Finding Novel Transcripts in High-Resolution Genome-Wide Microarray Data Using the GenRate Model

Brendan J. Frey[1], Quaid D. Morris[1,2],
Mark Robinson[1,2], and Timothy R. Hughes[2]

[1] Elec. and Comp. Eng., Univ. of Toronto,
Toronto ON M5S 3G4, Canada
http://www.psi.toronto.edu
[2] Banting and Best Dep. Med. Res., Univ. of Toronto,
Toronto ON M5G 1L6, Canada
http://hugheslab.med.utoronto.ca

Abstract. Genome-wide microarray designs containing millions to tens of millions of probes will soon become available for a variety of mammals, including mouse and human. These "tiling arrays" can potentially lead to significant advances in science and medicine, *e.g.*, by indicating new genes and alternative primary and secondary transcripts. While bottom-up pattern matching techniques (*e.g.*, hierarchical clustering) can be used to find gene structures in tiling data, we believe the many interacting hidden variables and complex noise patterns more naturally lead to an analysis based on generative models. We describe a generative model of tiling data and show how the iterative sum-product algorithm can be used to infer hybridization noise, probe sensitivity, new transcripts and alternative transcripts. We apply our method, called GenRate, to a new exon tiling data set from mouse chromosome 4 and show that it makes significantly more predictions than a previously described hierarchical clustering method at the same false positive rate. GenRate correctly predicts many known genes, and also predicts new gene structures. As new problems arise, additional hidden variables can be incorporated into the model in a principled fashion, so we believe that GenRate will prove to be a useful tool in the new era of genome-wide tiling microarray analysis.

1 Introduction

One of the most important current problems in molecular biology is the development of techniques for building libraries of genes and gene variants for organisms, and in particular higher mammals such as mice and humans. While an analysis of genomic nucleotide sequence data can be used to build such a library (c.f. [30]), it is the mRNA molecules that are transcribed from genomic DNA ("transcripts") that directly or indirectly constitute the library of functional elements. In fact, the many complex mechanisms that influence transcription of genomic DNA into mRNAs produces a set of functional transcripts that is

S. Miyano et al. (Eds.): RECOMB 2005, LNBI 3500, pp. 66–82, 2005.
© Springer-Verlag Berlin Heidelberg 2005

much richer than can be currently explained by analyzing genomic DNA alone. This richness is due to many mechanisms including transcription of non-protein-coding mRNA molecules that are nonetheless functional (c.f. [2]), tissue-specific transcriptional activity, alternative transcription of genomic DNA (*e.g.*, alternative start/stop transcription sites), and alternative post-transcriptional splicing of mRNA molecules (c.f. [3, 5, 4, 9]). Instead of attempting to understand these variants by studying genomic DNA alone, microarrays can be used to directly study the rich library of functional transcriptional variants. Previously, we used microarrays to study subsets of variants, including non-protein-coding mRNAs [6] and alternative-splicing variants [8]. Here, we describe a technique called "GenRate", which applies the max-product algorithm in a graphical model to perform a genome-wide analysis of high-resolution (many probes "per gene") microarray data.

In 2001, Shoemaker *et al.* demonstrated for the first time how DNA microarrays can be used to validate and refine predicted transcripts in portions of human chromosome 22q, using 8,183 exon probes [21]. By "tiling" the genome with probes, patterns of expression can be used to discover expressed elements. In the past 3 years, the use of microarrays for the discovery of expressed elements in genomes has increased with improvements in density, flexibility, and accessibility of the technology. Two complementary tiling strategies have emerged. In the first, the genome is tiled using candidate elements (*e.g.*, exons, ORFs, genes, RNAs), each of which is identified computationally and is represented one or a few times on the array [19, 21, 11, 14]. In the second, the entire genome sequence is comprehensively tiled, *e.g.*, overlapping oligonucleotides encompassing both strands are printed on arrays, such that all possible expressed sequences are represented [21, 13, 20, 23, 14, 15, 31]. Genome-wide tiling data using both approaches is currently becoming available [32, 1].

The above tiling approaches, as well as independent analysis by other methods [17, 11] have indicated that a substantially higher proportion of genomes are expressed than are currently annotated. The most recent genome-wide mammalian survey by Bertone *et al.* is based on expression in a single tissue (liver) and claims to have found 10,595 novel transcribed sequences over and above genes detected by other methods. However, since microarray data is noisy and since probe sensitivity and cross-hybridization noise can vary tremendously from one probe to another (a factor of 40 is not unusual), it is quite difficult to control the false detection rate using only one tissue. Most protein-coding transcripts are composed of multiple exons and most mammalian genes vary in expression between tissues, so tissue-dependent co-regulation of probes that are nearby in the genome provides evidence of a transcriptional unit [21]. We will refer to such a co-regulated transcriptional unit as a "CoReg".

Microarrays do not inherently provide information regarding the length of the RNA or DNA molecules detected, nor do they inherently reveal whether features designed to detect adjacent features on the chromosome are in fact detecting the same transcript. mRNAs, which account for the largest proportion of transcribed sequence in a genome, present a particular challenge. mRNAs

are composed only of spliced exons, often separated in the genome (and in the primary transcript) by thousands to tens of thousands of bases of intronic sequence. Each gene may have a variety of transcript variants, *e.g.*, due to alternative splicing [3, 4] and exons that are conserved across species (*e.g.* human and mouse) often undergo species-specific splicing [9]. Identifying the exons that comprise individual transcripts from genome- or exon-tiling data is not a trivial task, since falsely-predicted exons, overlapping features, transcript variants, and poor-quality measurements can confound assumptions based on simple correlation of magnitude or co-variation of expression.

Heuristics that group nearby probes using intensity of expression or co-regulation across experimental conditions can be used to approach this problem [21, 13, 23, 14, 32]. In [21], correlations between the expression patterns of nearby probes are used to merge probes into CoRegs. A merge step takes place if the correlation exceeds 0.5, but not if the number of non-merged probes between the two candidate probes is greater than 5. In [13], the density of the activity map of RNA transcription is used to verify putative exons. In [32], a single tissue is studied and two probes are merged if their intensities are in the top 90th percentile and if they are within 250nt of each other. In [14], principal components analysis (PCA) is first applied to probes within a window. Then, the distribution of PCA-based Mahalanobis distances of probes are compared with the distribution of distances for known intron probes, and each probe is merged into a CoReg if the distance of the probe to a selection of the PCA subspaces is low.

While the above techniques have been quite helpful in analyzing microarray data, an important disadvantage of the techniques is that they do not directly model various sources of noise and the noisy relationships between variables. For example, a highly-sensitive probe will indicate the presence of transcript, even if the true abundance is negligible. A poorly-designed probe will cross-hybridize to many other transcripts, again misleadingly indicating the presence of the transcript for which the probe was designed. By not optimizing a global scoring function derived from a model of the various processes, these techniques tend to make greedy local decisions that are not globally optimal. For example, while the assignment of a probe to a CoReg may be locally optimal, this decision removes the probe from consideration in other CoRegs, so the decision may not be globally optimal. Further, because these techniques do not clearly identify the probabilistic relationships between relevant hidden variables (*e.g.*, gene start/stop sites), it is not straightforward to modify them to account for new hidden variables or new data types. Also, because the separation between modeling assumptions and the optimization technique is not clear, it is difficult to improve performance in a principled manner.

Inspired by recent successes in using graphical probability models (*e.g.*, Bayesian networks) to analyze microarray data (c.f. [22, 24]), we have developed a generative probability model which jointly accounts for the stochastic nature of the arrangement of exons in genomic DNA, the stochastic nature of transcript expression, and the properties of probe sensitivity and noise in microarray data.

Inference in this model balances different sources of probabilistic evidence and makes a globally-optimal set of decisions for CoRegs, *after* combining probabilistic evidence. In contrast to our recent work [25] where exact inference is tractable, the model described in this paper accounts for more complex gene structures, such as alternative splicing isoforms, so exact inference is computationally burdensome. We describe how iterative application of the sum-product algorithm (a.k.a. "loopy belief propagation") can be used for efficient probabilistic inference. We compare the performance of our technique with a bottom-up threshold-based hierarchical clustering method [21], and we find that at low false positive rates, GenRate finds at least five times more exons. We also present new results showing that out of many novel mouse gene structures predicted by GenRate, the 9 highest-scoring structures that we tested are *all* confirmed by RT-PCR sequencing experiments.

2 Microarray Data

The microarray data set, a portion of which is shown in Fig. 1, is a subset of a full-genome data set to be described and released elsewhere [1]. Briefly, exons were predicted from repeat-masked mouse draft genome sequence (Build 28) using five different exon-prediction programs. Once Build 33 became available, we mapped the putative exons to the new genome. (While this data is based on putative *exons*, GenRate can be applied to any sequence-based expression data set, including genome tiling data.) A total of 48,966 non-overlapping putative exons were contained on chromosome 4 in Build 33. One 60-mer oligonucleotide probe for each exon was selected using conventional procedures, such that its binding free energy for the corresponding putative exon was as low as possible compared to its binding free energy with sequence elsewhere in the genome, taking into account other constraints on probe design. (For simplicity, we assume each probe has a unique position in the genome.) Arrays designs were submitted

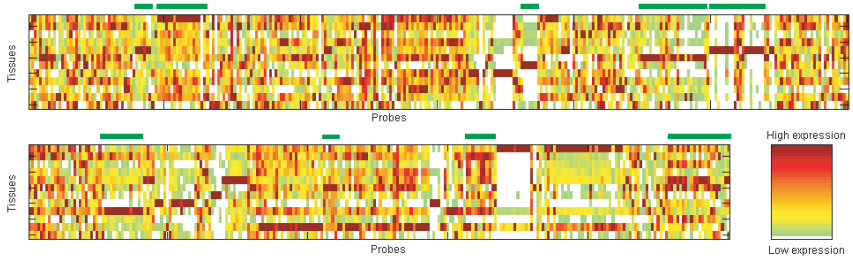


Fig. 1. A small fraction of our data set for chromosome 4, consisting of an expression measurement for each of 12 mouse tissue pools and 48,966 60-mer probes for repeat-masked putative exons arranged according to their order in Build 33 of the genome. Some "CoRegs" (co-regulated transcriptional units) were labeled by hand and are shown with green bars

to Agilent Technologies (Palo Alto, California) for array production. Twelve diverse samples were hybridized to the arrays, each consisting of a pool of cDNA from poly-A selected mRNA from mouse tissues (37 tissues total were represented). The pools were designed to maximize the diversity of genes expressed between the pools, without diluting them beyond detection limits [7]. Scanned microarray images were quantitated with GenePix (Axon Instruments), complex noise structures (spatial trends, blobs, smudges) were removed from the images using our spatial detrending algorithm [10], and each set of 12 pool-specific images was calibrated using the VSN algorithm [26] (using a set of one hundred "housekeeping" genes represented on every slide). For each of the 48,966 probes, the 12 values were then normalized to have intensities ranging from 0 to 1.

3 Generative Model

Our model accounts for the expression data by identifying a large number of CoRegs, each of which spans a certain number of probes. Each probe within a CoReg may correspond to an exon that is part of the CoReg or an intron that is not part of the CoReg. The probes for the tiling data are indexed by i and the probes are ordered according to their positions in the genome. Denote the expression vector for probe i by x_i, which contains the levels of expression of probe i across M experimental conditions. In our data, there are $M = 12$ tissue pools. To account for alternative primary and secondary transcripts, we allow CoRegs to overlap. So, we assume that when the genome sequence data is scanned in order, if a probe corresponds to an exon, the exon belongs to one of a small number of CoRegs that are currently active. Exons that take part in multiple transcripts are identified in a post-processing stage. This model enables multiple concurrent CoRegs to account for alternative splicing isoforms.

For concreteness, in this paper we assume that at most two CoRegs may be concurrently active, but the extension to a larger number is straightforward. So, each CoReg can be placed into one of two categories (labeled $q = 1$ and $q = 2$) and for probe i, the discrete variable e_i indicates whether the probe corresponds to an intron ($e_i = 0$) or an exon from the CoReg from category 1 ($e_i = 1$) or 2 ($e_i = 2$). At position i, ℓ_i^q is the remaining length (in probes) of the CoReg in category q, including the current probe. The maximum length is $\ell_i^q = L$ and $\ell_i^q = 0$ indicates that probe i is in-between CoRegs in category q.

To model the relationships between the variables $\{\ell_i^q\}$ and $\{e_i\}$, we computed statistics using confirmed exons derived from four cDNA and EST databases: Refseq, Fantom II, Unigene, and Ensembl. The database sequences were mapped to Build 33 of the mouse chromosome using BLAT [18] and only unique mappings with greater than 95% coverage and greater than 90% identity were retained. Probes whose chromosomal location fell within the boundaries of a mapped exon were taken to be confirmed. We model the lengths of CoRegs using a geometric distribution, with parameter $\lambda = 0.05$, which was estimated using cDNA genes. Importantly, there is a significant computational advantage in using the memoryless geometric distribution. Using cDNA genes to select the length prior will

introduce a bias during inference. However, we found in our experiments that the effect of this bias is small. In particular, the results are robust to up to one order of magnitude in variation of λ.

The "control knob" that we use to vary the number of CoRegs that GenRate finds is κ, the *a priori* probability of starting a CoReg at an arbitrarily chosen position. Combining the above distributions, and recalling that $\ell_i^q = 0$ indicates position i is in-between CoRegs in category q, we have

$$P(\ell_i^q | \ell_{i-1}^q \in \{0,1\}) = \begin{cases} 1-\kappa & \text{if } \ell_i^q = 0 \\ \kappa(0.05e^{-0.05\ell_i^q}) & \text{if } \ell_i^1 > 0, \end{cases}$$
$$P(\ell_i^q | \ell_{i-1}^q \in \{2,\ldots,L\}) = [\ell_i^q = \ell_{i-1}^q - 1],$$

where square brackets indicate Iverson's notation, *i.e.*, $[True] = 1$ and $[False] = 0$. Both 0 and 1 are included in the condition "$\ell_{i-1} \in \{0,1\}$", because a new CoReg may start directly after the previous CoReg has finished. The term $\kappa(0.05e^{-0.05\ell_i^q})$ is the probability of starting a CoReg with length ℓ_i^1.

From genes in the cDNA databases, we found that within individual genes, probes are introns with probability $\epsilon = 0.3$. Depending on whether one or two CoRegs are active, we use a multinomial approximation to the probability that a probe is an exon:

$$P(e_i = 0 | \ell_i^1, \ell_i^2) = \epsilon^{[\ell_i^1 > 0] + [\ell_i^2 > 0]},$$
$$P(e_i = 1 | \ell_i^1, \ell_i^2) = [\ell_i^1 > 0](1-\epsilon)(\frac{1+\epsilon}{2})^{[\ell_i^2 > 0]},$$
$$P(e_i = 2 | \ell_i^1, \ell_i^2) = [\ell_i^2 > 0](1-\epsilon)(\frac{1+\epsilon}{2})^{[\ell_i^1 > 0]},$$

where again square brackets indicate Iverson's notation. Note that under this model, $P(e_i = 0 | \ell_i^1 > 0, \ell_i^2 > 0) = \epsilon^2$, $P(e_i = 1 | \ell_i^1 > 0, \ell_i^2 = 0) = 1 - \epsilon$ and $P(e_i = 1 | \ell_i^1 > 0, \ell_i^2 > 0) = (1-\epsilon^2)/2$.

The similarity between the expression profiles belonging to the same CoReg is accounted for by a prototype expression vector. Each CoReg has a unique, hidden index variable and the prototype expression vector for CoReg j is μ_j. We denote the index of the CoReg at probe i in category q by c_i^q.

Different probes may have different sensitivities (for a variety of reasons, including free energy of binding), so we assume that each expression profile belonging to a CoReg is similar to a scaled version of the prototype. Since probe sensitivity is not tissue-specific, we use the same scaling factor for all M tissues. Also, different probes will be offset by different amounts (*e.g.*, due to different average amounts of cross-hybridization), so we include a tissue-independent additive variable for each probe. The expression profile x_i for an exon (where $e_i > 0$) is equal to the corresponding prototype $\mu_{c_i^{e_i}}$, plus isotropic Gaussian noise, we have

$$P(x_i | e_i = q, c_i^1, c_i^2, a_i, \{\mu_j\}) = \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi a_{i3}^2}} e^{-(x_{im} - [a_{i1}\mu_{c_i^q,m} + a_{i2}])^2 / 2a_{i3}^2},$$

where a_{i1}, a_{i2} and a_{i3} are the scale, offset and isotropic noise variance for probe i, collectively referred to as a_i. In the *a priori* distribution $P(a_i)$ over these variables, the scale is assumed to be uniformly distributed in $[1/30, 30]$, which corresponds to a liberal assumption about the range of sensitivities of the probes. The offsets are assumed to be uniform in $[-0.5, 0.5]$ and the variance is assumed to be uniform in $[0, 1]$. While these assumptions are simplistic, we find they are sufficient for obtaining high-precision predictions, as described below.

We assume that the expression profiles for false exons are independent of the identify of the CoReg. While this assumption is also simplistic and should be further researched, it simplifies the model and leads to good results, so we make it for now. Thus, the false exon profiles are modeled using a background expression profile distribution:

$$P(x_i|e_i = 0, c_i^1, c_i^2, a_i, \{\mu_j\}) = P_0(x_i)$$

Since the background distribution doesn't depend on c_i^1, c_i^2, a_i or $\{\mu\}$, we also write it as $P(x_i|e_i = 0)$. We obtained this background model by training a mixture of 100 Gaussians on the entire, unordered set of expression profiles using a robust split-and-merge training procedure, and then including a component that is uniform over the range of expression profiles.

The Bayesian network in Fig. 2 shows the dependencies between the random variables in this generative model. Often, when drawing Bayesian networks, the parameters (prototypes) are not shown. We include the prototypes in the Bayesian network to highlight that they induce long-range dependencies in the model. For example, if a learning algorithm uses too many prototypes to model CoRegs (gene structures) in the first part of the chromosome, not enough will be left to model the remainder of the chromosome. So, during learning, prototypes must somehow be distributed in a fair fashion across the chromosome. We address this problem in the next section.

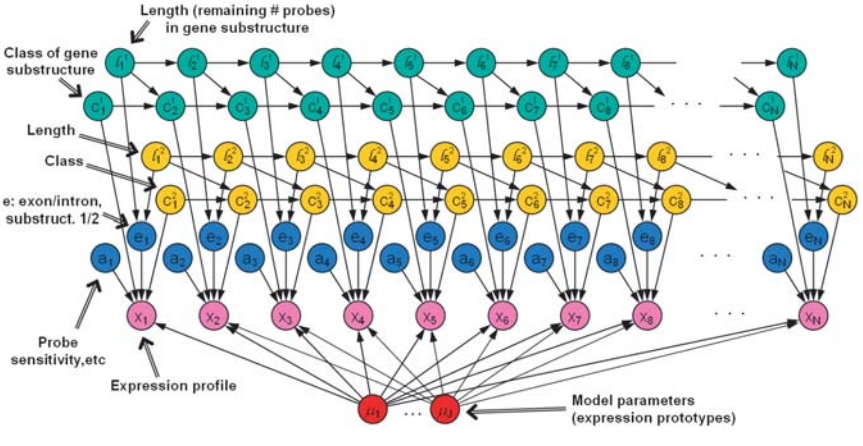


Fig. 2. A Bayesian network showing the variables and parameters in GenRate

Combining the structure of the Bayesian network with the conditional distributions described above, we can write the joint distribution as follows:

$$P(\{x_i\},\{a_i\},\{e_i\},\{c_i^1\},\{c_i^2\},\{\ell_i^1\},\{\ell_i^2\},\{\mu_j\}) = \Big(\prod_{i=1}^{N} P(x_i|e_i,c_i^1,c_i^2,a_i,\{\mu_j\})P(a_i)P(e_i|\ell_i^1,\ell_i^2) \Big(\prod_{q=1}^{2} P(c_i^q|c_{i-1}^q,\ell_{i-1}^q)P(\ell_i^q|\ell_{i-1}^q)\Big)\Big) \prod_{j=1}^{G} P(\mu_j),$$

where the appropriate initial conditions are obtained by setting $P(c_1^q|c_0^q,\ell_0^q) = [c_1^q = 1]$ and $P(\ell_1^q|\ell_0^q) \propto (1-\lambda)^{\ell_1^q}\lambda$, $\ell_q^1 = 1,\ldots,L$. The constant of proportionality normalizes the distribution (if ℓ were not bounded from above by L, the distribution not require normalization). Whenever a gene terminates, c_i^q is incremented in anticipation of modeling the next gene, so $P(c_i^q = n|c_{i-1}^q = n,\ell_{i-1}^q) = 1$ if $\ell_{i-1}^q > 1$ and $P(c_i^q = n+1|c_{i-1}^q = n,\ell_{i-1}^q) = 1$ if $\ell_{i-1}^q = 1$. We assume the prototypes are distributed according to the background model: $P(\mu_j) = P_0(\mu_j)$.

4 Inference and Learning

Exact inference of the variables *and* parameters in the above model is computationally intractable. Given the model parameters, the model has a chain-type structure, so a standard approach is to use the EM algorithm [27]. EM iterates between performing exact inference for the variables in the chain while holding the parameters constant, and then updating the parameters based on sufficient statistics computed during inference in the chain. However, the EM algorithm fails spectacularly on this problem, because it gets stuck in local minima where prototypes are used to model weakly-evidenced gene patterns in one part of the chromosome, at the cost of not modeling gene patterns elsewhere in the chromosome. In fact, the EM algorithm in long hidden Markov models is known to be extremely sensitive to initial conditions and tends to find poor local minima caused by suboptimal parsings of the long data sequence [33].

To circumvent the problem of poor local minima, we devised a computationally efficient scheme for finding good solutions in a discrete subspace of the parameter space, which can then be finely tuned using the EM algorithm. In our scheme, the prototypes are represented using examples from the data set (in a manner akin to using data points as cluster centers in "k-centers clustering"). In the original model, the prototype for x_i is derived from nearby expression patterns, corresponding to nearby exons in the genomic DNA. Thus, if x_i is part of a CoReg, there is likely another x nearby that is a good representative of the profile for the CoReg. In the new representation, we replace each pair of variables ℓ_i^q and c_i^q with a variable r_i^q that encodes the *relative location* of the prototype for x_i. r_i^q gives the distance, in indices, from x_i to the prototype x_j for the CoReg that x_i belongs to, *i.e.* $r_i^q = j - i$. For example, $r_i^q = -1$ indicates that the profile

immediately preceding x_i is the prototype for the gene to which x_i belongs. We limit the range of r_i^q to $-W, \ldots, W$, where W is a "window length". Within a CoReg, r_i^q decrements, indicating that the relative position of the prototype always decreases. We set aside a particular value of r_i^q, $r_0 = W + 1$, to account for the situation where i is in-between CoRegs.

Note that in this representation, the start of a CoReg corresponds to the condition $r_i^q = r_0$ and $r_{i+1}^q \neq r_0$, while the end of a CoReg corresponds to the condition $r_i^q \neq r_0$ and $r_{i+1}^q = r_0$. If a new CoReg starts directly after the previous CoReg, the boundary corresponds to the condition $r_{i+1}^q > r_i^q$. Since the r-variables are sufficient for describing CoReg boundaries, in fact, the ℓ variables need not be represented in the model. So, the new representation contains variables $\{r_i^q\}$, $\{e_i\}$, $\{a_i\}$, and $\{x_i\}$. If x_i is an exon in category q ($e_i = q$), the conditional distribution of x_i is

$$\prod_{m=1}^{M} \frac{1}{\sqrt{2\pi a_{i3}^2}} e^{-(x_{im} - [a_{i1} x_{i+r_i^q,m} + a_{i2}])^2 / 2a_{i3}^2},$$

except if x_i is the prototype for the gene ($r_i^q = 0$), in which case the distribution of x_i is $P_0(x_i)$. This model cannot be expressed as a Bayesian network, because constructing a Bayesian network from the above form of conditional distribution would create directed cycles. However, it can be expressed as a factor graph [28] or a directed factor graph [25].

The above model is a product of a Markov chain on $\{r_i^1\}$ and another Markov chain on $\{r_i^2\}$, coupled together by the switch e_i, which determines which chain is used to model the current expression vector, x_i. By combining the state spaces of the two Markov chains, exact inference can be performed using the forward-backward algorithm or the Viterbi algorithm. However, the combined state space has $4W^2$ states, where $2W$ is the maximum width of a CoReg, in probes. To enable our algorithm to find long CoRegs, we set $W = 100$, so the number of states in the combined chain would be $40,000$, making exact application of the Viterbi algorithm too slow. Instead, we apply the iterative sum-product algorithm to perform inference in the pair of coupled chains [28]. In each iteration, the algorithm performs a forward-backward pass in one chain, propagates probabilistic evidence across to the other chain, and then performs a forward-backward pass in the other chain.

We implemented the above inference algorithm in MATLAB, and for a given value of κ, our implementation takes approximately 10 minutes on a 2.4GHz PC to process the 48,966 probes and 12 tissue pools in chromosome 4 (with $W = 100$). The only free parameter in the model is κ, which sets the statistical significance of the genes found by GenRate.

5 Discussion of Computational Results

Fig. 3 shows a snapshot of the GenRate view screen that contains interesting examples of CoRegs found by GenRate. After we set the sensitivity control, κ,

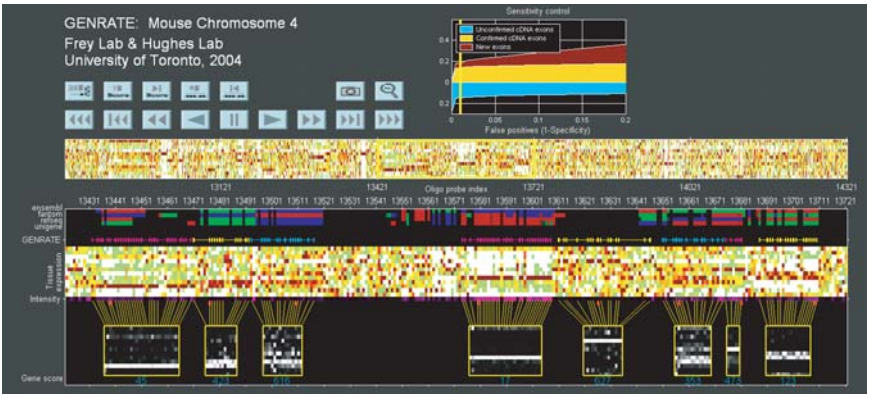


Fig. 3. The GenRate program (implemented in MATLAB) shows the genomic expression data and predicted CoRegs for a given false positive rate. The new genes, known genes and extensions of known genes that are found by GenRate are identified by shaded blocks, each of which indicates that the corresponding exon is included in the gene. Genes in cDNA databases (Ensembl, Fantom II, RefSeq, Unigene) are also shown. Each box at the bottom of the screen corresponds to a predicted gene structure and contains the normalized profiles for exons determined to be part of the gene. The corresponding raw profiles are connected to the box by lines. The score of each gene is printed below the corresponding box

to achieve a false positive rate of 1%, as described below, GenRate found 9,332 exons comprising 712 CoRegs. To determine how many of these predictions are new, we extracted confirmed genes derived from four cDNA and EST databases: Refseq, Fantom II, Unigene, and Ensembl. The database sequences were mapped to Build 33 of the mouse chromosome using BLAT and only unique mappings with greater than 95% coverage and greater than 90% identity were retained. Probes whose chromosomal location fell within the boundaries of a mapped exon were taken to be confirmed.

An important motivation for approaching this problem using a probability model is that the model should be capable of balancing probabilistic evidence provided by the expression data and the genomic exon arrangements. For example, there are several expression profiles that occur frequently in the data (in particular, profiles where activity in a single tissue pool dominates). If two of these profiles are found adjacent to each other in the data, should they be labeled as a gene? Obviously not, since this event occurs with high probability, *even if the putative exons are arranged in random order.*

To test the statistical significance of the results obtained by GenRate, we constructed a new version of the chromosome 4 data set, where the order of the columns (probes) is randomly permuted. For each value of κ in a range of values, we applied GenRate to the original data and the permuted data, and measured the number of positives and the number of false positives.

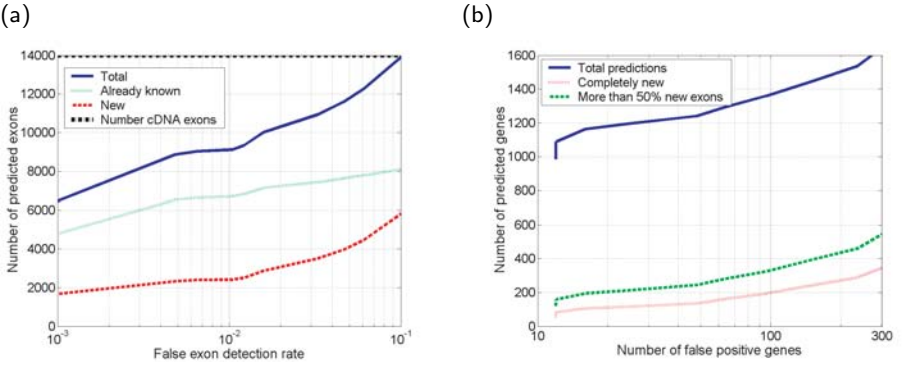


Fig. 4. (a) The number of exons in predicted CoRegs versus the exon false positive rate. The 3 curves correspond to the total number of predicted exons, the number of exons that are not in cDNA databases, and the number of known exons. The dash-dot line shows the number of exons in known genes from chromosome 4 (according to cDNA databases). (b) The number of predicted CoRegs versus the number of false positive predictions. The 3 curves correspond to the total number of predicted CoRegs, the number genes that are completely new (all exons within each structure are new), and the number of genes that contain at least 50% new exons

Fig. 4a shows the number of exons in CoRegs predicted by GenRate versus the false positive rate. Fig. 4b shows the number of predicted CoRegs versus the number of false positives. These curves demonstrate that GenRate is able to find CoRegs and associated exons with high precision. At an exon false positive rate of 1%, GenRate identifies 9,118 exons, 2,416 of which do not appear in known genes in cDNA databases, and GenRate identifies 65% of the exons in known genes in cDNA databases. This last number is a reasonable estimate of the proportion of genes that are expected to be expressed in the tissue pools represented in the data set. In Fig. 4b, when κ is set so GenRate finds 20 false positive CoRegs, GenRate identifies approximately 1,280 CoRegs, 209 of which contain at least 50% new exons, and 107 of which have no overlap with genes in cDNA databases.

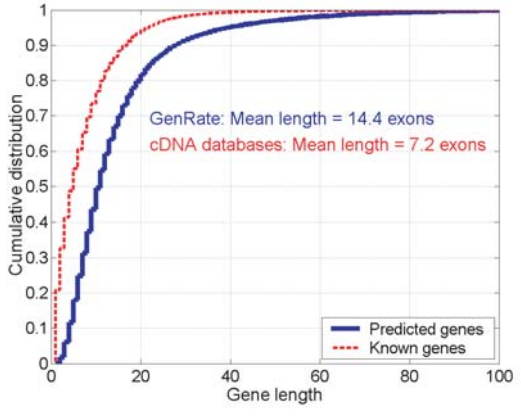
Interestingly, the genes found by GenRate tend to be longer than genes in cDNA databases, as shown in Fig. 5a. While some of this effect can be accounted for by the fact that GenRate tends to find longer transcripts because they have higher statistical significance than short transcripts (*e.g.*, those containing 1 or 2 exons), there are two other explanations that should be considered. First, neighboring genes that are co-regulated may be identified by GenRate as belonging to a single transcript. We found that 23% of pairs of neighboring genes in the RefSeq cDNA database that were both detected by GenRate were identified as a single CoReg by GenRate. However, it is possible that in many of these cases the neighboring pair of "genes" in the cDNA database are in fact a single gene and that GenRate is correctly merging the predictions together. This possibility is consistent with the latest revision of the human genome, which shows that

the number of genes is significantly lower than previously predicted [30], so that average gene length is longer than previously thought. In fact, as described below, we have shown using overlapping RT-PCR experiments that the longest, highest-scoring CoRegs identified by GenRate that consist of multiple cDNA "genes" exist as single, long transcripts.

5.1 Comparison to Hierarchical Clustering

A previously-described technique for assembling CoRegs from microarray tiling data consists of recursively merging pairs of probes into clusters, based on the correlation between the corresponding expression profiles and the distance between the probes in the genome [21]. In particular, if the correlation exceeds a threshold θ_1 *and* the genomic distance is less than another threshold θ_2, the probes are merged. In Fig. 5b, we compare the sensitivities of GenRate and this

(a)



(b)

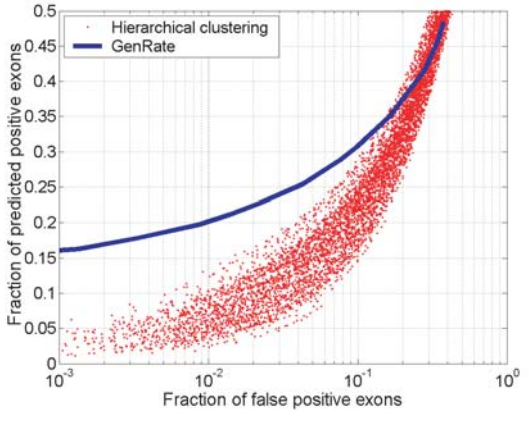


Fig. 5. (a) Cumulative distributions of gene lengths (in exons) for genes in cDNA databases and genes found by GenRate at an exon false positive rate of 1%. (b) A comparison between GenRate and correlation-based hierarchical clustering

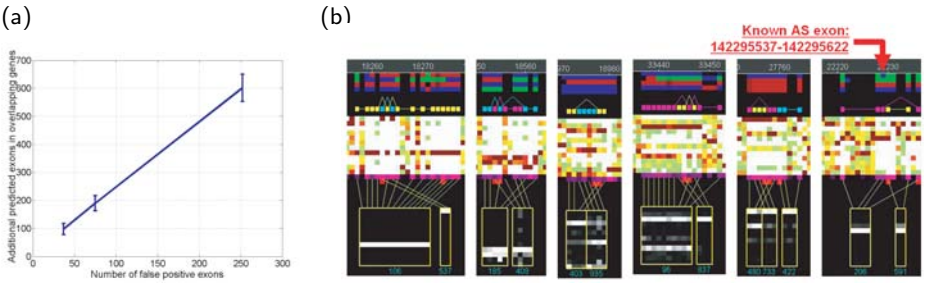


Fig. 6. (a) Additional exons predicted when overlapping CoRegs are taken into account. (b) Alternative splicing isoforms predicted by GenRate. The indicated structure corresponds to a known splicing event

recursive clustering procedure for a large number of different values of θ_1 and θ_2. For low false positive rates, GenRate detects at least five times more exons.

5.2 Detection of Splicing Isoforms

After inference, GenRate can list high-scoring examples of overlapping CoRegs, which may correspond to alternative primary and secondary transcripts. Fig. 6a shows the number of additional exons predicted in overlapping CoRegs versus the number of false positives. Fig. 6b shows six randomly-selected cases of overlapping CoRegs. We are still in the process of investigating the transcripts corresponding to these cases using RT-PCR. However, one of the predictions from GenRate (the last one shown in Fig. 6b) corresponds to a known alternative splicing isoform.

5.3 Non-monotonic Reasoning

Because GenRate combines different sources of probabilistic information in a global scoring (probability) function, for different settings of the sensitivity κ, GenRate produces different interpretations of the genome-wide structure of CoRegs. For example, two putative exons that are part of the same CoReg at one setting of κ may be re-assigned to different CoRegs at a different setting of κ. This type of inference, whereby decisions are changed as more evidence is considered, is called non-monotonic. (In contrast, simpler techniques, such as hierarchical clustering, produce monotonic inferences.) An important consequence of this is that for a given sensitivity, a low false positive rate can be achieved by re-running GenRate. Fig. 7 shows that by re-running GenRate, a much lower false positive rate is achieved at the same true positive rate. The red (dashed) curve was obtained by running GenRate with a high value of κ, scoring the CoRegs according to their local log-probabilities, and then applying a threshold to the scores to produce a predicted set of CoRegs. This was repeated with the randomly permuted data to obtain the plot of detected CoRegs versus false positives. The blue (solid) curve was produced by running GenRate with different

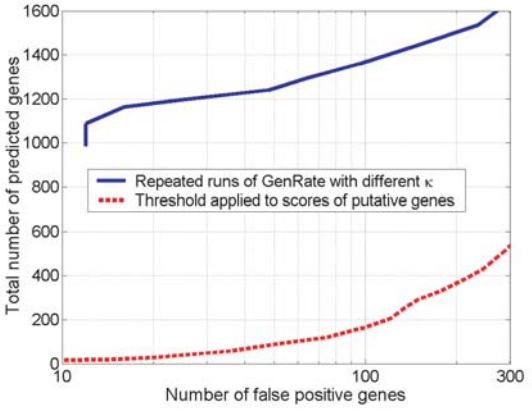


Fig. 7. When GenRate is re-run with a different sensitivity, κ, it may re-assign putative exons to other CoRegs. Compared to running GenRate once and using CoReg scores to rank CoRegs (red, dashed curve), running GenRate multiple times (blue, solid curve) leads to a significantly lower false positive rate at a given sensitivity

values of κ and retaining all predicted CoRegs (not applying a score threshold). By re-running GenRate, a much lower false positive rate is achieved for the same detection rate.

6 RT-PCR Experiments

Using RT-PCR, we have verified nine of the novel, high-scoring transcripts predicted by GenRate. In three cases, we selected predicted CoRegs that had high

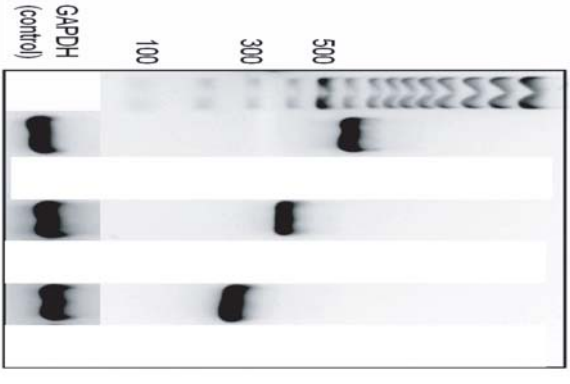


Fig. 8. RT-PCR results for three new transcripts identified by GenRate. The horizontal axis corresponds to the weight of the RT-PCR product and the darkness of each band corresponds to the amount of product with that weight

scores and no overlap with genes in the four cDNA databases. Fig. 8 shows the RT-PCR results for these predictions. The two PCR primers for each predicted transcript are from different exons separated by thousands of bases in the genome. For each predicted transcript, we selected a tissue pool with high microarray expression. We included the ubiquitously-expressed gene GAPDH to ensure proper RT-PCR amplification. The RT-PCR results confirm the predicted transcripts. Results on the other novel transcripts will be reported in another article [1].

7 Summary

GenRate is the first generative model that combines a model of genomic arrangement of putative exons with a model of expression patterns, for the purpose of discovering CoRegs in genome-wide tiling data. By balancing different sources of uncertainty, GenRate is able to achieve a significantly lower false positive rate than correlation-based hierarchical clustering methods. Applied to our microarray data, GenRate identifies many novel CoRegs with a low false-positive rate. We confirmed three of the predicted transcripts using RT-PCR experiments, and were able to recover known alternative splicing events and predict some new ones, albeit with high false positive rate. We have recently completed a genome-wide analysis of novel transcripts and this work has led us to a surprising conclusion, reported in [1], which appears to contradict recent results obtained by other researchers using microarrays to detect novel transcripts.

Because GenRate is based on a principled probability model, additional hidden variables can be incorporated in a straight-forward fashion. We believe GenRate will be a useful tool for analyzing other types of genome-wide tiling data, such as whole-genome tiling arrays.

Acknowledgments

This work was supported by a Premier's Research Excellence Award to Frey, a grant from the Canadian Institute for Health Research (CIHR) awarded to Hughes and Frey, and a CIHR NET grant awarded to Frey, Hughes and M. Escobar.

References

1. Frey B. J. *et al.* Genome-wide analysis of mouse transcription using exon-resolution microarrays and factor graphs. Under review.
2. Storz G. An expanding universe of noncoding RNAs. *Science* **296**, 1260-1263, 2002.
3. Mironov, AA *et al.* Frequent alternative splicing of human genes. *Genome Research* **9**, 1999.
4. Maniatis, T *et al.* Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 2002.
5. Burge C *et al.* Splicing of precursors to mRNAs by the spliceosomes. 2 edn, New York, Cold Spring Harbor Laboratory Press.

6. Peng WT *et al.* A panoramic view of yeast noncoding RNA processing. *Cell* **113:7**, 919-933, June 2003.
7. Zhang W *et al.* The functional landscape of mouse gene expression. *J. Biol.* **3:21**, Epub Dec 2004.
8. Pan Q. *et al.* Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16**, 929-941, December 2004.
9. Pan Q. *et al.* Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in Genet.* **21:2**, 73-77, Feb 2005.
10. Shai O *et al.* Spatial bias removal in microarray images. Univ. Toronto TR PSI-2003-21. 2003.
11. Hild M *et al.* An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. *Genome Biol.* 2003;5(1):R3. Epub 2003 Dec 22.
12. Hughes TR *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 2001 Apr;19(4):342-7.
13. Kapranov P *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002 May 3;296(5569):916-9.
14. Schadt EE *et al.* A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology* 5 (2004).
15. Stolc V *et al.* A gene expression map for the euchromatic genome of drosophila melanogaster. To appear in *Science* (2004).
16. Nuwaysir EF *et al.* Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Gen. Res.* 2002 Nov;12(11):1749-55.
17. FANTOM Consortium: RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Okazaki Y *et al. Nature.* 2002 Dec 5;420(6915):563-73.
18. Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.
19. Pen SG *et al.* Mining the human genome using microarrays of open reading frames. *Nat. Genet.* 2000 Nov;26(3):315-8.
20. Rinn JL *et al.* The transcriptional activity of human chromosome 22. *Genes Dev.* 2003 Feb 15;17(4):529-40.
21. Shoemaker DD *et al.* Experimental annotation of the human genome using microarray technology. *Nature* 2001 Feb 15;409(6822):922-7.
22. Friedman N *et al.* Using Bayesian networks to analyze expression data. *Jour Comp Biology* **7**, 2000.
23. Yamada K *et al.* Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome. *Science* **302**, 2003.
24. Segal E *et al.* Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19**, 2003.
25. Frey BJ *et al.* GenRate: A generative model that finds and scores new genes and exons in genomic microarray data. *Proc. PSB*, Jan. 2005.
26. Huber W *et al.* Variance stabilization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics* **18**, 2002.
27. Dempster AP *et al.* Maximum likelihood from incomplete data via the EM algorithm. *Proc. Royal Stat. Soc.* **B-39**, 1977.
28. Kschischang FR *et al.* Factor graphs and the sum-product algorithm. *IEEE Trans. Infor. Theory* **47**, 2001.
29. Frey BJ. Extending factor graphs so as to unify directed and undirected graphical models. *Proc. UAI*, Aug. 2003.

30. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 2004.
31. Berman P *et al.* Fast optimal genome tiling with applications to microarray design and homology search. *JCB* **11**, 766-785, 2004.
32. Bertone P *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **24**, Dec, 2004.
33. Ostendorf M. *IEEE Trans. Speech & Audio Proc.* **4**:360, 1996.