BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S337–S344

Identifying operons and untranslated regions of transcripts using Escherichia coli *RNA expression analysis*

Brian Tjaden[1,], David R. Haynor[2], Sergey Stolyar[3], Carsten Rosenow[4] and Eugene Kolker[3]*

[1]Department of Computer Science and Engineering, University of Washington, Box 352350, Seattle, WA 98195, USA, [2]Department of Radiology, University of Washington, Box 356004, Seattle, WA 98195, USA, [3]Institute for Systems Biology, 1441 North 34th St., Seattle, WA 98103, USA and [4]Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA

Received on January 22, 2002; revised and accepted on March 29, 2002

ABSTRACT

Microarrays traditionally have been used to assay the transcript expression of coding regions of genes. Here, we use *Escherichia coli* oligonucleotide microarrays to assay transcript expression of both open reading frames (ORFs) and intergenic regions. We then use hidden Markov models to analyse this expression data and estimate transcription boundaries of genes. This approach allows us to identify 5′ untranslated regions (5′ UTRs) of transcripts as well as genes that are likely to be operon members. The operon elements we identify correspond to documented operons with 99% specificity and 63% sensitivity. Similarly we find that our 5′ UTR results accurately coincide with experimentally verified promoter regions for most genes.

Contact: tjaden@cs.washington.edu

Keywords: untranslated regions; operons; intergenic; microarrays; *Escherichia coli*.

INTRODUCTION

Experimental identification of transcripts using Northern blot, reverse transcription-polymerase chain reaction (RT-PCR) and primer extension analysis is relatively costly and time consuming. Therefore, only a modest fraction of putative transcription boundaries have been documented for *Escherichia coli* genes. As a result, several computational approaches have been developed for predicting genome-wide transcription boundary elements such as promoters and transcription termination sites (Carafa *et al.*, 1990; Ermolaeva, 2000) as well as operons (Craven *et al.*, 2000; Ermolaeva *et al.*, 2001; Salgado *et al.*, 2000; Yada *et al.*, 1999). Here, we define an operon as a set of two or more adjacent genes in the same orientation that are co-transcribed into a single mRNA under particular conditions, and we define an operon element as exactly two adjacent genes that are part of the same operon. These computational approaches are based largely on sequence analysis, gene annotation information, and cross-species sequence comparison, although Craven *et al.* (2000) also incorporate gene expression data from PCR-spotted arrays into their operon prediction model. With the maturing of microarray technology, we anticipate mRNA expression data playing an increasingly important role in identifying the extent of an organism's transcriptome. In this study, we present a method for identifying transcript boundaries based on mRNA expression data. While microarrays typically assay transcript expression of coding regions only, here we use Affymetrix high-density oligonucleotide microarrays which assay transcript expression throughout the entire genome, including intergenic regions. We perform 28 such microarray experiments, and using this expression data we construct hidden Markov models (HMMs) to determine the most probable transcription boundaries for genes. In this way, we distinguish 5′ UTRs as well as operon elements based solely on the expression data. We then validate our results against experimentally verified transcripts (Salgado *et al.*, 2001). While we present our analysis as a stand-alone approach, we envision it most effectively utilized as a supplement to the aforementioned sequence analysis and gene annotation methods. Expression indices for all genes under all experiments along with all our results are available at http://www.cs.washington.edu/homes/tjaden/ismb2002/.

*To whom correspondence should be addressed.

SYSTEMS AND METHODS

Arrays

We obtained transcript expression data using Affymetrix high-density oligonucleotide arrays. The array design has been described elsewhere Selinger *et al.* (2000) and cell growth, RNA labeling, and hybridization protocols have been detailed in a previous study (Tjaden *et al.*, 2002). In summary, we obtained RNA from *E. coli* cells grown under 14 different conditions, and RNA for each condition was hybridized to two identical arrays for a total of 28 experiments. The conditions include cells grown aerobically, anaerobically, with glucose and glycerol as carbon sources, under carbon starvation, with rich both medium, at temperatures 20 and 42°C, transitioning from exponential growth to stationary phase, and transitioning from anaerobic to aerobic growth. Each array contains 295 936 spots, or probes, half of which, called Perfect Match (PM) probes, correspond exactly to 25mer oligonucleotide sequences from the *E. coli* genome, and half of which, called mismatch (MM) probes, correspond to the same 25mer sequences as the PM probes except the middle base (13th of 25) is complemented in the MM probe. These MM probes help to estimate probe-specific background effects in the expression measurements. Every *E. coli* ORF (Blattner *et al.*, 1997) is assayed by a set of PM and MM probe pairs, and each intergenic region at least 40 base pairs in length is assayed in both orientations by a set of probe pairs. With few exceptions, probe sets contain 15 probes. It is important to clarify that probe sets assaying an ORF correspond to oligonucleotides strictly within the coding sequence. Rigorous definition of a gene may include an ORF along with transcribed but untranslated regions as well as regulatory regions. In this analysis, however, we define an intergenic region as the genome segment between annotated ORFs.

Gene expression

One of the obstacles in using oligonucleotide probes to measure transcription expression is that different oligonucleotides have different hybridization affinities, so that different probes in a set that assay the same transcript may yield significantly different expression levels (Figure 1). To overcome this obstacle, we employed the approach of Li and Wong (2001), which captures the unique responsiveness of each probe. Their method can be viewed as an expectation-maximization algorithm that iteratively estimates the transcript expression level of each ORF in each experiment (expectation step), and then, based on these estimated expression levels, calculates optimal values for a set of parameters which capture the unique responsiveness of each probe (maximization step). The model determines not only estimates for the expression of each ORF in each

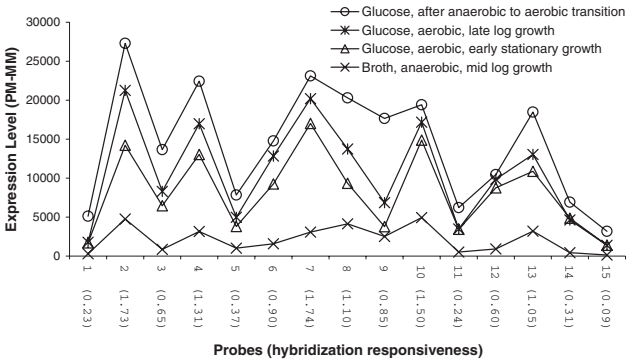

Fig. 1. Expression level of each of 15 probes assaying *dps* gene from 4 different experiments. The 4 plots demonstrate that the response of each probe is relatively consistent across experiments, but for a single experiment there is high variation between probes measuring the same transcript. The numbers in parentheses along the horizontal axis show the normalized hybridization responsiveness of each probe calculated using the approach of Li and Wong (2001).

experiment as well as responsiveness parameters for each probe (shown in parentheses along the abscissa for Figure 1), but it also allows quantification of standard errors for these terms which help to identify bad probes and, for each ORF, experiments which generate outlying expression values. Throughout our analysis, the expression vector for each ORF across our $m = 28$ experiments is calculated using this model (Li and Wong, 2001) and is denoted by $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ where θ_i is the expression index of the ORF in the ith experiment.

Intergenic expression

While the model of Li and Wong (2001) has met with considerable success, it requires sets of probes which are known *a priori* to assay the same transcript. This requirement is reasonable for ORFs since their annotations for *E. coli* are fairly well accepted. However for intergenic regions, transcription boundaries are largely unknown and our goal is to determine precisely which probes correspond to which transcripts. To achieve this, we begin with an expression vector θ for each ORF, and we look at probes assaying the intergenic regions immediately upstream and downstream of the ORF. Using hidden Markov models (Baum and Petrie, 1966; Krogh *et al.*, 1994) we then attempt to extend the ORF's transcript in both directions by associating intergenic probes with the ORF's transcript if the expression vector of an intergenic probe is highly correlated with the expression vector of the ORF.

It is important to note that the arrays only assay intergenic regions at least 40 base pairs in length. Since the majority of adjacent ORF pairs within documented

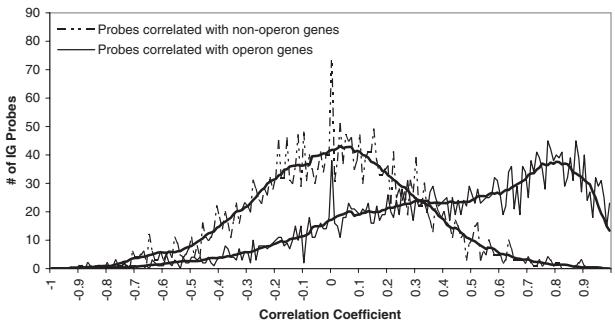

Fig. 2. Histograms showing the distribution of how the expression of intergenic probes is correlated with the expression of neighbouring ORFs. The dashed line gives the distribution of how the expression of each of 3040 intergenic probes which are known not to be within an operon correlate with the expression of a neighbouring ORF. The solid line gives the distribution of how the expression of each of 3040 intergenic probes which are known to be within an operon correlate with the expression of a neighbouring ORF. The bold lines show smoothed versions of the two distributions.

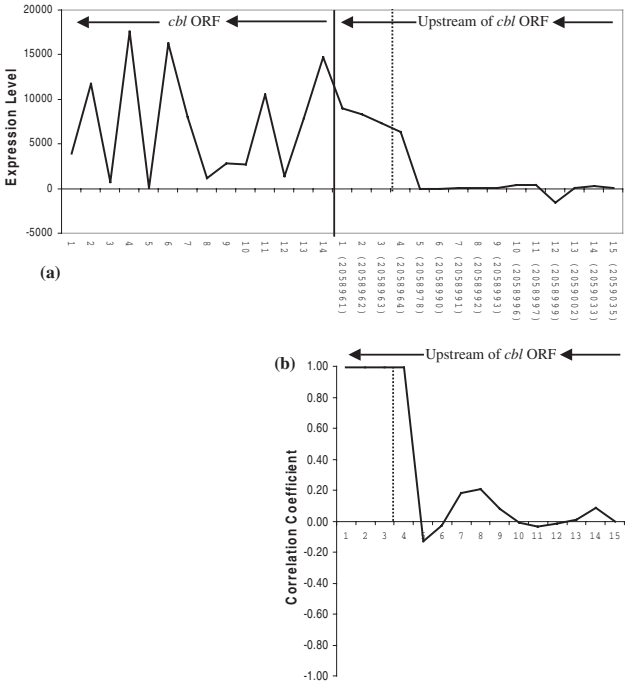

Fig. 3. (a) This graph shows the measured transcript expression of the *cbl* gene for a single experiment (starvation from glucose withdrawal at mid log phase growth). The left half of the graph shows the expression level (PM–MM intensity) of 14 probes assaying the *cbl* ORF. The right half of the graph shows the expression level (PM–MM intensity) of 15 probes assaying the region upstream of the *cbl* ORF. The number in parentheses beside each probe indicates the genomic coordinate of the start of the probe (e.g., probe 4 upstream of the *cbl* ORF corresponds to the 25mer oligonucleotide sequence at position 2058964-2058940 in the *E. coli* genome). The vertical dotted line shows the start position at 2058963 of the *cbl* promoter (Iwanicka-Nowicka and Hryniewicz, 1995). **(b)** The correlation, across all 28 experiments, of the expression of each of the 15 upstream probes with the expression of the *cbl* ORF.

operons are separated by less than 40 base pairs, we only assay a fraction of intergenic regions within operons, namely 154 such regions. Given a set of probes assaying the region upstream (or downstream) of a particular ORF, for each probe in the set we calculate the correlation of its expression vector to the expression vector of the corresponding ORF. The solid line in Figure 2 shows the distribution of correlation coefficients calculated from the expression vectors of probes assaying intergenic regions within known operons and the expression vectors of the immediately upstream (or downstream) ORF. The dashed line in Figure 2 shows the distribution of correlation coefficients between the expression vectors of probes in intergenic regions and the expression vectors of the immediately upstream (or downstream) ORF for the case when then the intergenic region is oppositely oriented from the upstream (or downstream) ORF. These two distributions are incorporated into our HMMs, as described in the next section, so that given the correlation between the expression of an intergenic probe and a neighbouring ORF, we can estimate whether or not the probe and ORF correspond to the same transcript.

Figure 3 provides an example of how correlation between expression of an intergenic region and its neighbouring ORF can elucidate transcription boundaries. Figure 3a shows the expression level from a single experiment as determined by 14 probes assaying the *cbl* ORF as well as 15 probes upstream of the ORF. From this expression data, we can make a reasonably accurate conjecture as to how far upstream of the *cbl* ORF its promoter lies. In general, this approach is problematic, however, since the responsiveness of individual probes assaying the same transcript can fluctuate dramatically, e.g., probes 4 and 5 assaying the *cbl* ORF show over a one hundred fold variation in their expression levels. Figure 3b shows how the expression vector of each of the 15 probes upstream of the *cbl* ORF correlates with the expression vector of the *cbl* ORF. Thus, we can deduce the location of the start of *cbl*'s $5'$ UTR, and we find our deduction corresponds almost exactly with the known promoter.

Hidden Markov models (HMMs)

Let us consider a set of $T = 15$ probes assaying an intergenic region upstream of an ORF. Let $O =$

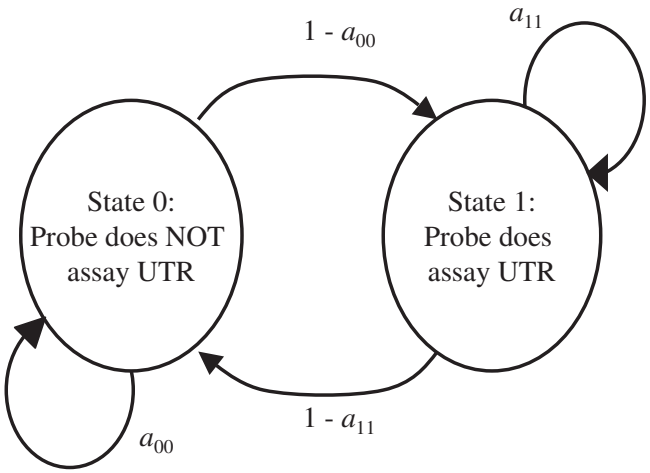

Fig. 4. A two state HMM. State 0 corresponds to a probe which assays a region which is not part of the gene's UTR. State 1 corresponds to a probe which assays a region which is part of the gene's UTR. The arrows represent the transitions between states.

$O_1, O_2, \ldots, O_T$ where O_j is the correlation coefficient of the expression vector of probe j with the expression vector θ of the ORF, and let $I = I_1, I_2, \ldots, I_T$ where I_j equals 1 if probe j assays the 5′ UTR of the gene and I_j equals 0 otherwise. If we knew the exact 5′ UTR for the gene, we would know which of the T probes assayed the gene transcript and which did not, i.e., we would know I, and we could use the appropriate distribution from Figure 2 to estimate the probability that the expression of probe j was correlated by O_j with θ. However, we do not know which probes assay UTR regions for the gene, rather we view this information as 'hidden' and our goal is to determine the sequence I which maximizes the joint probability $P(O, I)$. Further, we expect $P(I_j = 1 | I_{j-1} = 0)$ to be near zero because given that probe $j-1$ upstream of an ORF does not assay part of the gene's 5′ UTR, we do not expect probe j farther upstream to assay part of the gene's 5′ UTR. This dependency on previous 'states' motivates our use of HMMs.

Figure 4 shows a simple 2-state HMM which can be characterized by 3 sets of parameters, $\lambda = (A, B, \pi)$. Initial probabilities are represented by $\pi = \{\pi_x\}$ where π_x is the probability of starting in state x. Transition probabilities are represented by $A = \{a_{xy}\}$ where a_{xy} is the probability of being in state y given that the prior state was x. Emission probabilities are represented by $B = \{b_x(r)\}$ where $b_x(r)$ is the probability of generating or emitting the correlation coefficient $-1 \leqslant r \leqslant 1$ in state x. Our goal is to determine the most likely transcript boundary of a gene given the correlation data upstream or downstream of the ORF (as in Figure 3b). We illustrate how an HMM applies to our array data using the *cbl* gene and its upstream region as an example. From our array data, we first calculate the correlation coefficient of the expression vector for each of the 15 probes upstream of the *cbl* ORF with the expression vector θ for the *cbl* ORF to obtain the observation sequence $O^{cbl} = 1.00, 0.99, 0.99, 0.99, -0.13, -0.03, 0.18, 0.21, 0.08, -0.01, -0.04, -0.02, 0.01, 0.08, 0.00$ (Figure 3b). Our objective is to determine which probes assay UTR regions and which do not, i.e., we want to determine the most probable state sequence I. Thus, given a model λ, we want to find the sequence I which maximizes the joint probability of the observation sequence and state sequence given the model:

$$\begin{aligned}\arg\max_I [P(O, I|\lambda)] &= \arg\max_I [P(O|I, \lambda)P(I|\lambda)] \\ &= \arg\max_{I=I_1, I_2, \cdots, I_T} [\pi_{I_1} b_{I_1}(O_1) a_{I_1 I_2} b_{I_2} \\ &\quad \times (O_2) \cdots a_{I_{T-1} I_T} b_{I_T}(O_T)].\end{aligned}$$

Using a dynamic programming approach, the Viterbi Algorithm (Forney, 1973), we can determine this optimal state sequence. This approach, however, relies on having a model for the HMM with appropriate parameters for A, B, and π. For the emission probabilities B, we use the two smoothed distributions shown in Figure 2. For the initial probabilities π and transition probabilities A, we train our HMM from a set of observation sequences using the segmental k-means algorithm (Juang and Rabiner, 1990) so that A and π are chosen to maximize $P(O, I|\lambda)$. We implement two sets of HMMs: one for operon identification and one for 5′ UTR identification. Table 1 shows typical parameter values for A and π in the two HMM sets. Since we cannot identify transcription boundaries with our approach when a gene is not expressed in any experiment, we restrict all our training and test sets to intergenic regions upstream or downstream of ORFs where we observed at least minimal expression of the ORF transcript in one or more experiments. We deem a gene expressed in an experiment i if θ_i exceeds three times the standard error. Altering this threshold or employing other expression criteria did not significantly affect our results. Of the 154 intergenic regions within documented operons which we assay, we found evidence that the flanking operon genes were expressed in at least one experiment for 115 of these. For each of these 115 intergenic regions, we generate two observation sequences, one for correlation with the upstream ORF and one for correlation with the downstream ORF, for a total of 230 operon training sequences. We use leave-one-out-cross-validation (LOOCV) to train and classify these 115 operon intergenic regions. Similarly, we assay with exactly 15 probes, 274 intergenic regions upstream of ORFs with documented promoters and for which we observe at least minimal expression. Again, we use LOOCV to train our HMM and classify the upstream regions. As

Table 1 shows, the HMMs learned that the probability of transitioning out of state 0 is essentially nil, meaning that once we encounter a probe assaying a region upstream (or downstream) of an ORF which is not part of the UTR then we will not encounter a probe farther upstream (or downstream) which is part of the UTR.

Returning to our example with the *cbl* gene, using an HMM with parameters λ trained from 273 upstream observation sequences, we find the maximal value of $P(O^{cbl}, I|\lambda)$ occurs when $I = 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0$ and we hypothesized that the first 4 probes assaying regions upstream of the *cbl* ORF correspond to the 5′ UTR and the 11 probes farther upstream of these do not correspond to the UTR. Since the fourth probe upstream of the *cbl* ORF corresponds to the 25mer starting 28 base pairs upstream of the ORF and ending 3 base pairs upstream of the ORF, and since we have hypothesized that the expression of this probe is highly correlated with the expression θ observed across the ORF, we conjecture the 5′ UTR to begin around 28 base pairs upstream of the ORF. Or to be more precise, since the most probable path through the HMM indicates that the expression of the fifth probe upstream of the ORF is not correlated with θ and since the fifth probe corresponds to the 25mer starting 42 base pairs upstream of the ORF, we expect the 5′ UTR to begin between approximately 28 and 42 base pairs upstream of the ORF. The known transcription start site for *cbl* occurs ∼27 base pairs upstream of the ORF. In general, our resolution is quite high since we average one probe every 6 base pairs in intergenic regions, however, due to the error in array experiments and the stochastic nature of polymerase binding and transcription initiation, we are content to identify approximate transcription boundaries.

We proceed to apply our analysis genome-wide. For every ORF and corresponding upstream region which we assay, we train an HMM. Given the model and array data, we then identify the 5′ UTR by determining the most likely upstream continuation of probes which assay the same transcript as the ORF. For each intergenic region between two similarly oriented ORFs, we train an HMM from operon data (Table 1a) and we classify the intergenic region as part of an operon if the expression of probes assaying it correlates with the expression of both the upstream and downstream ORFs.

RESULTS AND DISCUSSION

As an initial study to determine whether correlation of expression is a reasonable indicator of co-transcription, we investigated how well expression correlation identifies co-transcribed genes (operons). As a positive test set, we used RegulonDB (Salgado *et al.*, 2001) to identify 463 pairs of genes which are adjacent on the genome, which are documented as belonging to the same operon, and for which we detect expression under at least one condition. As a negative test set, we randomly chose 463 pairs of genes which are adjacent on the genome but on opposite strands under the assumption that these gene pairs cannot be co-transcribed. After calculating the Pearson correlation coefficient between the expression vectors of the two genes in each pair, we classified a gene pair as part of an operon if the correlation coefficient exceeded some threshold. Figure 5 shows the receiver operating characteristic (ROC) curve as we vary the correlation threshold. The elbow of the curve occurs when our correlation threshold is set to 0.76 at which point we achieve 97% specificity and 57% sensitivity. The strength of these results suggests correlation of expression is at least a reasonable indication of co-transcription. In addition, a more careful investigation of our test sets provides insight into our false positive and false negative results. Firstly, expression correlation of two genes does not distinguish effectively between two genes which are co-transcribed and two genes which are not co-transcribed but are co-regulated or part of similarly regulated networks. For example, one of our few false positive classifications corresponds to the *argC* and *argE* neighbouring genes which we observed to be highly correlated ($r = 0.93$) but which are oppositely oriented and not co-transcribed. Both genes are regulated by the transcription factor ArgR and are involved in the arginine biosynthesis pathway. Our small false positive rate of 3%, however, suggests that it is relatively rare that mis-oriented neighbouring genes are tightly co-regulated across many conditions. When we apply our analysis to intergenic regions where probes are tiled on average one every six base pairs, we expect the high resolution largely to tease apart similarly regulated but separately transcribed elements. Our false negatives may be explained partially by the observation that genes corresponding to our false negative classifications tended to have lower expression levels and tended to be expressed under fewer conditions than those genes corresponding to our true positive classifications. As a result, the expression correlation of our false negative gene pairs was lower and had higher error. Further, several gene pairs which are documented as part of operons may be co-transcribed under some conditions but may have internal promoters and may be individually transcribed under other conditions. In these cases, the expression vectors of the gene pairs naturally will be less correlated. Examples include the *focA-pflB* operon (Sawers and Bock, 1989), *ftsQAZ* operon (Wang *et al.*, 1991), and *infB-rpsO* operon (Nakamura and Mizusawa, 1985; Sands *et al.*, 1988). It is unclear how common these occurrences are throughout *E. coli* operons. When we extended our analysis to incorporate expression of intergenic regions using our HMM approach, we correctly identified operon elements with 99% specificity

Table 1. Optimal parameters learned from training an HMM on **(a)** 230 observation sequences determined from intergenic regions within operons and **(b)** 274 observation sequences determined from intergenic regions upstream of genes with documented promoters. Each training observation sequence generally consists of 15 correlation coefficients which correspond to how the expression of each of 15 intergenic probes correlates with the expression of a neighbouring ORF

	Training Sequences	π_0	π_1	a_{00}	a_{01}	a_{11}	a_{10}
(a) Operon HMM	230	0.26	0.74	1.00	0.00	0.99	0.01
(b) 5′ UTR HMM	274	0.41	0.59	1.00	0.00	0.89	0.11

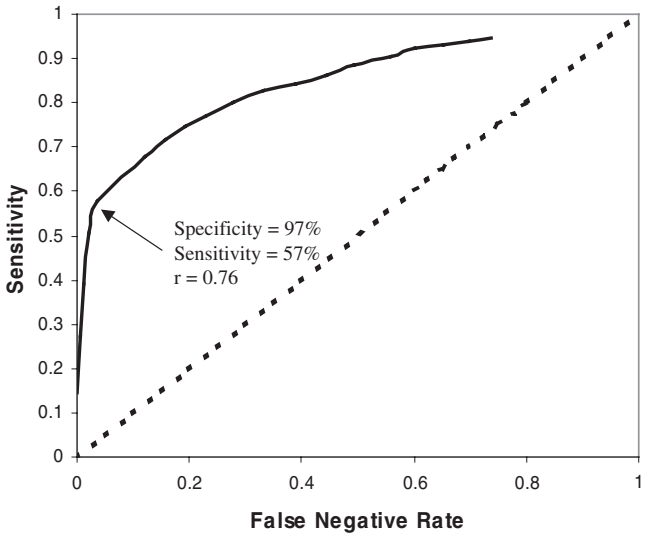

Fig. 5. We identify two neighbouring genes as part of an operon if the correlation coefficient of their expression vectors exceeds a threshold. The ROC curve shows the sensitivity/specificity trade-off as we vary our correlation threshold.

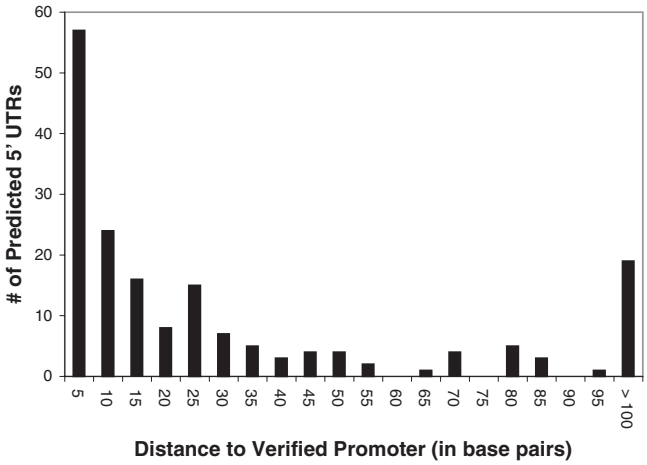

Fig. 6. Histogram of distances between verified promoters and the start of the 5′ UTR which we identify. For 71% of expressed genes which have a documented promoter, the 5′ UTR which we identify starts within 30 base pairs of a verified transcription initiation site.

and 63% sensitivity. These results are based on 115 positive examples of documented operon elements with intergenic regions greater than 40 base pairs and where we observe at least minimal transcript expression of the ORFs, and from 115 randomly chosen adjacent mis-oriented gene pairs as negative examples. Further, we identified an additional 227 new operon elements from our HMM analysis.

Validating our 5′ UTR results is less straightforward. We designate the start of a 5′ UTR to be the first base pair assayed by the farthest upstream probe which is part of the gene transcript as determined by our HMM. This designation is somewhat arbitrary since the oligonucleotide probes assay discrete transcript regions at non-uniform intervals so that we cannot hope to determine UTRs within a single base pair resolution. However, by using a single nucleotide as a surrogate for the start of the UTR, we can quantify the distance in base pairs between our prediction and an experimentally verified promoter. Figure 6 shows that for 71% of genes with known promoters for which we make 5′ UTR predictions, the start of the UTR which we identify lies within 30 base pairs of a documented transcription start site. While this metric allows us to approximate the locations of promoter regions, it doesn't necessarily help us understand how reliable our results are for UTR boundaries. Instead, we consider the following alternative which captures whether our data is completely consistent with verified promoter information for each gene. We say that our data *supports* a known promoter if our HMM determines that every intergenic probe downstream of the documented transcription start site assays part of the gene transcript and every intergenic probe upstream of the documented transcription start site does not assay part of the gene transcript. By this metric, 63% of the 5′ UTRs which we identify *support* a known promoter, i.e., for 63% of genes where we find evidence of upstream UTRs, we correctly identify all intergenic probes as assaying the UTR or not assaying the

UTR. We offer two explanations for the 37% of genes which our data does not *support*. Firstly, many genes have multiple promoters. While one promoter for a gene may be documented, the gene may have other unverified promoters which explain why our data analysis is not consistent with the documented promoter. Secondly, we find that approximately 4% of probes assaying ORF transcripts on the array are simply 'bad' probes. The response of bad probes never exceeds background or always has unusually high error. We identified these bad probes via two orthogonal approaches. The Li and Wong (2001) approach which we implemented distinguishes bad probes with relatively high standard errors. Also, we hybridized *E. coli* genomic DNA to a control microarray and identified probes whose PM intensity did not significantly exceed the MM intensity, indicating cross hybridization or other non-specific hybridization. We expect a similar percentage of bad intergenic probes, and further investigation is needed to determine effective methods for masking these probes. Since we are using a hidden Markov model, the bad probes in intergenic regions affect not only our UTR classification for the particular probe but also for probes farther upstream or downstream, thereby causing our estimates in these intergenic regions to be unreliable. Finally, we estimate a conservative lower bound on the specificity of our 5′ UTR results. We generate a negative test set by running our HMM analysis on the reverse complement of the upstream region of every *E. coli* ORF which is expressed at least minimally. We found no UTR evidence for 90% of the examples giving us a lower bound on our specificity. Our 10% false negative rate may be due to anti-sense gene regulation or to oppositely oriented neighbouring genes which are co-regulated.

We also applied our method to 3′ UTRs but the results are less reliable and require further investigation. Transcription termination is less well understood than transcription initiation. For training and testing our HMM, we only found 51 genes for which we observed sufficient expression and which have verified transcription termination sites. The transcription termination mechanism for these genes is almost exclusively rho-independent. Recent evidence suggests that gene transcripts may extend well beyond rho-independent hairpin-loop terminators only to be post-processed back to the termination site after transcription. Also, during coincident transcription and translation, the translation mechanism may variably regulate termination sites of transcripts (Abe *et al.*, 1999). As a result, our expression data for potential 3′ UTRs is less consistent.

CONCLUSIONS

We present a genome-wide approach using hidden Markov models on microarray data for identifying transcription boundaries. All results from this study along with the expression index for every ORF in all 28 experiments and the responsiveness of each probe (calculated via the method of Li and Wong (2001)) are available on our website. We validate our results against experimentally verified operons and 5′ UTRs, and we are in the process of confirming several of our new predictions with RT-PCR.

Previous approaches for genome-wide prediction of transcription boundaries have employed mostly sequence analysis and annotation data. Our approach is based on transcript expression experiments. Our method is not meant as a competing alternative to the sequence analysis approaches, but rather as a complement. Indeed, the most successful predictive methods (such as Craven *et al.* (2000)) are often those that combine data from orthogonal sources. As microarrays which tile not just coding sequences but whole genomes become more common, we foresee expression analysis becoming increasingly important in identifying not just operon and UTR boundaries but all elements of the transcriptome including small RNAs, new genes, and anti-sense regulatory transcripts.

ACKNOWLEDGEMENTS

This work was supported in part by the DOE Microbial Cell Program #8217 grant DE-FG08-01ER63218 to E.K. We also thank Affymetrix, Institute for Systems Biology and Leroy Hood for support, Alex Picone and Rini Mukherjee Saxena for technical assistance.

REFERENCES

Abe,H., Abo,T. and Aiba,H. (1999) Regulation of intrinsic terminator by translation in *Escherichia coli*: transcription termination at a distance downstream. *Genes to Cell*, **4**, 87–97.

Baum,L.E. and Petrie,T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, **37**, 1554–1563.

Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Carafa,Y.A., Brody,E. and Thermes,C. (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. *J. Mol. Biol.*, **216**, 835–858.

Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 116–127.

Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.

Ermolaeva,M.D., White,O. and Salzberg,S. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Forney,Jr,G.D. (1973) The Viterbi algorithm. *Proc. IEEE*, **61**, 263–278.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Iwanicka-Nowicka,R. and Hryniewicz,M.M. (1995) A new gene, cbl, encoding a member of the LysR family of transcriptional regulators belongs to *Escherichia coli* cys regulon. *Gene*, **166**, 11–17.

Juang,B.H. and Rabiner,L.R. (1990) The segmental *k*-means algorithm for estimating the parameters of hidden Markov models. *IEEE Trans. Accoust., Speech, Signal Processing*, **38**, 1639–1641.

Li,C. and Wong,H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Nakamura,Y. and Mizusawa,S. (1985) In vivo evidence that the nusA and infB genes of *E. coli* are part of the same multi-gene operon which encodes at least four proteins. *EMBO J.*, **4**, 527–532.

Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657. http://www.cifn.unam.mx/Computational_Genomics/regulondb/

Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Diaz-Peredo,E., Sanchez-Solano,F., Perez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.

Sands,J.F., Regnier,P., Cummings,H.S., Grunberg-Manago,M. and Hershey,J.W. (1988) The existence of two genes between infB and rpsO in the *Escherichia coli* genome: DNA sequencing and S1 nuclease mapping. *Nucleic Acids Res.*, **16**, 10803–10816.

Sawers,G. and Bock,A. (1989) Novel transcriptional control of the pyruvate formate-lyase gene: upstream regulatory sequences and multiple promoters regulate anaerobic expression. *J. Bacteriol.*, **171**, 2485–2498.

Selinger,D.W., Cheung,K.J., Mei,R., Johansson,E.M., Richmond,C.S., Blattner,F.R., Lockhart,D.J. and Church,G.M. (2000) RNA expression analysis using 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.*, **18**, 1262–1268.

Tjaden,B., Saxena,R.M., Stolyar,S., Haynor,D.R., Kolker,E. and Rosenow,C. (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. (submitted).

Wang,X.D., de Boer,P.A. and Rothfield,L.I. (1991) A factor that positively regulates cell division by activating transcription of the major cluster of essential cell division genes of *Escherichia coli*. *EMBO J.*, **10**, 3363–3372.

Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.