Statistical modelling of bacterial promoter sequences for regulatory motif discovery with the help of transcriptome data: application to Listeria monocytogenes

Ibrahim Sultan, Vincent Fromion, Sophie Schbath, Pierre Nicolas

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Statistical modelling of bacterial promoter sequences for regulatory motif discovery with the help of transcriptome data: application to *Listeria monocytogenes*

Ibrahim Sultan, Vincent Fromion, Sophie Schbath and Pierre Nicolas

Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

IS, 0000-0001-6084-529X; VF, 0000-0002-9194-5426; SS, 0000-0003-3574-8222; PN, 0000-0002-1133-0609

Automatic de novo identification of the main regulons of a bacterium from genome and transcriptome data remains a challenge. To address this task, we propose a statistical model that can use information on exact positions of the transcription start sites and condition-dependent expression profiles. The central idea of this model is to improve the probabilistic representation of the promoter DNA sequences by incorporating covariates summarizing expression profiles (e.g. coordinates in projection spaces or hierarchical clustering trees). A dedicated trans-dimensional Markov chain Monte Carlo algorithm adjusts the width and palindromic properties of the corresponding position-weight matrices, the number of parameters to describe exact position relative to the transcription start site, and chooses the expression covariates relevant for each motif. All parameters are estimated simultaneously, for many motifs and many expression covariates. The method is applied to a dataset of transcription start sites and expression profiles available for *Listeria monocytogenes*. The results validate the approach and provide a new global view of the transcription regulatory network of this important pathogen. Remarkably, a previously unreported motif is found in promoter regions of ribosomal protein genes, suggesting a role in the regulation of growth.

1. Introduction

Automatic de novo identification of the main regulons of an organism as 'simple' as a bacterium remains a challenge, despite motif discovery in DNA sequences being an old problem of bioinformatics for which many approaches have been developed [1,2]. The representation chosen for the motifs is at the heart of the methodology. Word-based representations usually consist of consensus strings written in an alphabet allowing degenerate symbols. Scoring then relies on a hypothesis testing framework to detect deviation from a null hypothesis, such as that of Markov sequence [3] or of equal occurrence frequencies between co-expression clusters [4,5]. Position-weight matrices (PWMs) allow more precise representations, typically accounting for frequencies of the four DNA nucleotides (A,C,G,T) at each position within the motif. Beside this probabilistic representation of nucleotide composition within motif occurrences, a full probabilistic model involves also a model for the background sequence (outside motif occurrences) and a model for the positions of motif occurrences in the sequence set. Motif discovery is then cast as the problem of estimating PWMs. The first algorithms implementing these ideas [6,7] remain among the most powerful and widely used tools to search for motifs based only on nucleotide composition properties. Due to motif degeneracy and limited number of occurrences, these approaches are usually successful only when it is possible to define datasets enriched for particular motifs. This is most often done based on experimental data. Hence, transcriptional regulatory network reconstruction tends to be an incremental process in which new components of the network are added one by one.

Cite this article: Sultan I, Fromion V, Schbath S, Nicolas P. 2020 Statistical modelling of bacterial promoter sequences for regulatory motif discovery with the help of transcriptome data: application to *Listeria monocytogenes*. *J. R. Soc. Interface* **17**: 20200600. http://dx.doi.org/10.1098/rsif.2020.0600

Received: 26 July 2020
Accepted: 10 September 2020

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
bioinformatics, computational biology, systems biology

Keywords:
DNA motifs, transcriptional regulatory network, Markov chain Monte Carlo, transcriptomics, bacteria

Authors for correspondence:
Ibrahim Sultan
e-mail: i.ibrahim.sultan@gmail.com
Pierre Nicolas
e-mail: pierre.nicolas@inrae.fr

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5127043.

THE ROYAL SOCIETY
PUBLISHING

Chromatin immunoprecipitation (ChIP) is the experimental technique that had the deepest impact on the field of motif discovery [2]. Nevertheless, its need for *a priori* selection of combinations of transcription factors (TFs) and biological conditions is an intrinsic limitation for de novo motif discovery at system level. Furthermore, ChIP experiments require either specific antibodies to recognize TFs or genetic engineering of functionally active tagged TFs. In parallel to ChIP, microarrays and RNA-Seq have been extensively applied to compare gene expression between growth conditions and/or between specific mutants and results have been used early for bacterial transcriptional network reconstruction [8,9]. With the goal to explore globally the transcription regulatory network of a bacterium, expression profiles across many conditions can be collected without focusing on specific TFs and without genetic engineering [10,11]. Genome-wide maps of transcription start sites (TSSs) constitute another type of transcriptome data which is increasingly available for bacteria [12,13] and can focus the search on regulatory motifs.

The development of specific methods for de novo motif discovery using expression profiles has attracted less attention than the use of ChIP data. The task is also more difficult because the data are not directly connected to a specific TF–DNA interaction. Nevertheless, diverse approaches have been proposed based on different methodological concepts such as: mutual information in FIRE [14], enrichment test in GEMS [4], or regression of expression data y given sequence data x in REDUCE [15] and MatrixREDUCE [16]. The first two approaches involve transforming the expression data into one-dimensional categorical values, while the third approach can directly accommodate multidimensional continuous expression data. The algorithm implemented in RED2 [5] intends to bypass the need for clustering of the first two approaches by computing mutual information or applying enrichment tests on overlapping sets of genes that are close in the expression space (neighbourhoods). These approaches face the difficulty of finding appropriate summary or probabilistic models for expression data.

The viewpoint adopted in this work steps back from the causality relationship by which TF binding sites should explain expression profiles. Instead, it sees the expression data only as a potential source of information on where motifs occur. In probabilistic terms, our approach consists of modelling sequence data x given expression data y (i.e. $x|y$). This choice makes it possible to build directly upon the powerful sequence modelling approaches for de novo motif discovery based on PWMs and full probabilistic modelling of the sequences, establishing a continuum between discovery of motifs related and unrelated to the available expression data. Hence, it becomes possible to envision the simultaneous use of the expression data and of all the statistical properties of the sequence. In [10], we previously proposed an approach for the discovery of sigma factor binding sites based on modelling $x|y$. This model was tailored for sigma factor binding sites whose specificity is to delineate and partition the promoter space [17]. Regulation by sigma factors tends thus to correspond to a preponderant and non-overlapping level of regulation that is particularly well captured by the structure of a hierarchical clustering tree and did not justify modelling the occurrence of more than one motif per sequence. Incorporation of positional information proved helpful in the case of sigma factor binding sites whose positions are strongly constrained with respect to the TSS [10].

Prompted by results obtained on sigma factors [10,11], this work develops a coherent probabilistic model of the DNA sequences to address the task of automatic de novo identification of the main regulons (not restricted to sigma factors) of a bacterium from genome and transcriptome data. For this purpose, the proposed model introduces two main novelties: overlaps between motif occurrences are allowed and covariates summarizing expression profiles are incorporated into the probability of occurrence in a given promoter region. These covariates can correspond to positions of the genes on axes such as obtained by PCA [18] or ICA [19,20] but we also show how to use positions in hierarchical clustering trees [8,21]. All the parameters are estimated in a Bayesian framework using a dedicated trans-dimensional Markov chain Monte Carlo (MCMC) algorithm. In order to validate the approach, we applied it to the food-borne pathogen *Listeria monocytogenes* on which a wealth of transcriptomics data have been collected owing to its status of model organism for the study of host–pathogen interactions and bacterial transcriptomics [22]. Sources of transcriptome data for this bacterium include a landmark study using RNA-Seq and high-density tiling arrays [23], an early use of genome-wide TSS mapping [24], and a comprehensive work done to aggregate available transcriptome datasets in a single database [25].

2. Methods and data

2.1. Probabilistic model of promoter sequences

2.1.1. Model overview

We consider a dataset composed of $\mathcal{N}$ DNA sequences, denoted by x, and $\mathcal{C}$ expression covariates, denoted by y. All the sequences have the same length $\mathcal{L}$ and are aligned with respect to experimentally determined TSSs. To carry out motif discovery from these data, we developed here the integrative probabilistic model of $x|y$ whose structure is illustrated in figure 1.

Briefly, DNA sequences are modelled as drawn from a collection of $\mathcal{M}$ motif models (PWMs), denoted by $(\theta_1, \ldots, \theta_{\mathcal{M}})$, and a Markov background, denoted by θ_0, conditionally on the positions of motif occurrences. These positions are encoded in a layer of latent variables $a = (a_{m,n})_{m=1:\mathcal{M},\, n=1:\mathcal{N}}$, with $a_{m,n} \in \{0, 1, \ldots, \mathcal{L}\}$ being the position of motif m in sequence n (0 encodes the absence of occurrence). Motif occurrences a are modelled as drawn taking into account possible preferential position with respect to TSS and covariates. For simplicity, the model assumes *z*ero or *o*ne *o*ccurrence *p*er *s*equence (ZOOPS, [6]) of each motif but occurrences of different motifs are allowed and may overlap. Statistical inference is carried out in a Bayesian framework and parameters are thus treated as random variables drawn from prior distributions. The complete mathematical description of the model is found in electronic supplementary material, S1 appendix section A1, whereas the presentation below focuses on the most salient points. A detailed directed acyclic graph (DAG) representing relationships between all variables (parameters, latent variables, observed data) is shown in electronic supplementary material, figure S1.

2.1.2. Incorporating expression data as covariates

Information from expression data summarized in y is incorporated into the probability of occurrence of a motif via a

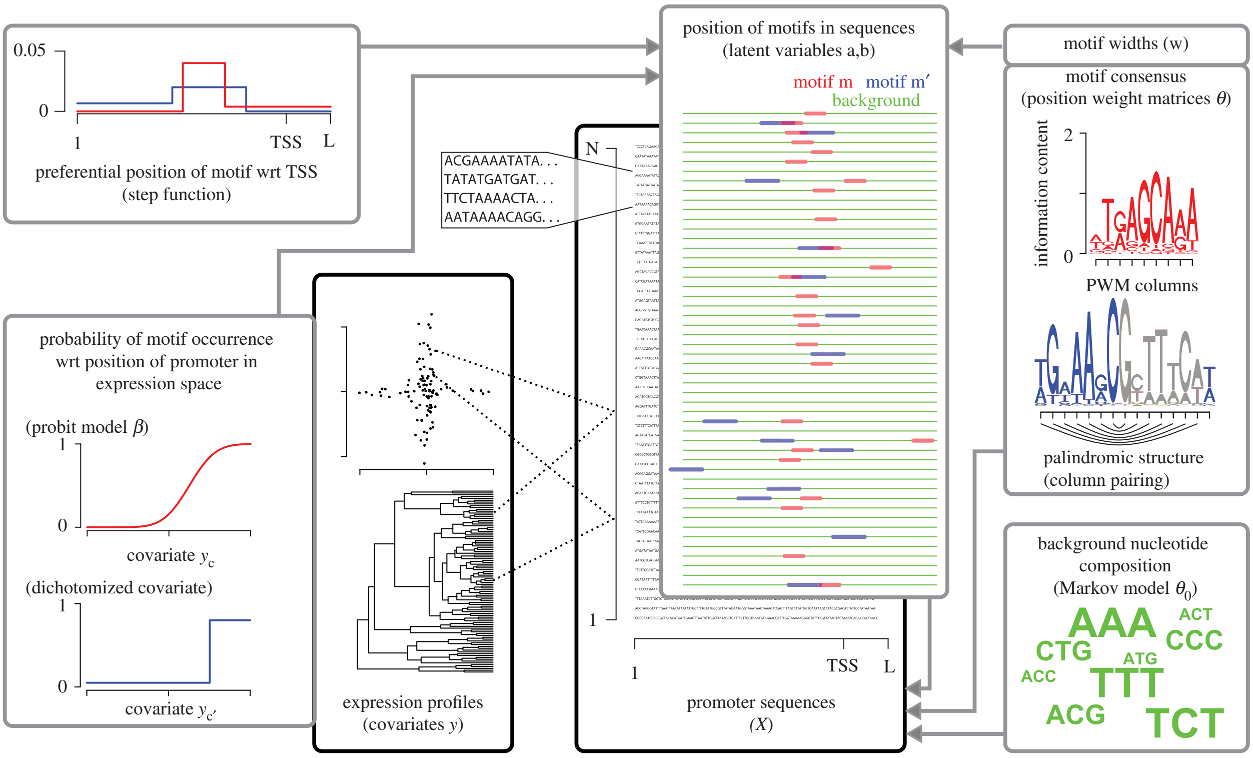

Figure 1. Hierarchical model of promoter sequences. The structure of the model is represented under the form of a simplified DAG: arrows reflect the factorization of the joint probability distribution. Text between brackets refer to notations and choices of specific modelling frameworks.

probit regression framework [26]. The standard version of this model would write the probability of occurrence of motif m in sequence n (i.e. the event $\{a_{m,n} > 0\}$) as

$$\pi(a_{m,n} > 0 | y, \beta) = \Phi(\beta_{m,0} + \sum_{c=1:\mathcal{C}} \beta_{m,c} y_{n,c}), \tag{2.1}$$

where $y_{n,c}$ is the numerical value of covariate c for sequence n, β contains the regression coefficients ($\beta_{m,c} \in \mathbb{R}$), and Φ is the cumulative distribution function of the standard normal distribution.

We developed an extension for this model that dichotomizes expression covariates according to automatically adjusted cut-offs (electronic supplementary material, S1 appendix A1.2); exactly as if co-expression clusters defined on the basis of $y_{\cdot,c}$ were also entered as covariates. This extension is very appealing because it can account for sharp changes in the probability of occurrence as a function of position in expression space. Importantly, it allows any amplitude of change, unlike sharp changes obtained by increasing $|\beta_{m,c}|$ in the standard probit which makes the probability to jump between 0 and 1.

To allow automatic selection and dichotomization (both choices affecting model dimension) of the covariates relevant for the occurrences of motif m, the model involves the following variables:

- $t_{m,c} \in \{0, 1\}$ which indicates whether covariate c should be taken into account;
- $\beta_{m,c} \in \mathbb{R}$ which is, when $t_{m,c} = 1$, a coefficient specific to covariate c; $\beta_{m,0}$ being the intercept parameter;
- $g_{m,c} \in \{0, 1\}$ which indicates, when $t_{m,c} = 1$, whether vector $y_{\cdot,c}$ is dichotomized;
- $h_{m,c} \in \{1, \ldots, \mathcal{N} - 1\}$ which corresponds, when $t_{m,c} = 1$ and $g_{m,c} = 1$, to the rank in the vector $y_{\cdot,c}$ of the cut-off value used for dichotomization.

In keeping with the probit regression framework, the probability of occurrence of motif m in sequence n writes then as

$$\pi(a_{m,n} > 0 | y, t, \beta, b, h) = \Phi\big(\beta_{m,0} + \sum_{c} \beta_{m,c} \mathbb{I}\{t_{m,c} = 1\} \tilde{y}_{m,n,c}\big), \tag{2.2}$$

where $\mathbb{I}$ is the indicator function ($\mathbb{I}\{z\} = 1$ if z is true, 0 otherwise) and $\tilde{y}_{m,n,c}$ corresponds to $y_{n,c}$ after possible dichotomization. This model can capture a great diversity of relationships between expression covariates and probability of motif occurrence (electronic supplementary material, figure S2 ABC).

Further extending the model described by equation (2.2), we also consider covariates that come in the form of trees. Indeed, the dichotomization proposed above makes it possible to incorporate whole tree structures in the regression model. In this case, dichotomization involves the choice of node in the tree instead of a cut-off value along an axis. Different probabilities of motif occurrence are then associated to inside and outside the sub-tree hanging under this node (electronic supplementary material, figure S2 D).

2.1.3. Modelling position with respect to TSS

DNA sequences are aligned with respect to experimentally determined TSSs. The position of motif occurrence encoded in $a_{m,n}$ then corresponds to a precise position relative to the TSS.

Given an occurrence of motif m in sequence n (event $\{a_{m,n} > 0\}$ modelled by the extended probit), the probability density function of its exact position is modelled as a step function (electronic supplementary material, S1 appendix A1.3.1). This model involves a parameter for the number of breakpoints (denoted by k_m) and two vectors (denoted by d_m and λ_m) which give the positions of the breakpoints and the probability of finding the occurrence in each of the corresponding segments.

2.1.4. Allowing motif occurrences to overlap

The model allows motif occurrences to overlap (electronic supplementary material, S1 appendix A1.3.1–2). This feature underpins the simplifying assumption of mutual independence between the occurrences of the $\mathcal{M}$ motifs made when modelling the position with respect to the TSS and the link with expression data. Another key benefit of allowing such overlaps is to permit updating motif width (w_m) without having to avoid collisions of motif occurrences.

When a number $o_{n,l} \geq 1$ of motifs overlap position l in sequence n, nucleotide $x_{n,l}$ is modelled as drawn from the arithmetic mean of the relevant PWM columns, namely

$$\pi(x_{n,l}|a, r, \theta, o_{n,l} \geq 1) = \frac{1}{o_{n,l}} \sum_{m \in O_{n,l}} \theta_{m,l-a_{m,n}+r_m,x_{n,l}}, \qquad (2.3)$$

where $O_{n,l}$ denotes the set of motifs that overlap position l, $\theta_{m,w,u}$ is the probability of nucleotide u at position w of motif m, and r_m is a variable recording which column of the PWM corresponds to the position referenced by $a_{m,n}$. The resulting density can be seen as the marginal of an equal-weight mixture.

2.1.5. Palindromic motifs

The dimeric nature and symmetry of many TFs explain that numerous known binding motifs are palindromic in the sense that symmetric positions with respect to the centre of the motif appear as mirrored according to Watson–Crick base pairing rule (A : T and C : G). Palindromic constraints on PWMs reduce the dimension of the model, thereby increasing the amount of data available to estimate each parameter and decreasing the size of the search space for parameter values.

Identifying as many motifs as possible simultaneously is however incompatible with imposing a strict palindromic structure to all PWMs. It is thus interesting to dynamically set and release palindromic constraints in the course of the algorithm but this implies important changes in the number of free parameters that cannot be implemented efficiently in the MCMC framework. For this reason, we developed a more flexible representation that allows intermediate states between non-palindromic and fully palindromic structures (electronic supplementary material, S1 appendix A1.3.3). Several variables are used to define the active constraints on $\theta_m = (\theta_{m,w,u})_{w=1:w_m, u\in\{\mathtt{A,C,G,T}\}}$: $p_m \in \{0,\ 1\}$, indicates if motif m has a (possibly partial) palindromic structure; c_m, records the position of the centre of symmetry; $q_{m,w} \in \{0, 1\}$, indicates whether columns w and $2c_m - w$ are paired. Intermediate states allow the number of free parameters to gradually increase or decrease. They also fit situations where a biological motif is only partially palindromic.

2.1.6. Bayesian inference

The model ingredients described above allow the complete data likelihood to be written,

$$\begin{aligned}\pi(x, a \mid y, r, \theta, \theta_0, \beta, h, d, \lambda)\\ = \pi(x \mid a, r, \theta, \theta_0)\pi(a \mid y, \beta, h, d, \lambda).\end{aligned} \qquad (2.4)$$

From equation (2.4) and the prior distributions for the parameters (electronic supplementary material, S1 appendix A1.4), Bayes' rule defines the joint posterior on which Bayesian inference is based

$$\pi(a, \underline{v}, \theta_0, \underline{w}, \underline{p}, \underline{c}, \underline{q}, \theta, r, \underline{k}, d, \lambda, \underline{t}, \beta, \underline{g}, h \mid x, y), \qquad (2.5)$$

where the parameters underscored (not found in equation (2.4)) determine the dimension of other parameters.

A MCMC algorithm was built to sample this joint posterior. To cope with the high dimension of the target, the algorithm is a block MCMC sampler [27] composed of 15 types of steps designed to update separate subsets (blocks) of variables. A sweep combines the different steps. Updates of the parameters of the probit models use the data augmentation scheme proposed by [26] (latent variable $z_{m,n}$). Similarly, in keeping with the usual treatment of mixture models [28], a latent variable (denoted by $b_{n,l}$) is introduced at each position of the sequence set to 'disambiguate' motif overlaps. The reversible-jump methodology [29] allows the changes of dimension needed to update the Markov order of the background (variable v), the active covariates and their possible dichotomization (block $t_{m,c}$, $g_{m,c}$, $h_{m,c}$), and the variables encoding the palindromic structure of the motif (block p_m, c_m, q_m). Under circumstances where the probability distribution of the variables whose dimension is modified can be integrated-out, the reversible-jump can be done in a Gibbs manner, i.e. by direct sampling from the conditional distribution. This is done for the joint update of ($t_{m,c}$, $g_{m,c}$, $h_{m,c}$) and the joint update of (p_m, c_m, q_m). The algorithm was implemented in a C++ program named `Multiple` whose correctness was carefully checked by successive conditional simulations to reveal analytical and coding errors [30]. Details of the MCMC algorithm are found in electronic supplementary material, S1 appendix A2.

2.2. Dataset for application to *Listeria monocytogenes*

2.2.1. Transcription start sites and expression profiles

Promoter sequences were defined as the 121 bp spanning from position −100 to +20 relative to each TSS based on a repertoire of 2299 TSSs mapped at 1 bp resolution on *L. monocytogenes* EGDe genome sequence [24]. The choice of −100 was in keeping with the size of the regions that we previously found to be enriched for the presence of known TF binding sites [11]. To remove sequence overlaps on the same strand, we used a simple greedy procedure that incorporated non-overlapping promoters one-by-one in the order of decreasing read-count (reflecting the level of experimental support and transcriptional activity for the TSS [24]). This led to a set of 1545 non-overlapping promoter regions (67% of the initial list of TSSs).

For the expression data, we relied on the compendium dataset established by aggregation of many different studies to build the Listeriomics website [25]. As downloaded, the data had dimensions 3159 × 254, where each row corresponds to a gene of *L. monocytogenes* EGDe and each column corresponds to the log of an expression ratio (log fold-change) comparing two samples (mutants, growth conditions, strains …) from a same study. Some columns and rows contained many missing values due to the heterogeneity of technologies. The number of columns was reduced from 254 to 165 after discarding the columns with a number of missing values higher than 1.5 times the median. In parallel, the number of rows was reduced from 3159 to 2825 based on the same criterion. Finally, the gene name associated with each TSS [24] permitted matching 1512 out of 1545 non-overlapping promoter sequences to one of these 2825 genes with expression data. This resulted in an expression matrix of dimension 1512 × 165 whose 165 columns represented 31 published expression studies, and included 17 RNA-Seq and 25 tiling array profiles.

2.2.2. Covariates for motif discovery

Both hierarchical clustering and projection methods were used as dimension reduction techniques to summarize the expression matrix. As projection methods, we applied PCA and ICA implemented in functions `prcomp` (package `stats`) and `fastICA` [31] of `R` without scaling the expression data. For PCA, we kept the 20 first components of the PCA (accounting for 68.8% of the total variance) after examining the rate of decrease of the residual variance. When applying ICA, the target dimension (number of components) is fixed before numerical optimization and the algorithm can converge to different projections that share only a subset of components. In keeping with the idea developed by [20], we chose the target dimension $K = 40$ after examining the stability of the components and only stable components were used. Here, we used average-link clustering based on absolute Pearson correlation coefficient (r) between columns of the source matrix (dimension 1512 × K) and a cut-off $|r| \geq 0.8$ to compare components between runs. The algorithm was run 100 times which led to 26 components found in at least 80% of the runs.

Two types of hierarchical clustering were applied corresponding to different options of `hclust` function (package `stats`): Ward and average-link based on Pearson distance ($1 - r$, where r is Pearson correlation coefficient). The distances between genes needed to build the trees were computed after centring and scaling rows (genes) of a symmetric expression matrix obtained by duplicating each column with a negative sign to remove the effect of the arbitrary orientation of the log fold-changes.

The final set of 50 covariates used in the motif discovery analysis consisted of the 20 first PCA components (covariates numbered 1–20), the 26 stable ICA components (covariates 21–46), and the hierarchical clustering trees obtained by Ward and average-link methods (covariates 47 and 48). The two trees were further duplicated (covariates 49 and 50) to make it possible for the model to use two nodes of the same tree.

3. Results

3.1. Exploration of the posterior landscape

Trajectories of the variables describing the $\mathcal{M}$ motif components during MCMC runs confirmed that the algorithm was able to adjust simultaneously the characteristics of many PWMs. To illustrate the behaviour of the algorithm, figure 2*a* depicts the parallel evolution in the MCMC run of two PWMs in terms of width and nucleotide composition. As shown in electronic supplementary material, figure S3, the algorithm uses gradual activation/deactivation of palindromic constraints to switch between non-palindromic and palindromic PWMs during its exploration of the posterior.

For de novo motif discovery, it is important to identify when stability of a motif component across tens of thousands of MCMC sweeps, as seen for the second motif shown in figure 2*a*, is not caused by slow-mixing but truly reflects attraction to peaks of the posterior density and should therefore be treated as a relevant motif prediction. Reaching similar motif components independently from different starting points is indicative of the second scenario. We thus performed 10 independent parallel runs of the MCMC algorithm. Each run consisted of 50 000 MCMC sweeps from a random starting point and $\mathcal{M}$ was fixed to 75 (12 to 14 days on Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30 GHz). Only the last 10 000 sweeps (with a thinning interval of 100 sweeps) were used in our analysis to characterize the posterior distributions.

The 10 runs produced information on 750 (10 × 75) motif components that were named from M1.0 (random seed 1, motif 0) to M10.74. We analysed and compared these motif components to extract distinct well-supported motifs that were not only stable across the last 10 000 sweeps but were also found in at least two runs. To declare that two motif components corresponded in fact to the same motif, we compared the sets of positions in the sequences that were predicted to be covered by each motif. These summarize the information carried by the different ingredients of the model (PWM, preferential position with respect to TSS, link with expression data). Namely, after listing positions with estimated posterior probability of coverage greater than 0.5, the pairwise distance between two motif components i and j was computed as

$$d(i, j) = 1 - \frac{2N(i, j)}{N(i) + N(j)}, \tag{3.1}$$

where $N(i)$ and $N(j)$ denote the total number of positions covered by each motif and $N(i, j)$ is the number of positions covered by both motifs. Accordingly, $d(i, j) = 0$ if these positions are exactly the same and $d(i, j) = 1$ if they differ completely. A special case of $d(i, j) = 1$ corresponds to $N(i) = 0$ which concerned 190 of the 750 motifs and typically reflected instability during the last 10 000 sweeps (e.g. M0.1 in figure 2*a*). In the same run, the minimal distance observed between two motif components was 0.59 and 92.7% of the motif components were at a distance greater than or equal to 0.9 of any other motif components. This confirmed that, within a run, motif components converge to distinct motifs.

The list of distinct well-supported motifs was established after average-link hierarchical clustering of the 750 motifs based on the distance defined in equation (3.1). As illustrated in figure 2*b*, this tree was cut at two different heights: 0.75 to define high-level clusters that separate well-distinct motifs; 0.25 to define low-level clusters containing very similar motifs. A representative motif was selected in each high-level cluster containing at least one low-level cluster (the member of the low-level clusters with the highest $N(i)$). This procedure identified the set of 40 representative motifs reported in table 1.

Figure 3 shows several examples of links between expression covariates and motif occurrences captured by the extended probit model, including covariates that were dichotomized and trees. A total of nine motifs were strongly linked (i.e. with posterior probability greater than or equal to 0.9) to a covariate of type vector (M20.1, M26.1, M33.8, M38.10, M58.7, M61.3, M62.3, M68.1 and M71.2) out of which three are represented in figure 3*a–c*. For M38.10 and M68.12, a marked preference for the dichotomized version was observed

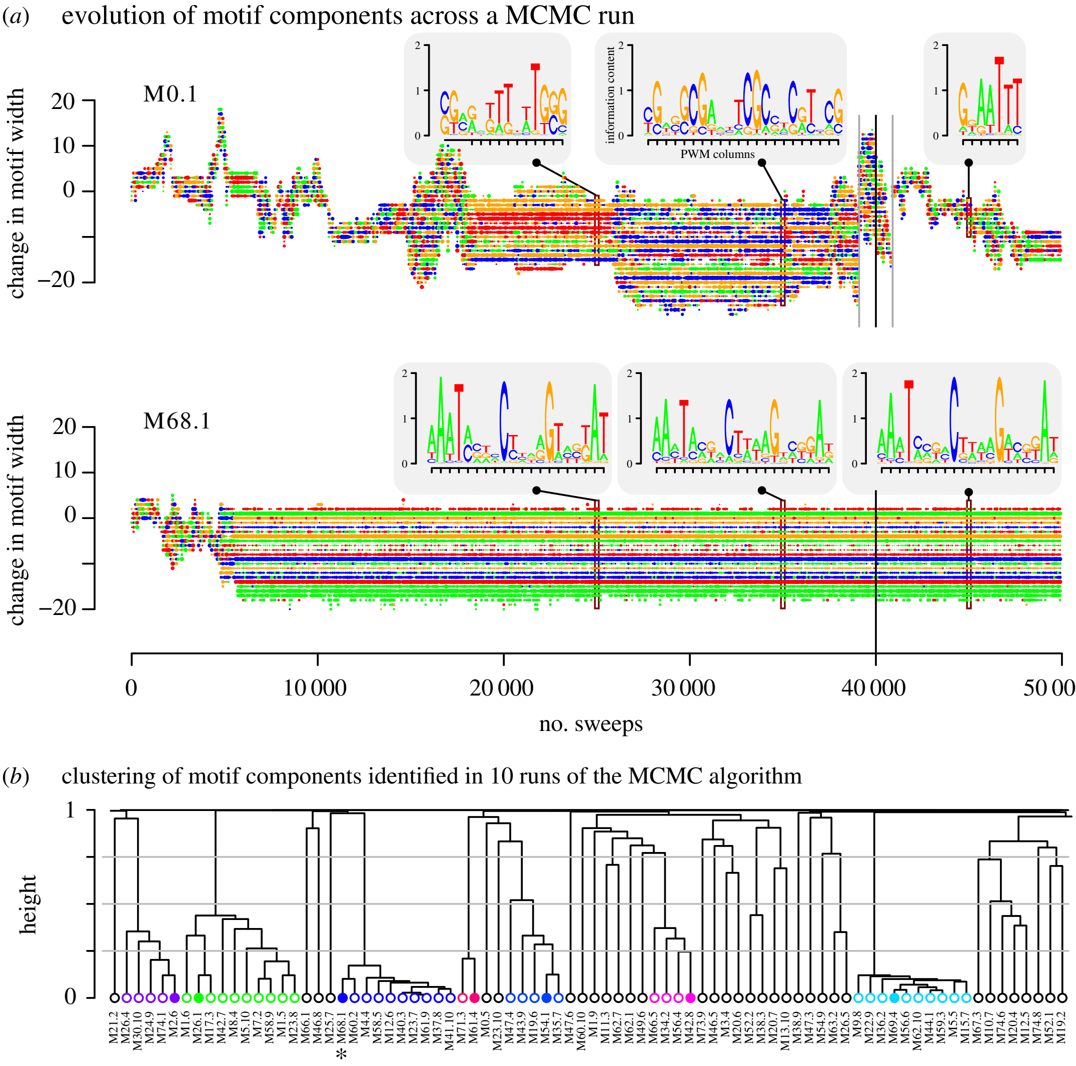

Figure 2. Exploration of the posterior landscape. (*a*) Evolution of two motif components across the 50 000 sweeps of the algorithm. Coloured dot points correspond to PWM columns, colour indicates the most likely nucleotide (green for A, blue for C, orange for G, red for T) and diameter reflects the information content of the nucleotide frequency. Increase or decrease of the width of the PWM on the 5′- and 3′-sides are represented by change in the number of coloured dots (on the lower and upper side, respectively). A black vertical line is drawn at sweep number 40 000 (end of our burn-in period). Grey vertical lines indicate sweeps at which the *y*-coordinate is re-centred for the purpose of the representation. Insert plots provide motif logo representations of the PWMs at selected sweeps (25 000, 35 000 and 45 000). (*b*) Hierarchical clustering of motif components to extract stable motifs. Only a portion of the tree is represented (80 motif components out of 750). Filled circles correspond to motifs belonging to the final list of 40 representative stable motifs. Their high-level clusters (defined by height 0.75) are represented by different colours. A star symbol indicates position of M68.1 whose evolution is represented in subplot *a*.

(figure 3*c*). Remarkably, all these nine associations concerned covariates defined by ICA (none defined by PCA). Of note, as expected given that we allowed some redundancy between the covariates, a strong link with a specific covariate did not cover all the cases of clear association between the presence/absence of a motif and the expression covariates (e.g. M29.8, figure 3*d*,*e*,*f*). Posterior probability greater than or equal to 0.9 of association with any of the trees concerned five motifs (M9.1, M12.4, M29.8, M69.4 and M38.10) but strong association with a specific tree (ward or average-link) was observed for only two of them, illustrating again some redundancy between covariates.

To further understand the contributions of different ingredients of our model to the identification of these 40 motifs, we examined the results of five submodels that did not incorporate one or several of three ingredients (position with respect to TSS, expression profiles and palindromic structures). The results show that all the motifs were not sensitive to the same model ingredients and many were sensitive to several ingredients (electronic supplementary material, S1 appendix A3) which attests to the interest of an integrative statistical model.

3.2. Validation by comparison with known motifs and regulons

The 40 distinct well-supported motifs exhibit considerable diversity in terms of abundance, preferred position with respect to TSS, link with expression data, and PWM characteristics (width, information content, palindromic structure). Table 1 summarizes the main characteristics of each motif. Figure 4 provides a graphical representation for two of these motifs. Similar figures are available for all motifs in electronic supplementary material, figure S1 and the numerical PWMs are found in electronic supplementary material, file S1. Lists of genes associated with each motif and precise positions of occurrences are reported in electronic supplementary material, files S2 and S3.

Table 1. Summary of the 40 stable motifs identified in *L. monocytogenes* EGD-e.

motif[a]	no.[b]	no. (0.8 – 0.2)[c]	W[d]	IC1[e]	Pal[f]	Pos[g]	FC[h]	runs[i]	comment[j]
M71.2	1325	1174–1435	9	3	2	−12 [1]	0.4	7 : 10	SigA -10
M71.8	1308	848–1506	6	2	2	−1 [1]	0.5	3 : 16	TSS (A)
M36.4	1033	552–1382	6	3	0	−36 [2]	0.5	4 : 10	SigA -35 (TTG)
M55.7	836	117–1486	5	0	1	−27 [38]	0.6	3 : 7	SigA -10 extension (CT)
M8.3	792	331–1273	5	1	0	0 [1]	0.5	2 : 15	TSS (G)
M9.10	735	447–1165	8	4	0	−17 [1]	0.5	4 : 8	SigA (extended -10, TG)
M61.5	561	171–1153	5	5	0	14 [10]	0.4	3 : 8	RBS (GGAGG)
M67.8	471	150–1159	14	6	0	−65 [35]	0.4	2 : 10	T-rich element
M26.1	252	107–640	13	5	1	−82 [20]	0.7	2 : 10	SigA -10 reverse strand
M27.6	240	153–590	6	5	0	−37 [2]	0.5	6 : 6	SigA -35 (TTGAC)
M29.8	122	91–150	12	4	2	−16 [2]	3.1	8 : 10	SigB -10
M9.1	116	43–368	22	2	6	−76 [35]	1.2	2 : 8	loose
M18.2	108	47–355	6	6	0	7 [12]	0.5	2 : 8	RBS (GAGGTG)
M12.4	97	71–109	7	4	0	−35 [3]	3.9	10 : 10	SigB -35
M69.4	87	59–128	15	8	**12**	−35 [51]	1.8	10 : 10	CcpA (CRE-box)
M42.8	62	14–295	17	0	0	−87 [56]	0.6	2 : 4	loose
M31.4	39	16–81	23	8	**22**	−44 [58]	1.9	2 : 2	Rex
M68.1	31	20–47	20	5	**16**	−53 [24]	3.0	10 : 10	LiaR
M58.7	27	14–40	19	6	**10**	−51 [53]	3.2	2 : 10	—
M62.3	26	18–34	20	13	**18**	−21 [60]	1.9	10 : 10	Fur
M38.10	22	18–38	15	8	**14**	−29 [23]	2.4	9 : 10	LexA
M53.9	20	8–50	20	8	**18**	−51 [55]	1.5	3 : 10	VirR
M2.6	19	12–53	9	4	0	−49 [2]	1.1	3 : 5	Spx
M13.7	19	7–48	21	5	0	−58 [68]	0.9	3 : 3	—
M54.1	14	6–40	23	8	0	−60 [56]	1.0	2 : 5	—
M61.6	14	9–16	25	14	**18**	−38 [2]	0.8	5 : 10	BglR2
M61.4	13	7–25	25	6	0	−45 [54]	1.2	2 : 2	—
M50.10	13	7–43	21	4	**12**	−83 [28]	0.9	2 : 2	—
M70.6	11	7–19	24	22	**22**	−51 [53]	1.3	10 : 10	—
M3.1	9	4–20	21	9	**20**	−47 [55]	1.5	6 : 10	—
M33.8	8	7–13	22	10	0	−31 [15]	2.8	9 : 10	SigL
M20.1	7	5–7	25	17	1	−75 [1]	6.6	10 : 10	—
M49.4	7	6–11	23	11	**20**	−52 [5]	10.5	10 : 10	PrfA
M17.4	7	5–14	24	8	**18**	−44 [57]	1.1	3 : 4	CcpB
M2.1	6	4–8	22	9	**16**	−38 [50]	1.1	2 : 5	—
M61.3	6	5–10	25	23	2	−51 [1]	3.1	10 : 10	—
M73.4	6	4–7	20	3	0	−49 [2]	5.0	6 : 6	—
M18.6	4	3–8	25	10	**23**	−61 [50]	1.2	6 : 7	—
M29.1	3	2–6	25	1	0	−46 [56]	1.6	2 : 10	loose
M59.2	2	0–13	23	0	0	−55 [56]	2.4	4 : 5	loose

[a] unique motif identifier build as Mxx.yy where yy identifies the run and xx the motif in the run;

[b] number of promoter regions where the motif is predicted to occur (estimated posterior probability ≥ 0.5), used to order the motifs;

[c] number of TSSs when the posterior probability cut-off is set to 0.8 (very likely) or 0.2 (possible);

[d] motif width corresponding to the number of columns included in the PWM with posterior probability ≥ 0.5;

[e] number of columns in the PWM with high information content, i.e. $2 + \sum_{u\in\{A,C,G,T\}} \theta_{m,w,u} \log_2 (\theta_{m,w,u}) \geq 1$;

[f] estimated number of paired columns in the PWM reflecting the degree of palindromness (in boldface when high);

[g] median position of occurrence for the middle of the motif with respect to the TSS (inter-quartile range reported between brackets), both numbers are derived from the estimated probability density function for the position described by the variables K_m, $\lambda_{m,.}$ and $d_{m,.}$;

[h] maximum across the 165 pairs of conditions for the median of the expression values (log fold-change) associated with the TSSs counted in the second column;

[i] number of parallel runs (out of 10) in which this motif was found as obtained by clustering based on amount of overlaps between occurrences, written in the format xx:yy where xx and yy are the numbers obtained with cut-offs 75% of overlap and 25% of overlap, respectively;

[j] link to known TFs if identified or other observations.

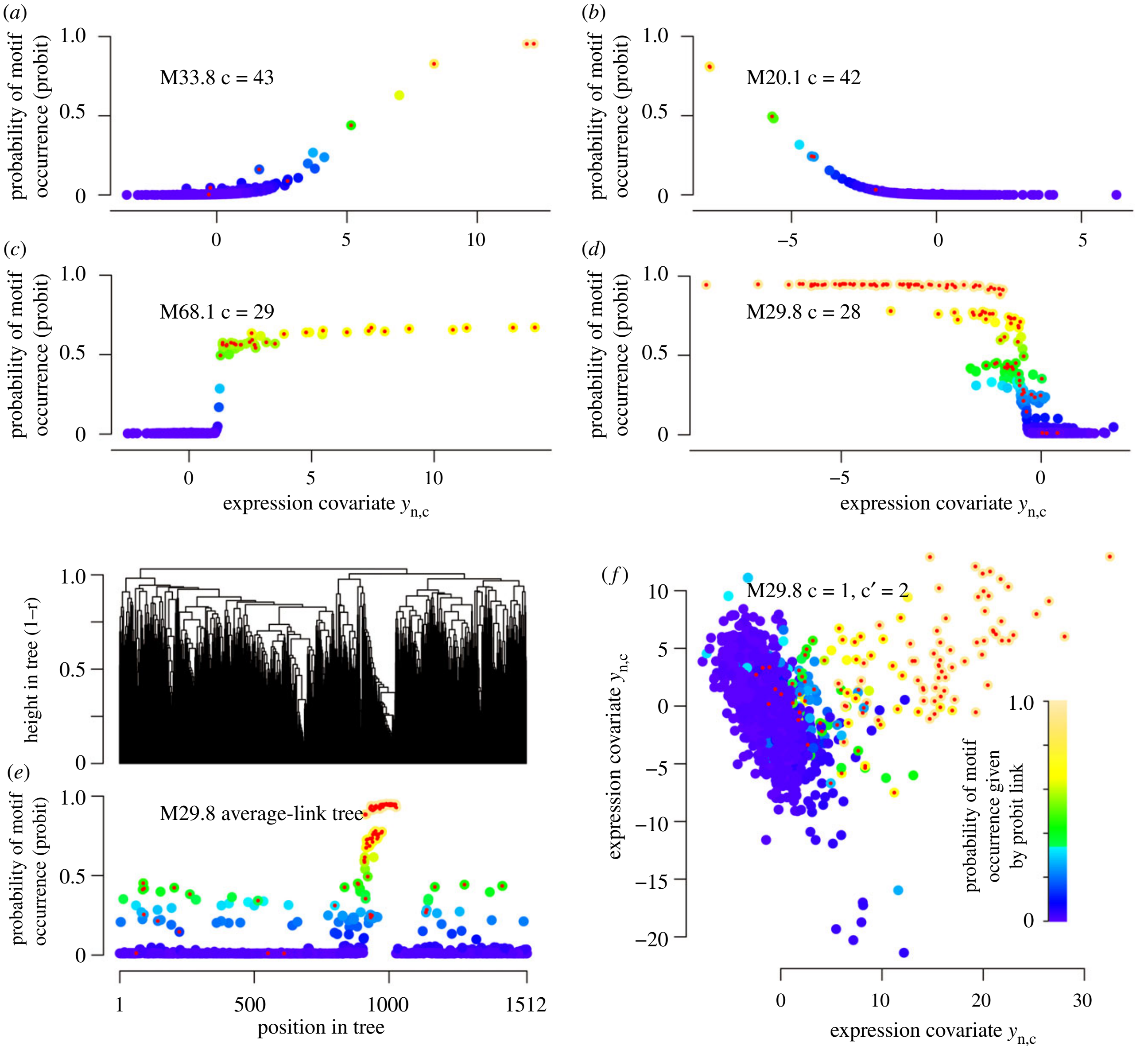

Figure 3. Estimated links between expression covariates and motif occurrences. Examples are shown for different motifs and covariates (indicated in subplot titles). Colour scale from blue to yellow reflects the estimated probability of motif occurrence given by the extended probit model (i.e. summarizing the information from the expression data). Red dots indicate sequences in which the motif is found (estimated posterior probability of motif occurrence greater than or equal to 0.5).

Three complementary approaches were used to connect the 40 motifs discovered by our de novo approach to known regulons. The first was comparison with lists of genes collected from tables published in several expression studies and positions of transcription factor binding sites recorded in the RegPrecise database [32]; the second approach was comparison with 188 reference PWMs derived from sequence alignments extracted from the propagated RegPrecise database (accessed in July 2018) for different taxonomic groups in the Firmicutes phylum: Listeriaceae (25 PWMs) and Staphylococcaceae (39 PWMs) and Bacillales (124 PWMs). The third approach consisted of dedicated literature searches associated with a careful manual examination of (i) genes downstream the promoters in which the motif was predicted to occur (ii) conditions in which the log fold-change deviated the most from 0 (iii) characteristics of PWMs and preferred positions of motif occurrences with respect to TSSs. These lines of observation provided convergent clues for many of the identified TFs.

Connections to known TFs are reported in the rightmost column of table 1. Most of the motifs with a high number of occurrences were found to describe general characteristics of promoter regions (variations on the following themes: SigA −10 and −35 boxes, nucleotide composition around TSS, Ribosome Binding Site). Systematic comparison with RegPrecise PWMs was particularly informative for the identification of BglR2, CcpA, CcpB, Fur, LexA, LiaR and Rex. Among the other identified TFs, SigB and SigL were identified based on position with respect to the TSS, comparison with sigma factor consensus defined for *Bacillus subtilis* [10], as well as overlap with previous experimental data on SigL (also known as RpoN or sigma54) and SigB regulons in *L. monocytogenes* [33–35]. Electronic supplementary material, figure S5 provides a graphical representation of the prediction of the SigB regulon from the joint occurrence of boxes −10 and −35 in comparison with previous studies. In brief, 89 promoter regions are predicted to contain both the −10 and −35 boxes, 88 out of them were previously reported as probable members of the SigB regulon in either [34] or [35]. Spx was identified based on (i) its position with respect to the TSS just upstream SigA −35 box, (ii) literature data on the Spx regulon of *L. monocytogenes* [36] and (iii) sequence properties reported for the Spx motif in *B. subtilis* described

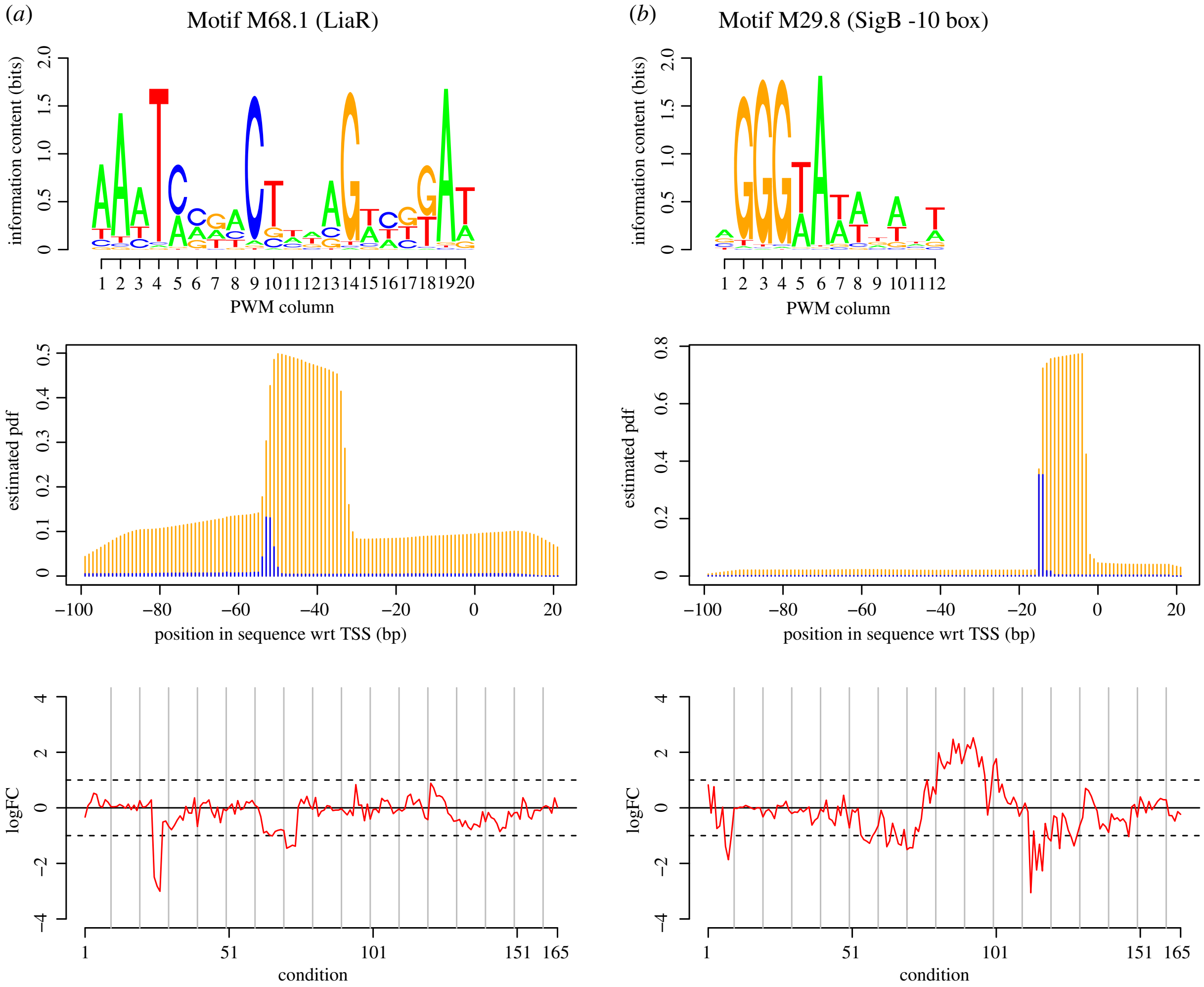

Figure 4. Two examples of motifs for known TFs rediscovered by our algorithm. These motifs were identified as corresponding to the binding sites of LiaR (M68.1 subplot *a*) and as the −10 box of the SigB binding sites (M29.8 subplot *b*). First row: sequence logo. Second row: estimated probability distribution function for the 5′-end of the occurrence (in blue) and probability of having the position covered by the occurrence (in orange). These probabilities are conditional on the presence of the motif in the promoter sequence. Third row: average log fold-change of expression level for downstream genes across the 165 pairs of conditions that were used to define the expression covariates.

as an AGCA element at position −44 [37]. PWMs found for PrfA and VirR, two key transcription regulators involved in *L. monocytogenes* virulence, were in line with previously described sequence properties [38,39]. Taken together, correspondences with known motifs validate our approach for de novo discovery.

3.3. New insights into the *L. monocytogenes* transcription network

The de novo re-identification of many known motifs suggests that new biologically relevant observations can be made from (i) new predictions of occurrences for known motifs and (ii) predictions of new motifs. The first aspect developed in electronic supplementary material, S1 appendix A4 encompasses cases of regulons that have been experimentally studied in *L. monocytogenes* (Fur, LiaR and LexA) and regulons that have not yet been mapped in this bacterium (CcpA, Rex and Spx). For instance, our results indicate that the LiaR regulon involved in the response to cell envelope stress may be approximately two-times larger than previously identified by differential expression analysis of the *liaS*-deletion mutant versus wild-type [40] and that the CcpA (Catabolite control protein A) regulon is almost as large as in *B. subtilis* [41], suggesting an overlooked role in *L. monocytogenes*.

Among the predicted regulons listed in table 1 that have not been linked to known motifs, the most spectacular by its size is associated with motif M58.7 whose central PWM columns correspond to the palindromic consensus ACGTAYYCGT (logo shown in electronic supplementary material, figure S4). The 27 predicted targets exhibit remarkable functional homogeneity, consisting almost exclusively of genes encoding the translation apparatus, including ribosomal proteins (electronic supplementary material, file S2). At the other end in terms of number of occurrences, M2.1 is a palindromic motif with four out of six occurrences found upstream of genes encoding oxidoreductases (electronic supplementary material, figure S2 and file S2).

3.4. Comparison with other methods

We compared our results to those of different well-established algorithms able to handle the dataset of 1512 sequences and 1512 × 165 expression matrix for de novo motif discovery. These algorithms are based on PWM estimation from the sole sequence dataset (MEME, [6]) or search specifically for motifs connected to the expression data using motif representations that can be either consensus string written in IUPAC degenerated symbols (FIRE and RED2, [5,14]) or PWMs (MatrixREDUCE, [16]). Each of these algorithms returns an internally computed significance score for each predicted motif. The significance cut-offs used here were either default cut-offs (thereby following the recommendation of their authors) or somewhat low or relaxed cut-offs (for MEME and RED2) when it seemed relevant to increase the number of motifs returned. Input data, parameter settings and results are detailed in electronic supplementary material, S1 appendix A5.

Motif discovery without auxiliary data using MEME with different settings (zero- or third-order Markov background, with or without palindromic constraints) returned only up to eight motifs with E-value less than or equal to 1 (see electronic supplementary material, S1 appendix A5.1). These motifs included motifs describing general properties of the *L. monocytogenes* promoter regions, such as SigA −10 box, the presence of a RBS and of T-rich elements (our M59.9), as well as more specific motifs that are relatively abundant and/or of high information content (IC1 in table 1): M70.6, CcpA and Fur. A variant of the sequence element described by M20.1 and M61.3 is also found. MEME which implements an expectation-maximization algorithm for maximum likelihood inference was the slowest of the four algorithms (still substantially faster than one run of our algorithm), taking between 18 h and 35 h depending on the settings.

The three algorithms (MatrixREDUCE, FIRE and RED2) searching for motifs related to expression data returned only short motifs (lengths up to 9) due to the use of k-mers as seeds in the initial steps of the searches (see electronic supplementary material, S1 appendix A5.2–4). With some parameter settings, they were all able to retrieve SigB −10 box (M29.8) which presents the particularity of being abundant and to exhibit a strong link to expression profile. It is also short and contains adjacent high information content PWM columns which makes it ideally suited for approaches based on k-mers (figure 4*b*). MatrixREDUCE also discovered a short motif that corresponds to the central part of PrfA. This motif presents the particularity of exhibiting, by far, the strongest link with expression data among all the motifs that we discovered (see column FC in table 1). Thus, comparison of our results with those of these four algorithms illustrates the utility of the statistical model described in this work: it combines the good behaviour of purely sequence-based approaches, like MEME, with the benefits of using expression data.

4. Discussion

The methodology developed in this work provides a new integrated framework for the discovery of regulatory motifs in bacteria with the help of transcriptome data (exact positions of TSSs and expression profiles).

By modelling the sequence and incorporating expression data as covariates, the method inherits the good behaviour of pure sequence-based approaches grounded on well-established statistical models of DNA sequences, such as implemented in MEME [6]. This can be connected to earlier works that have incorporated external data in sequence models for motif discovery via the definition of informative priors to favour motifs whose occurrences are found in regions of the genome that are more likely to contain binding sites; for instance because they are conserved between species or depleted in nucleosomes [42–44]. In our work, the relevant covariates that describe how the probability of occurrence of a motif differs between promoters are automatically selected together with the coefficients that specify their contributions and this is achieved simultaneously for many motifs. We also show how we can account for covariates with complex structures such as the positions of the sequences in a tree. The idea of using a tree whose topology and branch lengths reflect similarities between activity profiles is to provide an alternative to the use of a predefined set of co-expression clusters. It is also found in the previous model and algorithm that we specifically developed for de novo discovery of sigma factor binding sites [10]. However, probabilistic models are completely different. In [10], to give to sequences which are close in the tree (hence in the expression space) a greater chance to harbour binding sites for the same sigma factor, the occurrences of the different possible motifs are modelled as resulting from an ‘evolution’ process along the branches of the tree. The motivation for introducing here the use of a probit model was to handle more complex representations of the expression space than a single tree. With our extended probit model, several concurrent descriptions of the expression space can coexist (trees, continuous vectors) and the algorithm selects and combines the most relevant for each motif, with the possibility to dichotomize continuous vectors according to automatically adjusted cut-off values. This contrasts with selection of expression summary measures beforehand to which other approaches are subjected. It also makes the algorithm particularly well adapted to exploit compendium of expression data (the dataset considered in this work aggregates results from 31 published studies).

Another novel aspect of our sequence model is to allow overlaps between motif occurrences. This is a key ingredient to simplify the model and to facilitate MCMC updates when searching for multiple motifs. Importantly, we show here that in the same run the different PWM components converge to distinct motifs, thereby providing a alternative to the heuristic consisting of searching for motifs one-at-a-time and masking predicted occurrences in subsequent searches to avoid rediscovery of the same motifs (as implemented in MEME [6]). In this aspect, our model based on explicit modelling of motif occurrence overlaps appears as a proper statistical framework to implement what [45] named ‘repulsion’ and for which they proposed incorporating *ad hoc* repulsive forces between parallel MCMC runs. Binding sites do overlap in bacterial genomes [46] and our choice of modelling the contribution of different motifs that overlap by averaging the nucleotide emission probability density functions (PWM columns) is the simplest but probably not the most biologically realistic. Indeed, we also considered a more complicated model in which PWM columns contribute as a function of their information content. This satisfies the intuition that if two motifs overlap and one has a strong preference for a nucleotide at a particular position whereas the other has no or little preference, the motif with the strong preference tends to ‘impose’ its choice. In electronic supplementary material, S1 appendix A1–2, we refer to this model as the θ-dependent weight mixture model of motif overlap and we describe its implementation. Because of its drawbacks (a dependence structure making that θ can no longer be marginalized out) we have decided to use here only the simple model of motif overlap.

Each run of our algorithm takes ≈2 weeks on a dataset like the one studied in this study. High computational cost is inherent to the MCMC machinery, even if, to some extent, the code could probably be optimized. This cost is compensated by the purpose of the approach which is to integrate a large amount of heterogeneous information on sequences, TSSs, and expression profiles in order to retrieve, at once, a maximum number of motifs. In practice, most of the time is spent in the update of *a* where each possible position of occurrence are evaluated for each motif at each sweep. Thus, time complexity is approximately proportional to the total length of the sequences times the total width of the PWMs plus a small term roughly in the square of the total width for taking motif occurrence overlaps into account. Furthermore, the analysis that we conducted is based on several runs that serve to unambiguously identify peaks in the posterior landscape. These peaks are identified by the convergence of several runs to the same neighbourhood in a space of very large dimension. Here, we limited our analysis to 10 runs that were conducted in parallel. These 10 runs identified 40 stable motifs but it would not be surprising to find several other motifs by adding more runs, since some of the biologically known motifs have here only be found by two or three runs (Rex, Spx, CcpB). Future studies on other datasets may include more runs.

As mentioned in the introduction, *L. monocytogenes* was well suited for a proof-of-principle application of the method. Its regulatory network contains features shared with the related Gram-positive model bacterium *B. subtilis* and features that are specific, such as those involved in pathogenicity. Comparison of our list of 40 motifs with literature data validated the method by proving its ability to re-discover, in a pure de novo manner, regulons of many known TFs.

In essence, the method is based on the search for 'over-represented' motifs whose modelling significantly improves the probabilistic representation of the DNA sequence as measured in the likelihood. It is thus only adapted to identify the regulons of TFs playing the coordinating roles of regulating the expression of several to many transcription units. These so-called global transcription factors are opposed to local TFs that account for the vast majority of TFs in bacteria but regulate only one or very few targets in a specific biological pathway [47,48]. For regulons of TFs that have been previously subjected to analyses by transcriptomics (e.g. SigB, LexA, Fur, LiaR, PrfA, VirR), our results contain de novo predictions based on the presence of motif occurrences made in a unified framework incorporating experimentally determined TSS position and expression data. This is interesting since contributions of direct and indirect regulations have not always been fully disentangled in the literature by identification of transcription factors binding sites. Our results also contain the first published global predictions for the regulons of several TFs whose importance is suggested by knowledge in other bacteria such as *B. subtilis*, but which have not yet been experimentally studied in *L. monocytogenes* (CcpA, Rex and Spx). Detection of Spx binding sites is a particularly striking achievement since its consists of a very short motif with a constrained position of occurrence directly upstream of SigA −35 box which remained elusive until the use of dedicated ChIP experiments and regression analyses in *B. subtilis* [37].

A motivation of our work was to discover new important regulons. An interesting result, is the identification of the partially palindromic motif M58.7 whose occurrences upstream of genes encoding the translation apparatus suggest a role in control of growth rate. In the Gram-positive model bacterium *B. subtilis*, as well as in *Escherichia coli*, there exists a mechanism known as the stringent response [49,50]. In both bacteria, and probably also in *L. monocytogenes* [51], this regulatory mechanism acts by decreasing the production of the translation apparatus components when the resources in the medium become too scarce. Univocal coupling between available resources and growth rate is however not necessarily the most appropriate in all circumstances. The existence of dedicated regulatory mechanisms that control the growth rate, even in the presence of nutrients, may thus not seem unexpected, in particular for an intracellular pathogen like *L. monocytogenes*. The new regulon that we detected in *L. monocytogenes* might be an instance of such a mechanism whose biological role and regulatory molecules remain to be identified.

Data accessibility. The source code of `Multiple` is distributed under the GNU General Public License (https://forgemia.inra.fr/pierre.nicolas/multiple); expression covariates, sequence data and R code used for postprocessing MCMC output are made available in the same repository. Electronic supplementary materials provide a full description of the methods and results.

Authors' contributions. I.S., V.F., S.S. and P.N. designed the study, interpreted the results, contributed to the writing of the manuscript and gave final approval for publication. I.S. and P.N. implemented the algorithm and processed the data.

Competing interests. We declare we have no competing interest.

Funding. The PhD thesis of I.S. was funded by the EU H2020 Marie Skłodowska-Curie Actions, grant no. 641984 (project 'List_Maps').

Acknowledgements. We thank Birgitte H. Kallipolitis, Pascal Piveteau and the List_Maps consortium for helpful discussions; Guillaume Kon Kam King for comments on the manuscript; INRA MIGALE bioinformatics facility (doi:10.15454/1.5572390655343293E12) for providing computational resources.

References

1. Sandve GK, Drabløs F. 2006 A survey of motif discovery methods in an integrated framework. *Biol. Direct* **1**, 11. (doi:10.1186/1745-6150-1-11)
2. Zambelli F, Pesole G, Pavesi G. 2012 Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinform.* **14**, 225–237. (doi:10.1093/bib/bbs016)
3. Schbath S. 2000 An overview on the distribution of word counts in Markov chains. *J. Comput. Biol.* **7**, 193–201. (doi:10.1089/10665270050081469)
4. Young JA, Johnson JR, Benner C, Yan SF, Chen K, Le Roch KG, Zhou Y, Winzeler EA. 2008 *In silico* discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics* **9**, 70. (doi:10.1186/1471-2164-9-70)
5. Lajoie M, Gascuel O, Lefort V, Bréhélin L. 2012 Computational discovery of regulatory elements in a continuous expression space. *Genome Biol.* **13**, R109. (doi:10.1186/gb-2012-13-11-r109)
6. Bailey TL, Elkan C. 1995 Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* **21**, 51–80. (doi:10.1007/BF00993379)
7. Neuwald AF, Liu JS, Lawrence CE. 1995 Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618–1632. (doi:10.1002/pro.5560040820)
8. Mao L, Mackenzie C, Roh JH, Eraso JM, Kaplan S, Resat H. 2005 Combining microarray and genomic data to predict DNA binding motifs. *Microbiology* **151**, 3197–3213. (doi:10.1099/mic.0.28167-0)
9. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007

Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8. (doi:10.1371/journal.pbio.0050008)

10. Nicolas P *et al.* 2012 Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. *Science* **335**, 1103–1106. (doi:10.1126/science.1206848)
11. Mäder U *et al.* 2016 *Staphylococcus aureus* transcriptome architecture: from laboratory to infection-mimicking conditions. *PLoS Genet.* **12**, e1005962. (doi:10.1371/journal.pgen.1005962)
12. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010 A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141. (doi:10.1101/gr.100396.109)
13. Yan B, Boitano M, Clark TA, Ettwiller L. 2018 SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* **9**, 3676. (doi:10.1038/s41467-018-05997-6)
14. Elemento O, Slonim N, Tavazoie S. 2007 A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* **28**, 337–350. (doi:10.1016/j.molcel.2007.09.027)
15. Bussemaker HJ, Li H, Siggia ED. 2001 Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167. (doi:10.1038/84792)
16. Foat BC, Morozov AV, Bussemaker HJ. 2006 Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149. (doi:10.1093/bioinformatics/btl223)
17. Gruber TM, Gross CA. 2003 Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* **57**, 441–466. (doi:10.1146/annurev.micro.57.030502.090913)
18. Alter O, Brown PO, Botstein D. 2000 Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10 101–10 106. (doi:10.1073/pnas.97.18.10101)
19. Liebermeister W. 2002 Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18**, 51–60. (doi:10.1093/bioinformatics/18.1.51)
20. Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, Zinovyev A. 2017 Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics* **18**, 712. (doi:10.1186/s12864-017-4112-9)
21. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998 Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868. (doi:10.1073/pnas.95.25.14863)
22. Sorek R, Cossart P. 2010 Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* **11**, 9. (doi:10.1038/nrg2695)
23. Toledo-Arana A *et al.* 2009 The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950. (doi:10.1038/nature08080)
24. Wurtzel O, Sesto N, Mellin JR, Karunker I, Edelheit S, Bécavin C, Archambaud C, Cossart P, Sorek R. 2012 Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Mol. Syst. Biol.* **8**, 583. (doi:10.1038/msb.2012.11)
25. Bécavin C *et al.* 2017 Listeriomics: an interactive web platform for systems biology of *Listeria*. *MSystems* **2**, e00186-16. (doi:10.1128/mSystems.00186-16)
26. Albert JH, Chib S. 1993 Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **88**, 669–679. (doi:10.1080/01621459.1993.10476321)
27. Andrieu C, De Freitas N, Doucet A, Jordan MI. 2003 An introduction to MCMC for machine learning. *Mach. Learn.* **50**, 5–43. (doi:10.1023/A:1020281327116)
28. Robert CP, Casella G 2004 *Monte Carlo statistical methods*, 2nd edn. Springer texts in statistics. Berlin, Heidelberg: Springer.
29. Green PJ. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)
30. Geweke J. 2004 Getting it right: joint distribution tests of posterior simulators. *J. Am. Stat. Assoc.* **99**, 799–804. (doi:10.1198/016214504000001132)
31. Hyvarinen A. 1999 Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**, 626–634. (doi:10.1109/72.761722)
32. Novichkov PS *et al.* 2013 RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics* **14**, 745. (doi:10.1186/1471-2164-14-745)
33. Arous S, Buchrieser C, Folio P, Glaser P, Namane A, Hebraud M, Hechard Y. 2004 Global analysis of gene expression in an rpoN mutant of *Listeria monocytogenes*. *Microbiology* **150**, 1581–1590. (doi:10.1099/mic.0.26860-0)
34. Chaturongakul S, Raengpradub S, Palmer ME, Bergholz TM, Orsi RH, Hu Y, Ollinger J, Wiedmann M, Boor KJ. 2011 Transcriptomic and phenotypic analyses identify coregulated, overlapping regulons among PrfA, CtsR, HrcA, and the alternative sigma factors σB, σC, σH, and σL in *Listeria monocytogenes*. *Appl. Environ. Microbiol.* **77**, 187–200. (doi:10.1128/AEM.00952-10)
35. Palmer ME, Chaturongakul S, Wiedmann M, Boor KJ. 2011 The *Listeria monocytogenes* σB regulon and its virulence-associated functions are inhibited by a small molecule. *MBio* **2**, e00241–11. (doi:10.1128/mBio.00241-11)
36. Whiteley AT, Ruhland BR, Edrozo MB, Reniere ML. 2017 A redox-responsive transcription factor is critical for pathogenesis and aerobic growth of *Listeria monocytogenes*. *Infection and immunity*, pp. IAI–00978.
37. Rochat T *et al.* 2012 Genome-wide identification of genes directly regulated by the pleiotropic transcription factor Spx in *Bacillus subtilis*. *Nucleic Acids Res.* **40**, 9571–9583. (doi:10.1093/nar/gks755)
38. Mandin P, Fsihi H, Dussurget O, Vergassola M, Milohanic E, Toledo-Arana A, Lasa I, Johansson J, Cossart P. 2005 VirR, a response regulator critical for *Listeria monocytogenes* virulence. *Mol. Microbiol.* **57**, 1367–1380. (doi:10.1111/j.1365-2958.2005.04776.x)
39. Scortti M, Monzó HJ, Lacharme-Lora L, Lewis DA, Vázquez-Boland JA. 2007 The PrfA virulence regulon. *Microbes Infect.* **9**, 1196–1207. (doi:10.1016/j.micinf.2007.05.007)
40. Fritsch F, Mauder N, Williams T, Weiser J, Oberle M, Beier D. 2011 The cell envelope stress response mediated by the LiaFSRLm three-component system of *Listeria monocytogenes* is controlled via the phosphatase activity of the bifunctional histidine kinase LiaSLm. *Microbiology* **157**, 373–386. (doi:10.1099/mic.0.044776-0)
41. Zhu B, Stülke J. 2018 SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism *Bacillus subtilis*. *Nucleic Acids Res.* **46**, D743–D748. (doi:10.1093/nar/gkx908)
42. Bailey TL, Bodén M, Whitington T, Machanick P. 2010 The value of position-specific priors in motif discovery using MEME. *BMC Bioinf.* **11**, 179. (doi:10.1186/1471-2105-11-179)
43. Narlikar L, Gordân R, Hartemink AJ. 2007 A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.* **3**, e215. (doi:10.1371/journal.pcbi.0030215)
44. Klepper K, Drabløs F. 2010 PriorsEditor: a tool for the creation and use of positional priors in motif discovery. *Bioinformatics* **26**, 2195–2197. (doi:10.1093/bioinformatics/btq357)
45. Ikebata H, Yoshida R. 2015 Repulsive parallel MCMC algorithm for discovering diverse motifs from large sequence sets. *Bioinformatics* **31**, 1561–1568. (doi:10.1093/bioinformatics/btv017)
46. Hermsen R, Tans S. 2006 Transcriptional regulation by competing transcription factor modules. *PLoS Comput. Biol.* **2**, e164. (doi:10.1371/journal.pcbi.0020164)
47. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP. 2004 An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res.* **32**, 6643–6649. (doi:10.1093/nar/gkh1009)
48. Goelzer A, Bekkal Brikci F, Martin-Verstraete I, Noirot P, Bessiéres P, Aymerich S, Fromion V. 2008 Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst. Biol.* **2**, 20. (doi:10.1186/1752-0509-2-20)
49. Magnusson LU, Farewell A, Nyström T. 2005 ppGpp: a global regulator in *Escherichia coli*. *Trends Microbiol.* **13**, 236–242. (doi:10.1016/j.tim.2005.03.008)
50. Goelzer A, Fromion V. 2011 Bacterial growth rate reflects a bottleneck in resource allocation. *Biochim. Biophys. Acta.* **1810**, 978–988. (doi:10.1016/j.bbagen.2011.05.014)
51. Whiteley AT, Pollock AJ, Portnoy DA. 2015 The PAMP c-di-AMP is essential for *Listeria monocytogenes* growth in rich but not minimal media due to a toxic increase in (p)ppGpp. *Cell Host Microbe* **17**, 788–798. (doi:10.1016/j.chom.2015.05.006)