BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 18 2009, pages 2341–2347
doi:10.1093/bioinformatics/btp395

Gene expression

Transcriptional landscape estimation from tiling array data using a model of signal shift and drift

Pierre Nicolas[1,*], Aurélie Leduc[1], Stéphane Robin[2], Simon Rasmussen[3], Hanne Jarmer[3] and Philippe Bessières[1]

[1]INRA, Mathématique Informatique et Génome UR1077, 78350 Jouy-en-Josas, [2]AgroParisTech/INRA, Mathématiques et Informatique Appliquées UMR518, 16 rue Claude Bernard, 75005 Paris, France and [3]Technical University of Denmark, Center for Biological Sequence analysis, Building 208, 2800 Lyngby, Denmark

Received on January 27, 2009; revised on May 11, 2009; accepted on June 19, 2009

Advance Access publication June 26, 2009

Associate Editor: Trey Ideker

ABSTRACT

Motivation: High-density oligonucleotide tiling array technology holds the promise of a better description of the complexity and the dynamics of transcriptional landscapes. In organisms such as bacteria and yeasts, transcription can be measured on a genome-wide scale with a resolution >25 bp. The statistical models currently used to handle these data remain however very simple, the most popular being the piecewise constant Gaussian model with a fixed number of breakpoints.

Results: This article describes a new methodology based on a hidden Markov model that embeds the segmentation of a continuous-valued signal in a probabilistic setting. For a computationally affordable cost, this framework (i) alleviates the difficulty of choosing a fixed number of breakpoints, and (ii) permits retrieving more information than a unique segmentation by giving access to the whole probability distribution of the transcription profile. Importantly, the model is also enriched and accounts for subtle effects such as signal 'drift' and covariates. Relevance of this framework is demonstrated on a *Bacillus subtilis* dataset.

Availability: A software is distributed under the GPL.

Contact: pierre.nicolas@jouy.inra.fr

Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 INTRODUCTION

High-density oligonucleotide tiling arrays carry tightly spaced probes that provide uniform covering of the genomic sequence. By hybridization with RNA samples (cDNA), they have been used to query the transcriptional activity of the whole genome in an array of model organisms (Bertone *et al.*, 2004; Biemar *et al.*, 2006; He *et al.*, 2007; Stolc *et al.*, 2005). The approach is particularly attractive for organisms with small-sized genome such as bacteria and yeasts where a resolution >25 bp is more easily achieved (David *et al.*, 2006; S.Rasmussen *et al.*, submitted for publication). The generalization of the use of such arrays should provide unbiased and high-quality pictures of the complexity and the dynamics of transcriptional landscapes (Xu *et al.*, 2009). The great promise of these data justifies the improvement of the currently available statistical methods dedicated to their analysis.

From the methodological standpoint, the problem is naturally stated in terms of finding segments where the hybridization signal is relatively constant, delimited by breakpoints that are expected to correspond to biological features such as transcript start and stop sites or splicing sites. A variety of tools including local non-parametric smoothing (Royce *et al.*, 2007; Wang *et al.*, 2009) and iterative hypothesis testing (Olshen *et al.*, 2004) have been proposed to answer this question. Probably the most popular and best mathematically grounded methodology consists of seeking the piecewise constant model with Gaussian noise that best fits the signal (Huber *et al.*, 2006; Picard *et al.*, 2005). Namely, for a fixed number of segments S, fitting the model consists of finding the combination of breakpoints $1 < t_1 \leq \cdots \leq t_{S-1} \leq n$ that minimizes the sum of squared residuals:

$$G(t_1, \ldots, t_{S-1}) = \sum_{s=1}^{S} \sum_{k=t_{s-1}}^{t_s - 1} (x_k - \bar{x}_s)^2, \qquad (1)$$

where x_k is the signal at position k, $\bar{x}_s$ is the average signal level in segment s (i.e. between t_{s-1} and $t_s - 1$), $t_0 = 1$ and $t_S = n + 1$. In full generality, minimizing the sum of squared residuals in Equation (1) can be achieved by Dynamic Programming and requires time $O(n^2 S)$. Huber *et al.* (2006) fixed an upper bound l on the maximum length of each segment to reduce the time complexity to $O(nlS)$ with $l < n$. The problem of choosing the correct number of segments S was more specifically examined by Picard *et al.* (2005), but visual assessment and use of prior belief have also been advocated (Huber *et al.*, 2006) and have been useful in practice (David *et al.*, 2006, S.Rasmussen *et al.*, submitted for publication).

The simplicity of this approach is appealing but hinders a number of difficulties, the most important being the choice of the number of segments. In principle, this issue can be tackled by embedding the segmentation model in a probabilistic setting that includes not only the noise but also the evolution of the signal. This idea stimulated the development of hidden Markov models (HMMs) (Fridlyand *et al.*, 2004; Marioni *et al.*, 2006; Stjernqvist *et al.*, 2007) for the analysis of comparative genomic hybridization data. For transcriptomic data, a different approach consists of training HMMs to distinguish between transcribed and non-transcribed regions (Du *et al.*, 2006; Munch

*To whom correspondence should be addressed.

et al., 2006). When the quality of the data is good enough it is both more natural and more ambitious to try to recover the 'denoised' transcription signal instead of directly summarizing the data via a classification algorithm. Transcript level is, however, a continuous quantity and none of the available models is satisfactory for a continuous-valued underlying signal. An HMM that achieves this aim at a computationally affordable cost is described in the present article. The proposed model does also extend the piecewise constant model in two directions. First, it integrates the influence of covariates that serve to account for differential affinity between probes. This allows to achieve segmentation and within-array normalization in one step. Second, the proposed model relaxes the assumption of strictly constant transcript levels between abrupt 'shifts' by also allowing progressive 'drift' of the signal. Inference based on this model is examined and discussed.

2 METHODS

2.1 Experimental data

The main example dataset used here comes from pilot experiments conducted on *Bacillus subtilis* within the European Consortium BaSysBio (S.Rasmussen *et al.*, submitted for publication). This array consists of 383 149 probes starting every 22 nt on each strand of the *B.subtilis* genome (GenBank: AL009126). Probe lengths range between 45 nt and 65 nt and were adjusted to reduce melting temperature (TM) variations (isothermal design). Production of the tiling arrays, synthesis of labeled cDNA from the RNA samples with random priming, hybridization and signal acquisition were carried out by Nimblegen. Antisense artifacts were controlled by using actinomycin D during reverse transcription (Perocchi *et al.*, 2007). RNA was extracted from *B.subtilis* culture during exponential growth on rich medium. One out of four biological replicates gave a high-quality signal and is analyzed here (S.Rasmussen *et al.*, submitted for publication). For comparison with the algorithm of Huber *et al*. (2006), we also analyzed a dataset corresponding to the chromosome 1 of the yeast *Saccharomyces cerevisiae* (David *et al.*, 2006). This second array was produced by Affymetrix and uses shorter oligonucleotide (25 nt) tiled at intervals of 8 nt on each strand. The data from the three biological replicates were averaged after quantile normalization.

Both experimental settings included hybridization of genomic DNA (gDNA) preparations to assess variation of affinity between probes (four replicates for *B.subtilis* and three replicates for *S.cerevisiae*). Data were averaged across replicates after quantile normalization. *Bacillus subtilis* gDNA data varied smoothly between the replication origin and the replication terminus, presumably reflecting the chromosome dosage. Taking the residuals after median smoothing (window size 110 011 bp) removed this trend. For *S.cerevisiae* data, we preferred to compute the residuals as the distance to the mode rather than to the median to account for the highly skewed distribution of probe affinities. The formatted datasets are distributed with the software.

2.2 Shift and drift in an HMM framework

Like in previous approaches (Huber *et al.*, 2006; Olshen *et al.*, 2004; Picard *et al.*, 2005), the $\log_2$ of the observed intensity x_t is modeled as the sum of an unobservable signal u_t that is the focus of interest plus a Gaussian noise with SD σ. This general model can be written as:

$$x_t \,|\, u_t \sim \mathcal{N}(u_t, \sigma^2). \tag{2}$$

However, u_t is not seen in our model as a parameter but is itself a random variable. Correlation between probes that are adjacent on the chromosome is accounted for by a Markov transition kernel $\pi(u_t, u_{t+1})$ and $(x_t, u_t)_{1\le t\le n}$ is thus said to be an HMM (Durbin *et al.*, 1998; Rabiner, 1989). Compared with traditional use of HMMs, the complication comes from the continuous nature of u_t, whereas the efficient algorithmic machinery of the HMMs (Viterbi algorithm, forward–backward algorithm, expectation-maximization (EM) algorithm) works well for discrete and typically small number of hidden states (Rabiner, 1989). In general, with K hidden states, the time complexity of the algorithms is $O(nK^2)$.

Here, we propose a structure of the transition matrix $\pi(u_t, u_{t+1})$ accounting for abrupt shifts and progressive drifts in the unobservable signal u_t that allows to discretize the continuous range $U_{\min} \le u_t \le U_{\max}$ in K points spaced by a regular interval, $h = (U_{\max} - U_{\min})/(K-1)$. This particular structure warrants time complexity $O(nK)$ for the classical HMM algorithms and thus permits appropriately high resolution of discretization.

For values of u_t and u_{t+1} taken in the discretized hidden state space, the transition probability writes

$$\begin{aligned}
&\pi(u_t, u_{t+1}) = \\
&\quad \alpha_n \mathbb{I}_{\{u_{t+1}=u_t\}} + \alpha_s \eta_h(u_{t+1}) \\
&\quad + \alpha_u \mathbb{I}_{\{u_{t+1}>u_t\}} \lambda_u^{\frac{u_{t+1}-u_t}{h}-1} (1-\lambda_u)^{\mathbb{I}\{u_{t+1}\neq U_{\max}\}} \\
&\quad + \alpha_d \mathbb{I}_{\{u_{t+1}<u_t\}} \lambda_d^{\frac{u_t-u_{t+1}}{h}-1} (1-\lambda_d)^{\mathbb{I}\{u_{t+1}\neq U_{\min}\}} \\
&\quad + \alpha_u \mathbb{I}\{u_{t+1}=u_t=U_{\max}\} + \alpha_d \mathbb{I}\{u_{t+1}=u_t=U_{\min}\},
\end{aligned} \tag{3}$$

where the parameters verify $0 \le \alpha_n, \alpha_s, \alpha_u, \alpha_d \le 1$, $\alpha_n + \alpha_s + \alpha_u + \alpha_d = 1$ and $0 \le \lambda_u, \lambda_d < 1$, with $\mathbb{I}_{\{X\}}$ standing for 1 if X is true, 0 otherwise.

This transition kernel is best understood as a mixture of four types of moves with weights α_n, α_s, α_u and α_d. The parameter α_n accounts for unchanged u between successive probes. Shift moves have probability α_s and the distribution of the signal after the move is independent of the value of the signal before the move. This distribution is given by η_h and it approximates the marginal distribution of the signal. Namely, $\eta_h(u_{t+1}) = \int_{u_{t+1}-h/2}^{u_{t+1}+h/2} \eta(u)\mathrm{d}u$, where η is the kernel density estimate computed on x with a Gaussian kernel and Scott's bandwidth (Scott, 1992). The possibility of small drift, either upward or downward, is accounted for by α_u and α_d. Drift amplitudes are modeled by two geometric distributions of parameters λ_u and λ_d and average amplitudes write $h + h/(1-\lambda)$.

It can be verified that as $h \to 0$ and $h/(1-\lambda) \to \gamma$ the transition kernel of the discrete-valued Markov chain of Equation (3) converges in distribution toward the transition kernel of a continuous-valued Markov chain. In its continuous version, the kernel writes as a mixture of a point mass at u_t of weight α_n, a continuous-valued distribution of density η and weight α_s, and two shifted exponential distributions of rates γ_u and γ_d and weights α_u and α_d. With an appropriately high K it should thus be possible to approach, using the discrete-valued model of Equation (3), the results that one would obtain with the continuous-valued model.

The Supplementary Material available online gives a detailed presentation of the equations that allow $O(nK)$ implementations of the HMM classical algorithms, namely:

(1) likelihood computation ($\mathbb{P}(x_{1\ldots n})$),

(2) forward–backward algorithm (computation of $\mathbb{P}(u_t | x_{1\ldots n})$ for each t),

(3) Viterbi algorithm (finding the trajectory $u_{1\ldots n}$ that maximizes $\mathbb{P}(u_{1\ldots n} | x_{1\ldots n})$).

These algorithms are implemented in our software. All the parameters are estimated in the maximum likelihood (ML) framework with the EM algorithm, an iterative algorithm that alternates an E-step (forward–backward algorithm) and a M-step (parameter update). The output provides a detailed report on the 'denoised' signal based on the results of the Viterbi and forward–backward algorithms.

2.3 gDNA signal as a covariate

gDNA hybridization data were used in a preprocessing step by Huber *et al.* (2006) for the purpose of between-probe signal normalization and outlier

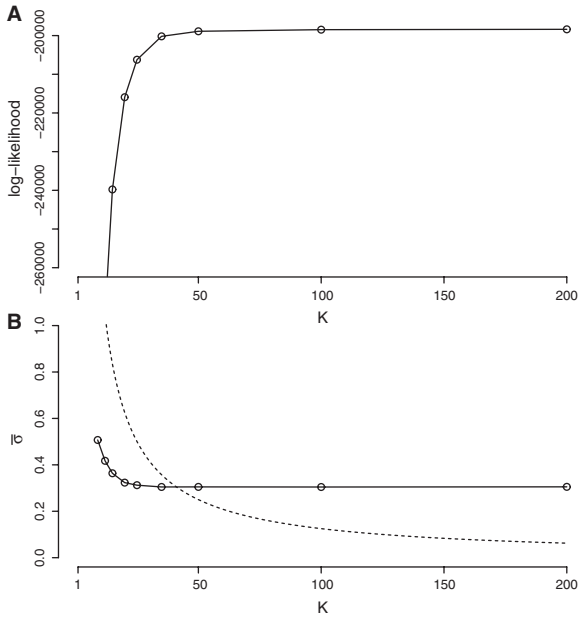

Fig. 1. Influence of the number of hidden states. (**A**) Log-likelihood (in natural log) as a function of the number of hidden states, K. (**B**) Estimated average variance of the noise $\bar{\sigma}$ as a function of K (plain line). The discretization step $h \propto 1/(K-1)$ is also shown (dotted line).

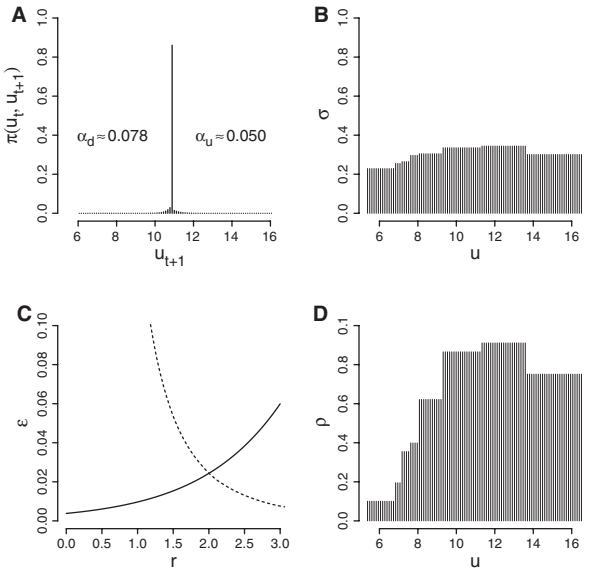

Fig. 2. Parameter estimates. (**A**) Transition matrix $\pi(u_t, u_{t+1})$. One row is represented. (**B**) SD of the noise σ as a function of the underlying signal level u_t. (**C**) Outlier probability ϵ as a function of the magnitude of the gDNA residuals r_t (plain line) and complementary cumulative distribution function of the gDNA residuals (dotted line). (**D**) Proportionality factor ρ applied to r_t as a function of the signal level u_t

trimming. The model proposed here accounts for these effects by modeling the gDNA hybridization intensities as a covariate.

The probability distribution for the observed variable x_t given the underlying signal u_t and the gDNA residuals r_t writes as a mixture model

$$x_t \,|\, u_t, r_t \sim (1-\epsilon(r_t))\mathcal{N}(u_t+\rho(u_t)r_t, \sigma(u_t)^2) + \epsilon(r_t)\mathcal{U}(U_{\min}, U_{\max}), \quad (4)$$

where $\epsilon(r_t)$ corresponds to the probability of outliers, $\mathcal{U}(U_{\min}, U_{\max})$ is the uniform distribution that models outlier data and $\mathcal{N}(u_t+\rho(u_t)r_t, \sigma(u_t)^2)$ is the Gaussian distribution modeling non-outlier data. This model is markedly richer than Equation (2). Notice (i) the non-constant proportionality factor $\rho(u_t)$ applied to r_t; (ii) the non-constant standard error $\sigma(u_t)$ of the Gaussian distribution; and (iii) the probability of outliers ϵ that depends on r_t. More precisely, ρ and σ are modeled as piecewise constant function of u_t with eight intervals, and ϵ is a two-parameter logistic function of the absolute value of r_t, $\epsilon(r_t) = 1/(1+e^{-(a+b|r_t|)})$. All the parameters are simultaneously estimated with the EM algorithm (see Supplementary Material).

Finally, left and right censoring are incorporated in the model to account for the experimental limitations that preclude exact measurements of extremely high and extremely low intensities. In practice, the lower and upper 5% of the original range of variation of the intensity x are considered as censored.

3 RESULTS AND DISCUSSION

3.1 Selecting the appropriate level of discretization

The model was designed with the explicit aim of modeling a continuous-valued underlying signal. In other words, discretization of the hidden state space is seen only as a necessary technicality and the step $h \propto 1/K$ should ideally be sufficiently small to have no impact on the results. Intuitively, the smaller the SD of the noise σ, the smaller the step h should be. The results obtained on the *B.subtilis* dataset and presented in Figure 1 confirm this intuition and thereby provide some form of validation for the model.

Figure 1 shows that increasing K (and thus decreasing h) actually increases the model adequation to the data as measured by the log-likelihood after ML estimation. Beyond a certain value of K the impact of this change becomes, however, almost unnoticeable. Figure 1 also reports the parallel evolution of h and σ. According to this plot, having h around 0.5σ seems more than sufficient. Indeed, with such a value of h, the 95% confidence interval (CI) of the distribution of the noise is about eight times as large as the discretization interval h. K was set to 100 for this particular dataset.

This choice of $K = 100$ corresponds to an acceptable running time for the algorithm. Our setting throughout this study consisted to explore 10 random starting points for the EM algorithm. Here, it resulted in a total of 885 iterations taking 5 h 6 min on an Intel(R) Xeon(TM) CPU 3.40 GHz CPU, less than the 5 h 36 min needed for the segmentation algorithm of Huber *et al.* (2006) with maximum segment length $l = 1000$ (22 000 bp) and segment number on each strand $S = 1500$.

3.2 Importance of modeling drift and covariates

Parameter estimates in model-based analyses are an invaluable source of information to understand both the behavior of the model and the data. The model contains a total of 23 parameters. Figure 2 is intended to provide an overview of their ML estimates on the *B.subtilis* data. The first row of Table 1 gives numerical values for a selection of parameters.

The shape of the transition matrix that describes the trajectory of the underlying signal is defined by the parameters in Equation (3), one row of this matrix is shown in Figure 2A. The sharp peak reflects the high value of α_n: it is estimated that the underlying signal remains

Table 1. Model comparison

Model	d[a]	ϵ[b]	α_s	α_u	α_d	$\bar{\rho}$	$\bar{\sigma}$	CV-LL[c]
$\mathcal{M}1$[d]	23	y	0.012	0.050	0.078	0.66	0.30	-1.986×10^5
$\mathcal{M}2$	21	n[e]	0.016	0.046	0.074	0.66	0.31	-2.022×10^5
$\mathcal{M}3$	21	y	0.029	0	0.039	0.64	0.33	-2.056×10^5
$\mathcal{M}4$	21	y	0.040	0.014	0	0.64	0.34	-2.109×10^5
$\mathcal{M}5$	19	y	0.046	0	0	0.63	0.34	-2.124×10^5
$\mathcal{M}6$	15	y	0.012	0.156	0.197	1	0.31	-2.775×10^5
$\mathcal{M}7$	15	y	0.014	0.036	0.053	0	0.46	-2.921×10^5
$\mathcal{M}8$	9	n	0.046	0	0	0	0.50	-3.021×10^5

[a]Model dimension as measured by the number of free parameters.
[b]'y' if the model accounts for outliers, 'n' otherwise.
[c]CV-LL: cross-validated log-likelihood.
[d]Full model.
[e]The parameter constraints that characterize each model are underlined.

unchanged between adjacent probes in >85% of the cases (α_s in Table 1). The narrow shoulders on both sides of the peak correspond to the upward and downward drift moves and reflect the value of the parameters (α_u, λ_u) and (α_d, λ_d), respectively. Close inspection reveals a small asymmetry, with upward moves being less frequent than downward moves (5.0% versus 7.8%). The small estimated proportion of abrupt shift moves between adjacent probes is almost invisible at this scale (1.2%).

As expected, the probability of outliers is estimated to increase with the magnitude of the residuals of the gDNA signal. The two-parameter logistic curve that models this relationship is shown in Figure 2C. Remarkably, the probability of outliers is found to be overall very small.

The parameters σ and ρ that model the observed intensity x_t are modeled as eight-parameter piecewise constant functions of the underlying signal level u_t. Figures 2B and D show these two functions. Whereas the SD of the noise σ is a relatively flat function of u_t, the parameter ρ that serves to account for the gDNA covariate varies by more than a factor of eight. An obvious characteristic of the latter is its sharp decrease for low values of the signal. This behavior probably reflects higher level of non-specific signal in the lower end of the intensity spectrum. It is also re-insuring to observe that the value of ρ in the middle of the spectrum is just slightly below unity, the value that we expect in an idealized situation [see the rationale behind the preprocessing step in Huber *et al.* (2006)].

As a whole, these results emphasize the importance of two specificities of our model: the modeling of drift moves as a complement to shift moves and the non-constant ρ that provides a simple adaptive method to account for the variation of affinity between probes.

To better understand the behavior of the model and the characteristics of the data, we carried out a comparative analysis of eight models. For the purpose of robust assessment of model fitness with respect to the *B.subtilis* dataset each model was fitted two times, once on each strand of the chromosome, and the likelihood was each time computed on the other strand. The sum of both log-likelihood terms is reported as the cross-validated log-likelihood in Table 1. Parameter values in Table 1 were estimated on the full dataset.

Sorted by decreasing value of adequacy with the data, the models ranged from $\mathcal{M}1$, the full 23-parameter model, to $\mathcal{M}8$, a nine-parameter model that does not account for drifts, outliers nor covariates. Not accounting for outliers has only a small impact on the overall model fitness ($\mathcal{M}2$ versus $\mathcal{M}1$), but the probability of shift moves is increased by >30% in this simpler model. This can have a non-negligible impact in practice given that these particular shift moves are indeed likely to be spurious. Not modeling drifts has a much more pronounced impact ($\mathcal{M}5$ versus $\mathcal{M}1$). Fitness is 6.5% better for $\mathcal{M}1$ than for $\mathcal{M}5$ and the estimated proportion of shift moves is about four times lower in $\mathcal{M}1$ (1.2% versus 4.6%), suggesting that a substantial fraction of the drift moves in $\mathcal{M}1$ are interpreted as shift moves in $\mathcal{M}5$. A closer examination underscores the importance of downward drift as compared with upward drift. Not accounting for downward drift has 74% more effect on the overall fitness that not accounting for upward drift ($\mathcal{M}3$ and $\mathcal{M}4$ versus $\mathcal{M}1$). More spectacularly, if a single drift direction is allowed, modeling downward drift improves the model ~4.5 times more than modeling only upward drift ($\mathcal{M}3$ versus $\mathcal{M}5$ and $\mathcal{M}4$ versus $\mathcal{M}5$). Setting ρ to either 1 or 0 were both found to result in a dramatic drop in fitness but with different specific effects. Setting ρ to 1 in $\mathcal{M}6$ results in estimation of high drift compared with original model, whereas setting ρ to 0 in $\mathcal{M}7$ results in estimation of high noise.

3.3 Estimation of transcriptional landscape: illustration on *B.subtilis* data

The ultimate goal of the use of the model is to infer the underlying signal supposed to reflect the actual transcriptional landscape.

The adoption of a probabilistic setting for the trajectory of the underlying signal allows for a considerably richer signal reconstruction than just 'optimal' trajectory reconstruction. Figure 3 gives an illustration of these possibilities by superimposing a number of results obtained with the model on a 10 000 bp region of the *B.subtilis* chromosome. Results include: (i) the prediction interval for the value of the signal u_t at each chromosome position; (ii) a point prediction for the signal value by the conditional mean of u_t (the best predictor in terms of quadratic error); (iii) the inferred position of the experimental point after correction for differential probe affinity [computed as $x_t - \hat{\rho}(\hat{u}_t)r_t$]; (iv) the exact position of each type of move in the best trajectory given by the Viterbi path (abrupt shift, upward drift and downward drift); and (v) the probability of having each type of move at each position. All these values can be read directly from the output of our software.

The biological pertinence of the distinction between shifts and drifts seems remarkable in Figure 3. Inferred shifts are found mostly in intergenic regions that a priori correspond to possible positions for transcriptional promoters and terminators.

The position of each move (2893 shifts and 13 460 drifts) was compared with sequence predictions for two biological features: Rho-independent (intrinsic) terminators predicted with the algorithm of d'Aubenton-Carafa *et al.* (1990); promoters dependent on Sigma-A predicted using an HMM whose structure was chosen according to the results of Nicolas *et al.* (2006). To fulfill the needs of an unbiased analysis, both categories of predictions were made without prior on the position of the genes and confidence cutoffs were set relatively low to increase sensitivity (a total 4164 Sigma-A predictions and 3492 terminator predictions are considered).

The results presented in Figure 4 confirm the practical relevance of the distinction between shift and drift moves. For upward moves, it shows the difference between shift and drift with respect to the distance between the breakpoint and the nearest promoter prediction.

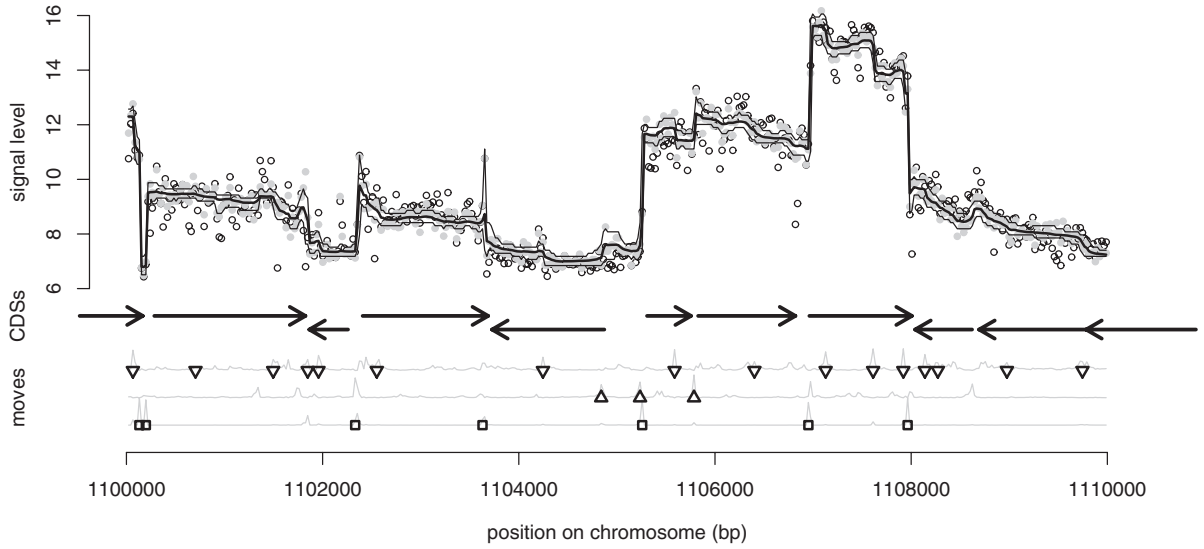

Fig. 3. Transcriptional landscape inference. Analysis of the signal on the (+)-strand of a 10 000 bp segment of the *B.subtilis* chromosome. Upper part: open circles show the original signal. Closed gray circles represent the signal after 'correction' with the gDNA covariate. The thick black line shows the expectation of the transcript level as computed with the HMM. Thin black lines correspond to the 95% CI. Middle part: horizontal arrows indicate GenBank CDSs. Lower part: shift moves along the most likely trajectory are shown as squares. Upward and downward drift moves are indicated by point-up and point-down triangles, respectively. Move probabilities are represented as gray lines.

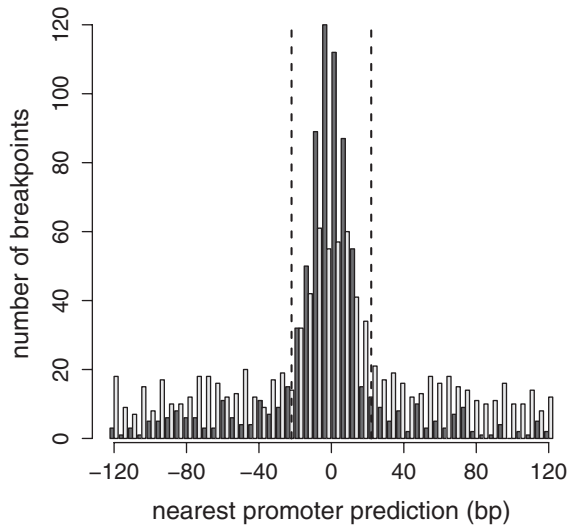

Fig. 4. Distance between breakpoints associated with upward signal changes and promoters predicted from the sequence alone. Black bars represent shift moves, light gray bars correspond to drift moves. Vertical dashed lines show the 22 bp cutoff. A negative value indicates that the promoter is upstream of the breakpoint.

Similar results for downward moves and terminator predictions are presented in the Supplementary Material (Fig. S1). Although shifts represent only 18% of all moves, a clear majority of the moves lying at <22 bp of a predicted biological feature are shifts. The proportion of shifts is 59% among the 977 upward moves near a predicted promoter, and 71% among the 1157 moves near a predicted terminator.

Drift might partly reflect local variations of labeled cDNA that result from technical artifacts such as random priming bias. Drift could also reflect biological differences in the amount of mRNA. In particular, Figures 4 and S1 leave no doubt that a fraction of the drifts correspond to promoters or terminators whose activity is too weak to be detected as shifts in this biological condition. A preliminary exploration of the patterns of drift is reported in the Supplementary Material. Figure S2 shows that downward drift is most pronounced after upward shifts and before downward shifts, near the $5'$ and $3'$ ends of transcriptionally active regions. An excess of upward drift is found before upward shifts, at the $3'$ end of regions with low-transcriptional activity. Random priming artifacts could most easily be invoked to explain downward drift at the $3'$ end of transcriptionally active regions (Xu *et al.*, 2009). Downward drift may also, for instance, be partly caused by molecules whose synthesis is still incomplete. Here, no single explanation could apparently account the patterns of upward and downward drift. Instead, drift is observed in a variety of chromosomal and transcriptional contexts that the landscape snapshots presented in Figures S3, S4 and S5 intend to illustrate. As an example, some spectacular cases of downward drift are found for transcription units apparently lacking a clear terminator. The intensity of the resulting downstream antisense transcription drifts downward progressively. In Figure S3, a pattern reminiscent of the bidirectional transcriptional activity recently described in *S.cerevisiae* (Xu *et al.*, 2009) can also be observed.

3.4 Benchmark comparisons

In addition to allow insightful reconstructions of the transcriptional landscape, good algorithms should identify breakpoints that match, as closely as possible, the position of the promoters and terminators. To compare different sets of breakpoints, promoter and terminator

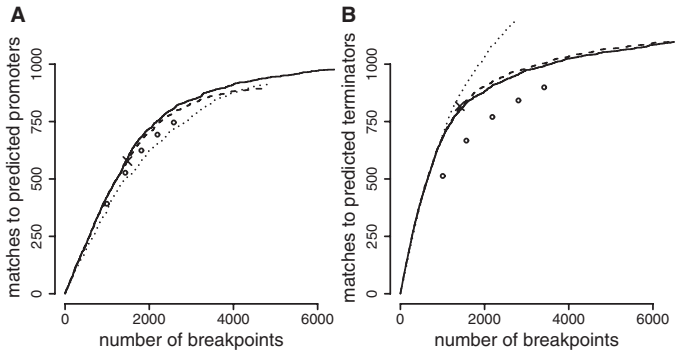

Fig. 5. Benchmark comparisons. The number of breakpoints matching promoter (**A**) and terminator (**B**) predictions (using a 22 bp distance cutoff) is reported as the number of breakpoints considered increases. Plain, dashed and dotted lines show the results obtained with the new HMM method, respectively, with the full model ($\mathcal{M}1$), the model without drift ($\mathcal{M}5$) and the model without drift, covariate and outliers ($\mathcal{M}8$). Open circles report the result of the segmentation by piecewise constant regression with the number of segment on each strand $S=(1000, 1500, 2000, 2500, 3000)$. The number of breakpoints detected by the HMMs were varied after ranking the moves according to the amplitude of the signal change. Crosses indicate the results for shift moves in $\mathcal{M}1$.

predictions were used as a proxy for the true (unknown) reference. Results are shown in Figure 5.

The results obtained with the HMMs, $\mathcal{M}1$, $\mathcal{M}5$ and $\mathcal{M}8$, give another confirmation of the biological pertinence of the distinction between shift moves and drift moves in $\mathcal{M}1$. It also revealed the deep impact of the correction for variation of affinity between probes using covariates, not implemented in $\mathcal{M}8$. The misbehavior of $\mathcal{M}8$ translates paradoxically in an apparent success at detecting terminators. This most likely does not reflect the transcription signal itself, but rather the low probe affinity due to the stem–loop secondary structure distinctive of the rho-independent terminators.

For the comparison of the new HMM segmentation method and the piecewise constant regression implemented in the algorithm of Huber *et al.* (2006), the later was run on the data after correction for difference of affinity between probes (as shown in Fig. 3) with maximum segment length $l=22000$ bp and number of segments on each strand S between 1000 and 3000. Results clearly demonstrate the benefit of the new HMM framework. For $S=1500$, the number of breakpoints matching promoter and terminator predictions were, respectively, 8.9% and 25% higher for the HMM.

3.5 Results on *S.cerevisiae* data

Examination of the segmentation produced by piecewise constant regression on Watson (+)-strand of *S.cerevisiae* yeast chromosome 1 leads to the choice of 152 (average segment size 1500 bp) as a sensible number of breakpoints (Huber *et al.*, 2006). A question was thus whether the automatic procedure presented here will identify a similar number of shift moves. The model was fitted on the mRNA and gDNA data of the 57 616 probes representing both strands of the chromosome 1.

The Viterbi path of our HMM on the (+)-strand contained 125 shift moves and 373 drift moves with a median distance of 60 bp between each of the 152 breakpoints of Huber *et al.* (2006) and the closest of the 125 shift moves. On this dataset, modeling drift can thus be useful to single out the most abrupt changes in the signal intensity.

Interestingly, further comparisons of models with and without drift indicated that drift improve the model fitness by only 1.4% on the *S.cerevisiae* data, much less than the 6.5% found on *B.subtilis* data. Biology and array technology are two sources of possible differences between *S.cerevisiae* and *B.subtilis* datasets. Our model of drift seems more relevant for prokaryotic data obtained using long isothermal probes.

4 CONCLUSIONS

This article describes a new methodology based on an HMM that embeds the segmentation of a continuous-valued signal in a probabilistic setting. For a computationally affordable cost, this framework alleviates the difficulty of choosing a fixed number of breakpoints and permits retrieving more information than a unique segmentation. Probabilistic modeling makes it straightforward to compute confidence measures on the estimated transcriptional landscape. This information should prove particularly useful to pinpoint the differences in large collections of arrays. Extension of the model could also be imagined to tackle the problem of the joint segmentation of datasets where transcript boundaries and expression level differ.

By accounting for gDNA hybridization data as a covariate, the model automatically corrects the data for the variation of affinity between probes. David *et al.* (2006) proposed for this purpose a preprocessing step to be carried out on the raw data, before log-transformation, and producing a significant fraction of negative values. The data could thus no longer be simply log-transformed and more complicated variance stabilization transformation, requiring multiple arrays, was used (Huber *et al.*, 2002). In comparison, the normalization carried out by the model needs only one array and it alters only minimally the overall distribution of the log of the original data.

The model is also enriched and accounts for subtle effects such as signal 'drift' and covariates. Interestingly, our results unambiguously document the existence of a drifts in the *B.subtilis* dataset. The interest of this observation is 2-fold. First, drift have not been accounted in the previous models and this may partially explain why selecting the number of breakpoints on real dataset proved so difficult (Huber *et al.*, 2006; Picard *et al.*, 2005). Second, the causes and the patterns of drift deserve to be investigated if we want to make the best use of tiling array expression data.

The software is distributed under the GNU Public License http://genome.jouy.inra.fr/∼pnicolas/hmmtiling/.

ACKNOWLEDGEMENTS

We thank Etienne Dervyn, Philippe Noirot and Franck Picard for constructive comments on the content of the manuscript.

Funding: BaSysBio project, European Commission research grant (LSHG-CT2006-037469).

Conflict of Interest: none declared.

REFERENCES

d'Aubenton Carafa,Y. *et al.* (1990) Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.*, **216**, 835–858.

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.

Biemar,F. *et al.* (2006) Comprehensive identification of *Drosophila* dorsal-ventral patterning genes using a whole-genome tiling array. *Proc. Natl. Acad. Sci. USA*, **103**, 12763–12768.

David,L. *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA*, **103**, 5320–5325.

Du,J. *et al.* (2006) A supervised hidden Markov model framework for efficiently segmenting tiling array data in transcriptional and ChIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, **22**, 3016–3024

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.

Fridlyand,J. *et al.* (2004) Hidden Markov model analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

He,H. *et al.* (2007) Mapping the *C. elegans* noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.*, **17**, 1471–1477

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), 96–104.

Huber,W. *et al.* (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, **e22**, 1963–1970.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Munch,K. *et al.* (2006) A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, **7**, e239.

Nicolas,P. *et al.* (2006) A reversible jump Markov chain Monte Carlo algorithm for bacterial promoter motifs discovery. *J. Comput. Biol.*, **13**, 651–667.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Perocchi,F. *et al.* (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.*, **35**, e128.

Picard,F. *et al.* (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, e27.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Royce,T.E. *et al.* (2007) An efficient pseudomedian filter for tiling microrrays. *BMC Bioinformatics*, **8**, e186.

Scott,D.W. (1992) *Multivariate Density Estimation. Theory, Practice and Visualization*. Wiley, New York.

Stolc,V. *et al.* (2005) Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *Proc. Natl Acad. Sci. USA*, **102**, 4453–4458.

Stjernqvist,S. *et al.* (2007) Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, **23**, 1006–1014.

Wang,L.-Y. *et al.* (2009) MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res.*, **19**, 106–117.

Xu,Z. *et al.* (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.