the case. Equally in expression only clustering fewer TFs were assigned roles, again emphasising that genes may be co-expressed with different regulation.

We measured functional coherence of the clusters using the average semantic similarity of Gene Ontology annotations of the clustered genes (Table 1). By this measure, most methods produce clusters that are significantly better than a random assignment of genes to clusters of the same size distribution. For neither AIC type objective function is there any strong evidence of a difference in results based on normalisation of the expression data. Finally, based on GO criteria and the implication of more TFs in regulatory roles, we marginally preferred AIC ($\lambda = 2.0$) with normalised expression and our subsequent analysis is based on these clusters.

The regulatory networks implied by the clustering are shown in Figure 3, first connecting TFs to their regulated clusters (upper panel) and then focusing on connections between all TFs and known cell cycle regulators (lower panel). It is notable that our choice of AIC as objective function produces a relatively large number of clusters, some of which are quite small. However, we note that even very small clusters, for example clusters 40, 42, 47, 48 and 49 containing 2–3 genes each, have clear regulation (see Supplemen-

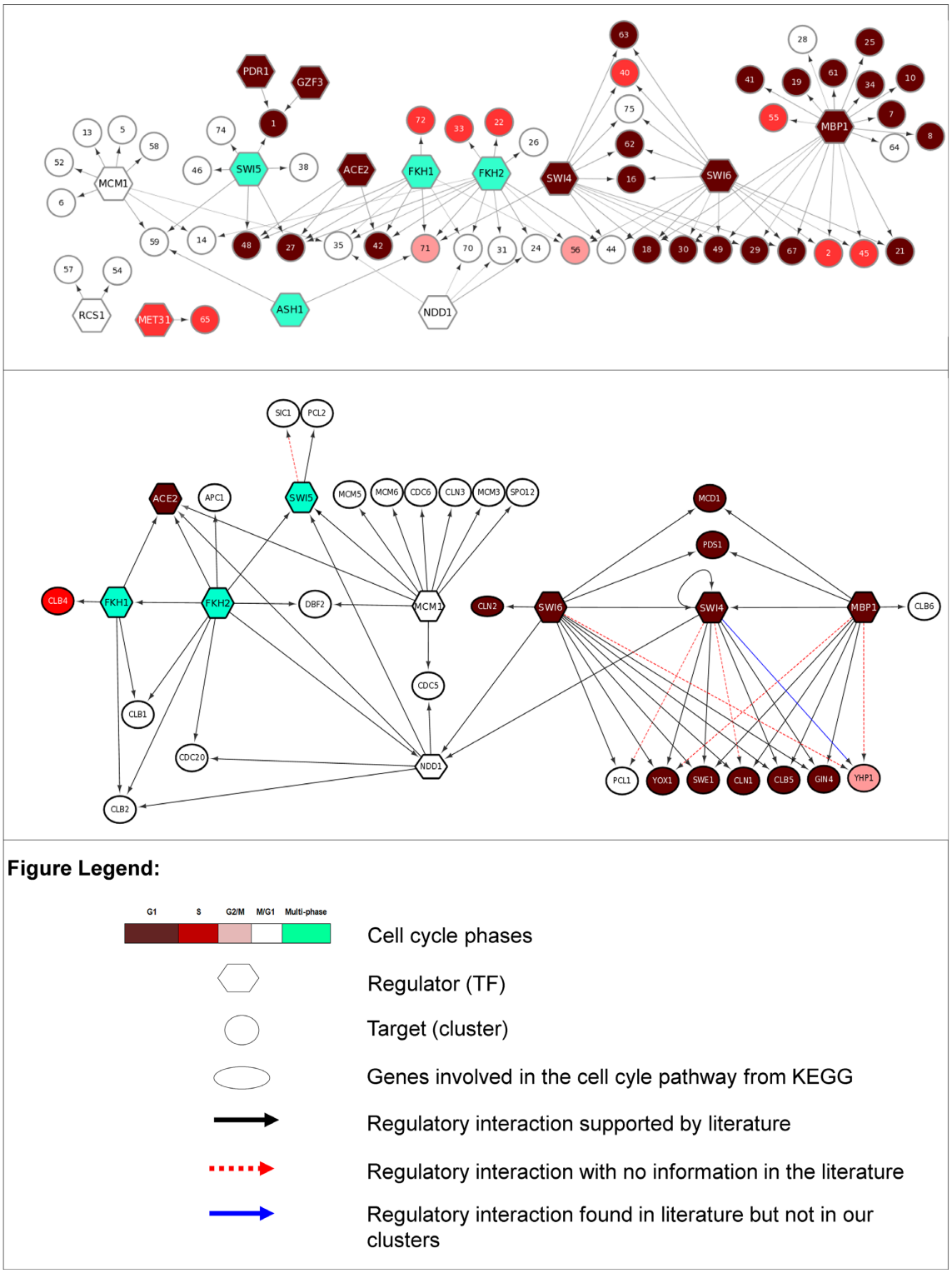

Figure 3. *Top*. The transcriptional regulatory network obtained from clusters with clear regulation and expression using the AIC objective function. The hexagonal nodes represent transcription factors and circular nodes regulated clusters (labeled 1–76, only clear clusters shown): colors represent cell cycle phases (peak expression phase for clusters, and the main phase of the regulated clusters for each transcription factor). *Bottom*. The regulatory network of transcription factors and other regulators extracted from the above network. Transcription factors shown are those associated by our algorithm to the regulation of clear clusters; other cell cycle regulators were identified in our gene set and overlapped with cell-cycle pathway map in KEGG (19).

Table 1. Statistics of clusters found by joint clustering of regulation and expression with different objective functions AIC ($\lambda = 2.0$) and AIC2.5 ($\lambda = 2.5$), with and without normalization of gene expression, compared to using LeTICE and using expression alone. Gene symbols in red are the nine well known yeast cell cycle transcription factors

	AIC normalized expression	AIC un-normalized expression	AIC2.5 normalized expression	AIC2.5 Un-normalized expression	LeTICE	Expression only
Number of clusters	76	91	23	33	14	19
Number of clear[1] clusters (total genes)	52 (236)	64 (252)	15 (225)	26 (272)	14 (136)	14 (252)
Average (+/- s.d.) size of clear[1] clusters	5 ± 3	4 ± 3	15 ± 6	11 ± 10	10 ± 5	34 ± 14
TFs found as candidate regulators for clear[1] clusters	ACE2 ASH1 FKH1 FKH2 GZF3 MBP1 MCM1 MET31 NDD1 PDR1 RCS1 SWI4 SWI5 SWI6	ACE2 ASH1 FKH1 FKH2 GZF3 MBP1 MCM1 MET31 NDD1 PDR1 RCS1 STB1 SWI4 SWI5 SWI6	ACE2 FKH1 FKH2 GZF3 MBP1 MCM1 NDD1 PDR1 SWI4 SWI5 SWI6	ACE2 FKH1 FKH2 GZF3 MBP1 MCM1 MET31 NDD1 PDR1 SWI4 SWI5 SWI6	BAS1 FKH2 HAP4 MBP1 NDD1 STP1 SWI4 SWI5 SWI6	ACE2 FKH1 FKH2 MBP1 MCM1 SWI5 SWI6
Average GO Semantic Similarity (Mean+/-s.d. for random clusters)	0.34* (0.27+/-0.02)	0.32* (0.25+/-0.02)	0.32* (0.26+/-0.02)	0.30* (0.25+/-0.02)	0.25 (0.24+/-0.01)	0.33* (0.24+/-0.02)

[1]'Clear' clusters have clear expression patterns (average pairwise Pearson correlation of expression > 0.5) and clear regulation (at least one transcription factor with binding probability > 0.5). Such TFs in clear clusters are considered candidate regulators for the cluster.

*Statistically significantly different from the values obtained by random assignment of genes to clusters with the same size distribution, $p < 0.01$.

tary Figure S2) and contain genes with related functions. These clusters have statistically significant functional enrichment in cellular budding (clus. 48), drug transport (clus. 47), chromatin assembly and disassembly, cell wall organisation (clus. 40), response to pheromones and sexual reproduction (clus. 42). Equally, there are often several clusters which are related in expression and regulation, for instance clusters 18, 29, 30, 49 and 67 whose expression patterns all peak in G1 phase and all show a high probability of regulation by the TFs SWI4, SWI6 and MBP1. These separate clusters have clearly different GO annotations: regulation of transcription (clus. 18), organelle fission and nuclear division (clus. 29), conjugation with cellular fusion (clus. 30), regulation of protein kinase activity/cell division/bud site selection (clus. 49) and deoxyribonucleotide biosynthetic processes (clus. 67), and their separation reflects differences of detail in the expression pattern and regulatory probabilities.

The regulation of the yeast cell cycle has been extensively studied both experimentally and in the context of algorithms aimed at reconstruction of the network from different sources of data (see (39) for a recent review). Here, we offer a very brief discussion of our results in that context. The regulatory relationships in Figure 3 are largely known, and most of the regulatory relationships in the lower panel are supported, as shown, by evidence from the literature. We have already commented on regulation of G1 phase genes by SWI4/SWI6/MBP1 which together form the heterodimeric transcription factors MBF and SBF, and note that our method also finds the known regulation of G1 phase cyclins CLN1, CLN2, CLB5, CLB6, SWE1 and GIN4 by these factors. Regulation of S phase genes also by SBF and MBF (40), particularly histones and genes associated with chromatin organization, is evident in clusters 2, 40 and 55, while other S phase clusters (clus. 33, 72) are regulated by FKH1 and FKH2. Moving to G2M and M phase, while SBF/MBF still participate in regulation, it becomes dominated by MCM1, NDD1, FKH1 and FKH2. In particular our algorithm finds regulation of the key cyclins CLB1 and CLB2 by FKH1/2 and NDD1, but does not discover known links to SBF or MCM1 (41,42). We note also the interesting disconnected component in Figure 3, cluster 65 (see Supplementary Figure S2) being regulated by MET31, comprising genes associated with S-adenosylmethionine metabolism which has been linked to cell cycle control (43,44).

Combinatorial regulation of genes by multiple TFs is known to be important and several of our clusters exhibited high probability binding by more than one TF (e.g. regulation by SWI4, SWI6 and MBP1 in Figure 2A). Such multiple regulation implies possible interaction between the factors concerned and in Supplementary Figure S4 we summarise genetic and physical interaction evidence supporting combinatorial interactions in our clusters. All but one identified combinatorial interaction is supported by some evidence from BioGRID (27), and most have extensive support. Finally we note that of the regulatory interactions predicted between TFs and genes within our clusters, only 34% are supported by significant correlation between those genes' expression patterns and the expression patterns of the regulating factors. Although this percentage increases if correlations off-set in time are considered, it shows that simple correlation of expression is not a good way of predicting regulation.

Application to acute myeloid leukaemia data

Based on our findings with simulated data, we investigated clustering of this mutation and expression data using the AIC related criteria with $\lambda = 2.0$ and 2.5. Again clustering with this real data set showed greater variability in results between these two penalty functions than was evident in simulations, with clusters predominantly very small (two samples) from $\lambda = 2.0$. Accordingly we chose $\lambda = 2.5$ in this case on biological grounds: the results are shown graphically in Figure 4A, and Supplementary Table S2 lists the up- and down-regulated genes in each cluster along with the associated gene mutations and probabilities. Figure 4B shows survival curves for larger clusters. Overall, it is clear that the method is able to discover clusters of samples where characteristic mutation patterns are associated with distinct patterns of gene expression; these are potentially related to different oncogenic mechanisms and show statistically significant differences in survival.

Cytogenetic abnormalities are well known in AML with several well-known translocations and associated gene fusions, and this is evident in several of the clusters our method produces. For example cluster 12 (Figure 4A) has been identified by our method because the fused PML and RARA genes are both classed as mutated in this data, and this pair of mutations is associated with a distinct and characteristic pattern of gene expression. We note a single sample in the middle of this cluster that has these mutations and gene expression pattern but is not annotated with the cytogenetic abnormality: this appears to be an annotation error. In a similar way, clusters 1, 8 and 9 are associated with cytogenetic abnormalities and MLL (alias KMT2a)/ELL, RUNX1/RUNX1T1 and CBFB/MYH11 translocations, respectively (29), and clusters 8, 9 and 12 are also associated with survival differences in this data (Figure 4B). These clusters stand in contrast to clusters 10, 15 and 18 which have distinct patterns of gene expression but are not associated with any mutations at high probability, illustrating that the method is robust to discover gene expression based clusters without an associated mutational pattern in the data.

Perhaps more interesting than cytogenetic abnormalities are clusters where other gene mutations are co-clustered with distinctive gene expression patterns: for example cluster 17 where a distinctive pattern of gene expression is associated with mutation of the transcription factor CEBPA, and the smaller clusters 11 and 14 which are associated with mutations in TP53 and STAG2 respectively. Clusters 4, 5, 13 and 16 are all larger clusters associated with mutation of NPM1 and to a greater or lesser extent the associated dysregulation of HOX genes (45) (Supplementary Table S2), but nevertheless have distinctively different patterns of gene expression. Cluster 4 is very strongly associated with HOX gene up-regulation and the NPM1 mutation is coupled with high probability of both FLT3 and DNMT3A mutation. Cluster 5 is similar, with a lower probability of DNMT3A mutation (<0.5), and clusters 13 and 16 are associated with more dysregulated genes and an absence of FLT3 mutations. It has been suggested that AML arises from three complementary classes of mutation (46,47): class I mutations in tyrosine kinases including FLT3; class II mutations in transcription factors including RUNX1, CEBPA and NPM1; class III mutations in genes associated with DNA methylation including DNMT3A; and, another class associated with tumor suppressor mutations. It is noteworthy that these clusters mix mutations from different classes, which along with the different patterns of gene expression emphasises mechanistic heterogeneity within the broad class of NPM1 mutated cases. The survival analysis (Figure 4B) shows that clusters 4 and 5 have much worse prognosis than 13 and 16, suggesting that the clusters may have clinical as well as mechanistic relevance and that appropriate biomarkers would combine both mutation and gene expression information.

DISCUSSION

The results above illustrate that a generic approach to clustering entities described by a mixture of binary and continuous variables is potentially useful in a wide range of applications with large data sets in molecular biology. While this could have been approached in several different ways, the choice of the probabilistic model has the advantage of identifying key variables in each cluster, for example associating mutations with high probability to gene expression patterns or identifying the most likely regulating transcription factors. It also provides a natural treatment of a low level of false positives and false negatives that can affect high-throughput data in both the examples given, and probably in many similar types of high-throughput biological data.

All clustering problems face the question of how many clusters. We approached this through model selection using well-known criteria and their variants, but mindful of their approximate nature and also of concerns about applicability in the case of mixture models (19) investigated these in detail using simulated data. It is interesting that this investigation concluded with a preference for criteria close to the well-known AIC, and that in both examples given this carried through to real data sets at least in generating clusters with clear biological meaning.

In the application to genetic regulation in yeast we note that the method produces results that to a large extent recapitulate existing knowledge. We chose to compare to LeTICE (25) as a recent method based on a similar premise

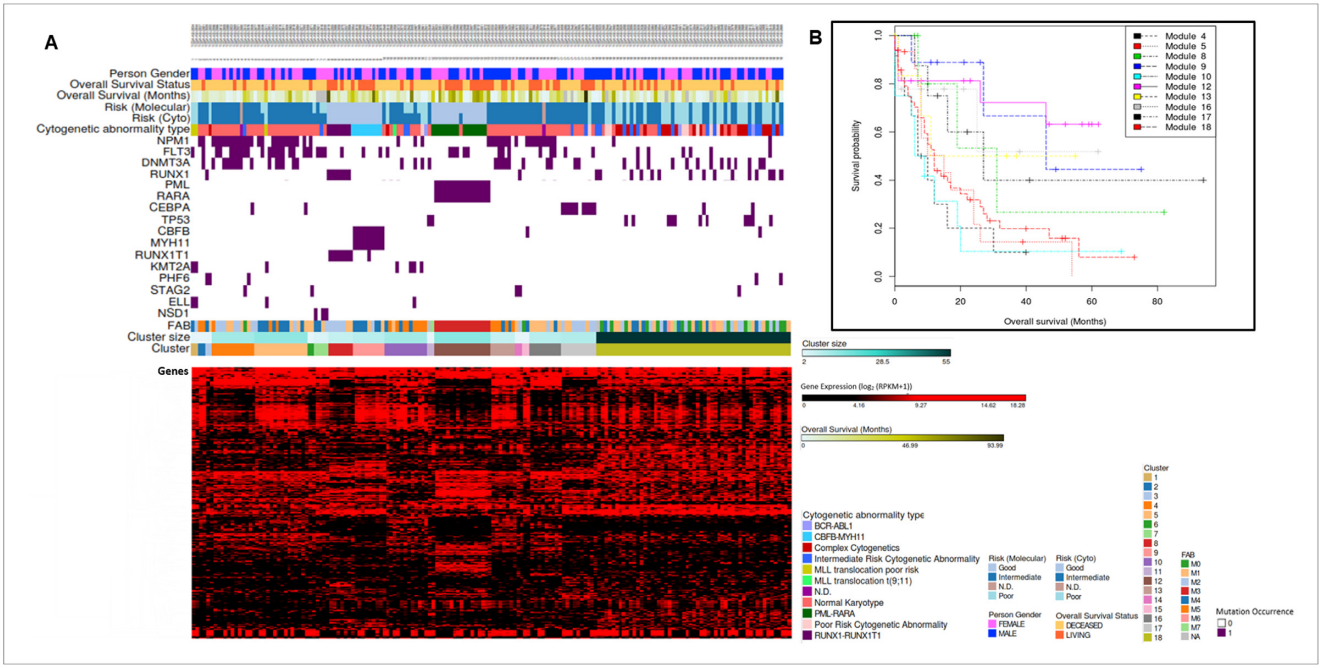

Figure 4. (**A**) Clustering of AML samples shown in columns of 170 samples using AIC ($\lambda = 2.5$) across the most variably expressed 500 genes (lower) and the 18 mutations (above, dark purple bar shows a mutation). Other relevant variables are also shown to aid interpretation (but were not used in clustering). (**B**) Kaplan–Meier estimators for the 10 clusters with more than two samples with survival information available in each cluster. The 10 Kaplan–Meier estimators perform differently with a significant *P*-value in the log-rank test, $P = 0.001$.

but otherwise methodologically very distinct. In our hands and on this data set, our method produced arguably better results. However, our view is that comparison of methods should be done independently of the authors of those methods. Accordingly we do not claim better performance, but simply take this as evidence that our method performs at least as well. In this application we suggest that limitations are not methodological, but associated with the limited nature of the data. The transcription factor binding data is not resolved by time or cell cycle phase, and this limits how well any method could perform.

We view clustering as a method of exploratory data analysis that can be used to generate hypotheses based on data, illustrated in this case as hypotheses about the regulation of groups of genes or mechanistic links between mutations and gene expression patterns in cancer samples. Our method is generic and we hope applicable beyond the examples given here. For example it could be applied without modification to the presence/absence of specific mutations in genes (rather than mutated or not), to time/cell cycle resolved transcription factor binding data and to other types of continuous data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge useful discussions with other members of the Leeds Bioinformatics group, and useful input from Drs J. Boyes and A. Peel.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC) [BB/I001220/1 to D.R.W.]; Bloodwise (formerly Leukaemia and Lymphoma Research [13052, 15002]; Universiti Kebangsaan Malaysia (to F.N.Z.A.). Funding for open access charge: BBSRC and Bloodwise.
Conflict of interest statement. None declared.

REFERENCES

1. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
2. Si,Y., Liu,P., Li,P. and Brutnell,T.P. (2014) Model-based clustering for RNA-seq data. *Bioinformatics*, **30**, 197–205.
3. Dunham,I., Kundaje,A. and Birney,E. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
4. Goode,D.K., Obier,N., Vijayabaskar,M.S., Lie,A.L.M., Lilly,A.J., Hannah,R., Lichtinger,M., Batta,K., Florkowska,M., Patel,R. *et al.* (2016) Dynamic gene regulatory networks drive hematopoietic specification and differentiation. *Dev. Cell*, **36**, 572–587.
5. Morlini,I. (2012) A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Adv. Data Anal. Classif.*, **6**, 5–28.
6. McParland,D. and Gormley,I.C. (2016) Model based clustering for mixed data: clustMD. *Adv. Data Anal. Classif.*, **10**, 155–169.
7. Browne,R.P. and McNicholas,P.D. (2012) Model-based clustering, classification, and discriminant analysis of data with mixed type. *J. Stat. Plan. Infer.*, **142**, 2976–2984.
8. Cai,J.H., Song,X.Y., Lam,K.H. and Ip,E.H.S. (2011) A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Comput. Stat. Data An.*, **55**, 2889–2907.
9. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with

application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
10. Kim,S., Herazo-Maya,J.D., Kang,D.D., Juan-Guardela,B.M., Tedrow,J., Martinez,F.J., Sciurba,F.C., Tseng,G.C. and Kaminski,N. (2015) Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics*, **16**, 924.
11. Mason,S.A., Sayyid,F., Kirk,P.D., Starr,C. and Wild,D.L. (2016) MDI-GPU: accelerating integrative modelling for genomic-scale data using GP-GPU computing. *Stat. Applic. Genet. Mol. Biol.*, **15**, 83–86.
12. Kirk,P., Griffin,J.E., Savage,R.S., Ghahramani,Z. and Wild,D.L. (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.
13. Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.
14. Akaike,H. (1974) New look at statistical-model identification. *IEEE Trans. Automatic Control*, **19**, 716–723.
15. Schwarz,G. (1978) Estimating Dimension of a Model. *Ann.Stat.*, **6**, 461–464.
16. Bozdogan,H. (1987) Model selection and akaike information criterion (Aic)—the general-theory and its analytical extensions. *Psychometrika*, **52**, 345–370.
17. Hannan,E.J. and Quinn,B.G. (1979) Determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B-Methodol.*, **41**, 190–195.
18. Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
19. Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach*. 2nd edn. Springer-Verlag, NY.
20. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
21. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
22. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
23. Simon,I., Barnett,J., Hannett,N., Harbison,C.T., Rinaldi,N.J., Volkert,T.L., Wyrick,J.J., Zeitlinger,J., Gifford,D.K., Jaakkola,T.S. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.
24. Tsai,H.K., Lu,H.H. and Li,W.H. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13532–13537.
25. Youn,A., Reiss,D.J. and Stuetzle,W. (2010) Learning transcriptional networks from the integration of ChIP-chip and expression data in a non-parametric model. *Bioinformatics*, **26**, 1879–1886.
26. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.
27. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
28. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
29. Cancer Genome Atlas Research Network (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.
30. Tomczak,K., Czerwinska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68–77.
31. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
32. Ross,M.E., Mahfouz,R., Onciu,M., Liu,H.C., Zhou,X., Song,G., Shurtleff,S.A., Pounds,S., Cheng,C., Ma,J. *et al.* (2004) Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, **104**, 3679–3687.
33. Valk,P.J., Verhaak,R.G., Beijen,M.A., Erpelinck,C.A., Barjesteh van Waalwijk van Doorn-Khosrovani,S., Boer,J.M., Beverloo,H.B., Moorhouse,M.J., van der Spek,P.J., Lowenberg,B. *et al.* (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *N. Engl. J. Med.*, **350**, 1617–1628.
34. Renneville,A., Roumier,C., Biggio,V., Nibourel,O., Boissel,N., Fenaux,P. and Preudhomme,C. (2008) Cooperating gene mutations in acute myeloid leukemia: a review of the literature. *Leukemia*, **22**, 915–931.
35. Becker,H., Maharry,K., Mrozek,K., Volinia,S., Eisfeld,A.K., Radmacher,M.D., Kohlschmidt,J., Metzeler,K.H., Schwind,S., Whitman,S.P. *et al.* (2014) Prognostic gene mutations and distinct gene- and microRNA-expression signatures in acute myeloid leukemia with a sole trisomy 8. *Leukemia*, **28**, 1754–1758.
36. Cerami,E., Gao,J.J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
37. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.
38. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
39. Haase,S.B. and Wittenberg,C. (2014) Topology and control of the cell-cycle-regulated transcriptional circuitry. *Genetics*, **196**, 65–90.
40. Eriksson,P.R., Ganguli,D., Nagarajavel,V. and Clark,D.J. (2012) Regulation of histone gene expression in budding yeast. *Genetics*, **191**, 7–20.
41. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
42. Zhu,G., Spellman,P.T., Volpe,T., Brown,P.O., Botstein,D., Davis,T.N. and Futcher,B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.
43. Mizunuma,M., Miyamura,K., Hirata,D., Yokoyama,H. and Miyakawa,T. (2004) Involvement of S-adenosylmethionine in G1 cell-cycle regulation in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6086–6091.
44. Su,N.Y., Ouni,I., Papagiannis,C.V. and Kaiser,P. (2008) A dominant suppressor mutation of the met30 cell cycle defect suggests regulation of the Saccharomyces cerevisiae Met4-Cbf1 transcription complex by Met32. *J. Biol. Chem.*, **283**, 11615–11624.
45. Alharbi,R.A., Pettengell,R., Pandha,H.S. and Morgan,R. (2013) The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia*, **27**, 1000–1008.
46. Dombret,H. (2011) Gene mutation and AML pathogenesis. *Blood*, **118**, 5366–5367.
47. Naoe,T. and Kiyoi,H. (2013) Gene mutations of acute myeloid leukemia in the genome era. *Int. J. Hematol.*, **97**, 165–174.