A Bayesian model that links microarray mRNA measurements to mass spectrometry protein measurements

Anitha Kannan[1] and Andrew Emili[2] and Brendan J. Frey[3]

[1] Microsoft Research, Cambridge, UK
http://www.research.microsoft.com/∼ankannan
ankannan@microsoft.com

[2] Banting & Best Department of Medical Research, University of Toronto, Canada
http://emililab.med.utoronto.ca
andrew.emili@utoronto.ca

[3] Electrical & Computer Engineering, University of Toronto, Canada
http://www.psi.toronto.edu
frey@psi.toronto.edu

Abstract. An important problem in biology is to understand correspondences between mRNA microarray levels and mass spectrometry peptide counts. Recently, a compendium of mRNA expression levels and protein abundances were released for the entire genome of the laboratory mouse, *Mus musculus*. The availability of these two data sets facilitate using machine learning methods to automatically infer plausible correspondences between the gene products. Knowing these correspondences can be helpful either for predicting protein abundances from microarray data or as an independent source of information that can be used for learning richer models such as regulatory networks. We propose a probabilistic model that relates protein abundances to mRNA expression levels. Using cross-mapped data from the above-mentioned studies, we learn the model and then score the genes for their strength of relationship by performing probabilistic inference in the learned model. While we gave a simplified outline of our technique in a publication aimed at biologists (Cell 2006), in this paper, we give a complete description of the Bayesian model and the computational technique used to perform inference. In addition, we demonstrate that the Bayesian technique achieves mappings with higher statistical significance, compared to standard linear regression and a maximum likelihood version of the proposed model.

1 Introduction

Proteins are macromolecules essential to the structure and function of all living cells. The biological process in which cells produce proteins from DNA involves an intermediate step where the DNA is transcribed into messenger RNA (mRNA), before being translated into a protein. An important problem in biology is to understand correspondences between levels of mRNA transcripts and

abundances of proteins that are produced. However, the biological processes underlying translational regulation are quite complex, so inferring correspondences between these two gene products is non-trivial. Addressing this correspondence problem would facilitate better understanding of cell functionality (c.f. [1, 2]). If we know that there is a direct relationship between the two gene products, we can determine protein abundance level at the genome level using simpler and more cost-effective microarray-based mRNA expression measurements. Alternatively, if we can ascertain no relationship between them, they can be treated as complementary independent sources of information that can be used in learning richer models, such as for predicting interaction networks.

In this paper, we seek to infer relationships between protein abundance and mRNA level using noisy high throughput expression profiles of mRNA obtained using microarrays and expression profiles of protein obtained using mass spectrometry. We define a probabilistic model that relates cross-mapped products from these data sets. After learning the parameters of the model using the cross-mapped data, we score the strength of the relationship between protein abundance level and mRNA expression level on a gene by gene basis. In addition, we perform permutation testing and assign a p-value to each gene, thereby obtaining a confidence measure of the significance of the inferred relationship. In [1], we provide a biologist's overview of some parts of the model described in this paper. Here, we provide thorough description of the computational method used to analyze the data. Further, we compare our method to linear regression (which has previously been proposed for this problem) and maximum likelihood estimation in our model, and show that the Bayesian approach recovers a larger number of statistically-significant relationships. The relationships thus detected provide a resource for potential new biological discoveries [1].

There have been a number of previous approaches to inferring correlations between mRNA and protein levels [3–5]. Almost all previous methods perform correlation analysis on a global scale, and report positive but weak association between transcript and translational levels. These methods suffer from three main problems. First, they usually summarize global relationships between the measured levels of the two gene products. However, it is an accepted fact that the processes involved in translating mRNA into protein product are quite complicated and vary between genes. Therefore, a more relevant goal is to infer correlation on a gene-by-gene basis [6]. Second, measurements obtained from existing technologies are prone to be quite noisy, but most previously proposed methods either do not explicitly account for noise or assume a strictly Gaussian form of noise. One way to account for non-Gaussian noise is to include additional hidden variables, such as unknown abundances of bio-molecules, but most previously-proposed methods do not incorporate such hidden variables. Methods reported in [3, 6] use robust correlation approach and were able to report stronger correlations, thereby conveying that we can better relate the two gene products by properly accounting for non-Gaussian noise. However, one problem with the approach suggested in [6] is that it uses a concordance test based on a presence or absence call, making it inappropriate when quantitative expression

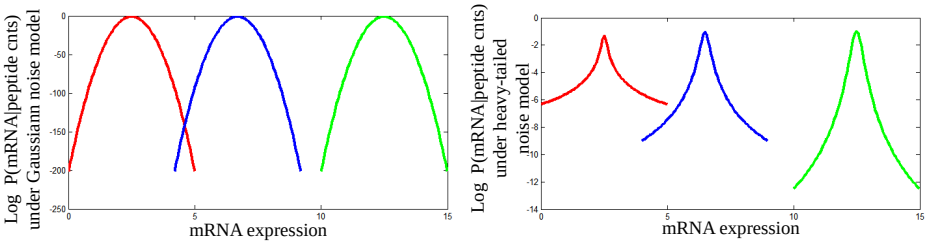

Fig. 1. A schematic figure illustrating differences between our approach and a Gaussian-noise-model approach. In (a), we show the probability of mRNA expression given different values for the peptide count, under a linear model with Gaussian noise of fixed variance. This noise model does not capture the fact that the variance of a sum of counts depends on the number of counts. Further, it does not account for outlying mRNA expression measurements, caused by spurious noise. In (b) we show the kind of probability model used in this paper – the variance decreases for larger values of the peptide count, and the noise model is heavy-tailed and thus the model is less sensitive to outliers.

values are to be analyzed. Following this line of thinking, we formulate a proper probability model that accounts for relevant noise processes, as shown schematically in Fig. 1. Third, previously, models that take into account multiple sources of variability that govern the relationships between the two gene products could not be directly learned from data mainly because of limited availability of data. Therefore, earlier approaches suffered from inability to be invariant to known sources of variation.

In this paper, we overcome these three problems in a principled fashion, allowing us to better understand relationships between mRNA and protein levels. We propose a method that analyzes the available data on a gene-by-gene basis. To obtain an understanding at the gene level, we introduce a probabilistic model that uses a Bernoulli switch variable, which when inferred, either explains microarray expression levels of mRNA as a noisy linear function of the hidden parameter of a Poisson distribution over observed peptide counts, or as being independent of peptide counts and accounted for by a background model learned using only microarray measurements. The probabilistic framework can account for both biological and experimental sources of uncertainty. In conjunction with our biology collaborators, we recently published comprehensive dataset of mRNA expression [8] and protein expression [1] across the entire genome of the laboratory mouse, *Mus musculus*. We make use of these large-scale data sets to learn the model. The probabilistic framework enables us to incorporate other possible kinds of data in a principled way.

The reminder of the paper is organized as follows: Sec. 2 describes the reliably cross-mapped dataset we used for our analysis. Sec. 3 describes the probability model we propose for inferring relationship between the two datsets, and the method used for inference and learning in this model. In Sec. 4, we provide the results of our analysis, and compare them with other existing standard tech-

niques. We draw conclusions in Sec. 5 and outline potential directions for future work.

2 Data and its representation

In this paper, we make use of the compendium of protein abundances reported in [1]. This molecular compendium provides the protein content of 4,768 proteins in four major organelle compartments (cytosol, membranes or microsomes, mitochondria and nuclei) in six organs (brain, heart, kidney, liver, lung and placenta) of the laboratory mouse, *Mus musculus*. Protein abundance is measured using a comprehensive comparative proteomic profiling procedure based on gel-free multidimensional protein identification technology (MudPIT). We performed 7 MudPIT experiments and summed their spectra to obtain a discrete measure of protein abundance known as peptide count. It is shown that peptide counts produced by this procedure are positively correlated with actual protein abundance and thus can plausibly be used as a quantitative measure of protein abundance [7].

Recently, two genome-scale surveys of mRNA transcripts levels in mouse tissues were published [8, 9]. While [8] uses high-density inkjet synthesized long-oligonucleotide microarrays, [9] uses custom short-oligonuleotide Affymetrix gene chips to study gene expression. We performed a three-way cross-mapping between these three data sets, and found 1,914 detected gene products to be in common across all three platforms. We used this cross-mapped set of genes to perform our analysis of inferring relationships between mRNA transcript levels from [8] and protein abundance from [1].

3 A probability model of mRNA expression and protein abundance

Figure 2 shows a Bayesian network for inferring relationships between mRNA expression levels and protein abundance levels on a gene-by-gene basis. We consider a set of G genes that are indexed by g. Let $\mathbf{m}$ and $\mathbf{y}$ be the measurements of its two gene products, mRNA expression level and protein abundance level. Both $\mathbf{m}$ and $\mathbf{y}$ are T-dimensional vectors corresponding to measurements of T tissues. We index the tissues by i so that i^{th} element, y_i, of the vector $\mathbf{y}$ is the peptide count corresponding to the i^{th} tissue.

As described in Sec. 2, the peptide count for a particular tissue is the sum of counts across multiple MudPIT experiments. When multiple MudPIT experiments record the presence of a particular protein, the final peptide count corresponding to the protein will be large. Therefore, for each tissue, we model the effect of multiple MudPIT experiments on its observed peptide counts using a latent variable x_i. We can view this variable as the true rate at which peptides are presented in all MudPIT experiments. We represent this unknown rate for all tissues by $\mathbf{x}$. With this, the distribution of the peptide counts for the protein

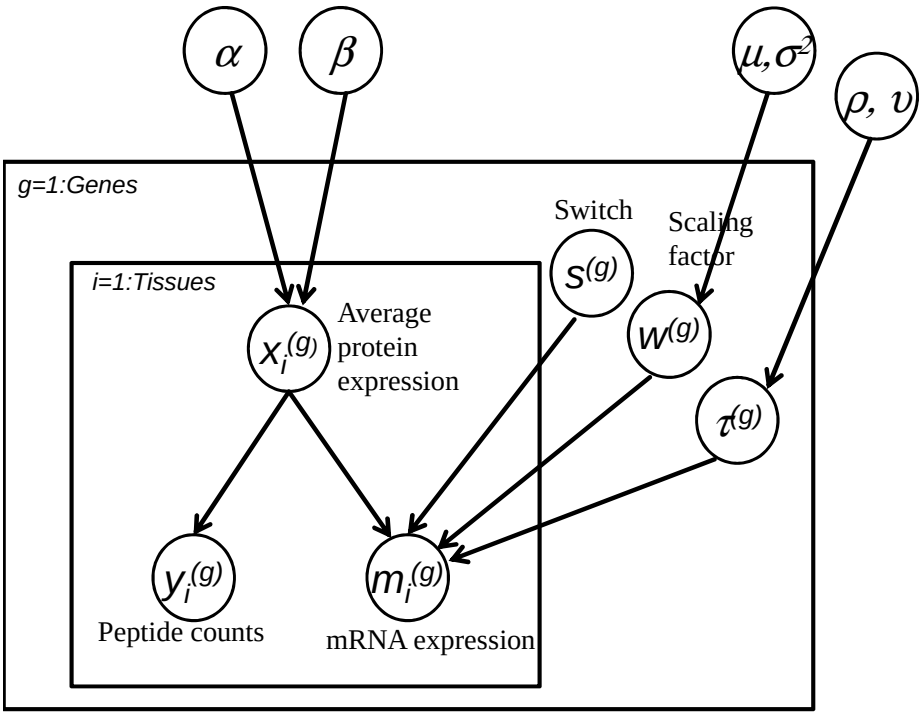

Fig. 2. A probability model for inferring the relationship between peptide counts measuring protein expression and microarray mRNA expression levels. This Bayesian network makes use of plate notation, where the sub-model within a rectangle, including edges entering the rectangle, is replicated; The nodes outside the rectangle and connected to the nodes in the rectangle are shared across all replications. For instance, the graph in the innermost rectangle corresponds to a single gene g and is replicated T times to match with T tissues. All the T tissues for a single gene g shares the same s, w and τ variable. As each gene has independent set of these variables, we use another sub model indexed by g. The variables that are shared across multiple genes are outside the outermost rectangular enclosing.

given the rate is modeled using an independent Poisson distribution for each tissue, with rate parameter x_i:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{T} p(y_i|x_i) = \prod_{i=1}^{T} \frac{e^{-x_i} x_i^{y_i}}{y_i!}. \tag{1}$$

We model the rate parameter $\mathbf{x}$ using a Gamma distribution,

$$p(\mathbf{x}) = \prod_{i=1}^{T} p(x_i) = \prod_{i=1}^{T} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i \beta}. \tag{2}$$

The Gamma distribution is suitable because it is the conjugate prior to the Poission distribution, so we can analytically compute the posterior distribution $p(\mathbf{x}|\mathbf{y})$ over the rate variables given the peptide counts. This is also a Gamma distribution given by

$$p(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{T} p(x_i|y_i) = \prod_{i=1}^{T} \frac{(\beta+1)^{\alpha+y_i}}{\Gamma(\alpha+y_i)} x_i^{(\alpha+y_i)-1} e^{-x_i(\beta+1)}. \quad (3)$$

This posterior distribution is concentrated on a small range of $\mathbf{x}$ when $\mathbf{y}$ is large; A large value for $\mathbf{y}$ indicates that the corresponding protein level is measured reliably by multiple MudPIT experiments. When the observed peptide count is small, it reflects uncertainty in the expression of a particular protein and therefore the prior distribution plays a more dominating role, making the posterior distribution more spread out.

We represent the expression profile of the mRNA abundance corresponding to a particular gene by a vector $\mathbf{m}$, so that the expression for the i^{th} tissue is m_i. We know that there are many genes for which mRNA expression and protein abundance levels do not agree due to either biological factors such as post-translational modifications or experimental factors such as noisy measurements and changes in conditions between measurements. We model the decision about whether or not the data can be mapped using a binary switch variable, s, with prior distribution denoted by $P(s)$. We typically fix $P(s = 1) = .95$, but this is only a prior on the decision and hence will play only a weak role and not force linear relationships where there aren't any. For a particular gene, if the switch is in the 'off' state (s=0), it indicates that the mRNA expression levels and the hidden peptide rates for that gene do not agree. In this case, the microarray measurements are assumed to be independent of the proteomics measurements, and the microarray measurements are accounted for by using a background model $p_o(\mathbf{m})$ learned using kernel density estimation (c.f. [10]). Kernel density estimation captures the underlying space of microarray expression profile by placing common variance Gaussian kernels on each profile, where variance is computed using leave-one-out cross validation. For our model, we use the entire available microarray data for each tissue to learn their corresponding independent background model.

If the switch variable is in the 'on' state (s=1), the microarray measurement for each tissue is explicitly modeled as a noisy linearly weighted function of the average peptide counts, given by $m_i = wx_i + \text{noise}$. We assume the noise in the microarray measurement is Gaussian with mean 0 and variance τ. While we assume Gaussian noise here, the model on the whole is multi-layer and hierarchical with many more random variables with different distributions including Poisson, Gamma, and discrete. This means that the resulting model is far from being Gaussian and therefore does not have the shortcomings of linear regression and correlation-based methods, which effectively assume Gaussian noise.

The scalar weight w is shared across all tissues for a particular gene and is interpreted as the scaling factor required to match x_i with m_i for all tissues, up to some noise level. As w models gene-specific effects, it accounts for technological

effects such as microarray probe sensitivity and microarray data normalization, to name a few. Also, w can capture biological effects such as translational efficiency; When the translational efficiency is higher for one gene compared to another, it will have a smaller value for w. The distribution of $\mathbf{m}$ conditioned on $s, \mathbf{x}$, w and τ is given by

$$p(\mathbf{m}|\mathbf{x}, \tau, s, w) = \begin{cases} \prod_i p_o(m_i) & \text{if } s = 0 \\ \prod_i \frac{1}{\sqrt{2\pi\tau}} \exp(-(m_i - wx_i)^2/2\tau) & \text{if } s = 1 \end{cases} \tag{4}$$

As described in Sec. 2, we have, for each gene, measurements from 6 tissues. This means that we need to infer the scaling factor w for each gene using only 6 measurements. Using only 6 measurements to determine w is likely to lead to overfitting, making the inferred relationship statistically insignificant. Instead, we take a Bayesian approach that enables us to integrate over all possible values of w. For this, we model the scaling factor w as a random variable with a Gaussian prior distribution so that the effect of w can be averaged out over its entire range of values. Similarly, we also treat the inverse variance τ^{-1} as a random variable with a Gamma distribution as the prior. The parameters of both these prior distributions are shared across all genes, and are learned using the entire data set of expression profiles.

Since the joint model is a Bayesian network, we can write the joint distribution over the variables modeled by it as the product of all the conditional distributions. With $\theta = \{w, \tau\}$,

$$p(\mathbf{x}, \mathbf{y}, \mathbf{m}, \theta, s) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})p(\mathbf{m}|\mathbf{x}, s, \theta)p(\theta)P(s). \tag{5}$$

The objective of the model is to account for relationship between microarray mRNA expression data and protein peptide count data. One approach to this is to learn the model so as to maximize $p(\mathbf{m}, \mathbf{y})$. However, if we trained the model to maximize the joint probability, $p(\mathbf{m}, \mathbf{y})$, the model will try to account for variability in protein and mRNA expression due to factors such as gene function and tissue-specificity. In fact, our intention for the model is not to explain the biological variability in gene expression but to infer the mapping between the two sources of data. Therefore, we learn the model $p(\mathbf{m}|\mathbf{y})$, which is conditioned on the observed peptide counts and thus need not explain gene- and tissue-specific variations, unless they pertain to the predictability of the mRNA abundance from the protein abundance. While one approach will be to model $p(\mathbf{y}|\mathbf{m})$, we choose to model $p(\mathbf{m}|\mathbf{y})$ because inference and learning is more straight-forward. In summary, given a data set of cross-mapped proteomics and microarray data, our goals are to

- Learn the parameters of the model that maximizes the probability $p(\mathbf{m}|\mathbf{y})$ of observing the mRNA expression profile given the peptide counts for the entire set of genes, and
- Infer, for each gene, the probability $P(s = 1|\mathbf{m}, \mathbf{y})$ that a linear relationship does exist between measurements of these gene products.

3.1 Learning the model

Given the mRNA expression levels and the protein abundance levels for a set of T tissues corresponding to G genes, the goal is to learn the parameters of the model that maximizes the conditional distribution $\prod_{g=1}^{G} p(\mathbf{m}^{(g)}|\mathbf{y}^{(g)})$. Using (5), we can write the joint distribution over the observations by integrating over all the hidden variables, $s, \mathbf{x}, \theta$:

$$p(\{\mathbf{m}^{(g)}, \mathbf{y}^{(g)}\}) = \int_{\theta} \prod_{g=1}^{G} \sum_{s} P(s^{(g)}) \int_{\mathbf{x}} p(\mathbf{x}^{(g)}) p(\mathbf{y}^{(g)}|\mathbf{x}^{(g)}) p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}, \theta) p(\theta) \tag{6}$$

As described above, we do not want our model to account for biological variability in gene expression. Therefore, we can approximate the marginal distribution over the peptide counts $p(\mathbf{y}^{(g)})$ by its empirical distribution. To achieve this, we replace $p(\mathbf{x}^{(g)})p(\mathbf{y}^{(g)}|\mathbf{x}^{(g)})$ with $p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)})p(\mathbf{y}^{(g)})$ as follows

$$p(\{\mathbf{m}^{(g)}, \mathbf{y}^{(g)}\}) = \int_{\theta} \prod_{g=1}^{G} \sum_{s} P(s^{(g)}) \int_{\mathbf{x}} p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)}) p(\mathbf{y}^{(g)}) p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}, \theta) p(\theta). \tag{7}$$

This expression gives us the desired conditional probability:

$$p(\{\mathbf{m}^{(g)}\}|\{\mathbf{y}^{(g)}\}) \approx \int_{\theta} \prod_{g=1}^{G} \int_{\mathbf{x}} \sum_{s} P(s^{(g)}) p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)}) p(\theta) p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}, \theta), \tag{8}$$

where $p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)})$ can be analytically computed as shown before.

The integration over θ cannot be computed analytically, and hence we cannot optimize the above quantity exactly. But, for each gene, we can lower bound it using an approximate posterior distribution $p(w, \tau^{-1}|\mathbf{y}^{(g)}, \mathbf{m}^{(g)}) \approx q^{(g)}(\theta) = q^{(g)}(w)q^{(g)}(\tau^{-1})$. Here, we use a factorized distribution because accounting for the joint distribution is computationally more difficult. For mathematical and computational convenience, we choose $q^{(g)}(w)$ as a Gaussian distribution and $q^{(g)}(\tau^{-1})$ as a Gamma distribution.

$$\begin{aligned} \log p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}) &\geq \log q(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}) \\ &= \int_{\theta} q^{(g)}(\theta) \log \frac{p(\theta) p(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}, \theta)}{q^{(g)}(\theta)} \end{aligned} \tag{9}$$

We optimize the bound by alternating between finding the posterior distributions and updating the model parameters. This guarantees that the bound becomes tighter with each iteration and becomes equal when the approximate posterior distribution is same as the true posterior distribution [11]. After computing

$q(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)})$, it can be substituted into (8) to give

$$p(\{\mathbf{m}^{(g)}\}|\{\mathbf{y}^{(g)}\}) \geq \prod_{g=1}^{G} \sum_{s} P(s^{(g)}) \int_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}^{(g)}) \int_{\theta} q^{(g)}(\theta) \log \frac{p(\theta)p(\mathbf{m}^{(g)}|\mathbf{x}, s, \theta)}{q^{(g)}(\theta)}$$

$$\approx \prod_{g=1}^{G} \sum_{s} P(s^{(g)}) \int_{\mathbf{x}} p(\mathbf{x}^{(g)}|\mathbf{y}^{(g)}) q(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)}). \tag{10}$$

Computing this integral over $\mathbf{x}$ is hard because it involves taking an expectation of $q(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)})$ with respect to $p(\mathbf{x}|\mathbf{y})$. An approach to compute this expectation is to sample from $p(\mathbf{x}|\mathbf{y})$ and average $q(\mathbf{m}^{(g)}|\mathbf{x}^{(g)}, s^{(g)})$ using the sample [12]. We resort to this approach as it is easy to sample from $p(\mathbf{x}|\mathbf{y})$, as given by (3). We draw N samples $\mathbf{x}^{(1)}, \cdots \mathbf{x}^{(N)}$ from $p(\mathbf{x}|\mathbf{y})$ and use them to approximate the true expectation using the sample average,

$$p(\mathbf{m}|\mathbf{x}) \approx \sum_{s} P(s) \sum_{n=1}^{N} q(\mathbf{m}|\mathbf{x}^{(n)}, s). \tag{11}$$

At this juncture, one may wonder why not collect samples from θ as opposed to using variational inference to integrate over θ. The reason we choose to do it this way is that it is hard to sample from the distribution governing θ, and would require Markov-chain monte-carlo methods. In contrast, sampling from $p(\mathbf{x}|\mathbf{y})$ is exact as it is a known distribution that is easy to sample from. Our goal of learn a model that maximizes $p(\{\mathbf{m}^{(g)}\}|\{\mathbf{y}^{(g)}\})$ lends itself to a simple inference algorithm.

3.2 Inferring strength of relationships

For a gene under consideration, the strength of the relationship between its pair of mRNA expression level and protein abundance is given by the probability, $P(s|\mathbf{m}, \mathbf{y})$. We can compute this quantity by applying Bayes rule:

$$P(s|\mathbf{m}, \mathbf{y}) = \frac{\int_{\mathbf{x}} p(\mathbf{m}|s, \mathbf{x}) p(\mathbf{x}|\mathbf{y}) P(s)}{\sum_{s} \int_{\mathbf{x}} p(\mathbf{m}|s, \mathbf{x}) p(\mathbf{x}|\mathbf{y}) P(s)} \approx \frac{\frac{1}{N} \sum_{n=1}^{N} P(s) q(\mathbf{m}|s, \mathbf{x}^{(n)})}{\sum_{s} \frac{1}{N} \sum_{n=1}^{N} P(s) q(\mathbf{m}|s, \mathbf{x}^{(n)})}. \tag{12}$$

For this computation, we evaluated the integral using sampling; We used N samples $\mathbf{x}^{(1)}, \cdots \mathbf{x}^{(n)}$ from $p(\mathbf{x}|\mathbf{y})$, in combination with (11). To compute this probability, we first fix a value for s. When $s = 1$, for each tissue, we obtain 1000 samples from $p(x_i|y_i)$. We then compute $p(\mathbf{m}|s, \mathbf{x}^{(n)})$ for each of the samples and then average the probabilities. When $s = 0$, we do not need to sample from $p(x_i|y_i)$ as we use the background model for explaining the microarray measurements. We can normalize the two to obtain the desired distribution, $P(s = 1|\mathbf{m}, \mathbf{y})$.

4 Results

We learned the model described above with the data from Sec. 2. Then, for each of the 1,914 genes, we computed the probability of a linear relationship between the mRNA expression profile and the corresponding peptide counts as given by (12). We used a permutation test to examine whether our probability calculation is well-calibrated. Since the model parameters are shared across all genes, we used 720 permutations where each permutation involved independently permuting the peptide counts of the tissues within the gene, while making sure that the permutation did not result in the observed data (this is possible because many peptide counts are 0). After each set of permutations, we re-learned the model and scored the genes. Then, we computed the p-value based on how many times we observed a particular probabilistic score by chance.

A plausible alternative approach to our proposed Bayesian way of integrating over parameters θ, is to replace the parameters with point estimates, known as maximum likelihood (ML) estimation. We learned the model by performing ML estimation on w (using an EM-type algorithm) and scored the genes based on p-value obtained using permutation testing as described above. We also studied another much simpler approach, which is to fit a linear regressor (LR) for each gene individually, assuming a fixed variance. We used permutation testing in this case as well to obtain a p-value for each gene. After obtaining p-values using these three methods, we chose to look at the 568 genes that had p-values less than .05 for all the three methods. From Fig. 3(a), it is clear that for a large number of genes, the p-values achieved by the Bayesian method is much lower than the p-value achieved by ML or LR, with ML doing slightly better than LR. Further, for a given p-value threshold, the number of genes satisfying the threshold is larger for the Bayesian method than for either the ML or LR methods. The advantage of using the Bayesian approach is that when under uncertainty (here, we need to estimate w from only 6 measurements), it integrates over all possible ranges of values for the parameters, as opposed to ML and LR, where only a single value for the parameters is used. This is another advantage of our approach which allows us to obtain relationships that are much more reliable as it can better reason under uncertainty, taking into account all the modeled sources of variability. Fig. 3(b)-(c) shows each of these 568 genes as a point in a scatter plot comparing the Bayesian p-values to the ML or LR p-values.

We also analyzed scenarios when one of the methods infers a much stronger relationship (or lack thereof) than the other methods. For this, we performed three experiments. We selected all genes that had p-values less than .005 under our proposed Bayesian method, and had p-values greater than .05 under the other two methods. We found 158 such genes, 10 of which are shown in Fig. 4a. We see that these genes have expression values that are large and would require paying a huge penalty under LR and ML to match the mRNA expression with the average peptide count. The Bayesian method, in contrast, can average over all values of w, appropriately weighted, and thus is less sensitive to the large deviations. In Fig. 4b, we show 10 of the 44 genes where ML performs better than the other methods under the same threshold criterion described above. For

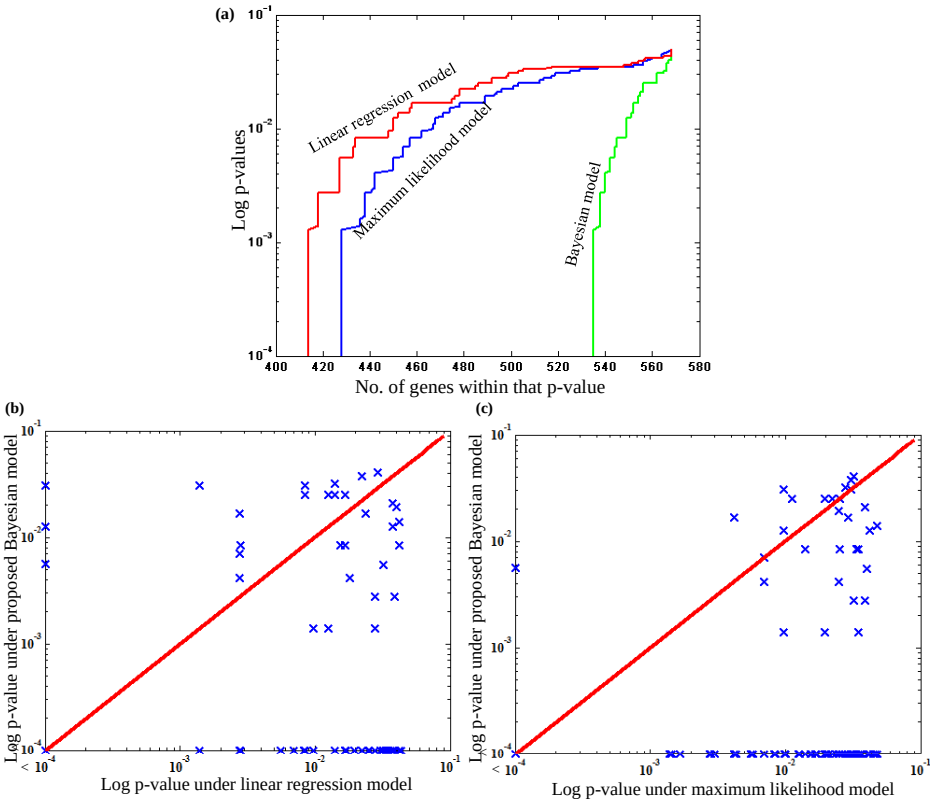

Fig. 3. Comparison between our Bayesian method, maximum likelihood (ML) estimation in the same model, and linear regression (LR). We consider only 568 genes that have p-values (computed from a permutation test) less than .05 in all three methods. (a) shows that for a large number of genes, the Bayesian mapping achieves much lower p-values than the ML or LR methods and for a given p-value threshold detects a larger number of cases with p-values below this threshold. (b) shows the scatter plot of the p-values obtained by the LR and Bayesian method while (c) shows the scatter plot of the p-values obtained by the ML and Bayesian method. In (b) and (c), markers below the 45· red line correspond to genes whose Bayesian p-value is lower (indicating higher statistical significance for the Bayesian method) than the ML or LR methods respectively, and vice versa.

these genes, the expression profiles of both gene products are relatively flat and small, indicating quite weak signal values. Fig. 4c shows 10 of the 61 examples in which the p-value under LR is smaller than the other two methods. In this case, the peptide counts are quite tissue-specific and the mRNA expressions are similar, thereby allowing LR to fit the data exactly.

We can also use this model to partition the data into three groups of biological interest: inliers, borderline cases and outliers. The outliers correspond to the set of genes that have $P(s = 1|\mathbf{m}, \mathbf{y}) \leq .33$ and have p-values that within .05.

(a).

Peptide counts in 6 tissues						mRNA expression in 6 tissues						Bayesian method	Maximum Likelihood	Linear Regression
222	7	17	7	68	10	258067.0	36733.9	62589.4	46911.9	51498.5	54361.5	0.0000	0.0641	0.0557
56	41	42	76	41	62	119394.5	53294.2	224569.2	63075.5	68085.9	106413.1	0.0000	0.5780	0.6546
13	9	30	158	6	12	13896.1	8959.5	295560.2	252050.4	3838.9	26542.9	0.0000	0.0960	0.1419
29	8	36	88	12	12	9993.5	11318.2	113171.8	109753.4	8689.5	17042.4	0.0000	0.0543	0.0501
68	9	15	3	51	18	57206.0	35429.3	28944.2	59111.0	139675.8	77963.1	0.0000	0.1989	0.2267
33	9	6	15	20	67	41335.5	28128.8	59399.3	33697.4	30087.3	57800.1	0.0000	0.2531	0.2281
31	7	5	50	14	35	25363.7	38761.3	34774.5	57689.4	44649.5	92432.4	0.0000	0.1669	0.1892
6	2	15	5	55	49	32725.1	40117.5	102303.0	70609.5	49975.7	85342.5	0.0000	0.2754	0.3185
63	6	0	0	30	33	65343.7	44448.6	27950.5	50716.9	98760.1	36788.7	0.0000	0.1978	0.2145

(b).

Peptide counts in 6 tissues						mRNA expression in 6 tissues						Bayesian method	Maximum Likelihood	Linear Regression
6	59	58	41	174	560	863.2	635.3	1032.2	674.5	589.6	810.9	0.7775	0.0000	0.4743
22	98	38	22	306	59	877.1	440.4	462.1	431.7	432.5	579.8	0.2730	0.0000	0.8858
0	23	43	0	217	0	852.4	1046.7	873.8	692.3	855.8	1184.5	1.0000	0.0000	0.5882
26	57	53	7	114	19	788.9	540.1	825.7	611.4	517.5	897.4	0.4604	0.0000	0.8748
188	17	17	3	24	7	1960.8	1187.2	887.8	1142.0	1142.9	1301.1	0.6769	0.0000	0.1059
15	61	16	2	83	15	421.5	2087.1	1097.3	1422.5	1806.1	1827.0	0.5237	0.0000	0.1365
7	24	7	0	57	38	923.7	785.4	569.1	936.6	824.0	938.5	0.4373	0.0000	0.4819
8	2	10	27	40	30	1335.5	440.4	462.1	485.2	1158.1	598.4	0.2170	0.0000	0.3004
0	8	0	5	55	12	421.7	500.0	690.6	462.1	443.1	550.1	0.1393	0.0000	0.7688
0	0	0	0	79	0	1151.9	928.1	1119.7	861.1	1036.0	862.7	1.0000	0.0000	0.4000

(c).

Peptide counts in 6 tissues						mRNA expression in 6 tissues						Bayesian method	Maximum Likelihood	Linear Regression
322	0	0	0	4	0	58229.3	4614.8	1816.2	1810.9	13710.9	2340.1	0.1379	0.1379	0.0000
0	0	0	128	0	0	421.5	466.9	8370.2	8573.9	978.3	622.6	0.6000	0.2000	0.0000
0	5	0	8	77	20	1024.9	8538.7	2883.7	9706.3	13624.5	11467.5	0.3175	0.0641	0.0000
0	0	0	0	102	0	2467.6	2848.1	1740.6	2926.6	3198.1	2449.6	1.0000	0.8000	0.0000
69	0	0	0	0	0	2262.4	1319.9	1940.4	895.0	2079.2	2116.0	1.0000	0.4100	0.0000
0	0	0	0	50	2	1321.4	1424.3	2167.1	1648.5	6201.2	2880.3	0.1379	0.1379	0.0000
0	0	0	0	33	2	1217.6	2176.6	2395.0	1354.7	5926.5	2775.7	0.1379	0.1379	0.0000
0	0	0	3	6	23	1710.5	4019.5	2968.3	4200.3	5247.1	5922.9	0.3445	0.0980	0.0000

Fig. 4. Comparison of the genes products when (a) Proposed Bayesian method (b) Maximum likelihood (c) Linear regression estimates a significant relationship than the other two methods. See accompanying text for details.

These correspond to genes where there is significant disagreement between the measurements of the two gene products. We found that 503 pairs of gene products were in this class. Several of these were blood-borne factors that showed highest mRNA probe signal intensity in liver (where they are primarily synthesized), whereas the corresponding proteins are preferentially detected in the lung and placenta (which are rich in blood vessels). We were able to uncover 409 genes in which the measurements of the two gene products were significantly correlated ($P(s = 1|\mathbf{m}, \mathbf{y}) \geq .66$ and had p-values that were within .05). We call these genes inliers. We group all genes that do not belong to either inliers or outliers as borderline. Figure 5 depicts a breakdown of the genes into the three categories.

We further performed analysis to find if these categorizations have relationships to biological functions. We found that inlier genes were significantly enriched (p-value $< 10^{-3}$) in gene ontology (GO) function annotations such as cell adhesion and central nervous system. The outlier genes were enriched for GO function annotations including embryogenesis and transport, while borderline genes were enriched with function annotations such as mitochondrion, and functional and skeletal developmental anomalies.

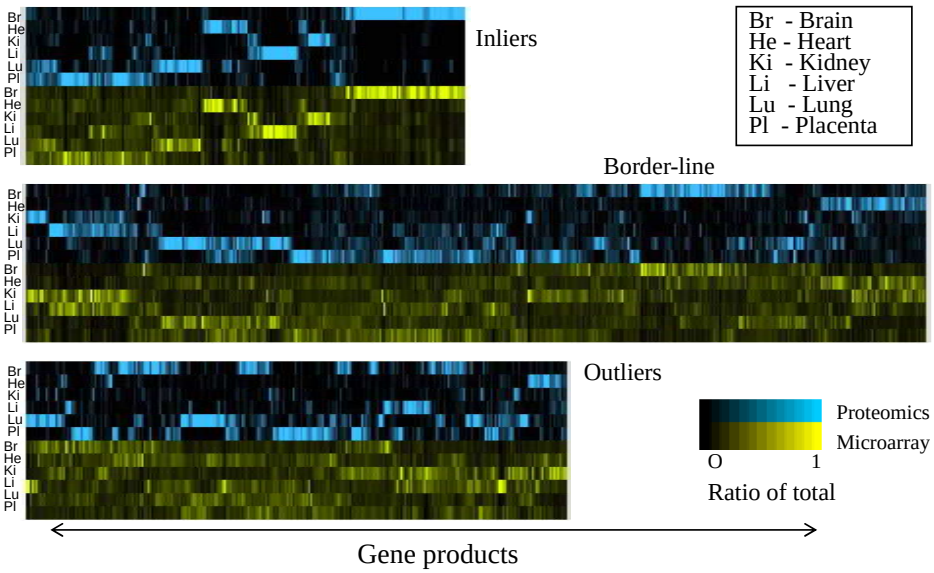

Fig. 5. Clustering of combined peptide and microarray gene profiles into three categories: inliers, borderline and outliers. We can see that for inliers, the genes are typically tissue specific and have similar expression patterns between the two data sets. For outliers, there is significant disagreement in expression profiles

5 Conclusions

We introduced a probabilistic model for inferring relationships between mRNA expression levels and protein abundance measured as peptide counts. Our model enables probabilistic scoring of the strength of the relationship between the gene products on a gene-by-gene basis. In addition, the same model can be used to test the significance of the relationship. Our model provides a principled framework for including various hidden variables and sources of uncertainty, both biological and experimental, that can affect the measured protein abundance and mRNA expression levels. Importantly, we showed that a Bayesian treatment of our model yields a larger number of statistically significant predictions, in comparison to a maximum-likelihood treatment of our model and linear regression (correlation). We studied experimental variability in the measurement of protein and mRNA levels and we were able to partition a set of genes into inliers, outliers and borderline, based on whether their gene products significantly agree or disagree or unknown. We can further augment the model to incorporate various other information such as protein functions, protein-protein interactions and temporal measurements. As an example, for learning regulatory networks, we can augment our proposed model for inferring correlation between mRNA of a gene and all its protein products, or infer RNA interference by finding relationships between several mRNAs and a single protein.

Acknowledgments

We thank Clement Chung, Brian Cox, Thomas Kislinger, Quaid Morris and Timothy Hughes for helpful discussions. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada, Genome Canada/Ontario Genomics Institute, and the Canadian Institutes for Health Research. B.J. Frey is a Fellow of the Canadian Institute for Advanced Research.

References

1. Kislinger, T., Cox, B., Kannan, A. *et al.*, *Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling* **Cell**,Apr 7;125(1):173-86(2006).
2. Greenbaum, D., Colangelo, C., Williams, K., Gerstein, M. *Comparing protein abundance and mRNA expression levels on a genomic scale.* **Genome Biol.** ,4(9):117(2003)
3. Gygi, SP., Rochon, Y., Franza, BR., Aebersold, R. *Correlation between protein and mRNA abundance in yeast* **Mol Cell Biology**, Mar;19(3):1720-30,(1999).
4. Griffin, TJ., Gygi, SP., Ideker, T., Rist, B., Eng, J., Hood, L., Aebersold, R., *Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae* **Mol Cell Proteomics**, Apr;1(4):323-33.,(2002).
5. Lian, Z., Kluger, Y., Greenbaum, DS., Tuck, D., Gerstein, M., Berliner, N., Weissman, SM. and Newburger, PE. *Genomic and proteomic analysis of the myeloid differentiation program: global analysis of gene expression during induced differentiation in the MPRO cell line* **Blood**, Nov 1;100(9):3209-20, (2002).
6. Mootha, VK., Bunkenborg, J., Olsen, JV., Hjerrild, M., Wisniewski, JR., Stahl, E., Bolouri, MS., Ray, HN., Sihag, S., Kamal, M., Patterson, N., Lander, ES., Mann, M. *Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria.* **Cell**, Nov 26;115(5):629-40, (2003).
7. Liu, H., Sadygov, RG., Yates, JR. *A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics* **Anal. Chem.**,76 (14), 4193 -4201(2004).
8. Zhang, W. , Morris, Q., *et al.*, *The functional landscape of mouse gene expression.* **Journal of Biology**,3(5):21(2004).
9. Su, A., Wiltshire, T. , Batalov, S. *et al.*, *A gene atlas of the mouse and human protein-encoding transcriptomes* **PNAS**, 101(16): 60626067 (2004)
10. Duda,RO. and Hart,PE. *Pattern Classification and Scene Analysis* **Wiley-Interscience**, (2000).
11. Jordan,MI. , Ghahramani,Z., Jaakkola, TS. and Saul, LK, *An Introduction to Variational Methods for Graphical Models* **Machine Learning**, 37-2, (1999).
12. R.M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods* **Technical Report, University of Toronto**, CRG-TR-93-1, (1993).