Iterative analysis of microarray data

Delbert Dueck[1], Jim Huang[1], Quaid D. Morris and Brendan J. Frey

([1] indicates co-first-authors)

Electrical and Computer Engineering
University of Toronto
10 King's College Road, Toronto, ON M5S 3G4
www.psi.toronto.edu

Abstract

In the past decade, technologies have been developed that enable researchers to quantify the levels of specific DNA transcripts present in cultures and tissues. While the signal-to-noise ratios of these technologies has consistently improved over the years, the signals provide only an observable projection of hidden, complex underlying processes. A variety of interesting information processing problems aimed at uncovering these hidden processes have emerged, and in this paper we survey some of the problems being studied in our group.

1 Introduction

One of the most significant discoveries of the twentieth century is that the bio-chemical processes used to build and maintain most living organisms are controlled by discrete-symbol genetic programs stored in bio-molecules. Examples of such genetic programs include DNA, RNA and proteins. The "central dogma" of molecular biology is that DNA encodes genes, which are transcribed from the DNA into RNA sequences, which are translated into proteins, which carry out cellular activities. Advances in bio-molecular sensing technologies have enabled researchers to probe the genetic states of organisms and determine the presence of specific sequences. These technologies can be applied to a variety of problems, ranging from determining the entire DNA sequence that is specific to an organism (e.g., the "human genome") to detecting the activation of a specific gene in a tissue sample (e.g., for the purpose of detecting disease or monitoring drug response). Although researchers now have digital representations of the actors in cellular processes and the ability to probe these actors, it is not yet known how the information is digitally encoded and how the digital components interact to regulate cellular activities.

One technology for probing the amounts of various RNA sequences in a tissue sample is the "mRNA expression array" [1]. An expression array is a $1'' \times 2''$ biology slide containing tens of thousands of probes, each of which is specific to a particular nucleic acid sequence (e.g., 40 nucleotides). After many mRNA sequences are extracted from a tissue sample, they are tagged with a fluorescent molecule and washed over the array. mRNA sequences from the tissue that match each probe will tend to hybridize (stick to the probe), so when the slide is scanned by a laser, the level of florescence at each probe indicates the amount of the corresponding mRNA sequence in the tissue sample. If the mRNA probes are selected properly, each probe can also indicate the degree to which a corresponding gene is expressed in the tissue, which is why these arrays are

sometimes called "gene expression arrays" or "gene microarrays". By varying the type of sample (e.g., heart versus brain) or by varying the conditions under which the sample is taken (e.g., healthy versus diseased), multiple arrays can be used to construct a vector of expression levels for each gene, where the elements of the vector correspond to the different experimental conditions. Such a vector is called an "expression profile".

By viewing expression profiles as points in a vector space, researchers have been able to use well-known vector-space data analysis techniques to identify new patterns of biological significance, and quickly make predictions based on previous studies that required a large amount of time and resources. In particular, the construction of these profiles has enabled the large-scale prediction of gene function for genes with unknown function, based on genes with known function [2, 13]. Because many biological functions depend upon the coordinated expression of multiple genes, similarity of expression profile often implies similarity of function [4]. This relationship has been used by authors employing discriminative techniques to predict the function of uncharacterized genes [5]. These schemes make use of annotation databases, which assign characterized genes to one or more predefined functional categories.

A variety of interesting information processing problems emerge in the analysis of microarray data, ranging from problems of data denoising to problems of finding new gene structures (c.f. [3]) and functional annotation of known genes. In this paper, we describe two projects currently underway in our group. The first project deals with clustering unannotated genes into multiple, simultaneous categories, which can be viewed as a sparse matrix factorization problem. (See [9] for a detailed description of our method.) The second project deals with removing the effects of cross-hybridization, where the probe designed for a specific mRNA hybridizes with a different mRNA, and the goal is to remove this cross-hybridization noise from the microarray data.

2 Probabilistic sparse matrix factorization

Often, data vectors lie in a low-dimensional subspace. One approach to analyzing data vectors lying in a low-dimensional linear subspace is to stack them to form a data matrix, $\mathbf{X}$, and then find a low-rank matrix factorization of the data matrix. Given $\mathbf{X} \in \mathcal{R}^{G \times T}$, matrix factorization techniques find a $\mathbf{Y} \in \mathcal{R}^{G \times C}$ and a $\mathbf{Z} \in \mathcal{R}^{C \times T}$ such that $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$. Interpreting the rows of $\mathbf{X}$ as input vectors $\{\mathbf{x}_g\}_{g=1}^{G}$, the rows of $\mathbf{Z}$ (i.e. $\{\mathbf{z}_c\}_{c=1}^{C}$) can be viewed as vectors that span the C-dimensional linear subspace, in which case the g^{th} row of $\mathbf{Y}$ contains the coefficients $\{y_{gc}\}_{c=1}^{C}$ that combine these vectors to explain the g^{th} row of $\mathbf{X}$.

A variety of techniques have been proposed for finding matrix factorizations, including non-probabilistic techniques such as principal components analysis (PCA), and probabilistic techniques that account for noise, such as factor analysis. In many situations, we anticipate that the data consists of additive combinations of positive sources. In these cases, non-probabilistic techniques have been proposed for finding factorizations under non-negativity constraints [6],[7]. For many kinds of data, we expect the low-dimensional representation for each input vector to be sparse This type of problem was called "sparse matrix factorization" in [8], and is related to independent component analysis [10]. In their model, Srebro and Jaakkola augment the $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$ matrix factorization setup with the sparseness structure constraint that each row of $\mathbf{Y}$ has at most N non-zero entries. They then describe an iterative algorithm for finding a sparse matrix factorization that makes hard decisions at each step.

We extend previous work on sparse matrix factorization and derive a method that

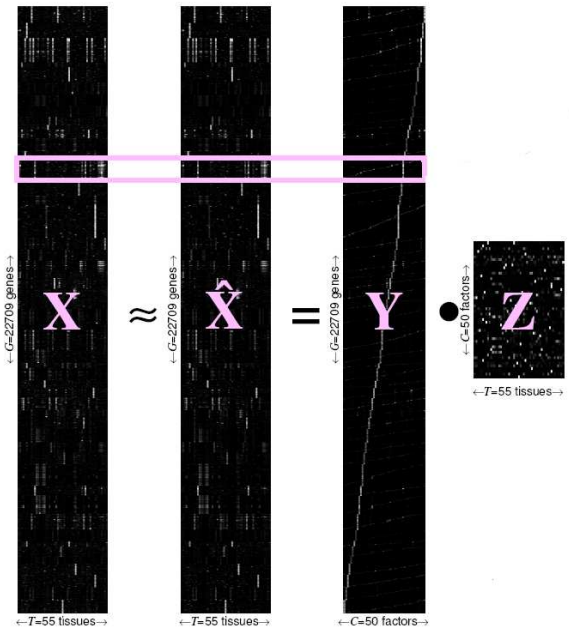

Figure 1: Data matrix $\mathbf{X}$, approximated by the product of a sparse matrix $\mathbf{Y}$ and a low-rank matrix $\mathbf{Z}$. Gene expression profiles appear as row vectors in $\mathbf{X}$ and are sorted by primary class (s_{g1}), secondary class (s_{g2}), etc.

finds such sparse factorizations while accounting for uncertainties due to (1) different levels of noise in the data, (2) different levels of noise in the factors used to explain the data, and (3) uncertainty as to which hidden prototypes are selected to explain each input vector. An example of such a factorization is shown in Figure 1.

Let $\mathbf{X}$ be the matrix of gene expression data such that rows correspond to each of G genes and columns to each of T tissues (i.e. entry x_{gt} represents the amount by which gene g is expressed in cells of tissue type t.) We denote the collection of unobserved transcription factor profiles as a matrix, $\mathbf{Z}$, with rows corresponding to each of C transcription factors and T columns corresponding to tissues, as before. Each gene expression profile, $\mathbf{x}_g$, can be approximated by a linear combination of a small number (r_g) of these transcription factor profiles, $\mathbf{z}_c$:

$$\mathbf{x}_g \approx \sum_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}} \tag{1}$$

The transcription factor profiles contributing to the g^{th} gene expression profile are indexed by $\{s_{g1}, s_{g2}, \ldots, s_{gr_g}\}$, with corresponding weights $\{y_{gs_{g1}}, y_{gs_{g2}}, \ldots, y_{gs_{gr_g}}\}$. This is identical to the $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$ matrix factorization with $\{\mathbf{S}, \mathbf{r}\}$ representing the sparseness structure constraint. We account for varying levels of noise in the observed data by assuming the presence of isotropic Gaussian sensor noise (with variance ψ_g^2) for each gene's expression profile, so the likelihood of $\mathbf{x}_g$ is as follows:

$$P(\mathbf{x}_g | \mathbf{y}_g, \mathbf{Z}, \mathbf{s}_g, r_g, \psi_g^2) = \mathcal{N}\Big(\mathbf{x}_g;\, \sum_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}},\, \psi_g^2 \mathbf{I}\Big) \tag{2}$$

We complete the model with prior assumptions that the transcription factor profiles ($\mathbf{z}_c$) are normally distributed and that the transcription factor indices (s_{gn}) are uniformly distributed. The number of causes, r_g, contributing to each gene's profile is multinomially distributed such that $P(r_g = n) = \nu_n$, where ν is a user-specified N-vector. We make no assumptions about $\mathbf{Y}$ beyond the sparseness constraint, so $P(\mathbf{Y}) \propto 1$.

Multiplying these priors by (2) forms the following joint distribution:

$$P(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\Psi) = P(\mathbf{X}|\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r},\Psi)\cdot P(\mathbf{Y})\cdot P(\mathbf{Z})\cdot P(\mathbf{S}|\mathbf{r})\cdot P(\mathbf{r}) \propto \tag{3}$$
$$\prod_{g=1}^{G}\mathcal{N}\Big(\mathbf{x}_g;\, \textstyle\sum_{n=1}^{r_g} y_{gs_{gn}}\mathbf{z}_{s_{gn}},\, \psi_g^2\mathbf{I}\Big)\cdot\prod_{c=1}^{C}\mathcal{N}(\mathbf{z}_c;\,\mathbf{0},\,\mathbf{I}) \quad \cdot\prod_{g=1}^{G}\prod_{c=1}^{C}\prod_{n=1}^{N}\big(\tfrac{1}{C}\big)^{\delta(s_{gn}-c)}\cdot\prod_{g=1}^{G}\prod_{n=1}^{N}(\nu_n)^{\delta(r_g-n)}$$

Exact inference with (3) is intractable so we utilize a factorized variational method [11] and approximate the posterior distribution with a mean-field decomposition:

$$P(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\mathbf{X},\Psi) \approx \prod_{g=1}^{G} Q(\mathbf{y}_g)\cdot\prod_{c=1}^{C} Q(\mathbf{z}_c)\prod_{g=1}^{G}\prod_{n=1}^{N} Q(s_{gn})\cdot\prod_{g=1}^{G} Q(r_g) \tag{4}$$

We parameterize the Q-distribution as follows:

$$Q(\mathbf{y}_g) = \overbrace{\prod_{n=1}^{r_g}\delta\left(y_{gs_{gn}}-\lambda_{gs_{gn}}\right)}^{\lambda_{gc}\text{ is a point estimate of }y_{gc}}\cdot\overbrace{\prod_{\substack{c=1\\ c\notin\{s_{g1},s_{g2},\ldots,s_{gr_g}\}}}^{C}\delta\left(y_{gc}\right)}^{\text{but for }\{\mathbf{s}_g,r_g\}}$$

$$Q(z_{ct}) = \mathcal{N}(z_{ct};\,\zeta_{ct},\;,\phi_c^2),\quad Q(s_{gn}{=}c) = \sigma_{gnc},\quad Q(r_g{=}n) = \rho_{gn}.$$

Using this approach, factorization corresponds to bringing the Q-distribution as close as possible to the P-distribution by varying the Q-distribution parameters (hence, λ, ζ, ϕ, σ, and ρ are referred to as variational parameters). To ensure normalization, the variational parameters are constrained to satisfy $\sum_{c=1}^{C}\sigma_{gnc}{=}1$ and $\sum_{n=1}^{N}\rho_{gn}{=}1$ — these become Lagrange multipliers in the later optimization problem.

In order to closely approximate the P-distribution, we seek to minimize the relative entropy, $\mathrm{D}(Q\|P)$, between it and the Q-distribution:

$$\min_{\substack{\{\lambda,\zeta,\phi,\sigma,\rho\}\\ \mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}}}\int Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})\cdot\log\frac{Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})}{P(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\mathbf{X},\Psi)} \tag{5}$$

There is no closed-form expression for the posterior (denominator in (5)), but we can subtract $\log P(\mathbf{X})$ inside the integral (it is independent of the variational parameters) to form the readily-minimized free energy, $\mathcal{F}$. The free energy can be minimized sequentially with respect to each variational parameter (λ, ζ, ϕ, σ, ρ) by analytically finding zeros of the partial derivatives with respect to them. This coordinate descent represents the E-step in variational EM [11] that alternates with a brief M-step, where the global sensor noise is learned by similarly solving $\partial\mathcal{F}/\partial\Psi = 0$. For a more detailed treatment of the algorithm and parameter update equations, see [12].

2.1 Experimental Results

To experimentally analyze the performance of our algorithm, we use the expression data from Zhang et al. in [13]. Each gene's profile is a measurement of its expression in a set of 55 mouse tissues. Expression levels are rectified at the median expression level, so that the data set is entirely non-negative, consistent with the notion of light intensities also being non-negative.

The functional category labels for the genes with known biological function were taken from [13]. An example of a category label is "cell wall biosynthesis", which indicates the gene expresses a protein that is involved in building the cell wall. These labels are derived from Gene Ontology Biological Process (GO-BP) category labels [14] assigned to genes by EBI and MGI.

Among the 22,709 genes in the Zhang et al. database, 9499 have annotations in one or more of 992 different GO categories. The category sizes range from 3 to 456 genes, with more than half the categories having fewer than 20 genes.

We present results for the 22,709 genes $\times$ 55 tissues data set shown in Figure 1. The data matrix, $\mathbf{X}$, is shown alongside the model's approximation, $\hat{\mathbf{X}}$, also expressed as $\hat{\mathbf{X}} = \mathbf{Y} \cdot \mathbf{Z}$. A total of $C = 50$ factors were used, and the user-specified prior on the number of factors (r_g) explaining each expression profile was set to $\nu = \begin{bmatrix} .55 & .27 & .18 \end{bmatrix}$, making $N = 3$.[1] Gene expression profiles (row vectors) are first sorted by 'primary' transcription factor (s_{g1}); next, within each s_{g1} grouping, genes are sorted by 'secondary' transcription factor (s_{g2}), and so on. This organization is easily visualized in the hierarchical diagonal structure of $\mathbf{Y}$.

The overall objective for this research is to develop a model accurately capturing biologically significant factors behind gene expression data. Success can be measured by determining how closely each of the transcription factors learned by the model are aligned with Gene Ontology (GO) categories (a.k.a. the "enrichment") [15]. We use p-values calculated from the hypergeometric distribution to quantify the probability that categorizations could have resulted by mere chance [16]. Specifically, we select the subset of genes whose 'primary' factor is transcription factor c (i.e. $s_{g1} = c$), and compare this cluster alongside the most similar GO category, assigning the significance of transcription factor c to be the resulting hypergeometric p-value. This process is repeated for 'secondary' (s_{g2}) and 'tertiary' (s_{g3}) factors.

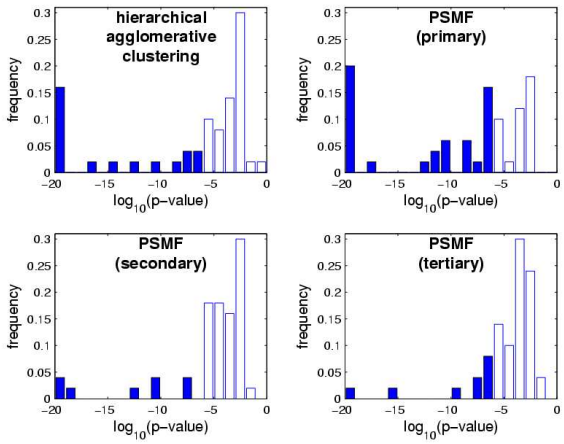

Figure 2: P-values for hierarchical agglomerative clustering and probabilistic sparse matrix factorization (C=50, N=3). Significant factors/clusters (after Bonferroni correction of threshold $\alpha = .05$) are shown as solid bars.

[1] A uniform prior ν (reflecting no knowledge about the distribution of $\mathbf{r}$) would give equal preference to all values of a particular r_g. For any given $r_g < N$, a factor can almost always be found that, if present with infinitesimal weight (y_{gc}), will imperceptibly improve the cost function ($\mathcal{F}$), with the end result that almost all r_g would then equal N. Weighting the prior towards lower values ensures that factors will only be included if they make a noteworthy difference. We only choose $\nu_n \propto 1/n$, $\forall n < N$ for simplicity.

Histograms of these p-values, as well as for industry-standard hierarchical agglomerative clustering (by Pearson correlation) into C classes [17], are shown in Figure 2.

A p-value of $\alpha = .05$ may be used as a threshold of significance, however this threshold must first be corrected to account for the fact that the most favorable GO category was chosen for comparison with each cluster. The Bonferroni correction conservatively accounts for this by dividing the significance threshold by the total number of comparisons (number of GO categories multiplied by number of clusters examined). In [9], we show that the soft-decision factorized variational method (PSMF) performs better than the hard-decision iterated conditional modes technique (SMF) [8, 18], in terms of finding statistically significant clusters.

2.2 Conclusions

Many kinds of data vectors can most naturally be explained as an additive combination of a selection of prototype vectors, which can be viewed as computational problem of finding a sparse matrix factorization. While most work on biological gene expression arrays has focused on clustering techniques and methods for dimensionality reduction, there is recent interest in performing these tasks jointly, which corresponds to sparse matrix factorization. Like [8], our algorithm computes a sparse matrix factorization, but instead of making point estimates (hard decisions) for factor selections, our algorithm computes probability distributions. We find that this enables the algorithm to avoid local minima found by iterated conditional modes and obtain clusters that have higher statistical significance than other methods.

3 Iterative cross-hybridization compensation

While in the previous section we described progress on the high-level task of functional annotation of genes based on microarray data, in this section, we describe progress on a data-preprocessing problem: removing cross-hybridization signals from microarray data.

One of the important problems in DNA microarray analysis and gene functional prediction is that the data generated from microarray experiments is typically prone to many types of noise from various sources. A particular class of noise that has received little attention so far is the biological effect of gene cross-hybridization on microarray data [19]. Here, probes on a microarray will measure not only the mRNA level of their target gene, but also the unwanted mRNA levels of other, non-specific genes. We propose a signal model to represent the effect of cross-hybridization in microarray data, in which the expression profiles measured by probes are to be explained by taking a weighted linear sum of a relatively small number of latent gene expression profile variables. To compensate for the effect of cross-hybridization, these hidden gene expression profiles are iteratively inferred from the data via a method based on the Iterated Conditional Modes (ICM) [18] inference algorithm which will then use the inferred quantities to perform cross-hybridization compensation (XHC). Microarray data in which this cross-hybridization effect has been removed will allow for increased specificity and sensitivity in predicting the function of genes from microarray data produced from microarray experiments.

To gauge the impact of cross-hybridization on the observed expression data, we performed a statistical test in which each observed expression profile vector from the data described below was normalized in the range $[0, 1]$. Taking the oligonucleotide sequences of the probes on the array and the corresponding sequences of their target genes, we performed a sequence-based search [21] to find the possible probe-to-gene mappings. We

then computed the Pearson correlation coefficient between randomly-paired profiles and sequence-matched profiles and found that for a particular correlation threshold, 33% of the sequence-matched profiles have high correlation, while only 2% of randomly-paired profiles have high correlation. This validates our hypothesis that the data contains a significant level of cross-hybridization signal.

To remove cross-hybridization noise, we model each probe in the array as being capable of hybridizing to multiple genes. Conversely, each gene can have multiple probes that can hybridize to it. We assume a model for the microarray data in which the set of N observed expression profile vectors X can be explained by a smaller set of M latent gene expression profiles Z via an affine transformation represented by the *hybridization matrix* Λ plus independent Gaussian noise with covariance $\sigma^2 I$. Let the microarray contain N probes and let T be the number of measurements per probe. For a given probe, each of these T measurements corresponds to a scalar mRNA expression level in a particular tissue pool. Let $X = [x_1, x_2, \cdots, x_T]$ denote a matrix in which the t^{th} column denotes the measurement across all N probes in the microarray for tissue pool $t, t = 0, 1, \cdots, T$. Accordingly, let $Z = [z_1, z_2, \cdots, z_T]$ denote the matrix of *hidden gene expression profiles* to be inferred over the T tissue pools; the t^{th} column of Z therefore denotes the true, latent mRNA expression levels across all M probed genes for tissue pool t. We thus model the relationship between X and Z as follows: $x_t \sim \mathcal{N}(\Lambda z_t, \sigma^2 I)$, $z_t \sim \delta(z - z_t)$, where $x_t \in \mathbb{R}_+^N, z_t \in \mathbb{R}_+^M, \Lambda \in \mathbb{R}_+^{(N \times M)}$. Thus, we seek to iteratively estimate the hybridization matrix Λ and the set of latent gene expression profiles Z over all T measurements by maximizing the log-likelihood of the observed data X under the assumption of uniform random sensor noise level across all probes.

The log-likelihood $\mathcal{L}(X, Z)$ of the observed data X and the latent data Z can be written (omitting constant terms) as

$$\begin{aligned} \mathcal{L}(X, Z) &= \log \prod_{t=1}^{T} p(x_t|z_t)p(z_t) = \sum_{t=1}^{T} \log p(x_t|z_t)p(z_t) \qquad (6) \\ &= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (x_t - \Lambda z_t)^T (x_t - \Lambda z_t) = -\frac{1}{2\sigma^2} \sum_{t=1}^{T} \sum_{n=1}^{N} (x_{tn} - z_{tm(n,1)} - \lambda z_{tm(n,2)})^2, \end{aligned}$$

where $m(n, 1)$ and $m(n, 2)$ denote the gene indices corresponding to the primary and secondary gene respectively for probe n. Optimizing the above equations by taking derivatives with respect to the hidden profiles Z and the free $\lambda_2^{(n)}$ parameters and setting them to zero yields the following ICM update equations for the latent profiles and the $\lambda_2^{(n)}$ parameters:

$$z_{tk} = \frac{\sum_{n:m(n,1)=k} (x_{tn} - \lambda_2^{(n)} z_{tm(n,2)}) + \sum_{n:m(n,2)=k} \lambda_2^{(n)} (x_{tn} - z_{tm(n,1)})}{M_1^{(k)} + \sum_{n:m(n,2)=k} (\lambda_2^{(n)})^2} \qquad (7)$$

and

$$\lambda_2^{(n)} = \frac{\sum_t z_{tm(n,2)} (x_{tn} - z_{tm(n,1)})}{\sum_t z_{tm(n,2)}^2} \qquad (8)$$

where $M_1^{(k)}$ is the number of array probes that can hybridize to gene k. Thus, we iterate between equations (7) and (8) until convergence in the log-likelihood. The latent variable and paramater estimates z_{tn} and $\lambda_2^{(n)}$ are therefore the pointwise *maximum a posteriori* (MAP) values corresponding to the maxima of the joint distribution $p(X, Z)$.

3.1 Experimental results

We pre-processed the data by removing genes that exhibited a large variance in their corresponding probes' measured expression levels. The possible probe-to-gene sequence mappings were determined via a BLAST search [21] (default settings) between the probe oligonucleotide sequences and the sequences of the targeted genes. The output of the search is a ranked list of matching gene nucleotide sequences for each probe sequence, each of which is ranked by its number of mismatches to the probe. The number of mismatches is a good measure of the hybridization binding potential between a probe and its target gene according to the thermodynamic properties of hybridization [20]. In our initial experiments, we limit the number of genes per probe to 2. The 2 genes to which each probe was mapped were selected according to the 2 best matching genes for the probe in the output of the BLAST search. In addition, each gene was assumed to have at least one *primary probe* associated with it: that is, a probe which hybridizes to that gene with a unity hybridization coefficient. We removed any genes that hybridized to less than 3 probes on the array; this was done to remove gene profiles that would be uninformative for the purposes of inferring the hidden gene expression profiles based on the effect of cross-hybridization.

We tested the algorithm for two cases: A probe hybridizes to a primary and secondary gene according to its BLAST matches; A probe hybridizes to its primary probe according to its best BLAST match and to a randomly-selected secondary gene. The objective is thus to verify that the ICM algorithm is not simply fitting coefficients to noisy gene expression profiles which do not have biological meaning. In addition, we ran the ICM algorithm in the cases that each probe hybridizes to only one gene and then two genes. This allows us to verify that the use of 2 gene expression profiles to explain an observed expression profile allows for a significant reduction in reconstruction error than if only 1 gene profile is used. We report results using the Signal-to-Noise Ratio (SNR) over the entire array, as a measure of how much error is incurred in estimating the observed expression profiles using the inferred gene profiles. We define the microarray SNR as $SNR_{array} = 10 \log \left(\frac{\sum_t \|x_t\|^2}{\sum_t \|x_t - \hat{x_t}\|^2} \right)$, where $\hat{x_t} = \Lambda z_t$ is the estimate of the observed expression profile x_t given Λ and z_t.

Figure 3 shows the cumulative distribution plot of the resulting SNRs. A higher proportion of probes have high SNR figures with XHC using the pre-processed data than the proportion obtained from optimizing on the full measured data set or a subset thereof that is thresholded at the 33^{rd} percentile in expression level. The resulting gain in microarray SNR with respect to the case in which a single gene is assigned to a probe suggests that performing XHC produces an increase in SNR by increasing the precision with which the observed data can be estimated by taking into account multiple sources of hybridization for each probe. In addition, the significant gain in SNR with respect to the case in which each probe's secondary gene is randomly assigned further enhances the biological validity and significance of the inferred gene expression profiles.

3.2 Conclusions

We have shown that cross-hybridization has a significant impact on a non-negligible proportion of microarray measurements and we have introduced a model and iterative algorithm that partially compensates for cross-hybridization. Possible improvements to the current approach include allowing for more than 2 hybridizing genes per probe and allowing the number of cross-hybridizing genes to be randomly distributed; allowing for

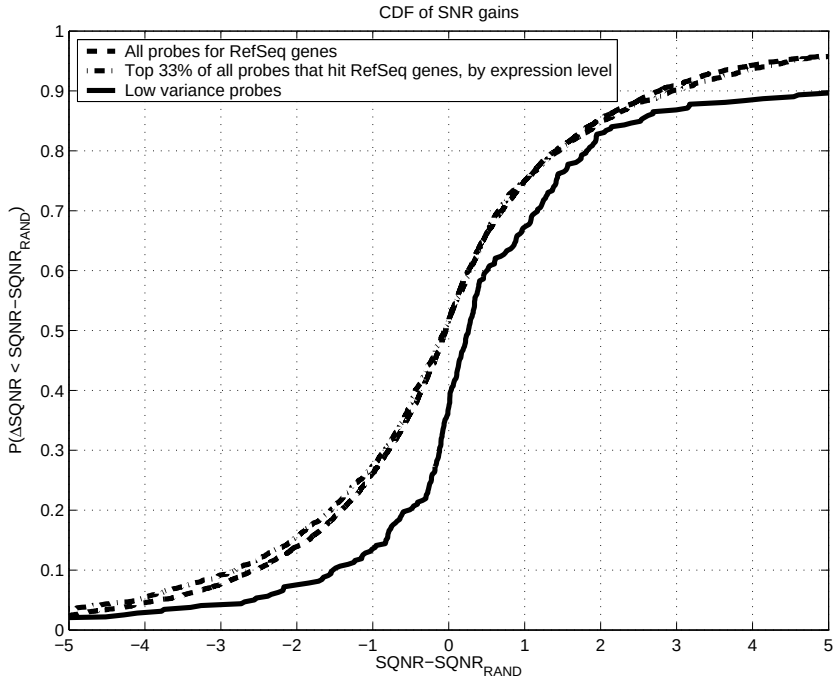

Figure 3: Cumulative distributions of SNR gains for cross-hybridization compensation.

a probability distribution to model uncertainty in the hidden gene profiles; introducing a prior distribution for modelling the sparse structure of the parameter matrix Λ; allowing for a more robust estimation of gene expression profiles with respect to noisy expression data and outliers.

Acknowledgements

We thank Tim Hughes from the University of Toronto for making the data available, and Sam Roweis, Tommi Jaakkola and Nati Srebro for helpful discussions. We also gratefully acknowledge the financial support of NSERC and a grant to M. Escobar from CIHR.

References

[1] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467–470.

[2] Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.

[3] Frey, B.J. *et al.* (2005) GenRate: A generative model that finds and scores new genes and exons in genomic microarray data. *Proceedings of the Pacific Symposium on BioComputing*, Hawaii, January, 2005.

[4] Marcotte, E.M., et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**: 751–753.

[5] Brown, M.P. et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences* **97**: 262–267.

[6] Lee, D. and Seung, H. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791.

[7] Lee, D. and Seung, H. (2001) Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 7*, pp. 556–562.

[8] Srebro, N. and Jaakkola, T. (2001) Sparse Matrix Factorization of Gene Expression Data. Unpublished note, MIT Artificial Intelligence Laboratory. Available at www.ai.mit.edu/-research/abstracts/abstracts2001/genomics/01srebro.pdf

[9] Dueck, D., Morris, Q.D., Frey, B.J. (2004) Probabilistic sparse matrix factorization with an application to microarray data. Submitted to *Artificial Intelligence and Statistics 2005*.

[10] Bell, A. J. and Sejnowski, T. J. (1995) An information maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**: 1129-1159.

[11] Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1998) An introduction to variational methods for graphical models. In M.I. Jordan (ed.), *Learning in Graphical Models*. Norwell, MA: Kluwer Academic Publishers.

[12] Dueck, D., and Frey, B. (2004) Probabilistic Sparse Matrix Factorization. University of Toronto technical report PSI-2004-23.

[13] Zhang, W., Morris, Q., et al. (2004) The functional landscape of mouse gene expression. Submitted for publication.

[14] Ashburner, M. et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**: 25–29.

[15] Bar-Joseph, Z. et al. (2003) K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data. *Bioinfomatics* **19**: 1070–1078.

[16] Tavazoie, S. et al. (1999) Systematic determination of genetic network architecture. *Nature Genetics* **22(3)**, 213–215.

[17] Eisen, M.V., Spellman, P.T., Brown, P.O., and Botstein, D. (1999) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**: 14863–14868.

[18] Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B* **48**: 259–302.

[19] J.D. Wren, A. Kulkarni, J. Joslin, R.A. Butow, and H.R. Garner. "Cross-Hybridization on PCR-Spotted Microarrays," *IEEE Engineering in Medecine and Biology, Vol. 21, pp. 71-75, March 2002.*

[20] J.J. SantaLucia, H.T. Allawi, and P.A. Seneviratne. "Improved Nearest-Neighbor Parameters for predicting DNA duplex stability," *Biochemistry, Vol. 35, pp. 3555-3562, 1998.*

[21] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. "Basic local alignment search tool," *J Mol Biol 215(3):403-10, 1990.*

[22] B.J. Frey, N. Mohammad, W. Zhang, Q.D. Morris, M.D. Robinson, R. Chang, O. Shai, S. Mnaimneh, Q. Pan, J. Rossant, J. Aubin, B.J. Blencowe, and T.R. Hughes. "Full-genome exon profiling in mus musculus " *(in preparation).*