Vol. 00 no. 00 2005
Pages 1–8

Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models - Supplementary Information

Simon Rogers [1,*], Mark Girolami[1],Walter Kolch[2,3],Katrina M. Waters[4],Tao Liu[4],Brian Thrall[4] and H. Steven Wiley[4,5] *

[1]Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK
[2]Beatson Institute for Cancer Research, Signalling and Proteomics Laboratory, Garscube Estate, Glasgow, G61 1BD, UK
[3]Institute of Biomedical and Life Sciences, Sir Henry Wellcome Functional Genomics Facillity, University of Glasgow, G12 8QQ, UK [4]Systems Biology Program, Pacific Northwest National Laboratory, Richland, WA 99352, USA [5]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation: Modern transcriptomics and proteomics enable us to survey the expression of RNAs and proteins at large scales. While these data are usually generated and analysed separately, there is an increasing interest in comparing and co-analysing transcriptome and proteome expression data. A major open question is whether transcriptome and proteome expression is linked and how it is coordinated.
Results: Here we have developed a probabilistic clustering model that permits analysis of the links between transcriptomic and proteomic profiles in a sensible and flexible manner. Our coupled mixture model defines a prior probability distribution over the component to which a protein profile should be assigned conditioned on which component the associated mRNA profile belongs to. By providing probabilistic assignments this approach sits between the two extremes of concatenating the data on the assumption that mRNA and protein clusters would have a one-to-one relationship, and independent clustering where the mRNA profile provides no information on the protein profile and vice-versa. We apply this approach to a large dataset of quantitative transcriptomic and proteomic expression data obtained from a human breast epithelial cell line (HMEC) stimulated by epidermal growth factor (EGF) over a series of timepoints corresponding to one cell cycle. The results reveal a complex relationship between transcriptome and proteome with most mRNA clusters linked to at least two protein clusters, and vice versa. A more detailed analysis incorporating information on gene function from the gene ontology database shows that a high correlation of mRNA and protein expression is limited to the components of some molecular machines, such as the ribosome, cell adhesion complexes and the TCP-1 chaperonin involved in protein folding.
Conclusions: The dynamic regulation of the transcriptome and proteome in mammalian cells in response to an acute mitogenic stimulus appears largely independent with very little correspondence between mRNA and protein expression. The exceptions involve a few selected multi-protein complexes that require the stoichiometric expression of components for correct function. This finding has wide ramifications regarding the understanding of gene and protein expression including its control and evolution. It also shows that transcriptomic and proteomic expression analysis are complementary and non-redundant.

1 ADDITIONAL MODEL DETAILS, PARAMETER INFERENCE AND IMPLEMENTATION DETAILS

1.1 Covariance function parameterisation

We do not work with full covariance matrices. Rather we assume diagonal covariance matrices with a single variance parameter ($\mathbf{\Sigma} = \sigma^2\mathbf{I}$). We are thus assuming that there is no time-correlation and that the level of noise remains unchanged throughout the time series. Incorporating correlation over time is clearly an interesting area for future development. However, fitting full covariance matrices with the quantity of data available was not feasible and so investigating alternative approaches is necessary (for example, one of the methods discussed above). The assumption that the level of noise is constant through time seems rather less limiting. It might be argued that we should expect noise to increase during the experiment as cells lose their synchronisation - however, we found that parameterising the covariance matrix with a separate variance parameter for each time point or weighting the variance such that it increases over time ($\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_T^2)$ or $\mathbf{\Sigma} = diag(\sigma^2, a_1\sigma^2, \ldots, a_{T-1}\sigma^2)$ with $a_{T-1} \geq a_{T-2} \geq \cdots \geq a_1$) made no qualitative difference to the results obtained.

[*]to whom correspondence should be addressed

1.2 Parameter inference

The expectation-maximisation (EM) algorithm Dempster *et al.* (1977) can be used to find model parameters and cluster assignments corresponding to a local maximum of the model likelihood. The log likelihood is given by

$$\log p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\Delta}_1^m, \ldots, \boldsymbol{\Delta}_K^m, \boldsymbol{\Delta}_1^p, \ldots, \boldsymbol{\Delta}_J^p)$$
$$= \sum_{g=1}^{G} \log \left[\sum_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}_g^m|\boldsymbol{\mu}_k, \sigma_k^2) \sum_{j=1}^{J} \theta_{jk} \mathcal{N}(\mathbf{x}_g^p|\boldsymbol{\mu}_j, \sigma_j^2) \right) \right] \quad (1)$$

where we have omitted the m and p superscripts from the model parameters - $\boldsymbol{\mu}_k$ and σ_k^2 will always refer to parameters from the mRNA mixture and vice-versa. To derive the necessary EM update equations, we first introduce the variational distribution Q_{gk} such that $\sum_k Q_{gk} = 1$ that can be interpreted as the mRNA cluster membership probabilities for the gth gene. Introducing these distributions and applying Jensen's inequality produces the lower bound on the log likelihood given in equation 2.

Because of the summation within the log in the final term of equation 2, we must introduce a second set of variational distributions, γ_{jkg} such that $\sum_j \gamma_{jkn} = 1$ that can be interpreted as the probability that the protein profile for gene g is in cluster j given that its mRNA profile is in cluster k. Our final bound is given in equation 3. . Taking partial derivatives of this bound with respect to the variational parameters produces the following two updates that make up the E-step of our algorithm

$$\gamma_{jkg} = \frac{\theta_{jk}\mathcal{N}(\mathbf{x}_n^p|\boldsymbol{\mu}_j, \sigma_j^2)}{\sum_{j'} \theta_{j'k}\mathcal{N}(\mathbf{x}_n^p|\boldsymbol{\mu}_{j'}, \sigma_{j'}^2)} \quad (4)$$

$$Q_{gk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n^m|\boldsymbol{\mu}_k, \sigma_k^2)\left\{\sum_j \theta_{jk}\mathcal{N}(\mathbf{x}_n^p|\boldsymbol{\mu}_j, \sigma_j^2)\right\}}{\sum_{k'} \pi_{k'}\mathcal{N}(\mathbf{x}_n^m|\boldsymbol{\mu}_{k'}, \sigma_{k'}^2)\left\{\sum_j \theta_{jk'}\mathcal{N}(\mathbf{x}_n^p|\boldsymbol{\mu}_j, \sigma_j^2)\right\}} \quad (5)$$

It is important to note that in deriving 5, we have used 4 and so γ_{jkg} should always be updated before Q_{gk} in the E-step. The M-step consists of the following updates, derived by taking partial derivates of the bound with respect to the model parameters

$$\pi_k = \frac{1}{G}\sum_{g=1}^{G} Q_{gk} \quad (6)$$

$$\boldsymbol{\mu}_k = \frac{\sum_g Q_{gk}\mathbf{x}_g^m}{\sum_g Q_{gk}} \quad (7)$$

$$\sigma_k^2 = \frac{\sum_g Q_{gk}\sum_t (x_{gt}^m - \mu_{kt})^2}{T\sum_n Q_{gk}} \quad (8)$$

$$\theta_{jk} = \frac{\sum_g Q_{gk}\gamma_{jkg}}{\sum_g Q_{gk}} \quad (9)$$

$$\boldsymbol{\mu}_j = \frac{\sum_{g,k} Q_{gk}\gamma_{jkg}\mathbf{x}_g^p}{\sum_{g,k} Q_{gk}\gamma_{jkg}} \quad (10)$$

$$\sigma_j^2 = \frac{\sum_{g,k} Q_{gk}\gamma_{jkg}\sum_t (x_{gt}^p - \mu_{jt})^2}{T\sum_{g,k} Q_{gk}\gamma_{jkg}}, \quad (11)$$

where the subscript t denotes the tth element of a vector and T is the total number of timepoints (i.e. the length of $\mathbf{x}$).

To summarise, the EM algorithm for the coupled mixture model is

1. Initialise the model parameters (more details provided in the next section)
2. Perform the E-step described by equations 4 and 5
3. Perform the M-step described by equations 6 to 11
4. Compute the change in the value of the lower bound on the log likelihood given in 3. If this is less than a predetermined threshold, stop, otherwise return to 2.

Initialisation and re-starting

The EM algorithm finds a local maximum of the likelihood function and as such is sensitive to initial conditions. To overcome this problem, we ran the algorithm with 100 different random initialisations and kept the one that gave the highest value of the lower bound on the log likelihood.

Reproducibility

Testing the reproducibility of the results is not easy due to the symmetry of the likelihood to permutations of the component labels (j and k). To ensure that the results generated by the model were reproducible, we tested the consistency of the GO terms that were found to be enriched. Using 100 random initialisations, we extracted GO terms that were significantly ($p <= 0.1$) enriched in either mRNA clusters, protein clusters or both and then computed the frequency of occurrence of each term. Approximately 50 terms were enriched for each initialisation which, when combined, results in 473 unique terms. Assuming independence across terms we can model the number of times we might expect a term to appear with the binomial distribution with success probability $50/473$. In figure 1 we show the observed frequency of occurence along with a curve produced from the binomial cdf. We can clearly see that a vast number of terms (~ 100) are observed more frequently than one would expect at random. Under the binomial assumption, the probability of a particular term appearing in all 100 initialisations (we observe 8 such terms) is approximately 3×10^{-98}.

2 THE DATA

2.1 Cells

The human mammary epithelial cell line HMEC, strain 184A1, Stampfer and Yaswen (1993) was used in this study. This cell line is non-tumorigenic and resembles normal breast epithelial cells Stampfer and Yaswen (1993). Cells were routinely cultured in DHFB-I medium supplemented with 12.5 ng/ml EGF as described Band and Sager (1989). All other reagents were of cell culture grade or higher quality For experiments the cells were placed in culture medium without serum and growth factors for 48 hours in order to induce growth arrest and synchronise the population in the quiescent state G0. Proliferation and mitogenesis was induced by treatment with 10ng/ml EGF. This results in cells progressing into S-phase (doubling of DNA content) and mitosis, 12 and 18 hours, respectively, after EGF stimulation. Samples were taken at the 0, 0.25, 1, 4, 8, 13, 18 and 24 hrs and profiled for RNA and protein expression as described below.

$$
\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Delta}) &\geq \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\left(\frac{\pi_k}{Q_{gk}}\mathcal{N}(\mathbf{x}_g^m|\boldsymbol{\mu}_k,\sigma_k^2)\sum_{j=1}^{J}\theta_{jk}\mathcal{N}(\mathbf{x}_g^p|\boldsymbol{\mu}_j,\sigma_j^2)\right)\\
&\geq \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\pi_k - \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log Q_{gk} + \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\mathcal{N}(\mathbf{x}_g^m|\boldsymbol{\mu}_k,\sigma_k^2)\\
&+ \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\sum_{j=1}^{J}\theta_{jk}\mathcal{N}(\mathbf{x}_g^p|\boldsymbol{\mu}_j,\sigma_j^2). \qquad (2)
\end{aligned}
$$

$$
\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\Delta}) &\geq \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\pi_k - \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log Q_{gk} + \sum_{g=1}^{G}\sum_{k=1}^{K} Q_{gk}\log\mathcal{N}(\mathbf{x}_g^m|\boldsymbol{\mu}_k,\sigma_k^2)\\
&+ \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J} Q_{gk}\gamma_{jkg}\log\theta_{jk} - \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J} Q_{gk}\gamma_{jkg}\log\gamma_{jkg}\\
&+ \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{j=1}^{J} Q_{gk}\gamma_{jkg}\log\mathcal{N}(\mathbf{x}_g^p|\boldsymbol{\mu}_j,\sigma_j^2). \qquad (3)
\end{aligned}
$$

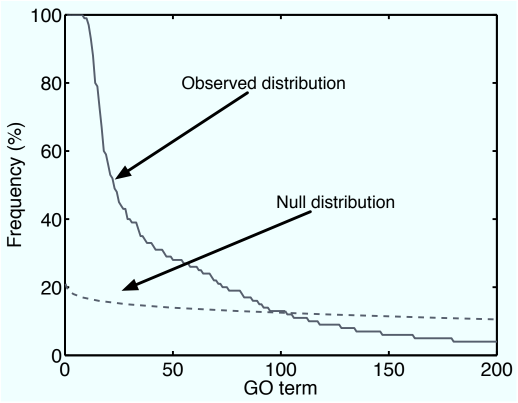


Fig. 1. Evaluating significance of GO enrichment. The solid curve shows the number of terms found in a particular proportion of restarts. The dashed curve depicts the null distribution assuming a binomial distribution (enriched v non-enriched).

2.2 Transcriptome Analysis

Cells were lysed and total RNA harvested using RNeasy (Qiagen, Valencia, CA). RNA expression profiles were generated using Nimblegen whole genome 60-mer oligonucleotide arrays (Design Version 2003_10_27) which contains 38,108 features (Nimblegen, Madison, WI), with hybridizations carried out in triplicates (3 technical replicates). Raw intensity data were processed by quantile normalisation Bolstad *et al.* (2003), pairwise analysis of variance using a significance level of $p < 0.01$ Kerr *et al.* (2004) and calculation of false discovery rate Benjamini and Hochberg (1995).

2.3 Proteomics Analysis

Protein expression was profiled using the accurate mass and time (AMT) tag approach Smith *et al.* (2002). Capillary liquid chromatography (LC) was used to separate peptides and Fourier-transform ion cyclotron resonance-mass spectrometry (FTICR-MS) was used for high mass accuracy measurements for peptide identification and quantification. An existing AMT tag database encompassing the monoisotopic masses and normalized chromatographic elution times of peptides identified from previous LC-MS/MS analyses of HMEC proteins under a range of experimental conditions Chen *et al.* (2003); Jacobs *et al.* (2004); Liu *et al.* (2005, 2004) was used as a base reference database for the LC-FTICR measurements in this study. The existing HMEC database was further enriched by conducting additional two dimensional LC-MS/MS analysis as described previouslyLiu *et al.* (2004); Qian *et al.* (2005), using $375\mu g$ of HMEC protein pooled from each of time point samples in this study. Proteins from each of the eight lysates were digested separately and differentially labeled using post-digestion trypsin-catalyzed 16O-to-18O exchange Liu *et al.* (2004). The control sample (0 hr) was labeled with 16O and all the other samples were labeled with 18O. Samples were analyzed using an Apex III 9.4-T FTICR mass spectrometer (Bruker Doltonics, Billerica, MA) and the LC-FTICR data analysis was conducted as previously described Qian *et al.* (2005). Briefly, the initial analysis of raw LC-FTICR data involved a mass transformation or deisotoping step using in house software (ICR2LS). The ICR2LS analysis generates a text file report for each LC-FTICR data set which includes the monoisotopic masses and the corresponding intensities for all detected species for each spectrum. In-house software (VIPER) was used to detect LC-MS features (i.e., a peak with unique mass and elution times) and assign them to peptides in the AMT database. Data processing steps included filtering data based on isotopic fitting, finding features and pairs of features, computing abundance ratios for pairs of features

(16O:18O), normalizing LC elution times, and matching the accurate measured masses (+ 5 ppm) and NET (+ 2%) values of each feature to the corresponding mass and time tag in the database to identify peptide sequences. All identified peptides were assigned an identical probability of 1.0 and entered into ProteinProphet software Nesvizhskii *et al.* (2003) to remove redundant proteins. Protein abundance ratios were calculated as an average of the peptide isotopic ratios after removing outliers using Grubb's test.

3 EXPERIMENTAL DETAILS

3.1 Data preparation and normalisation

mRNA and protein profiles were aligned by merging the REFSEQ mRNA accession with the protein PPI ID and REFSEQ accession using MATCHMINER Bussey *et al.* (2003). Positive matches were found for 1458 out of the 1687 proteins measured by LC-FTICR. These 1458 proteins correspond to 1595 mRNAs. Proteins with missing values were discarded, leaving 542 mRNA-protein pairs. One benefit of probabilistic clustering is the ability to handle missing values, however for this analysis, we decided to restrict ourselves to the complete data. The values at $t = 0$ were omitted from both datasets and, following Waters *et al.* (2008), the $t = 15$ minutes data was removed for the mRNA data as they were outside acceptable normalisation standards. Both datasets were normalised so that for each gene, the mean mRNA and protein levels were zero with standard deviation 1.

3.2 Calculating GO enrichments

Each gene was labeled with Gene Ontology (GO) annotations using the REFSEQ mRNA accession and gene names using GeneTools Beisvag *et al.* (2006). Enrichment significance was calculated using the one-sided mid p-value (see Rivals *et al.* (2007), p.403). Each test was corrected for multiple testing by multiplying by the number of tests performed. A p-value cutoff of 0.1 was used throughout.

3.3 Determining the number of clusters

Determining a good value for the number of components in a mixture model is non-trivial. The model likelihood will increase indefinitely as the number of components increases. Out of sample likelihood (from, for example, cross validation) will not increase indefinitely but leaving data out when we have a limited quantity and some of the clusters may be quite small is likely to give a large variance in the resulting value. An alternative is to use the Bayesian information criteria, (BIC) defined as

$$BIC = -2 \log L + P \log(G)$$

where $\log L$ is the log-likelihood of the fitted model (given by equation 1), G is the sample size (number of genes) and P is the total number of parameters being estimated. Note that this is the value of the model likelihood and not the bound. Searching over the $K \times J$ space for the coupled model is somewhat unwieldy. Therefore, we will evaluate the number of components of mixture models - this is an approximation but, as the BIC score is only going to give an approximate number of components anyway it shouldn't be too limiting. The number of parameters is therefore $KT + 2K$ where the first term corresponds to the component mean vectors and the second term to the component variances and the prior component probabilities. A plot of the BIC versus K (and J in the protein case) can be seen in figure 2. For each number of components, ten restarts were used with different random initial conditions. We are looking for the model with lowest BIC and whilst we can see that there is no clear number of components in either case, it seems that there are slightly more components in the protein model than mRNA and that the numbers look to be somewhere in the region of $K = 15$ and $J = 20$. These are the values that we will use in our analysis.

In addition to using BIC, we experimented with Dirichlet Process (DP) priors for the individual cluster models. The posterior distribution over the number of components when using a DP prior is highly sensitive to the base measure and concentration parameter (Medvedovic and Sivaganesan, 2002, for example). We used the standard Normal-Inverse-Wishart prior as our base measure as its conjugacy to the Gaussian means we were able to marginalise out the model parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ to improve convergence. Under a wide range of hyper-parameter and concentration parameter settings, we found that the mode of the posterior over number of components was always between $K = 14$ and $K = 20$ for the mRNA data and a similar range for the proteomic data with J generally slightly higher than K. This is in broad agreement with the values chosen via BIC. One particular posterior distribution can be seen in figure 3.

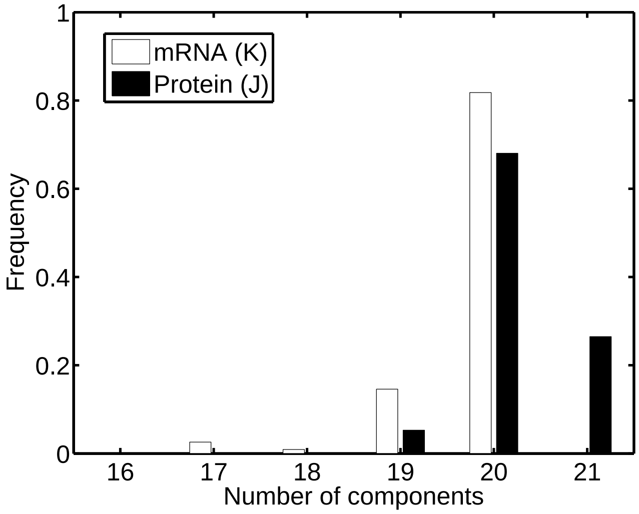


Fig. 3. Posterior distribution over number of components for individual Gaussian mixtures over the mRNA and proteomic data with a Dirichlet Process prior for one particular setting of the base measure hyper-parameters. The distributions over a wide range of hyper-parameter and concentration parameter values are roughly consistent with the results obtained using the BIC.

3.4 Comparing clusterings

To compare the clusterings obtained by individually clustering the data types separately, we used the modified Rand index (for

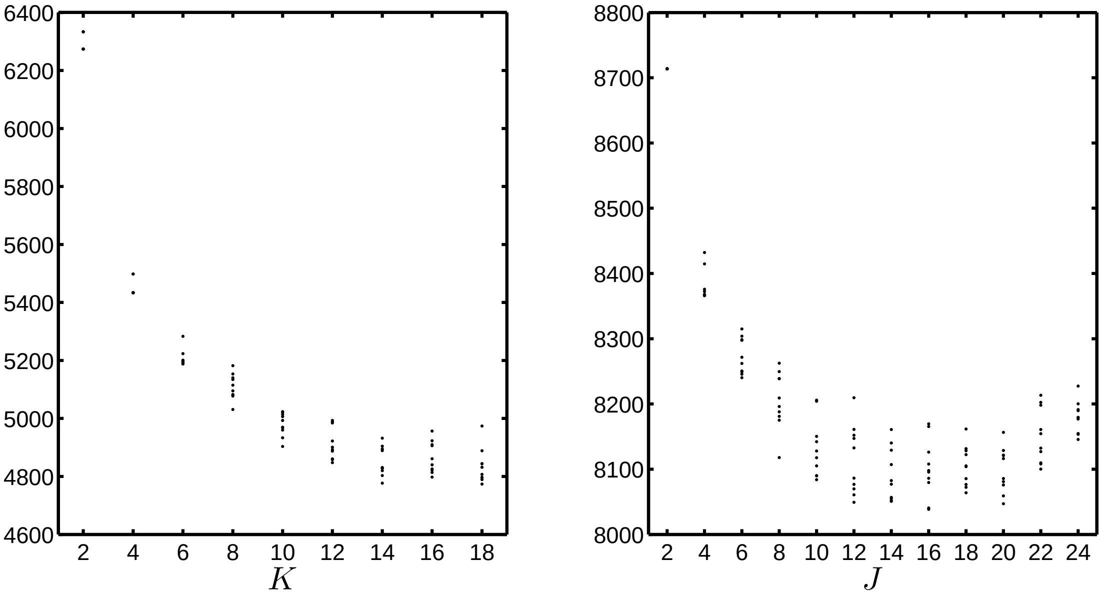


Fig. 2. Bayesian Information Criterion as a function of the number of components for a Gaussian mixture model trained on the mRNA data (left) and proteomic data (right). The results suggest values of approximately $K = 15$ and $J = 20$.

example, Meila (2007)). The standard Rand index is given as

$$r = \frac{a+b}{\binom{G}{2}}$$

where a is the number of pairs of genes that are assigned to the same cluster in both the mRNA and protein clusterings and b is the number of pairs of genes that are assigned to different clusters in both the mRNA and protein clusterings. The modified Rand index is a slightly modified version of this statistic that transforms the statistic to have an expected value of zero. A more comprehensive discussion can be found in Meila (2007).

4 ADDITIONAL RESULTS

4.1 Preliminary experiments - basic analysis

Before looking at results from the coupled model, we begin with some illustrative examples that demonstrate the problems with simply concatenating the data together and motivate a combined analysis such as that seen in this paper.

4.1.1 Cluster similarity Firstly, it is illuminating to cluster the genes by their mRNA profiles and their protein profiles separately to see how similar the resulting partitions are. To measure the similarity between two clusterings we use the modified Rand index (see for example Meila (2007), details given in methods section). The higher the value, the more similar are the partitions created by the clustering, with a maximum of 1 if the two are identical. In figure 4 (left panel, labeled *true*) we show the values of the modified Rand index for comparisons between a Gaussian mixture model on the mRNA data ($K = 15$ clusters - the number of clusters present was determined using the Bayesian Information Criterion (BIC), see methods section for details) and one on the proteomics data ($J = 20$ clusters). Uncertainties are obtained by running the Expectation Maximisation (EM) algorithm for Gaussian mixtures with twenty different random initialisations. Next to this, labeled *permuted* is the modified Rand index if we randomise the order of the genes in the protein clustering. Hence, this lower box provides a measure of similarity if the partitioning is randomised whilst retaining the same cluster sizes. We can see that the true clusterings are more similar than the true mRNA one is with the randomised protein ones, suggesting the presence of some structure. To put these values of the Rand index into perspective, the right panel of figure 4 shows the value when the mRNA clustering is compared against itself with an increasing number of gene swaps (i.e. switching the assignments of two randomly selected genes where genes can be selected more than once). As more swaps are made, we would expect the similarity to decrease. With 0 swaps, the modified Rand index is 1 as the mRNA clustering is clearly identical to itself. Around 800 swaps are required to achieve a value similar to that observed between partitions created by mRNA and protein clusterings, suggesting a high diversity between expression at the mRNA and protein levels. Therefore, there is only a small similarity between a clustering of genes obtained from their mRNA profiles and one obtained from their protein profiles. These results suggest that concatenating the data would be rather foolish. If the clusterings were very similar, we could hypothesize that when concatenated, there would be approximately the same number of clusters as in the original model (i.e. approximately 20). However, the large dissimilarity between the partitions produced by the two data types suggests that a large number of the mRNA cluster, protein cluster combinations would be present. Feasibly, the number of clusters could be as large as $K \times J$.

Given the number of genes, $G \approx 500$, this is impractical. Despite the computational problems that this would represent, we will show in the next section that a lot of the individual clusters have some biological relevance (as measured by enrichment of gene ontology (GO) terms) that would be lost if we subdivided them to such a large extent.

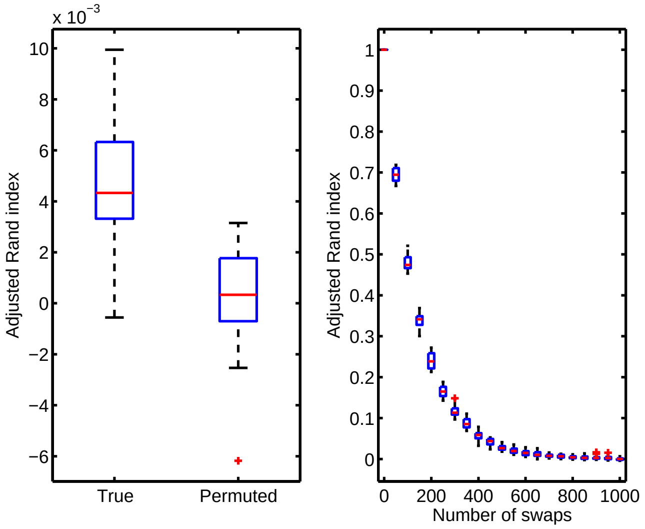


Fig. 4. Comparison of clusterings achieved with mRNA data and proteomic data with a null distribution obtained by re-ordering genes in protein clustering (left panel). Comparison of mRNA clustering with itself after a number of random gene swaps. The true values seen in the left panel require of the order of 800 swaps.

4.1.2 Gene ontology enrichment Computing the number of enriched GO terms in the individual and a concatenated clustering allows us to see if anything is lost through concatenation. p-values for enrichment of GO terms are calculated as described in the methods section. Figure 5(a) shows the number of enriched terms for individual clusterings and a concatenated clustering. These results are interesting in themselves. For example, very few cellular location GO terms are enriched in the mRNA clustering compared to the protein clustering. It also looks as though the concatenated method performs favorably compared with the two individual methods. However, the terms found in the two individual clusterings are not the same. In figure 5(b), we show the number of unique terms found when combining the terms from the two individual models. This is far higher than the number found for the concatenated model. Hence, clustering the two data types separately produces different biologically meaningful clusters, some of which are lost if the data are concatenated.

4.2 The coupled mixture model - high level observations

Figure 6 shows the mRNA and protein clusters produced by the joint model and links between them for which $p(j|k) > 0.1$. The complexity is immediately apparent - most mRNA clusters are very strongly linked to at least two protein clusters. In the methods section, we described how the coupled mixture model could be thought to exist on the spectrum between two extremes corresponding to concatenating the data (mRNA and protein clusters would have a one-to-one relationship) and independent clustering (mRNA profile provides no information about protein profile and vice-versa). We can see from these results that we are far from the concatenated scenario - i.e. knowledge of the mRNA profile of a gene (i.e. which cluster it belongs to) does not tell us which cluster its protein profile it belongs to. Rather, our approach reveals a complex network of associations that seems to be characterised by multiple and probabalistic relationships. While as we show below there is a recognizable structure behind many of these relationships, there is no universal principle that links transcription to translation. This has implications in a large volume of biological research where the mRNA profile is often taken to be a suitable proxy for the activity of its respective protein. For the genes studied here, the relationship between mRNA and protein profile appears highly complex, and in the most part not directly correlated. It is important to bear in mind that despite its volume exceeding that of most studies this data spans only a small subset of the genome and proteome. As such due care should be taken when extrapolating any observations.

Another way in which we can see the complexity of the relationship between the mRNA and protein profiles is by computing the entropy of the coupling distribution $p(j|k)$ and that of the reversed coupling distribution $p(k|j) = p(j|k)p(k)/(\sum_{k'} p(j|k')p(k'))$, computed from Bayes theory. We average the entropy for each $p(j|k)$ over k to give

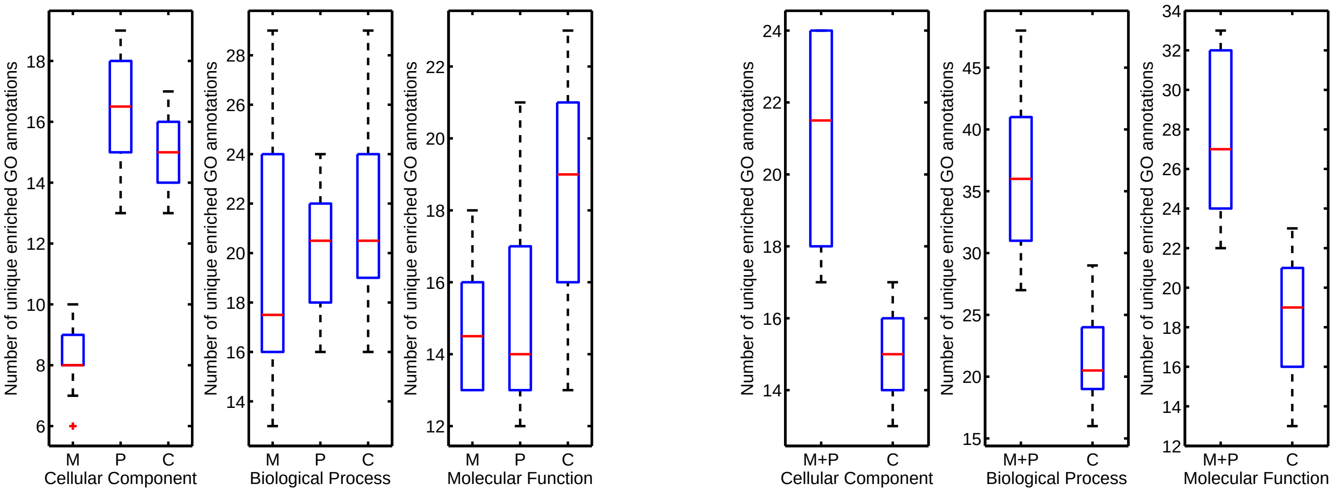
$$E = -\frac{1}{K}\sum_{k}\sum_{j} p(j|k)\log_2 p(j|k)$$

Particularly, we can compare the entropy for the inferred values with that for a coupled model trained with the order of genes in the protein data permuted. If there are strong relationships between mRNA and protein components, we would expect a low entropy ($E = 0$ corresponds to the concatenated model, the maximum value of E for $p(j|k)$ is 4.32 and for $p(k|j)$ is 3.91). Performing this comparison over 100 random restarts of the algorithm (each of which had a different protein order) we can see the density of entropy values for $p(j|k)$ and $p(k|j)$ in figure 7. In both cases, we see that the true entropy is slightly higher than that for the permuted data although the difference is not as high as one might expect. This can partly be explained by the low number of clusters in the two data sets taken individually - if we reorder the proteins, we will be swapping quite a few for others with similar profiles. The entropy significantly deviates from the least informative end of the scale ($E = 4.32$ and $E = 3.91$ respectively) although the models are somewhat closer to this than to complete order ($E = 0$).

It is important to point out that the small decrease in entropy between the true model and a permuted model in each case does not mean that no structure is present. It is more indicative of the fact that genes are organised into large clusters at the mRNA and protein levels but much smaller clusters (typically 10 genes and fewer) when the data is considered together.

REFERENCES

Band, V. and Sager, R. (1989). Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a



(a)

(b)

Fig. 5. (a) Number of enriched GO terms for individual mRNA clustering (M), individual proteomic clustering (P) and concatenated clusterings (C). (b) Number of unique terms when we combine terms from individual clusterings (M+P) versus those for concatenated clustering C.

medium that supports long-term growth of both cell types. *Proc Natl Acad Sci USA*, **86**(4), 1249–53.

Beisvag, V., Junge, F., Bergum, H., Jolsum, L., Lydersen, S., Gunther, C.-C., Ramampiaro, H., Langaas, M., Sandvik, A., and Laegreid, A. (2006). Genetools - application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, **7**(1), 470.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B*, **57**(1), 289–300.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–193. 10.1093/bioinformatics/19.2.185.

Bussey, K., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W., Zeeberg, B., Ajay, and Weinstein, J. (2003). Matchminer: a tool for batch navigation among gene and gene product identifiers. *Genome Biology*, **4**(4), R27.

Chen, W.-N. U., Yu, L.-R., Strittmatter, E. F., Thrall, B. D., Camp, D. G., and Smith, R. D. (2003). Detection of in situ labeled cell surface proteins by mass spectrometry: application to the membrane subproteome of human mammary epithelial cells. *Proteomics*, **3**(8), 1647–51.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, B*, **39**, 1–38.

Jacobs, J. M., Mottaz, H. M., Yu, L.-R., Anderson, D. J., Moore, R. J., Chen, W.-N. U., Auberry, K. J., Strittmatter, E. F., Monroe, M. E., Thrall, B. D., Camp, D. G., and Smith, R. D. (2004). Multidimensional proteome analysis of human mammary epithelial cells. *Journal of Proteome Research*, **3**(1), 68–75.

Kerr, M., Martin, M., and Churchill, G. (2004). Analysis of variance for gene expression microarray data. *J Comput Biol*, **7**, 819–837.

Liu, T., Qian, W.-J., Strittmatter, E. F., Camp, D. G., Anderson, G. A., Thrall, B. D., and Smith, R. D. (2004). High-throughput comparative proteome analysis using a quantitative cysteinyl-peptide enrichment technology. *Anal Chem*, **76**(18), 5345–53.

Liu, T., Qian, W., Chen, W., Jacobs, J., Moore, R., Anderson, D., Gritsenko, M., Monroe, M., Thrall, B., II, D. C., and Smith, R. (2005). Improved proteome coverage by using high efficiency cysteinyl peptide enrichment: The human mammary epithelial cell proteome. *Proteomics*, **5**(5), 1263–1273.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**(9), 1194–1206.

Meila, M. (2007). Comparing clusterings - an information based distance. *J Multivariate Anal*, **98**(5), 873–895.

Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, **75**(17), 4646–58.

Qian, W.-J., Monroe, M. E., Liu, T., Jacobs, J. M., Anderson, G. A., Shen, Y., Moore, R. J., Anderson, D. J., Zhang, R., Calvano, S. E., Lowry, S. F., Xiao, W., Moldawer, L. L., Davis, R. W., Tompkins, R. G., Camp, D. G., and Smith, R. D. (2005). Quantitative proteome analysis of human plasma following in vivo lipopolysaccharide administration using 16o/18o labeling and the accurate mass and time tag approach. *Molecular Cellular Proteomics*, **4**(5), 700–709. 10.1074/mcp.M500045-MCP200.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, **23**(4), 401–407. 10.1093/bioinformatics/btl633.

Smith, R., Anderson, G. A., Lipton, M., Pasa-Tolic, L., Shen, Y., Conrads, T., Veenstra, T., and Udseth, H. (2002). An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**, 513–523.

Stampfer, M. and Yaswen, P. (1993). Culture systems for study of human mammary epithelial cell proliferation, differentiation and transformation. *Cancer Surv*, **18**, 7–34.

Waters, K., Liu, T., Quesnberry, R., Qian, W., Willse, A., Bandyopadhyay, S., Kathmann, L., Weber, T., Smith, R., Wiley, H., and Thrall, B. (2008). Systems analysis of response of human mammary epithelial cells to egf by integration of gene expression and proteomic data. *Under Submission*.

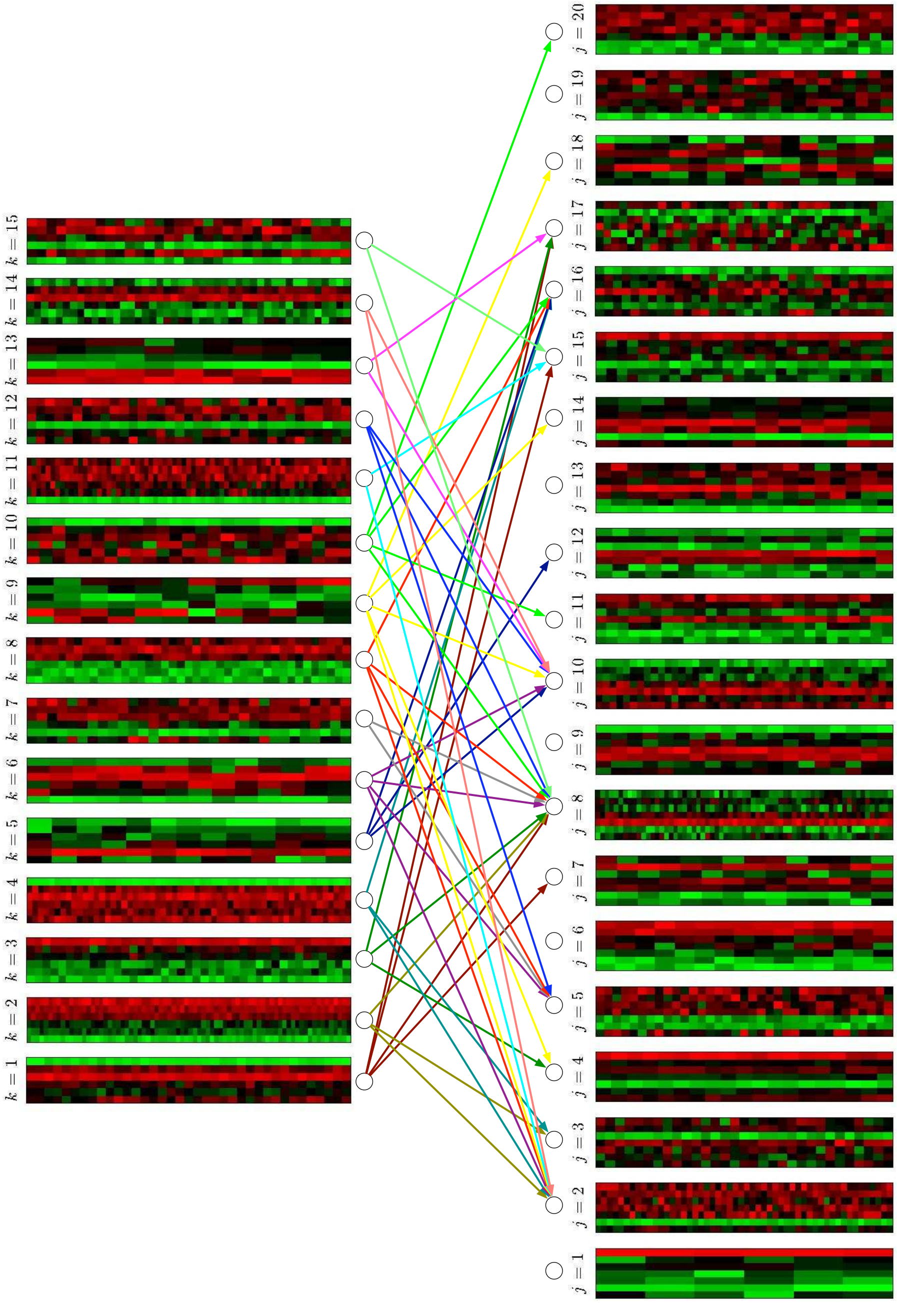
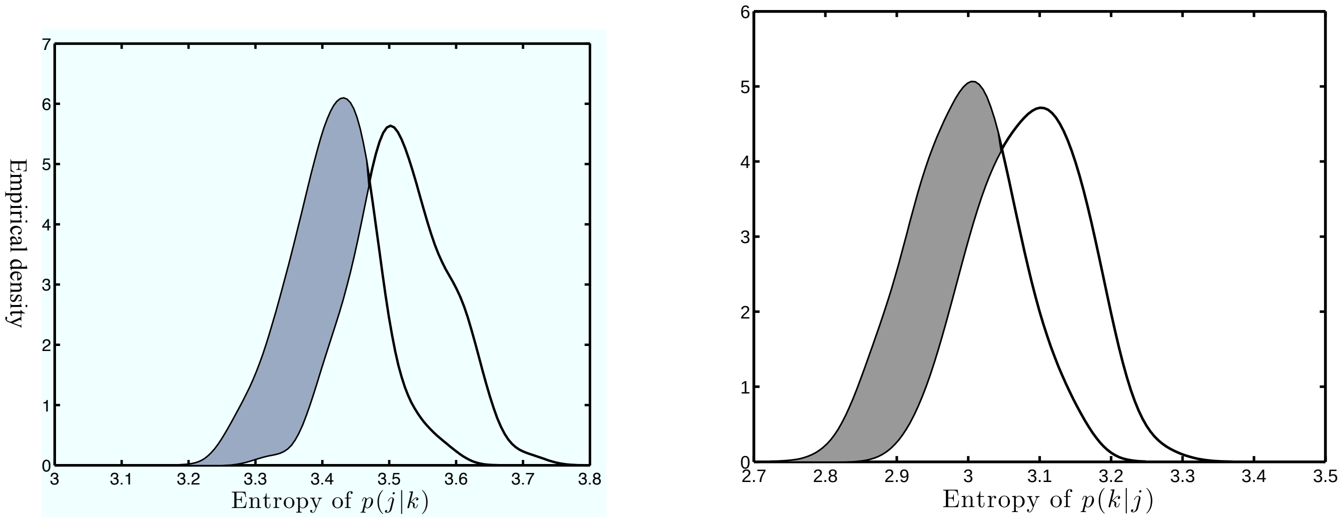


Fig. 6. mRNA (left/top) and protein clusters (right/bottom) with links shown where $p(j|k) > 0.1$.



(a)

(b)

Fig. 7. Distribution of mean entropy values of $p(j|k)$ (a) and $p(k|j)$ (b). In both cases the shaded (left hand) curve corresponds to the true value and the value from permutations. A value of 0 would correspond to concatenated clustering and maximum values, corresponding to $p(j|k) = 1/J \;\; \forall J$ and $p(k|j) = 1/K \;\; \forall K$ are 4.32 and 3.91 respectively.