

GENRATE: A GENERATIVE MODEL THAT FINDS AND SCORES NEW GENES AND EXONS IN GENOMIC MICROARRAY DATA ^a

BRENDAN J. FREY^{1,2}, QUAID D. MORRIS^{1,2}, WEN ZHANG²,
NAVEED MOHAMMAD², TIMOTHY R. HUGHES²

¹ *Dept. of Electrical and Computer Engineering*

² *Banting & Best Dept. of Medical Research*

University of Toronto, Toronto, ON, M5S 3G4, Canada

E-mail: frey@psi.toronto.edu

Recently, researchers have made some progress in using microarrays to validate predicted exons in genome sequence and find new gene structures. However, current methods rely on separately making threshold-based decisions on intensity of expression, similarity of expression profiles, and arrangements of exons in the genome. We have taken a Bayesian approach and developed GenRate, a generative model that accounts for both genome-wide expression data taken from multiple conditions (*e.g.* tissues) and co-location and density of probes in DNA sequence data. GenRate balances probabilistic evidence derived from different sources and outputs scores (log-likelihoods) for each gene model, enabling the estimation of false-positive and false-negative rates. The model has a number of local minima that is exponential in the length of the DNA sequence data, so direct application of the EM learning algorithm produces poor results. We describe a novel way of parameterizing the model using examples from the data set, so that good solutions are found using an efficient algorithm. We apply GenRate to a subset of mouse genome-wide expression data that we have created, and discuss the statistical significance of the genes found by GenRate. Three of the highest-ranking gene structures found by GenRate, each containing thousands of bases from the genome, are confirmed using RT-PCR experiments.

1 Introduction

The use of DNA microarrays for the discovery of expressed elements in genomes is increasing with improvements in density, flexibility, and accessibility of the technology. Two general strategies have emerged. In the first, candidate elements (*e.g.* ORFs, genes, exons, RNAs) are identified computationally, and each is represented one or a few times on the array^{10,12,4}. In the second, the entire genome sequence is "tiled"; for example, overlapping oligonucleotides encompassing both strands are printed on arrays, such that all possible expressed sequences are represented^{12,6,11,13}. Both approaches, as well as independent analyses by other methods^{8,4} have indicated that a substantially higher proportion of genomes are expressed than are currently annotated, underscoring the shortcomings of current sequence-based gene prediction algorithms and emphasizing the need for empirical analysis.

Microarrays do not inherently provide information regarding the length of the RNA or DNA molecules detected, nor do they inherently reveal whether

^aThis work was supported by a Premier's Research Excellence Award to Frey and a grant from the Canadian Institute for Health Research awarded to Hughes and Frey.

features designed to detect adjacent features on the chromosome are in fact detecting the same transcript. Co-expression (*i.e.* co-detection) of adjacent features can be taken as evidence supporting that the corresponding probes are indeed detecting the same molecular species. However, mRNAs, which account for the largest proportion of transcribed sequence in a genome, present a particular challenge in this paradigm. 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. 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.

We describe a generative model that jointly accounts for the stochastic nature of the arrangement of exons in genomic DNA and the noise properties in microarray data. The generative model, called GenRate, uses expression data taken from multiple conditions, accounts for co-location statistics of probes in DNA sequence data, and finds and scores gene structures. While the version of GenRate described here does not model expression variability introduced by alternative splicing, overlapping genes, and alternative transcription sites, in a future paper we will describe an extension, which does account for these effects.

2 Microarray data

The microarray data are a subset of a full-genome data set to be described elsewhere¹. Briefly, exons were predicted from Repeat-masked mouse draft genome sequence (Build 28) using five different exon-prediction programs. (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 63,041 non-overlapping exons were contained on chromosome 4. 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 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 limits². 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 algorithm³, and each set of 12 pool-specific images was calibrated using the VSN algorithm¹⁵ (using a

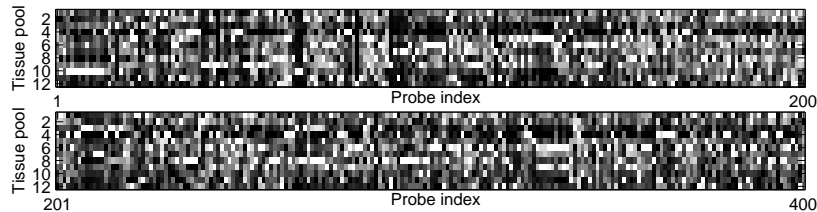


Figure 1: A small fraction of our data set, which consists of an expression measurement for each of 12 mouse tissue pools and 63,041 60-mer probes for putative exons arranged according to the order in Build 28 of the genome.

set of one hundred "housekeeping" genes represented on every slide). For each of the 63,041 probes, the 12 values were then normalized to have intensities ranging from 0 to 1.

Fig. 1 shows a portion of the data from chromosome 4. Because the probes are arranged according to their order in the genome, a consecutive sequence of similar expression profiles (columns) provides evidence of co-regulation of the corresponding putative exons, and thus provides evidence of a gene structure. For example, probes 4 to 14 have similar expression profiles (with high expression in the 10th tissue pool), which provides evidence of a gene. However, such visually obvious examples are relatively rare. More common examples that GenRate finds include complex gap patterns and noisy, albeit statistically significant patterns that are hard to identify visually.

3 Previous Work

Heuristics that group nearby putative exons using intensity of expression or co-regulation across experimental conditions can be used to approach this problem^{12,6,13}. In⁶, a dense activity map of RNA transcription is used to verify putative exons. A disadvantage of this approach is that it cannot detect weakly-expressed exons that have a large biological impact, due to high translational efficiency. In addition to detecting high levels of transcriptional activity, our approach finds correspondences in patterns of activation across multiple tissue pools, so even weakly-expressed exons that have tissue-dependent activity can be detected.

In¹², correlations between the expression patterns of nearby probes are used to merge probes into putative gene structures. 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. Our approach differs from this approach in two ways. First, our algorithm doesn't make a sequence of threshold-based decisions, but instead uses distributions over gene lengths, gap lengths and probe similarity to jointly make decisions and compute maximum a

posteriori gene structures. So, for example, extraordinarily similar expression profiles may be merged even if there is a large gap between the corresponding exon. Second, since GenRate uses a principled generative probability model, decisions on gene structures are based on an automatic comparison of the likelihood of seeing the gene structure under the gene model and the background expression profile model. So, if the profiles for two probes are quite unusual compared to typical profiles in the data (*i.e.*, they have unusual patterns of tissue-specificity), the two probes may be merged into a gene structure even if their expression profiles are only weakly similar.

In the above previous work, it is not clear how to properly balance evidence provided by similarity of expression profiles with that provided by sequence features (*e.g.* gene length, intron lengths, intra-gene gap length). Recently researchers have successfully shown that complex probability models can be used effectively to combine different sources of information in genomics data (*c.f.* ¹⁴). In a probability model, combining different sources of information is realized by a computationally efficient application of Bayes rule to combine sources of information in a probabilistic manner.

4 Generative Probability Model

GenRate can be applied to any genome-based expression data set, since it works on the assumption that the expression data is arranged in order on the genome. In our model, the probes are indexed by i and the probes are ordered according to their locations in the genome. Denote the expression vector for probe i by \mathbf{x}_i , which contains the levels of expression of probe i across K experimental conditions. In our data, there are $K = 12$ tissue pools. Since probe i is selected from putative exon sequence data, it may in fact correspond to a false exon either between genes or within a gene. e_i is a binary variable, where $e_i = 1$ indicates a true exon and $e_i = 0$ indicates a false exon. If probe i is within a gene, the remaining length of the gene (in probes) including probe i is ℓ_i . $\ell_i = 0$ indicates that probe i does not belong to the gene.

To model the relationships between the variables $\{\ell_i\}$ 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 28 of the mouse chromosome using BLAT ⁹ 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. The genes in these databases are obviously subject to selection bias, so statistics based on these genes will be biased, an effect we ignore for now.

We model the lengths of genes using a geometric distribution, with parameter $\lambda = 0.05$, which was estimated using cDNA genes. This approximation is accurate for lengths greater than 5. For shorter lengths, the accuracy of

the prior is not critical, because the prior probability of starting a gene (see the next paragraph) dominates. Importantly, there is a significant computational advantage in using the memory-less geometric distribution. Using cDNA genes to select the length prior will introduce a bias, so other priors should be investigated, but in this paper we report results using the geometric prior.

The “control knob” that we use to vary the number of genes that GenRate finds is κ , the *a priori* probability of starting a gene at an arbitrarily chosen position. Combining the above distributions, and recalling that $\ell_i = 0$ indicates an inter-gene region, we have

$$\begin{aligned} P(\ell_i = 0 | \ell_{i-1} = 0 \text{ or } 1) &= 1 - \kappa \\ P(\ell_i | \ell_{i-1} = 0 \text{ or } 1) &= \kappa \cdot 0.05 \exp(-0.05\ell_i), \quad \text{if } \ell_i > 0 \\ P(\ell_i = n - 1 | \ell_{i-1} = n) &= 1, \quad \text{if } \ell_{i-1} > 1. \end{aligned} \tag{1}$$

The expression “ $\ell_{i-1} = 0$ or 1” occurs because a new gene may start immediately after the previous gene has finished.

From the data on confirmed genes, we found that within genes, each probe has a probability of 0.7 corresponding to a correct predicted exon. We assume that within genes, putative exons are true exons independently and with probability

$$P(e_i = 1 | \ell_i > 0) = \epsilon, \tag{2}$$

where from the above data we estimated $\epsilon = 0.7$. Although we have not verified this assumption directly using the data, we find that the results obtained using this assumption give high sensitivity with high specificity (see below). Between genes, all putative exons are false, so $P(e_i = 1 | \ell_i = 0) = 0$.

The similarity between the expression profiles belonging to the same gene is accounted for by a gene-specific prototype expression vector. While this model does not properly take into account effects introduced by alternative splicing, overlapping genes and alternative transcription sites, as shown in the experimental section, this model is sufficient for finding a large number of new exons and genes. In the gene model, each gene has a unique, hidden index variable and the prototype expression vector for gene j is $\boldsymbol{\mu}_j$. We denote the index of the gene at probe i by c_i . 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 gene is similar to a scaled version of the prototype. Since probe sensitivity is not tissue-specific, we use the same scaling factor for all K 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. Assuming the expression profile \mathbf{x}_i for a true exon ($e_i = 1$) is equal to

the corresponding prototype $\boldsymbol{\mu}_{c_i}$, plus isotropic Gaussian noise, we have

$$P(\mathbf{x}_i|e = 1, c_i, \{\boldsymbol{\mu}_j\}, a_i) = \prod_{k=1}^K \frac{1}{\sqrt{2\pi a_{i3}^2}} \exp(-(x_{ik} - [a_{i1}\mu_{c_i k} + a_{i2}])^2 / 2a_{i3}^2), \quad (3)$$

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 *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 $[0, 1]$. These assumptions are naive and require further research, but we find they are sufficient for obtaining good results.

False exons are modelled using a background expression profile distribution,

$$P(\mathbf{x}_i|e_i = 0, c_i, a_i, \{\boldsymbol{\mu}_j\}) = P_0(\mathbf{x}_i) \quad (4)$$

Since the background distribution doesn't depend on c_i , a_i or $\{\boldsymbol{\mu}_j\}$, we also write it as $P(\mathbf{x}_i|e = 0)$. We obtained values for these probability densities by training a mixture of 100 Gaussians on the entire, unordered set of expression profiles, and including a component that is uniform over the range of expression profiles.

The structural relationships between the variables described above are indicated by the Bayesian network in Fig. 2. Often, when drawing Bayesian networks, the prototypes are considered as parameters and not shown. We include the prototypes in the Bayesian network to show that they induce long-range dependencies in the GenRate model. For example, if all of the prototypes are used to model gene structures in the first part of the chromosome, none 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.

Combining the structure of the Bayesian network with the conditional distributions described above, we have a joint distribution,

$$P(\{\mathbf{x}_i\}, \{a_i\}, \{e_i\}, \{c_i\}, \{\ell_i\}, \{\boldsymbol{\mu}_j\}) = \prod_{i=1}^N P(\mathbf{x}_i|e_i, c_i, a_i, \{\boldsymbol{\mu}_j\})P(a_i)P(e_i|\ell_i)P(c_i|c_{i-1}, \ell_{i-1})P(\ell_i|\ell_{i-1}) \prod_{j=1}^G P(\boldsymbol{\mu}_j), \quad (5)$$

where in this expression $P(c_1|c_0, \ell_0)$ and $P(\ell_1|\ell_0)$ are equal to $P(\ell_1)$ and $P(c_1)$. Most of the components in the above model are described above. As for the gene indices, c_i , we assume that c_i is ordered, starting at 1: $P(c_i = 1) = 1$.

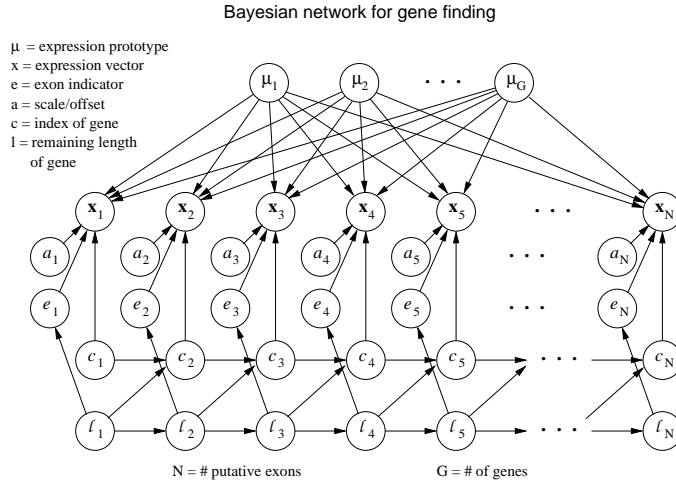


Figure 2: A Bayesian network showing the variables and parameters in GenRate.

Whenever a gene terminates, c_i is incremented in anticipation of modelling the next gene, so $P(c_i = n | c_{i-1} = n, l_{i-1}) = 1$ if $l_{i-1} > 1$ and $P(c_i = n + 1 | c_{i-1} = n, l_{i-1}) = 1$ if $l_{i-1} = 1$. We assume the prototypes are distributed according to the background model: $P(\mu_j) = P_0(\mu_j)$.

5 Inference and Learning

Exact computation of the marginal probabilities or the MAP configuration in the above model is computationally intractable. A standard way of coping with this intractability in this form of model is to use the EM algorithm¹⁶. The EM algorithm fails spectacularly on this problem. This is not surprising, since the EM algorithm in long hidden Markov models is known to find poor local minima caused by suboptimal parsings of the long data sequence¹⁹. In the GenRate model, the EM algorithm 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 modelling strongly-evidenced gene patterns in another part of the chromosome.

To circumvent the problem of very poor local minima, we devised a novel scheme where the parameters (μ 's) are represented using examples from the data set. An additional advantage of this approach is that since learning consists of identifying nearby profiles as prototypes, learning can be performed in a single forward-backward pass (Viterbi pass).

The scheme is based on the fact that the model for each x_i is derived from nearby expression patterns, corresponding to nearby exons in the genomic

DNA. So, if \mathbf{x}_i is an exon, there ought to be another \mathbf{x} nearby that is a good representative of the profile for the gene. To accomplish this we replace each variables c_i with a variable r_i , which indicates the distance, in indices, from \mathbf{x}_i to the prototype \mathbf{x}_j for the gene that \mathbf{x}_i is in, *i.e.* $r_i = j - i$. For example, $r_i = -1$ indicates that the profile immediately preceding \mathbf{x}_i is the prototype for the gene to which \mathbf{x}_i belongs. The new conditional distribution for \mathbf{x}_i is

$$P_e(\mathbf{x}_i | e_i = 1, r_i, \mathbf{x}_{i+r_i}, a_i) = \begin{cases} \prod_{k=1}^K \frac{1}{\sqrt{2\pi a_{i3}^2}} \exp(-(x_{ik} - [a_{i1}x_{i+r_i,k} + a_{i2}])^2 / 2a_{i3}^2), & \text{if } r_i \neq 0 \\ P(\mathbf{x}_i | e_i = 1, r_i, \mathbf{x}_{i+r_i}, a_i) = P(\mathbf{x}_i | e = 0), & \text{if } r_i = 0. \end{cases} \quad (6)$$

Here, r_i acts as a switch to select the parent of \mathbf{x}_i . To ensure the r_i 's take on appropriate values, the conditional distribution for r_i is given by $P(r_i = n - 1 | r_{i-1} = n, \ell_{i-1}, \ell_i) = 1$ if $\ell_{i-1} > 1$ and $P(r_i | r_{i-1}, \ell_{i-1}, \ell_i) = \text{Unif}(0, \dots, \ell_i)$ if $\ell_{i-1} = 1$. This ensures that when a new gene starts, r_i will be drawn randomly from within the length of the gene and that r_i will decrement throughout the duration of the new gene. This model can be described using a factor graph ¹⁸.

The above model is a product of terms that has a Markov chain structure with tractable state complexity, so the forward-backward algorithm or Viterbi algorithm can be used for exact inference. Some readers may be concerned about the presence of the continuous variables a_i . However, these variables do not have parents, so they can be integrated or maximized for each configuration of the (discrete) variables in their Markov blankets (r_i and e_i). We use the Viterbi algorithm to find the MAP configuration of the model. Our MATLAB implementation of GenRate takes approximately 3 minutes to process the 63,041 probes and 12 tissue pools in chromosome 4, with a restriction on the gene length of $W = 200$ probes.

The only free parameter in the model is κ , which sets the statistical significance of the genes found by GenRate. The score of each gene found by GenRate can be computed by taking the log-ratio of the probability under GenRate of the MAP path through the gene and the probability of the path corresponding to non-gene probes (intra-gene probes).

6 Experimental Results

Fig. 3 shows a snapshot of the GenRate view screen that contains interesting examples. After we set the probability of starting an exon at an arbitrary position (κ) to achieve a false positive rate of 1%, as described below, GenRate found 16,082 exons in chromosome 4, comprising 1,477 genes.

To determine how many of these predictions are new, we extracted confirmed genes derived from four cDNA and EST databases: Refseq, Fantom II,

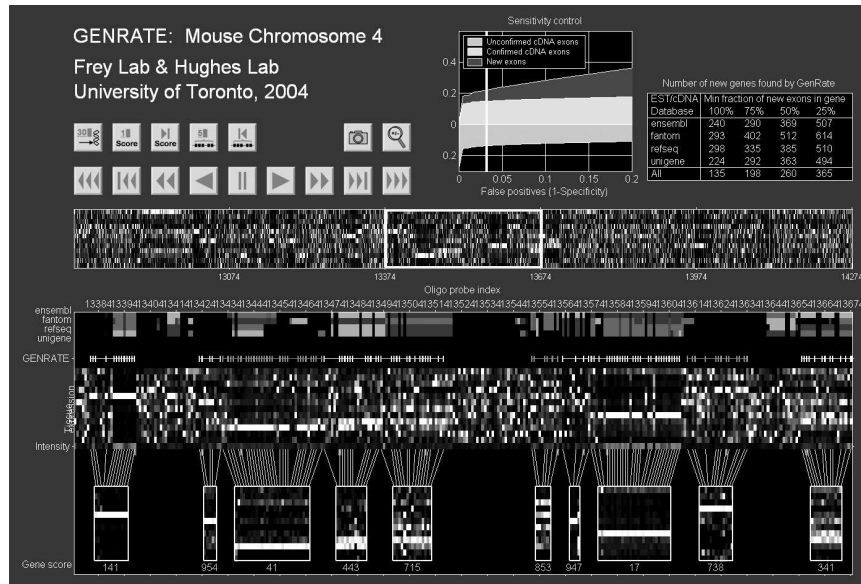


Figure 3: The GenRate MATLAB program shows the genomic expression data and predicted gene structures for a given false positive rate. Genes found by GenRate and genes in cDNA databases (Ensembl, Fantom II, RefSeq, Unigene), are identified by shaded blocks, each of which indicates that the corresponding exon is included in the gene. Each box at the bottom of the screen corresponds to a predicted gene 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 rank of each gene is printed below the corresponding box.

Unigene, and Ensembl. The database sequences were mapped to Build 28 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.

The following table shows the number of new genes found by GenRate, relative to the 4 databases and the combined set of databases. Different levels of strictness are used to label each gene as new, ranging from 100% to 25% of the predicted exons as “unconfirmed” by these four cDNA databases. Note that 783 of the 1,477 genes found by GenRate have at least 50% exon overlap with the confirmed genes in all cDNA databases, but that 427 of the genes found by GenRate have no exon overlap with the confirmed genes in all cDNA databases. In the set of genes that are completely new, the minimum, median and maximum gene lengths (in number of exons) are 2, 15 and 66.

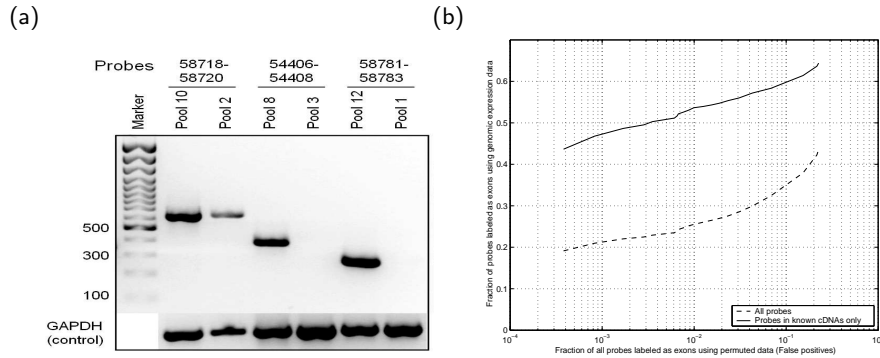


Figure 4: (a) RT-PCR results for three new genes identified by GenRate. The vertical axis corresponds to the weight of the RT-PCR product and the darkness of each band corresponds to the amount of product with that weight. (b) Fraction of predicted positives and positives confirmed by cDNA databases, versus fraction of positives predicted using randomly permuted data (false positives).

EST/cDNA Database	Number of new genes predicted by GenRate				
	Minimum fraction of exons in each gene that are new	100%	75%	50%	25%
Ensembl		1030	1157	1224	1284
Fantom II		606	776	902	1032
RefSeq		793	853	929	1032
Unigene		620	724	813	944
All		427	557	656	783

We are currently performing an extensive set of RT-PCR and Northern blotting experiments to verify the tissue-specific expression and exon structure of novel genes discovered by GenRate. Results on the first three genes tested (selected to have high scores and to overlap with no genes in the four cDNA databases) are shown in Fig. 4a. The two PCR primers for each predicted gene are from different exons separated by thousands of bases in the genome. For each predicted gene, we selected 1 tissue pool with high microarray expression, and 1 tissue pool with low expression. We included the ubiquitously-expressed gene GAPDH to ensure proper RT-PCR amplification. The RT-PCR results confirm the predicted genes and their tissue-specific expression.

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 columns (putative exons) are placed in random order. We then applied GenRate to the permuted data and compared the results to the results obtained on the original data, for varying levels of κ . Fig. 4b summarizes the results. The x -axis shows the fraction of the 63,041 probes that are labeled as exons *in the permuted data*. These can be viewed as false positives, since without knowing the order of the putative exons in the genome, we do not expect to be able to find gene structures with any statistical significance.

The plot shows two curves. The dashed curve is the fraction of all 63,041 probes that are labeled by GenRate as exons in the original (unpermuted) data. The solid curve is the fraction of exons from known genes (see above) that are labeled by GenRate as exons in the original (unpermuted) data. These curves demonstrate that GenRate is able to find predicted gene structures and gene structures compatible with cDNA databases with high statistical significance. For example, at a false positive rate of 1%, 25% of the probes in the original data are labeled as exons, and 54% of the probes in known cDNAs are correctly labeled as exons. This is a reasonable estimate of the proportion of genes that are expected to be expressed in the tissue pools represented in the data set.

7 Summary and future directions

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 exon and gene discovery. Applied to our microarray data set, GenRate identifies many new genes with a very low false-positive rate. Using RT-PCR to verify 3 new genes predicted by GenRate, we found that all 3 predicted exon sequences (each containing thousands of bases from the genome), are indeed transcriptionally active. We are developing an extension to GenRate, which accounts for alternative splicing, overlapping genes and alternative transcription sites, and applying it to our genome-wide data set of over 12,000,000 measurements. GenRate can be applied to any sequence-based data set, such as whole-genome tiling data.

Bibliography

1. Frey BJ *et. al.* Full-genome exon profiling in *mus musculus*. In preparation.

2. Zhang W *et. al.* The functional landscape of mouse gene expression. Under review.
3. Shai O *et. al.* Spatial bias removal in microarray images. Univ. Toronto TR PSI-2003-21. 2003.
4. 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.
5. Hughes TR *et. al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 2001 Apr;19(4):342-7.
6. Kapranov P *et. al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002 May 3;296(5569):916-9.
7. Nuwaysir EF *et. al.* Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Gen. Res.* 2002 Nov;12(11):1749-55.
8. 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.
9. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.
10. Pen SG *et. al.* Mining the human genome using microarrays of open reading frames. *Nat. Genet.* 2000 Nov;26(3):315-8.
11. Rinn JL *et. al.* The transcriptional activity of human chromosome 22. *Genes Dev.* 2003 Feb 15;17(4):529-40.
12. Shoemaker DD *et. al.* Experimental annotation of the human genome using microarray technology. *Nature* 2001 Feb 15;409(6822):922-7.
13. Yamada K *et. al.* Empirical Analysis of Transcriptional Activity in the *Arabidopsis* Genome. *Science* 2003 October 31;302:842-6.
14. Segal E *et. al.* Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 19:273-82, 2003.
15. Huber W *et. al.* Variance stabilization applied to microarray data calibration and to quantification of differential expression. *Bioinformatics* 18:S96-S104, 2002.
16. Dempster AP *et. al.* Maximum likelihood from incomplete data via the EM algorithm. *Proc. Royal Stat. Soc.* B-39:1–38, 1977.
17. Kschischang FR *et. al.* Factor graphs and the sum-product algorithm. *IEEE Trans. Infor. Theory* 47(2):498–519, February 2001.
18. Frey BJ. Extending factor graphs so as to unify directed and undirected graphical models. *Proc. UAI* 2003 August.
19. Ostendorf M. *IEEE Trans. Speech & Audio Proc.* 4:360, 1996.