TRANSCRIPT NORMALIZATION AND SEGMENTATION OF TILING ARRAY DATA

GEORG ZELLER

Friedrich Mieschер Laboratory of the Max Planck Society &
Max Planck Institute for Developmental Biology, Dept. for Molecular Biology
Spemannstr. 35 & 39, 72076 Tübingen, Germany
E-mail: Georg.Zeller@tuebingen.mpg.de

STEFAN R. HENZ, SASCHA LAUBINGER & DETLEF WEIGEL

Max Planck Institute for Developmental Biology, Dept. for Molecular Biology
Spemannstr. 35, 72076 Tübingen, Germany
E-mail: {Stefan.Henz,Sascha.Laubinger,Detlef.Weigel}@tuebingen.mpg.de

GUNNAR RÄTSCH

Friedrich Mieschер Laboratory of the Max Planck Society
Spemannstr. 39, 72076 Tübingen, Germany
E-mail: Gunnar.Raetsch@tuebingen.mpg.de

For the analysis of transcriptional tiling arrays we have developed two methods based on state-of-the-art machine learning algorithms. First, we present a novel transcript normalization technique to alleviate the effect of oligonucleotide probe sequences on hybridization intensity. It is specifically designed to decrease the variability observed for individual probes complementary to the same transcript. Applying this normalization technique to *Arabidopsis* tiling arrays, we are able to reduce sequence biases and also significantly improve separation in signal intensity between exonic and intronic/intergenic probes. Our second contribution is a method for transcript mapping. It extends an algorithm proposed for yeast tiling arrays to the more challenging task of spliced transcript identification. When evaluated on raw versus normalized intensities our method achieves highest prediction accuracy when segmentation is performed on transcript-normalized tiling array data.

Datasets, software and the appendix are available for download at `http://www.fml.mpg.de/raetsch/projects/PSBTiling`

1. Introduction

Tiling arrays on which oligonucleotide probes are spotted at high density have made it feasible to study whole genomes in an unbiased and cost-effective way. They have been used for experiments as diverse as transcriptome analysis, ChIP-chip and DNA sequence variation detection.[4,6,7,8,9,11].

The analysis of tiling array data, however, is not straightforward since intensity measurements are known to be influenced by many factors. In order to allow direct comparisons between arrays potentially hybridized under slightly differ-

ent experimental conditions, the measurements are typically first normalized as a whole, e.g. by array quantile normalization.[3] Another major reason for variability in hybridization intensity are divergent sequence properties of oligonucleotide probes that have not been optimized due to constraints on tiling array design. In this work we compare a newly developed *transcript normalization* technique for the removal of sequence-specific effects to the recently proposed *sequence quantile normalization*.[16] Our approach particularly aims at reducing the variability around mRNA transcript levels which are ideally assumed to be constant across all exon probes of the same transcript. We have therefore developed a regression model that estimates the deviation between the observed intensities of individual probes and the *transcript intensity* taking probe sequences as input. Such a normalization is expected to be beneficial particularly for transcript mapping approaches attempting to segment the genome into transcriptional units of approximately constant hybridization intensity.

The monitoring of known genes and especially the identification of novel transcripts with whole-genome tiling arrays has received increasing attention over the last years. For the analysis of *S. cerevisiae* tiling arrays, Huber et al.[11] proposed a method that segments the yeast chromosomes such that the sum of squared differences of signal intensities to their mean within a given segment is minimized. To solve this mathematical problem, known as Structural Change Model Segmentation (SCM), they adapted the dynamic programming algorithm proposed by Bai and Perron.[2] While this relatively simple approach has been successfully applied to yeast tiling array data, the segmentation problem is considerably more challenging for the genomes of higher eukaryotes that are capable of (alternative) splicing. Here, gene density is typically lower, exon segments are much shorter and interrupted by potentially very long intron sequences. A more sophisticated model, called GenRate, has been proposed by Frey et al.[9] It explicitly models coregulated units (CoRegs) such as exons of the same gene exhibiting the same expression level. However, the generative model for sequences of hybridization measurements that constitutes the core of their method is based on several assumptions on the structure of a transcript and the distribution of hybridization measurements (e.g. Gaussian distribution of intensity differences from a designated reference probe, geometrically distributed distance of the reference probe from the transcript start, etc.).

Building on this work we propose a novel method that is able to accurately recognize transcripts from tiling array measurements. Our approach is based on a discriminative learning technique closely related to Hidden Markov (HM) Support Vector Machines (SVMs)[1] which combine the advantages of HM models[5] for label sequence learning with those of the discriminative SVM framework. A

precursor method can be seen as a reformulation of the SCM method modeling interruptions of active regions (exons) with inactive regions (introns). For this model we still assume Gaussian noise for the deviation of exon probe intensities from their average. Since this assumption is typically not satisfied, we augment the method with more flexible *scoring functions* replacing the squared error terms. Their shapes are estimated from data in order to optimally segment the sequence of intensity measurements. As a supervised learning approach, our algorithm is trained on hybridization intensities together with segmentations determined from known mRNA transcripts.

2. Normalization of Transcriptional Tiling Arrays

2.1. *Array Data and Preprocessing*

We analyzed data from *A. thaliana* tiling arrays manufactured by Affymetrix. For hybridization, total RNA of 21 day-old inflorescences was amplified using oligo-dT-T7 primers. Resulting RNA was converted into double-stranded cDNA, fragmented, labeled and hybridized to Affymetrix *TilingR* arrays following standard protocols (see Appendix A for details).

In a first normalization step, measurements affected by artifacts already apparent from the scanned image of the array were removed using a software called *Harshlight*.[19] To facilitate inter-array comparisons *quantile-normalization* was applied, which involves computing the mean over the empirical intensity distributions of all considered arrays. This mean distribution is then re-assigned to each of the arrays, thus effectively removing differences in intensity distribution between arrays.[3] All intensity measurements were $\log_2$ transformed for the subsequent normalization steps.

2.2. *Sequence Quantile Normalization (SQN)*

Sequence quantile normalization (SQN) has been proposed as an extension of the above described quantile-normalization to remove probe sequence effects.[16] For each 25-mer probe having nucleotide $j \in A, C, G, T$ at position $k = 1, \ldots, 25$ the rank $r_{i,j,k}$ of its intensity y_i among all other probes with the same nucleotide at position k is calculated and normalized by the number of such probes $C_{j,k}$. These position-wise contributions are then averaged: $\hat{S}_i = \frac{1}{25} \sum_{k=1}^{25} \frac{r_{i,j,k}}{C_{j,k}}$. Since the sequence bias is not uniform across positions and summands are not independent, the multivariate regression problem is solved iteratively; in each step the above average is computed and afterwards intensities y_i are replaced by $\hat{S}_i$ which is repeated until convergence.[16]

As a side effect, intensities are substituted by relative ranks that are uniformly distributed between zero and one. In order to obtain normalized intensity values comparable to the original measurements from the array, we modified the averaging as follows. Intensity distributions were approximated by piece-wise linear functions $g_k(r_{i,j,k}) \approx y_i$. In our case, g is parametrized by 200 supporting points with uniformly spaced x-values s_x between zero and one. The corresponding y-values s_y are estimated by linear interpolation between y_m and y_n having ranks $r_{m,j,k} = \max_{m'}\{r_{m',j,k} \mid r_{m',j,k}/C_{j,k} \leq s_x\}$ and $r_{n,j,k} = \min_{n'}\{r_{n',j,k} \mid r_{n',j,k}/C_{j,k} \geq s_x\}$, respectively. Instead of averaging relative ranks, we then calculated the mean $\hat{g} = \frac{1}{25}\sum_{k=1}^{25} g_k$ of the supporting points s_y. From this averaged $\hat{g}$ we reconstructed the normalized intensities by linear interpolation between the supporting points of $\hat{g}$.

2.3. *Transcript normalization techniques*

Ideally, one would expect constant hybridization intensity for all probes measuring the same transcript. Similarly, the background signal of probes in untranscribed or intronic regions of the genome would ideally be constant. However in practice, this is generally not the case (see e.g. Royce et al.[15] for a discussion).

Here we propose a method to reduce within-gene variability caused by probe sequence effects. In a first step we estimate constant transcript and background intensities $\overline{y}_i$ based on the TAIR7 genome annotation,[14] in the following simply referred to as *transcript intensities*: If a probe i is annotated as exon, $\overline{y}_i$ is the median of the intensities y_i of probes in exons of the same gene. Similarly for intron probes, we compute $\overline{y}_i$ as the median over intronic probes of the same gene and for intergenic regions $\overline{y}_i$ is the median of all probes mapped to regions annotated as intergenic. (Probes that were mapped to intron / exon boundaries, more than one splice-form or overlapping genes are excluded from training and evaluation.)

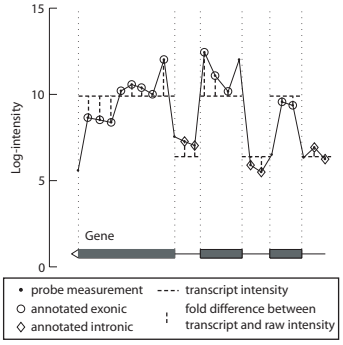


Figure 1.: Illustration of raw and transcript intensities for part of a transcript

Assuming that the concentration of mRNA hybridized to all exon probes of a gene is constant, the differences between the raw intensities and the transcript intensities $\hat{y}_i := y_i - \overline{y}_i$ are mainly due to probe sequence-specific effects (ignoring cross-hybridization, experimental artifacts and thermodynamic noise). Furthermore, it is conceivable that probe effects also depend on the mRNA concentration, and hence the differences $\hat{y}_i$ may also depend on the transcript intensities $\overline{y}_i$ of the exons of the gene. Since

it is not *a priori* clear how this dependence should be modeled, it appears reasonable to non-parametrically model the difference by a function of the form $f(\boldsymbol{x}_i, \overline{y}_i) \approx y_i - \overline{y}_i$ that depends both on sequence features $\boldsymbol{x}_i$ of the probe as well as its transcript intensity. However, in order to use this correction, one would have to know in advance whether a certain probe is exonic, intronic or intergenic which is not generally the case. We therefore propose to estimate the function depending not on the transcript intensity, but instead on the raw intensities, i.e. $f(\boldsymbol{x}_i, y_i) \approx y_i - \overline{y}_i$.

Given the large amounts of available data for estimating $f(\boldsymbol{x}, y)$, we can discretize the parameter y into Q quantiles and estimate Q independent functions $f_q(\boldsymbol{x})$. Then $f(\boldsymbol{x}, y)$ is given by

$$f(\boldsymbol{x}, y) = \begin{cases} f_1(\boldsymbol{x}) & \text{for } y \in (-\infty, y_1) \\ \cdots & \cdots \\ f_i(\boldsymbol{x}) & \text{for } y \in [y_i, y_{i+1}) \\ \cdots & \cdots \\ f_Q(\boldsymbol{x}) & \text{for } y \in [y_Q, \infty) \end{cases}$$

As input $\boldsymbol{x}_i$ to the regression function f_q the sequence $\boldsymbol{s}_i$ of probe i was provided together with additional features derived from the sequence: sequence entropy $-\sum_{i=1}^{4} f_i \times \log(f_i)$, where f_i is the frequency of the nucleotide $i \in \{A, C, G, T\}$ in the probe sequence and GC content. Furthermore, two hairpin scores were used: One is the maximum number of base pairs over all possible hairpin structures that a probe can form, the other one is equal to the maximum number of consecutive base pairs over all possible hairpin structures (similarly used for intensity modelling in Zhan et al.[22]).

Based on these sequence features, we considered two methods for learning the functions f_q based on Q sets of n training examples $(\boldsymbol{x}_i^q, \hat{y}_i^q)$, where $\hat{y}_i = y_i - \overline{y}_i$, $i = 1, \ldots, N$ and $q = 1, \ldots, Q$:

Support Vector Regression (SVR) For regression, we applied Support Vector Machines[17] with a kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ that computes the "similarity" of two examples $\boldsymbol{x}$ and $\boldsymbol{x}'$. Here we used a sum of the Weighted Degree (WD) kernel[13,12] and a linear kernel. The WD kernel has been developed to model sequence properties taking the occurrence and position of substrings up to a certain length d into account.[13] We considered substrings up to order $d = 3$ and allowed a shift of 1 bp between positions of the substring,[12] which can be efficiently dealt with using string indexing data structures.[18] The linear kernel computed the scalar product of the sequence-derived features described above. We used the freely available implementations from the *Shogun toolbox*.[18]

Ridge regression (RR) For every training example we explicitly generated a feature vector from the sequence $\boldsymbol{s}$ having an entry for every possible mono-, di- and tri-nucleotide at every position in the probe (one if present at a po-

sition, zero otherwise; similar to the implicit representation in the WD kernel). The resulting feature vector was augmented with the sequence derived features to form $\boldsymbol{x}_i$. In training, the λ-regularized quadratic error is minimized:[10] $\min\ \lambda||w||^2 + \sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i - \hat{y}_i)^2$ with $\boldsymbol{w} = \left(\lambda I + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\sum_{i=1}^{n}\hat{y}_i\boldsymbol{x}_i$ being its solution. Then $f_q(\boldsymbol{x}) = \boldsymbol{w}_q^T\boldsymbol{x}$ is the resulting regression estimate.

Ridge regression is straightforward to implement in any programming language supporting matrix operations and linear equation solvers. In terms of computation time it is much less demanding than both SVR and SQN.

3. Transcript Identification

In this section we describe a novel segmentation algorithm for transcriptional tiling array data. It is based on ideas similarly presented before,[9,11] but uses a different strategy for learning and inference (cf. Section 1).

The goal is to characterize each probe as either intergenic (not transcribed) or as part of a transcriptional unit (either exon or intron). Instead of predicting the label of a probe (intergenic, exonic or intronic) directly, we learn to associate a state with each probe given its hybridization measurements and the local context. From the state sequence we can easily infer the label sequence (see Figure 2).

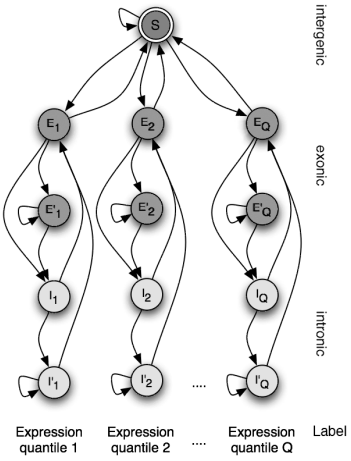


Figure 2.: State model with a subset of states for each expression quantile (columns). The label corresponding to each state is indicated on the right.

For learning we first defined the target state sequence, i.e. the "truth" that we attempted to approximate. It was generated from known transcripts and hybridization measurements. We then applied HMSVMs[1] for label sequence learning to build a discriminative model capable of predicting the state and hence the label sequence given the hybridization measurements alone.

State Model The simplest version of the state model had only three states: intergenic, exonic & intronic. It was extended in two ways: (a) by introducing an intron/exon start state that allowed modeling of the start and the continuation of exons & introns separately and (b) by repeating the exon and intron states for each expression quantile which allowed us to model discretized expression levels separately (see below). The resulting state model is outlined in Figure 2. Finally, to compensate

for the 3' intensity bias described in Appendix E, we also allow transitions from the exon states of one level to the ones of the next higher or lower level.

Generation of Labelings For genomic regions with known transcripts we considered the sense direction of up to 1 kbp flanking intergenic regions while maintaining a distance of at least 100 bp to the next annotated gene. Within this region we assigned one of the following labels to every probe: *intergenic*, *exonic*, *intronic* and *boundary*. In a second step we subdivided genes according to the median hybridization intensity of all exonic probes into one of $Q = 20$ expression quantiles. For each probe a state was determined from its label and expression quantile. (The boundary probes were excluded in evaluation.)

Parametrization and Learning Algorithm Our goal was to learn a function $f : \mathbb{R}^\star \to \Sigma^\star$ predicting a state sequence $\boldsymbol{\sigma} \in \Sigma^\star$ given a sequence of hybridization measurements $\boldsymbol{\chi} \in \mathbb{R}^\star$, both of equal length T. This was done indirectly via a $\boldsymbol{\theta}$-parametrized discriminant function $F_{\boldsymbol{\theta}} : \mathbb{R}^\star \times \Sigma^\star \to \mathbb{R}$ that assigned a real-valued score to a pair of observation and state sequence.[1,20] Knowing $F_{\boldsymbol{\theta}}$ allowed to determine the maximally scoring state sequence by dynamic programming,[5] i.e. $f(\boldsymbol{\chi}) = \operatorname*{argmax}_{\boldsymbol{\sigma} \in \mathcal{S}^\star} F_{\boldsymbol{\theta}}(\boldsymbol{\chi}, \boldsymbol{\sigma})$.

For each state $\tau \in \Sigma$, we employed a *scoring function* $g_\tau : \mathbb{R} \to \mathbb{R}$. $F_{\boldsymbol{\theta}}$ was then obtained as the sum of the individual scoring contributions and the *transition scores* given by $\phi : \Sigma \times \Sigma \to \mathbb{R}$:

$$F_{\boldsymbol{\theta}}(\boldsymbol{\chi}, \boldsymbol{\sigma}) = \sum_{t=1}^{T} \sum_{\tau \in \Sigma} [[\sigma_t = \tau]] \, g_\tau(\chi_t) + \phi(\sigma_{t-1}, \sigma_t)$$

where [[.]] denotes the indicator function. We modeled the scoring functions g_τ as piecewise linear functions[13] (PLiF) with $L = 20$ supporting points $s_1, \ldots, s_L$. Together with the transition scores ϕ, the y-values at the supporting points $\theta_{\tau,l} =: g_\tau(s_l)$ constituted the parametrization of the model, collectively denoted by $\boldsymbol{\theta}$.

During discriminative training a large margin of separation between the score of the correct path and *any* other wrong path was enforced. (For details on the optimization problem see Appendix C and Altun et al.[1])

4. Results and Discussion

4.1. *Probe Normalization*

The *A. thaliana* genome was partitioned into ≈300 regions while avoiding splits in annotated genes. Mapping perfect match (PM) probes to genome locations resulted in ≈10000 probes per region. We randomly chose 40% of these regions for

training, 20% for hyper-parameter tuning and the remaining 40% as a test set for performance assessment. The test regions were further used for the segmentation experiments in Section 4.3.

Removal of Sequence Effects Figure 3 shows that hybridization intensity is strongly correlated with the GC content of the probe causing more than 4-fold changes in median intensity. This sequence effect was reduced by all normalization methods. However, Figure 3 also indicates that the effect is (in part) explained by GC-richness of coding regions.[21] Position-specific sequence effects were further investigated with so-called quantile plots.[16] The strongest reduction of first-order sequence effects was achieved with SQN, although positional sequence effects were reduced by all normalization methods (see Appendix D).

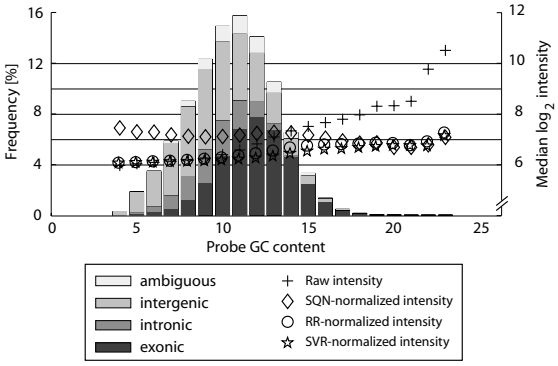


Figure 3. Median hybridization intensity depends on GC content of oligonucleotide probes. The histogram obtained by partitioning probes according to their GC content is shown as bar plots. In each bin the frequency of exonic, intronic and intergenic probes is indicated by different gray-scales, and the median log-intensity is shown before and after the application of normalization methods (see inset).

Reduction of Transcript Intensity Variability For the assessment of transcript variability, i.e. the deviation of individual probe intensities y_i from the constant transcript or background intensity $\overline{y}_i$, we introduced two metrics, T_1 and T_2. Both relate the variability of normalized intensities $y_i - f(\boldsymbol{x}_i, y_i)$ to the variability of raw intensities, and values smaller than 1 indicate a reduction. We defined $T_1 := \frac{\sum_i |y_i - f(\boldsymbol{x}_i, y_i) - \overline{y}_i|}{\sum_i |y_i - \overline{y}_i|}$ as the normalized absolute transcript variability and $T_2 := \frac{\sum_i (y_i - f(\boldsymbol{x}_i, y_i) - \overline{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2}$ as the normalized squared transcript variability. SVR minimizes the so-called ϵ-insensitive loss closely related to the absolute error, while Ridge regression minimizes the squared loss. Therefore, we expected and observed smaller T_1 values for SVR and smaller T_2 values for RR (see Figure 4). With both methods transcript variability was reduced to approximately half the values of raw intensities. For SQN we observed both T_1 and T_2 greater than 1 indicating increased transcript

Method	T_1	T_2
SQN	1.83	3.16
SVR	0.54	0.47
RR	0.58	0.44

Figure 4.: Within-gene variability after normalization.

variability. One may argue that SQN is therefore not well-suited as a preprocessing routine for transcript mapping (see also Figures 5 and 6). However, as SQN does not directly attempt to reduce transcript variability, this comparison should been interpreted with caution.

4.2. *Exon Probe Identification*

In a simple approach to identify transcribed exonic regions we used a threshold model on the hybridization measurements. Probes with intensities above the threshold were classified as exonic and below the threshold as untranscribed or intronic. We compared the resulting classification of probes with the TAIR7 annotation.[14] For every threshold we calculated precision and recall, defined as the proportion of probes mapped to exons among all probes having intensities greater than the threshold and the proportion of probes with intensities greater than the threshold among all probes that are annotated as exonic, respectively. Thresholding was applied to raw intensity values as well as the normalized intensities from SQN, SVR and RR. The resulting precision-recall curves (PRCs) are displayed in Figure 5 A. We observed that the two transcript normalization methods, SVR and RR, consistently improved exon probe identification compared to raw intensities. For SQN the recognition deteriorated. However, when probes were sub-samples prior to thresholding and evaluation such that the set of exonic probes had the same GC-content the background set (as reported in Royce et al.[16]), the performance of SQN recovered, but was still below SVR and RR (cf. Figure 5 B). Note that the sub-sampling strategy changes the distributions and can not easily be applied to identify exon probes in the whole genome.

In a second experiment we only considered the transcribed regions of the genes in the test regions (exons and introns). We now allowed a threshold to be chosen separately for each gene. Note that this problem is much easier compared to a single global threshold. However, this approach cannot be directly applied when the transcript boundaries are not already known. For each gene we estimated the Receiver-Operator-Characteristic (ROC) curve separately and averaged them over all genes.[a] In Figure 6 we display the area under the averaged ROC curves for genes in different transcript intensity quantiles. As expected, exons could be more accurately identified in highly expressed transcripts. Again, we observed a superior performance of the transcript normalization techniques.

[a] We considered ROC curves instead of PRCs, since the class sizes varied among genes making PRCs incomparable.

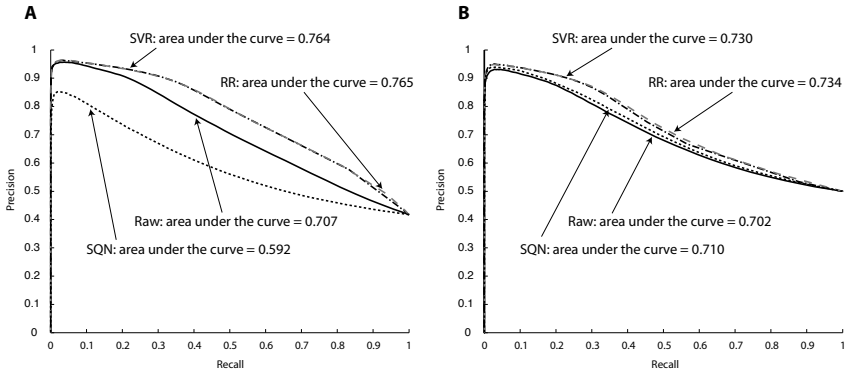


Figure 5. Separation in intensity between probes mapped to known exons and probes in regions annotated as untranscribed or intronic improved after normalization with SVR as well as after normalization with RR. **A** By varying the cutoff value, we calculated the precision-recall curve from all probes in the test regions. **B** Prior to thresholding and precision-recall estimation, probes were sub-sampled to obtain the same GC-content among exonic and intergenic / intronic probes.

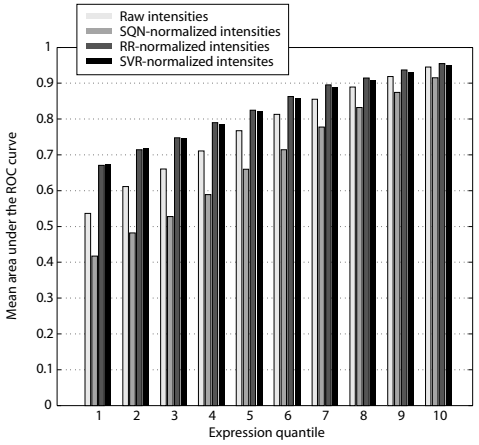


Figure 6. Separation in intensity between intron and exon probes broken down by expression quantiles and normalization methods. Expression values were calculated based on the median intensity of probes annotated as exonic. For each gene the area under the ROC curve (auROC) was obtained by local thresholding and for each expression quantile, auROC values were averaged over all genes in that quantile.

4.3. *Identification of Transcripts*

In a final experiment we show a proof of concept for our transcript identification algorithm. For this we considered genomic regions (from the test set described in Section 2) with known transcripts including 1 kbp of their flanking intergenic regions. We truncated intergenic regions at the boundaries of adjacent, known transcripts. For training, we took 100 randomly chosen regions, containing a single gene each, 500 such regions for model selection and 500 other regions for evaluation. We compared our method with the two simple thresholding approaches described in the previous section. In the first one we used a global

	Global threshold	Local threshold	HMSVMs
Raw intensities	70.4%	79.3%	77.1%
Sequence quantile normalization	65.5%	75.3%	70.9%
Support vector regression	73.5%	82.1%	82.9%
Ridge regression	73.9%	82.1%	82.5%

Figure 7. Accuracy of transcript identification in test regions with exactly one gene. Accuracy is defined as the sum of true positive and true negative exon probes over the total number of probes in a gene.

threshold which could be realistically applied for exon probe identification. In the second one an individual threshold was chosen for each gene to maximize classification accuracy. Note that this method has an advantage in the comparison because the threshold is determined based on expression levels of (unknown) test genes to be identified. Moreover, it cannot be straightforwardly applied to genome-wide detection of exon probes. As input we provided raw as well as normalized hybridization intensities discussed in Section 2 to our segmentation and the two thresholding methods. This resulted in a mapping of probes to exons, introns or intergenic regions. The accuracies of these predictions are summarized in Figure 7. In this comparison our segmentation method was considerably better than global thresholding, and even slightly better than the locally optimal threshold when trancript-normalized intensities were give as input. Moreover, we re-confirmed the findings of the previous section that transcript normalization significantly improved discrimination between exonic and untranscribed / intronic regions not only for thresholding on a per-probe basis, but in particular for a considerably more complex segmentation algorithm.

References

1. Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov Support Vector Machines. In *Proc. 20th Int. Conf. Mach. Learn.*, pages 3–10, 2003.
2. J. Bai and P. Perron. Computation and analysis of multiple structural change models. *J. Appl. Econom.*, 18:1–22, 2003.
3. B.M. Bolstad, R.A Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.
4. R.M. Clark, G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T.T. Hu, G. Fu, D. Hinds, H. Chen, K. Frazer, D. Huson, B. Schölkopf, M. Nordborg, G. Rätsch, J. Ecker, and D. Weigel. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317(5836), July 2007.
5. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, 7th edition, 1998.

6. J.S. Caroll et al. Chromosome-wide mapping of estrogen receptor binding. *Cell*, 122:33–43, 2005.
7. L. David et al. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA*, 103:5320–5325, 2006.
8. P. Bertone et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306:2242–2246, 2004.
9. B.J. Frey, Q.D. Morris, and T.R. Hughes. Genrate: A generative model that reveals novel transcripts in genome-tiling microarray data. *Journal of Computational Biology*, 13(2):200–214, 2006.
10. A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.
11. W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(6):1963–1970, 2006.
12. G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21:i369–i377, 2005.
13. G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R.J. Sommer, and B. Schölkopf. Improving the Caenorhabditis elegans genome annotation using machine learning. *PLoS Computational Biology*, 3(2):e20, 2007.
14. S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucl. Acids Res.*, 31(1):224–8, 2003.
15. T.E. Royce, J.S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics*, 21(8):466–475, 2005.
16. T.E. Royce, J.S. Rozowsky, and M.B. Gerstein. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23(8):988–997, 2007.
17. B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.
18. S. Sonnenburg, G. Rẗsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.
19. M. Suarez-Farinas, M. Pellegrino, K. Wittkowski, and M. Magnasco. Harshlight: a "corrective make-up" program for microarray chips. *BMC Bioinformatics*, 6(1):294, 2005.
20. B. Taskar, C. Guestrin, and D. Koller. Max margin markov networks. In *Advances in Neural Information Processing Systems 13*, 2003.
21. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.
22. Y. Zhan and D. Kulp. Model-P: A Basecalling Method for Resequencing Microarrays of Diploid Samples. *Bioinformatics*, 21(suppl_2):ii182–189, 2005.