3442–3453 *Nucleic Acids Research, 2015, Vol. 43, No. 7* *Published online 16 March 2015*
doi: 10.1093/nar/gkv170

Distinguishing between productive and abortive promoters using a random forest classifier in *Mycoplasma pneumoniae*

Verónica Lloréns-Rico[1,2], Maria Lluch-Senar[1,2,*] and Luis Serrano[1,2,3,*]

[1]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Dr Aiguader 88, 08003 Barcelona, Spain and [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08010 Barcelona, Spain

Received November 26, 2014; Revised February 09, 2015; Accepted February 22, 2015

ABSTRACT

Distinguishing between promoter-like sequences in bacteria that belong to true or abortive promoters, or to those that do not initiate transcription at all, is one of the important challenges in transcriptomics. To address this problem, we have studied the genome-reduced bacterium *Mycoplasma pneumoniae*, for which the RNAs associated with transcriptional start sites have been recently experimentally identified. We determined the contribution to transcription events of different genomic features: the –10, extended –10 and –35 boxes, the UP element, the bases surrounding the –10 box and the nearest-neighbor free energy of the promoter region. Using a random forest classifier and the aforementioned features transformed into scores, we could distinguish between true, abortive promoters and non-promoters with good –10 box sequences. The methods used in this characterization of promoters can be extended to other bacteria and have important applications for promoter design in bacterial genome engineering.

INTRODUCTION

The breakthroughs in bacterial transcriptomics technologies in recent years have altered our simplistic perspective on transcriptional regulation of prokaryotic genomes, revealing novel and complex layers of transcriptional regulation (1–6). These discoveries have been accompanied by the computational challenge of predicting novel features for the subsequent interpretation of transcriptomic experiments. Promoter prediction is a key computational challenge, necessary for characterizing the transcriptional units of bacterial cells, traditionally known as operons (7). The complexity of these units is greater than expected, usually showing more than one transcription start site (TSS) and therefore more than one associated promoter, as well as different transcription termination sites (5).

A number of algorithms employ sequence features to identify promoter sites in prokaryotic genomes (8–19). Some of these work using position-weight matrices (PWMs) of the different promoter motifs to scan the genome (8,9), whilst others implement hidden Markov models to identify promoters (10). Another group of algorithms has applied machine-learning techniques to promoter recognition, such as support vector machines (SVMs) (11) or artificial neural networks (12–16). Lastly, some algorithms apply a combination of some of the methods above (17). Most of these algorithms rely on the sequence motifs recognized by sigma factors, which guide the RNA polymerase complex to TSSs. In many bacteria, the housekeeping transcription factor sigma 70 binds to two regions upstream of the transcription start site: one region located 10 bp upstream the TSS (–10 box, or Pribnow box) with the consensus motif TANAAT (where N is any base) (20) and another region around 35 bp upstream the TSS bearing the motif TTGACA (–35 box) (21). The spacer between these two boxes may span from 15 to 21 bases (22–24). However, sigma 70 binding motifs are not fully conserved in all bacteria. For example, some *Mycoplasma* species lack the consensus sigma 70 –35 motif and have a degenerate –35 box instead, which is dispensable in some promoters (25,26). In this respect, it has been shown mainly in Gram-positive bacteria like *Bacillus subtilis* that the lack of a –35 motif can be compensated by the presence of an extended –10 box (TG-N-Pribnow) (27,28), which is sufficient to trigger transcription initiation (29,30). This sigma 70 variability limits a broad applicability of the aforementioned methods. Many bacteria have additional sigma factors that bind to different motifs in response to external conditions or perturbations (31,32). A further motif, called the UP element, was later found located upstream of the –35 box; this element consists of an AT-rich tract and interacts with the alpha subunit of the RNA polymerase (33).

*To whom correspondence should be addressed. Tel: +34933160247; Fax: +34933160099; Email: luis.serrano@crg.eu
Correspondence may also be addressed to Maria Lluch-Senar. Tel: +34933160101; Fax: +34933969983; Email: maria.lluch@crg.es

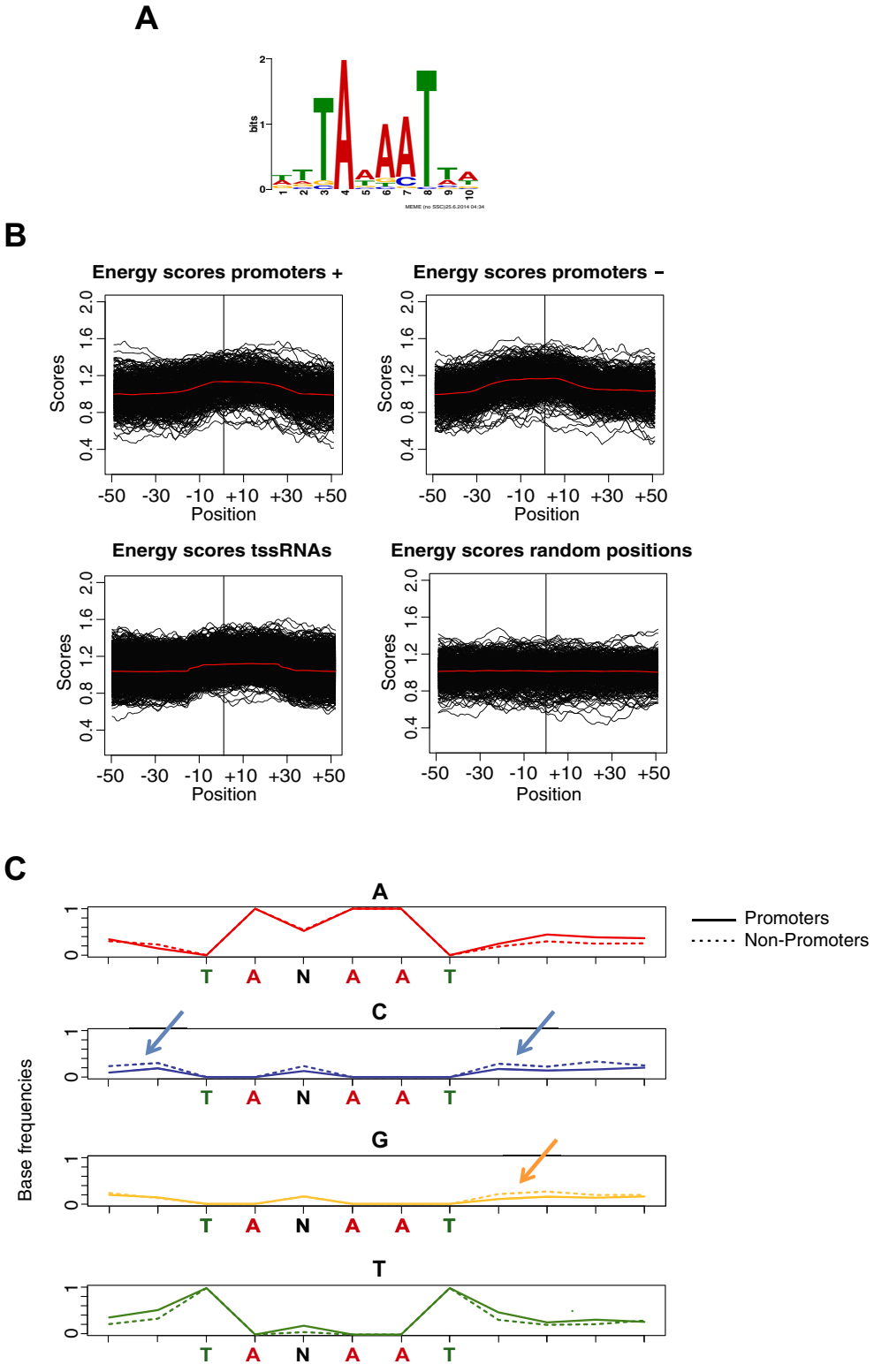

Figure 1. Promoter features. (**A**) Web logo of the Pribnow motif (also termed the –10 box) as determined by MEME analysis (http://meme.nbcr.net/meme/) of experimentally determined TSSs. (**B**) Nearest-neighbor free energy scores. The nearest-neighbor free energy scores of the regions surrounding the experimentally determined promoters in the plus and minus strands are represented (left and central panel, respectively). The right panel corresponds to the scores of random sequences in the genome of *M. pneumoniae*. (**C**) Frequencies of the four different nucleotides on both sides of the –10 box in promoter-like elements (see 'Materials and Methods' section). Dashed lines represent non-promoter sequences bearing the Pribnow motif, while solid lines represent true promoter sequences. Colored arrows indicate the positions in which a nucleotide is significantly overrepresented ($P < 0.001$) in the group of non-promoter sequences.

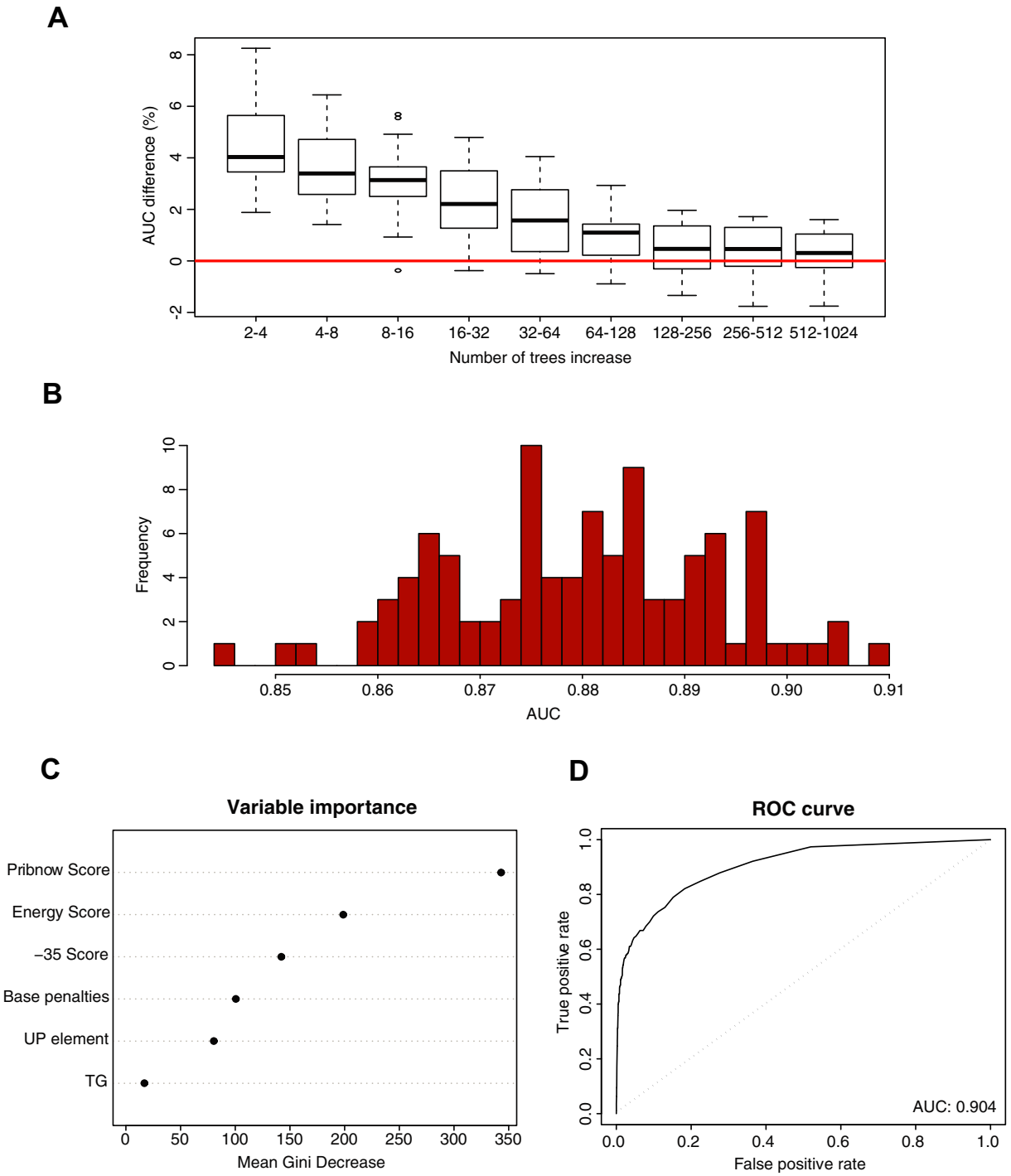

Figure 2. Random forest results. (**A**) Number of trees for the random forest classifier. The boxplots represent the AUC increase when doubling the number of trees, selecting 20 different training and test sets. From 256 trees onward, the average gain in the AUC is close to zero and thus not significant. (**B**) Random forest AUCs. The histogram represent the AUC values obtained with the random forest using 100 different training and testing sets to test the robustness of the method. Values range from 0.85 to 0.92. (**C**) Variable importance as given by the mean decrease in the Gini coefficient, which represents the contribution of each variable to the homogeneity of the results of the random forest. The Pribnow score and the free energy score are the variables with the highest importance in our study. (**D**) ROC curve of the random forest, obtained after training the random forest and testing its performance on a group of promoter-like sequences. The area under the curve obtained is 0.904.

Table 1. Results from the promoter prediction in *M. pneumoniae*

Prediction results	Initial assessment	After re-annotation
709 predicted promoters	133 false positives	44 TSSs or tssRNAs
		89 false positives
	498 TSSs	498 TSSs
	35 tssRNAs	35 tssRNAs
	43 close to an ATG start codon	43 close to an ATG start codon
146 false negatives	146 false negatives	146 false negatives

Results are classified according to the different stages of the analysis, and separate columns for the initial assessment and the re-annotation are shown.

In addition to the specific-sequence motifs at promoters, several studies have shown that the DNA double helix is less stable or more conformationally flexible at promoter regions, facilitating the opening of the double-stranded DNA to accommodate the transcription machinery (34–39). Indeed, it has recently been possible to correlate some of these physical properties to functional features of promoters, such as regulation by transcription factors (40). Based on the above information, some algorithms identify promoters by considering: (i) relative stability of the DNA, as measured by its free energy (41–44); (ii) stress-induced duplex destabilization (SIDD) of the DNA (45) or (iii) DNA bendability or curvature (40,46). As in the sequence-based prediction methods, machine-learning approaches such as neural networks have been used in combination with these properties to identify promoter sequences (47,48). Presumably, structure-based methodologies have a broader scope than sequence-based approaches, as they avoid the limitation of the latter regarding interspecies variability in sigma 70 factors. However, it has been shown that these physical properties are highly correlated with the GC content of the genomes analyzed (49,50), and that the majority of available methods are specific to bacteria with medium-to-high GC content, with lower sensitivity or specificity in other scenarios (44).

A few methods have combined sequence motifs and structural properties of the DNA to predict promoters, mostly in eukaryotes (51) but also for some prokaryotes (52). Although the combination of both types of features improves the classification of promoter sequences, the same restrictions and species-specificity stated above apply.

In general, sequence-based and structure-based methods find that promoter-like motifs usually outnumber the true promoters in bacterial genomes, resulting in large numbers of false positives from these predictions (15,18). Indeed, it has been observed that true promoters tend to be found in clusters of promoter-like motifs that compete for the binding of the sigma 70 factor (19). Surprisingly, it is not uncommon to find that in these clusters, there is at least one promoter-like sequence that holds a better score than the true positive promoter (19). This is an indicator that the sole presence of the consensus motifs is not sufficient to initiate transcription.

We have recently uncovered the existence of a class of small RNAs, termed tssRNAs, found at the TSSs of genes, but also in isolation ('abortive promoters') in both Gram-negative (*Escherichia coli*) and Gram-positive (*Mycoplasma pneumoniae*) bacteria (53). tssRNAs found in isolation present Pribnow box promoter sequences but result in short transcripts (of around 40 bases) that are not associated to transcription of longer RNAs. Furthermore, merging a tssRNA to a gene encoding for a GFP protein does not result in expression of the gene (53). This shows that besides having an appropriate Pribnow box, other requirements have to be fulfilled to accomplish productive transcription (note that we consider promoters that result in RNAs longer than 80 bases as true promoters). So far, there are no methods that distinguish between promoters that result in productive or abortive transcription.

Here, we have constructed a random forest (54) classifier based on both structural and sequence features of *M. pneumoniae* promoters. A random forest is a machine learning technique to perform classification and regression. The major difference with other machine learning approaches used in promoter prediction, such as artificial neural networks or support vector machines, is that random forests are ensemble methods. This means that they compile the output of individual predictors with an overall performance better than the performance of any of the individual predictors (55). In the case of random forests, the individual predictors are decision trees. Each decision tree uses a randomly-drawn subset of the training data (two-thirds), in a process termed 'bagging', and a subset of the variables to construct an individual classifier. Then, the test samples are classified by each individual tree, and the random forest returns the mode of all the trees as the output of the prediction. One of the major advantages of the random forests is that they are highly robust to noise in the data and less prone to overfitting as compared to other machine learning techniques (54). The main disadvantage of the random forests is that they cannot extrapolate values in regression tasks, as they average the output of a number of decision trees.

Mycoplasma pneumoniae is a genome-reduced bacterium (with 816 kb and 689 ORFs) (56) that belongs to the *Mollicutes* class and is characterized by the lack of a cell wall and by having low GC-content genomes. The low GC content results in a genome-wide spread of promoter-like elements, which pose the intriguing question of what determines if a sequence is transcribed. We show that in this bacterium, the majority of experimentally determined TSSs have promoters with a consensus Pribnow box motif of TANAAT (where N is any base) with almost no degeneration. However, having a good Pribnow box is not enough to drive productive transcription. We have looked at different features that could affect transcription like the –35 box, the UP element, the extended –10 motif, the energy of the DNA region containing the Pribnow box and the nature of the bases surrounding the Pribnow, all of which we included in a random forest classifier for promoter site prediction. To our knowledge, this is the first method that uses random forests for

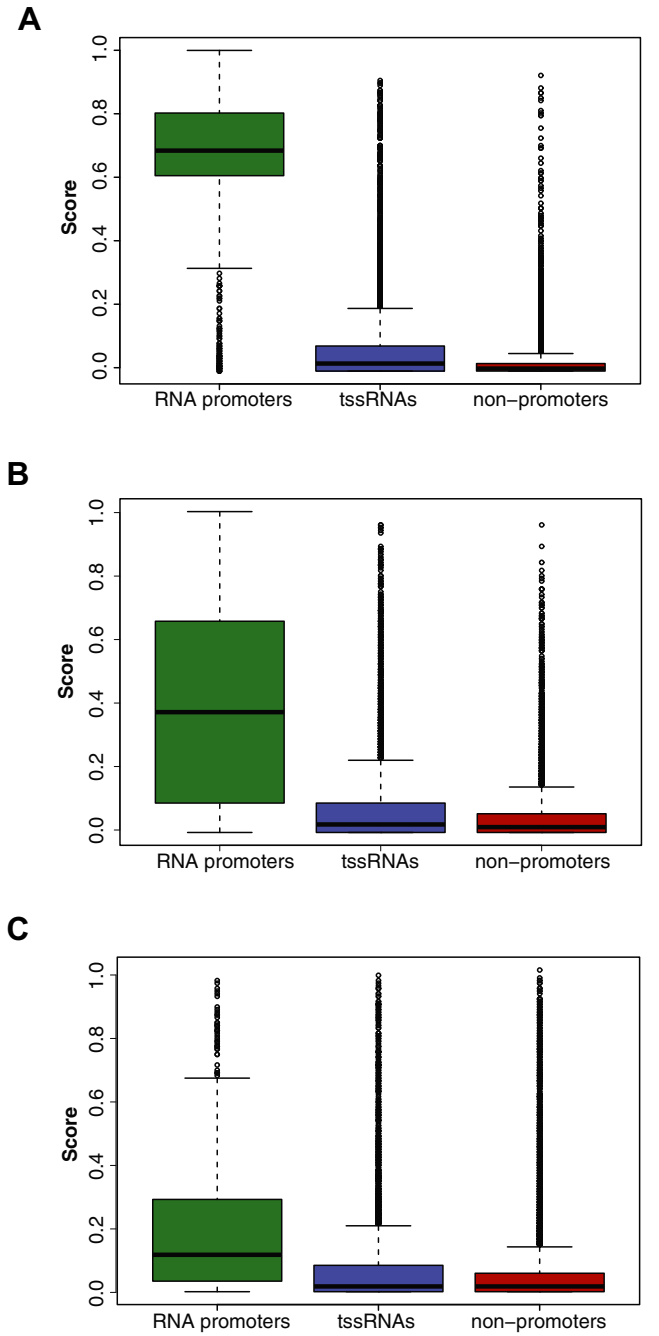

Figure 3. Promoter prediction in *M. pneumoniae*. (**A**) Random forest scores for promoter-like sequences in *M. pneumoniae*. There is a clear separation between true promoters and the other two groups (tssRNA promoters and non-promoter sequences). tssRNAs, despite triggering transcription of short sequences (∼45 bp), are more similar to non-promoter sequences than to true promoters. (**B**) Random forest scores for promoter-like sequences in *M. pneumoniae*, after shuffling the energy scores of all sequences. (**C**) Random forest scores for promoter-like sequences in *M. pneumoniae*, after shuffling the Pribnow scores of all sequences. The removal of structural or Pribnow parameters worsens the prediction of the classifier, resulting in the scores of the three categories being much closer and overlapping with each other.

promoter prediction, and it can distinguish between productive promotors, abortive promoters and promoter-like sequences with no transcriptional activity (non-promoters). We propose that a similar methodology could be applied to other bacteria to classify promoters.

MATERIALS AND METHODS

We constructed a random forest classifier to identify promoter sequences in *M. pneumoniae*. Random forests are ensemble classifiers that use decision trees as individual predictors (Supplementary Figure S1). Each of these decision trees uses a subset of the training data and a subset of the variables to generate an independent classifier. The random forest compiles the results of each individual tree to produce an ensemble output. This random forest classifier uses the following criteria in order to discern between promoter and non-promoter sequences: (i) the –10 box (Pribnow box) motif; (ii) the –35 box motif; (iii) the UP element; (iv) the nearest-neighbor DNA duplex free energies of the bases surrounding the promoter; (v) the presence of G and C bases adjacent to the Pribnow box and (vi) the presence of an extended –10 box (TG-TANAAT). All of these criteria were independently measured and transformed into scores that were used for the RF training.

Pribnow motif score

Six hundred forty-seven TSSs from productive promoters that had been experimentally determined to a single base resolution (53) and manually curated (Supplementary Table S1) were used as the input for the MEME motif finding software (57) to identify the canonical Pribnow box in *M. pneumoniae*. Only TSSs from protein-coding genes and ncRNAs larger than 100 bases were considered. For the analysis, sequences of 16 bp upstream of the TSSs were selected. A unique TANAAT motif was found for 511 of the sequences analyzed (Figure 1A), representing the Pribnow box (e-value of 1.5e-158). The Position Probability Matrix (PPM), which indicates the frequency of each nucleotide at each position of the motif, was therefore used to scan the *M. pneumoniae* genome according to the following equation:

$$-10_{\text{score}}(P) = -\log_2\left(\prod_{i=P}^{i=P+l} N_i\right)$$

where P is the position of the genome being scanned, l is the length of the motif ($l = 6$ for the –10 box) and N_i is the frequency of the nucleotide N in the ith position of the motif, according to the corresponding PPM. The score given to each hexamer was divided by the maximum possible score and then log-normalized. After this normalization, lower values of the normalized Pribnow score will correspond to better promoter sequences.

–35 motif

The –35 box of TTGACA that is observed in other well-studied Gram-positive and Gram-negative bacterial species is rarely found in *M. pneumoniae* promoters (25). A search of the 647 TSSs (19– 25 bp upstream the Pribnow box,

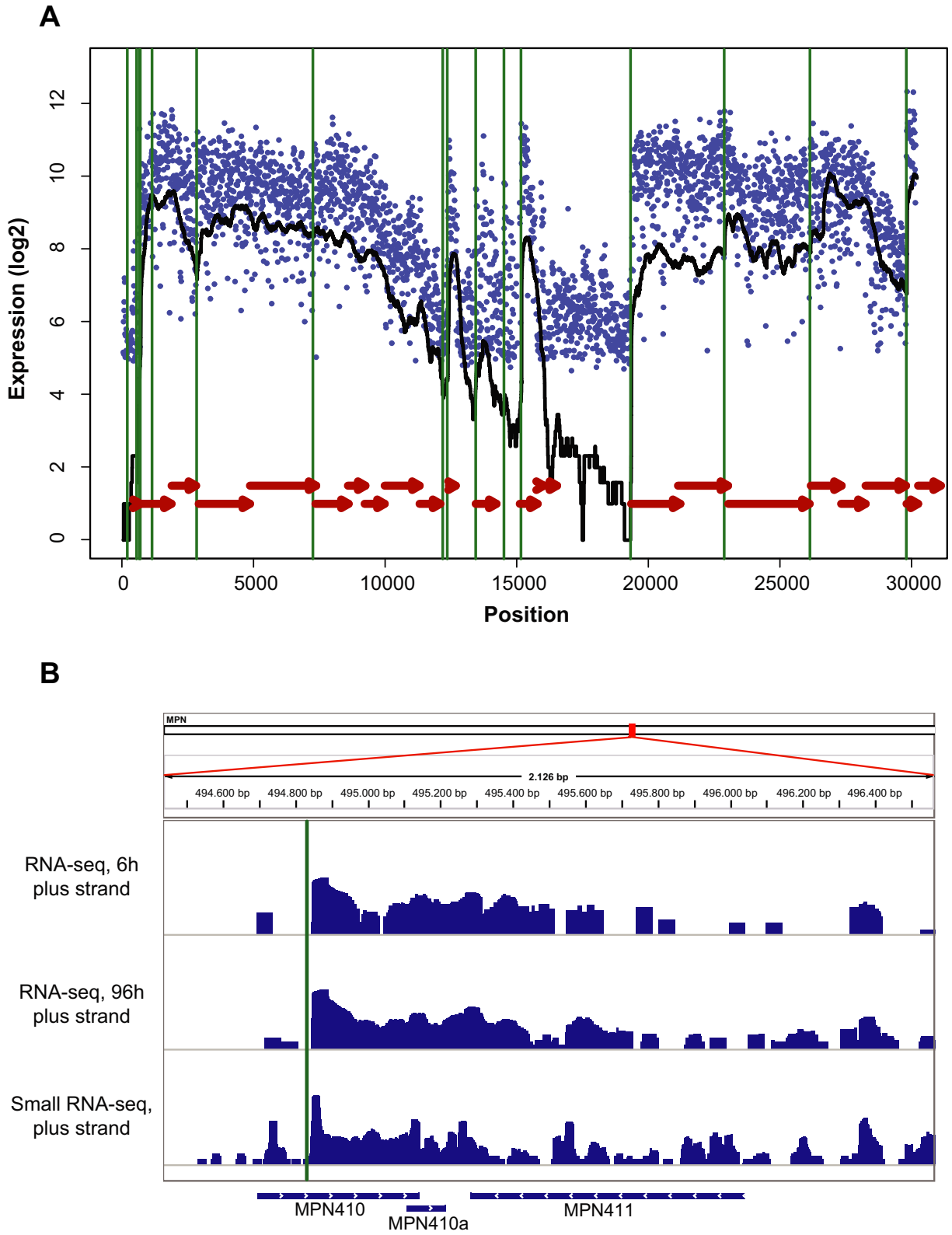

Figure 4. Promoter re-annotation in *M. pneumoniae*. (**A**) Promoters detected in a region of 30 kb of the *M. pneumoniae* genome. Blue dots represent data from RNA tiling arrays in *M. pneumoniae*, the black line represents RNA-seq data, red arrows represent the annotated genes in the plus strand and vertical green lines represent the promoters found by the random forest classifier. These promoters coincide with sharp increases in the values of expression, both in the RNA-seq and the tiling data, validating the prediction. (**B**) Manual curation and re-annotation of promoters in the genome of *M. pneumoniae*. A promoter was found on the positive strand at the position 494 837 (vertical green line), which did not coincide with any annotated TSS (65). The predicted promoter, which is inside gene MPN410, coincides with a sharp increase in the RNA-seq data in the three different experiments represented.

with a spacer of 15– 21 bp (24)) identified the canonical TTGACA motif in only 18 promoters, while a search across all promoter regions did not reveal a significant motif in a MEME. However, by visual inspecting the promoter regions in *M. pneumoniae*, we identified a degenerated –35 box with the consensus sequence TTGANN in 107 promoter regions. Other Gram-positive bacteria, such as *B. subtilis,* also present a similarly degenerated –35 element. Therefore, we used the sigma 70 promoters from *B. subtilis* (58) to derive a PPM for interrogating the genome of *M. pneumoniae* as follows:

$$-35_{\text{score}}(P) = -\log_2\left(\prod_{i=P}^{i=P+l} N_i\right)$$

where P corresponds to the genome position being scanned, l is the length of the motif and N_i is the frequency of the nucleotide N at the position i within the motif. As the spacer between the –35 and the –10 boxes in *M. pneumoniae* promoters is variable, we assigned the best –35 box for each instance of the –10 motif in the possible range (e.g. 19–25 bp upstream the Pribnow box, with a spacer of 15–21 bp (24)). As noted above, lower values of this score will correspond to better –35 boxes.

UP element

The presence of the UP element was considered for the random forest as the fraction of ATs located in the –45 region. For this purpose, we considered the fraction of ATs in the region between 30 and 45 bp upstream of the Pribnow box. This fraction was log-normalized to convert it into a score for the random forest classifier.

Free energy

The matrix of nearest-neighbor free energies of the different pairs of consecutive bases (obtained from (59)) was scaled by dividing each value by the maximum absolute energy, to have all of them ranging from 0 to 1 (Supplementary Table S2). The free energy score is the $-\log_2$ of the product of the scaled nearest-neighbor free energies of a promoter 60 bases window (located between 35 bases upstream the –10 box and 25 bases downstream the –10 box), and it was then normalized over the mean of all the genome (Figure 1B).

$$\Delta G^0_{\text{score}}(P) = -\log_2\left(\prod_{i=P-35}^{i=P+25} \Delta G^0_{\text{i,i+1}}\right)$$

where $\Delta G^0_{i,i+1}$ represents the nearest-neighbor free energy of the bases ith and $(i + 1)$th and P corresponds to the genome base being scanned. Higher scores for this parameter represent more favorable energies for the separation of the double-stranded DNA.

Base penalties

A preliminary analysis of all the instances of the TANAAT motif in the *M. pneumoniae* genome was performed, comparing the sequences that gave rise to promoters versus the ones that were not associated to any transcriptional event. This analysis showed that, in experimentally determined promoters, there are certain biases toward the exclusion of G and C bases in the immediate vicinity of the –10 box (e.g. two bases before and three bases after; Figure 1C). Therefore, we included the log-normalized fraction of G and C nucleotides in the vicinity of the Pribnow box as an independent criterion for the random forest.

Extended Pribnow box

There is one remarkable exception regarding the biases against guanidine and cytosine nucleotides next to the –10 box: the presence of a guanidine nucleotide before the Pribnow box is not depleted, as it can exist as a part of the so-called 'extended Pribnow motif' of TG-N-Pribnow (60). The TG pair located immediately upstream of the Pribnow box was regarded as a categorical variable in the RF, represented as 0 or 1 for the presence or absence of each of the pairs next to the Pribnow motif, respectively.

Construction of the RF classifier

Once all the criteria had been defined, we scanned all positions of the *M. pneumoniae* genome to score them according to each criterion. We scanned the genome of *M. pneumoniae* using the Pribnow sequence probability matrix defined above using FIMO (61). This tool allows all the occurrences of a given motif in a DNA or protein sequence to be identified by evaluating their similitude with the consensus one. 81 087 sequences were found to match the Pribnow element with a $P < 0.05$. We mapped each known, experimentally-determined TSS to its corresponding Pribnow box and found that only 28 of 647 TSSs did not have any associated –10 box. To remove some of the 81 087 sequences, we looked at the top 90% of the 619 sequences that passed the filter of $P < 0.05$. These sequences had an associated Pribnow motif with a $P \leq 0.00824$. We used this value as a threshold to filter non-promoter sequences and obtained a list of 14 663 sequences (14 374 plus the true promoters that did not pass the filter of $P < 0.00824$). These promoter-like sequences fall into three different categories: RNA promoters that give rise to RNAs longer than 80 bp (647); tssRNA promoters that produce abortive transcripts shorter than 45 bp (5379; (53)); and non-promoter sequences, promoter-like sequences not associated to any transcriptional event (8438).

From this set of 14 663, abortive promoters giving rise to independent tssRNAs (5379; (53)) were not included in the random forest classifier as they may have features of both RNA promoters and non-promoter sequences. Of the remaining 9284 sequences, we selected a random set using 70% of the hits, formed by 453 true promoters (positive set) and 5906 non-promoter sequences (negative set). In order to further constrain the method, we added 5000 sequences selected randomly from the *M. pneumoniae* genome to the negative set.

The selection of the number of trees in the random forest was performed empirically. We followed the approach described in Oshiro *et al.* (62) and generated random forests of 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024 trees. For each number of trees, we generated 20 different random forests

by choosing different subsets to train the method, and we calculated the receiver-operating characteristic (ROC) curves specifying the sensitivity and the specificity of the classifier. ROC curves display the true positive rate of the method (or sensitivity, calculated as the number of true positives divided by the total number of positives in the sample) versus the false positive rate (1 − specificity, calculated as the number of false positives divided by the total number of negatives in the sample). These values are plotted using different thresholds and thus form a curve. The performance of the classifier is assessed by determining the area under the curve (AUC). AUC values can range from 0.5 (the method does not perform better than random) and 1 (the method classifies all the samples perfectly with no mispredictions). Accordingly, larger areas signify better classifiers.

RESULTS

Promoter features

Taking advantage of the experimentally-determined 647 TSSs, we derived a promoter-scoring matrix for the Pribnow box after a motif search using the MEME suite (Figure 1A; see 'Materials and Methods' section) (63). Furthermore, we determine the nearest-neighbor free energy for these TSSs by using a 60-bp window (of 35 bases upstream to 19 bases downstream of the Pribnow box). We observed that true promoters have a specific average profile, while tssRNAs have a less pronounced one and non-promoter sequences are flat (Figure 1B). From these analyses, we derived two scoring functions (see 'Materials and Methods' section) and used them to scan the whole genome of *M. pneumoniae*. Only using the Pribnow score to identify promoters led to significant overprediction (14 374 sequences with a Pribnow box of $P < 0.00824$). Apart from experimentally-determined promoters, we found that some of the predicted hits coincided with non-productive promoters that produce independent short abortive transcripts (tssRNAs, (53)), while for others, we could not find any evidence of transcriptional activity in the previous-published microarray and deep sequencing data obtained under many different perturbations (non-promoters (1,53)) (Supplementary Figure S2A). Using a set of 14 663 promoter-like sequences (comprising the 14 374 sequences plus the TSSs above the threshold, see 'Materials and Methods' section), we found that RNA promoters have on average slightly better Pribnow box scores (Supplementary Figure S2B) and energy scores (Supplementary Figure S2C) than tssRNAs and non-promoter sequences. However, although the differences were statistically significant between the real promoters and the two other categories (Mann–Whitney U-test, $P < 2.2 \times 10^{-16}$ for both the Pribnow and energy scores), there is a large overlap among the three classes described.

Sequence properties of the bases surrounding the promoter stabilize the polymerase complex

To test whether promoter prediction could be improved and the number of false positives reduced, we analyzed four other different criteria known to be involved in promoter strength in bacteria: the –35 box, the frequency of AT bases in the –45 region (also termed the UP element), the biases toward excluding certain nucleotides next to the –10 motif and the presence of a TG dinucleotide that generates an extended Pribnow box (see 'Materials and Methods' section). We then compared the discrimination against G and C bases before and after the Pribnow box in productive promoters versus non-promoters (Figure 1C). For the extended –10 box motif, no significant enrichment was observed in true promoters (it was associated 139 times to real promoters, and 3438 times to tssRNAs and non-promoter sequences). Nonetheless, we decided to keep this parameter, as it has been described that the presence of the TG upstream the Pribnow box renders the –35 motif unnecessary in Gram-positive bacteria (27,28). For the UP element, we observed an AT enrichment in real promoters (Supplementary Figure S2D). Finally, the –35 scores of true promoter sequences are slightly better than those of non-promoters (Mann–Whitney U-test, $P = 2.41 \times 10^{-5}$), deeming this factor important for the construction of the random forest despite the degeneration of the consensus –35 motif. Therefore, it seems that true promoters have other features in addition to a good Pribnow box and a favorable energy profile as compared with non-promoter sequences with similar Pribnow boxes.

Random forest classifier

To determine if the discrimination between RNA promoters, tssRNA promoters and non-promoter sequences with good Pribnow boxes could be improved, we built a random forest classifier using the four criteria discussed above (e.g. the extended –10 bp, the –35 and the UP elements and the biases against G and C bases close to the Pribnow motif) in addition to the Pribnow and DNA duplex. To choose an adequate number of trees, we first analyzed the increase of the AUC versus the increase in the number of trees (Figure 2A) and set the number of trees when we reached a plateau (128 trees), as any further increase in the number of trees did not result in a significant AUC gain. To determine the specificity and sensitivity of the generated random forest, we used the promoter-like sequence set, from which we excluded the sequences used to train the random forest classifier (see 'Materials and Methods' section). We tested the random forest by choosing 100 different subsets for the training and obtained very similar results in terms of sensitivity and specificity (Figure 2B).

From the six parameters used, the most important one is the Pribnow box score, followed by the nearest-neighbor free energy score (Figure 2C). Other sequence features, such as the –35 box, the UP element and the biases toward excluding G and C nucleotides in the vicinity of the –10 box, had smaller contributions to the predictor. The presence of the TG motif extending the –10 box had little value for the classifier.

The performance of the random forest classifier was determined twofold. Firstly, we determined the out-of-bag (OOB) error estimate. The random forest classifier uses two-thirds of the training data (sampled with replacement) to build each tree ('bagging'). Therefore, on average, each sample will be excluded from one-third of the trees. The random forest uses the set of trees in which a sample has been left 'out of the bag' to classify it. The total proportion of mis-

predictions is the OOB error estimate. The reported OOB error was 3.43%. Secondly, we analyzed the predictive value of the random forest classifier on a test set formed by the promoter-like sequences not used for training the classifier and random sequences (see 'Materials and Methods' section). For the test set, we calculated the ROC curve to assess the performance of the method (Figure 2D). The scores used to calculate the ROC curve represent the frequency of trees in the RF that predict a sequence as a promoter and were used as the different thresholds to construct the ROC curve. The AUC corresponding to the test set was 0.904.

Promoter prediction in *M. pneumoniae*

In order to choose for a score threshold to reliably determine promoters in *M. pneumoniae*, we studied the distributions of the output scores of the random forest prediction. For this purpose, we analyzed the scores of the 14 463 promoter-like sequences (see 'Materials and Methods' section). These sequences were grouped into three categories: real promoters (experimentally determined and annotated in (53)), independent tssRNAs and non-promoter sequences (e.g. not associated to transcription events). The results showed that experimentally-determined promoters are separate from the other two groups, whilst tssRNAs, despite having –10 boxes similar to those of promoters, are more similar to non-promoter sequences (Figure 3A). Since the energy and Pribnow scores are the more important parameters, we determined if removing one of them would still allow the random forest to efficiently discriminate between the three promoter categories. For this, we randomly shuffled the energy or Pribnow values of the 14 463 promoter-like sequences and tested the performance of the random forest classifier with this new dataset. The separation of the different categories worsened after the resampling of energy and Pribnow values, indicating that both are essential for productive promoters (Figure 3B and C).

With these results, we set the score threshold for promoter identification to 0.6 in order to minimize the number of false positives obtained (211 non-annotated promoters) while retaining the largest number of real promoters possible (498 of 647).

Re-annotation of promoter sequences in *M. pneumoniae*

Using the 0.6 cutoff, the random forest classifier was able to find 709 putative promoters in the genome of *M. pneumoniae* (Supplementary Table S3). Of these, 576 coincide with steep changes both in RNA-seq and tiling data, indicating a TSS in the vicinity of the predicted promoter (Figure 4A).

Out of the 709 promoters predicted, 498 were found to be at a distance closer than 25 bp to an annotated TSS (70.15%, Table 1). Of the remaining 211 hits, 35 corresponded to annotated independent tssRNAs and 43 were located at less than 200 bp from a starting ATG codon. Ten of these predicted promoters were intragenic, while the remaining 33 were found in intergenic regions. This latter group could represent internal promoters at operons with non-annotated TSSs.

The remaining 133 false positives were manually curated by manual inspection of different RNA-seq experiments (53) and visualizing them on the Integrative Genomics Viewer (IGV) (64). We found that 44 of these are associated to non-annotated TSSs (either corresponding to tssRNAs or longer transcripts; Figure 4B) (65). With this re-annotation of new promoter sites, only 89 of the total 709 predicted promoters (12.55%) are not associated to any transcriptional event (false positives) under the experimental conditions tested (Table 1).

We also analyzed the TSSs of known full-length transcripts for which the promoter was not predicted by our approach (false negatives). Out of the initial 647 curated TSSs, 501 had one associated promoter (and three had two promoters and TSSs). We studied the properties of the remaining 146 TSSs (false negatives). For each of these TSSs, we selected the putative promoter as the position with the highest random forest score up to 25 bp upstream of the TSS. The general scores for these promoters followed a uniform distribution between 0 and 0.6. The false negatives could be divided in two groups: one group with good –10 element scores (<15; 63 sequences) and one group with bad –10 scores (≥ 15; 83 sequences). It is very likely that the latter group was not identified by the random forest classifier because of their poor Pribnow scores (note that this parameter is the most important for the classifier). For the former group, we investigated the possible reasons why these sequences were not identified as real promoters. We found that energies were slightly worse than energies of real promoters in this group (t-test, $P = 1.77 \times 10^{-4}$). Also, the presence of G and C nucleotides next to the Pribnow box was higher in these sequences (t-test, $P = 0.01$). Regarding the set of false positives with worse Pribnow scores there was not significantly difference to the set of true positives regarding this parameter (t-test, $P = 0.54$). This could point to a compensatory mechanism for promoters with worse Pribnow, favoring the stabilization of the open loop.

Out of the 146 false negatives, 70 (47.94%) are ncRNAs, representing a large enrichment related to the number of ncRNAs in the genome of *M. pneumoniae* (Fisher's exact test, $P = 1.184 \times 10^{-5}$). This could suggest that some of them could be tssRNAs and therefore non-productive promoters. Finally, while we observed no significant differences in the gene length of the true positives and false negatives (Mann–Whitney U-test, $P = 0.07$), we did see differences in their gene expression: false negatives had lower expression levels than true positive promoters in the exponential phase (Mann–Whitney U-test, $P = 0.001$) (Supplementary Figure S2E).

DISCUSSION

We constructed a random forest classifier for promoter prediction based on six different parameters, both structural and sequence-based. We trained this classifier with a subset of known promoters from the low GC-content bacterium *M. pneumoniae*, as well as with negative sets of non-promoter sequences. From the six parameters used to evaluate promoter sequences, the –10 motif score and the nearest-neighbor free energy score were the most important ones to discern between true promoters and non-promoter sequences. Therefore, we conclude that both sequence-based

and biophysical parameters are determinants of an optimal promoter prediction.

By applying this random forest to the whole set of promoter-like sequences in *M. pneumoniae*, which comprises true promoter sequences, promoters of independent tssRNAs (not associated to full-length transcripts) and non-promoters (i.e. promoter-like sequences that are not associated to any transcriptional event), we observed that the scores reported for tssRNAs were more similar to those of false promoters and deviated from the values of real promoters. Previous studies on tssRNAs in bacteria (53) found that tssRNAs that are associated to full-length transcripts had sequence features similar to independent tssRNAs. By adding the parameter of DNA structural properties, we were able to make a significant distinction between both types of sequences. The fact that promoters of independent tssRNAs present lower scores than promoters of full-length transcripts points to a failure in one of the steps of transcription initiation. It was shown that while these sequences are able to recruit the RNA polymerase holoenzyme, they are for some reason unable to produce long transcripts (53). Together, these results suggest that the higher stability of the sequences surrounding these promoters prevents the double helix from correctly unwinding to facilitate transcription elongation. In the case of promoter-like sequences that are not associated to any transcription event, these sites must be unable to recruit the RNA polymerase and initiate transcription.

Given the elevated AT content of *Mycoplasma* genomes, it is not rare that non-productive promoter-like sequences arise randomly due to point mutations, especially in the core –10 box, which has a low information content (66). If these sequences are not deleterious, they will not be selected against and will rather accumulate in the genome. Some sequences will not be able to recruit the RNA polymerase to initiate transcription, but others will, giving rise to independent tssRNAs or even in some cases to longer transcripts such as ncRNAs (67).

By using a combined approach, we now show that it is possible to distinguish promoters of full-length transcripts from abortive and non-productive promoters. These findings highlight that an adequate structural context is essential for the complete assembly of the RNA polymerase complex and for DNA unwinding to initiate transcription; thus, the structural properties also need to be considered. Furthermore, the methods used here can be applied to other bacteria, provided that appropriate training sets are available. Therefore, this work may aid in the species-specific design of synthetic promoters, allowing the researcher to predict beforehand whether or not the designed sequence will give rise to a productive transcription event.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Union Seventh Framework Programme (FP7/2007–2013), through the European Research Council [232913]; Fundación Botín, the Spanish Ministry of Economy and Competitiveness [BIO2007-61762]; National Plan of R + D + i; ISCIII – Subdirección General de Evaluación y Fomento de la Investigación [PI10/01702]; European Regional Development Fund (ERDF) (to the ICREA Research Professor L.S.]; Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017′ [SEV-2012-0208]. Funding for open access charge: European Union Seventh Framework Programme (FP7/2007–2013), through the European Research Council [232913]; Fundación Botín, the Spanish Ministry of Economy and Competitiveness [BIO2007-61762]; National Plan of R + D + i; ISCIII – Subdirección General de Evaluación y Fomento de la Investigación [PI10/01702]; European Regional Development Fund (ERDF) (to the ICREA Research Professor L.S.]; Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017′ [SEV-2012-0208].

Conflict of interest statement. None declared.

REFERENCES

1. Guell,M., van Noort,V., Yus,E., Chen,W.H., Leigh-Bell,J., Michalodimitrakis,K., Yamada,T., Arumugam,M., Doerks,T., Kuhner,S. *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–1271.
2. Nicolas,P., Mader,U., Dervyn,E., Rochat,T., Leduc,A., Pigeonneau,N., Bidnenko,E., Marchadier,E., Hoebeke,M., Aymerich,S. *et al.* (2012) Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science*, **335**, 1103–1106.
3. Li,S., Dong,X. and Su,Z. (2013) Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. *BMC Genomics*, **14**, 520.
4. Koonin,E.V. and Wolf,Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, **36**, 6688–6719.
5. Guell,M., Yus,E., Lluch-Senar,M. and Serrano,L. (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.*, **9**, 658–669.
6. Passalacqua,K.D., Varadarajan,A., Ondov,B.D., Okou,D.T., Zwick,M.E. and Bergman,N.H. (2009) Structure and complexity of a bacterial transcriptome. *J. Bacteriol.*, **191**, 3203–3211.
7. Jacob,F., Perrin,D., Sanchez,C. and Monod,J. (1960) Operon: a group of genes with the expression coordinated by an operator. *C. R. Hebd. Seances Acad. Sci.*, **250**, 1727–1729.
8. Li,Q.Z. and Lin,H. (2006) The recognition and prediction of sigma70 promoters in Escherichia coli K-12. *J. Theor. Biol.*, **242**, 135–141.
9. Todt,T.J., Wels,M., Bongers,R.S., Siezen,R.S., van Hijum,S.A. and Kleerebezem,M. (2012) Genome-wide prediction and validation of sigma70 promoters in Lactobacillus plantarum WCFS1. *PLoS One*, **7**, e45097.
10. Jarmer,H., Larsen,T.S., Krogh,A., Saxild,H.H., Brunak,S. and Knudsen,S. (2001) Sigma A recognition sites in the Bacillus subtilis genome. *Microbiology*, **147**, 2417–2424.
11. Gordon,J.J., Towsey,M.W., Hogan,J.M., Mathews,S.A. and Timms,P. (2006) Improved prediction of bacterial transcription start sites. *Bioinformatics*, **22**, 142–148.
12. Demeler,B. and Zhou,G.W. (1991) Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res.*, **19**, 1593–1599.
13. de Avila,E.S.S., Echeverrigaray,S. and Gerhardt,G.J. (2011) BacPP: bacterial promoter prediction–a tool for accurate sigma-factor specific assignment in enterobacteria. *J. Theor. Biol.*, **287**, 92–99.
14. Burden,S., Lin,Y.X. and Zhang,R. (2005) Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences. *Bioinformatics*, **21**, 601–607.
15. Horton,P.B. and Kanehisa,M. (1992) An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucleic Acids Res.*, **20**, 4331–4338.
16. Kalate,R.N., Tambe,S.S. and Kulkarni,B.D. (2003) Artificial neural networks for prediction of mycobacterial promoter sequences. *Comput. Biol. Chem.*, **27**, 555–564.

17. de Jong,A., Pietersma,H., Cordes,M., Kuipers,O.P. and Kok,J. (2012) PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics*, **13**, 299.
18. Hertz,G.Z. and Stormo,G.D. (1996) Escherichia coli promoter sequences: analysis and prediction. *Methods Enzymol.*, **273**, 30–42.
19. Huerta,A.M. and Collado-Vides,J. (2003) Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.*, **333**, 261–278.
20. Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 784–788.
21. Schaller,H., Gray,C. and Herrmann,K. (1975) Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 737–741.
22. Stefano,J.E. and Gralla,J.D. (1982) Spacer mutations in the lac ps promoter. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 1069–1072.
23. Aoyama,T., Takanami,M., Ohtsuka,E., Taniyama,Y., Marumoto,R., Sato,H. and Ikehara,M. (1983) Essential structure of E. coli promoter: effect of spacer length between the two consensus sequences on promoter function. *Nucleic Acids Res.*, **11**, 5855–5864.
24. Hawley,D.K. and McClure,W.R. (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res.*, **11**, 2237–2255.
25. Weiner,J. 3rd, Herrmann,R. and Browning,G.F. (2000) Transcription in Mycoplasma pneumoniae. *Nucleic Acids Res.*, **28**, 4488–4496.
26. Halbedel,S., Eilers,H., Jonas,B., Busse,J., Hecker,M., Engelmann,S. and Stulke,J. (2007) Transcription in Mycoplasma pneumoniae: analysis of the promoters of the ackA and ldh genes. *J. Mol. Biol.*, **371**, 596–607.
27. Sabelnikov,A.G., Greenberg,B. and Lacks,S.A. (1995) An extended -10 promoter alone directs transcription of the DpnII operon of Streptococcus pneumoniae. *J. Mol. Biol.*, **250**, 144–155.
28. Djordjevic,M. (2011) Redefining Escherichia coli sigma(70) promoter elements: -15 motif as a complement of the -10 motif. *J. Bacteriol.*, **193**, 6305–6314.
29. Voskuil,M.I. and Chambliss,G.H. (1998) The -16 region of Bacillus subtilis and other gram-positive bacterial promoters. *Nucleic Acids Res.*, **26**, 3584–3590.
30. Voskuil,M.I., Voepel,K. and Chambliss,G.H. (1995) The -16 region, a vital sequence for the utilization of a promoter in Bacillus subtilis and Escherichia coli. *Mol. Microbiol.*, **17**, 271–279.
31. Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.*, **57**, 441–466.
32. Kazmierczak,M.J., Wiedmann,M. and Boor,K.J. (2005) Alternative sigma factors and their roles in bacterial virulence. *Microbiol. Mol. Biol. Rev.*, **69**, 527–543.
33. Ross,W., Gosink,K.K., Salomon,J., Igarashi,K., Zou,C., Ishihama,A., Severinov,K. and Gourse,R.L. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*, **262**, 1407–1413.
34. Benham,C.J. (1992) Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.*, **225**, 835–847.
35. Zhabinskaya,D. and Benham,C.J. (2012) Theoretical analysis of competing conformational transitions in superhelical DNA. *PLoS Comput. Biol.*, **8**, e1002484.
36. Margalit,H., Shapiro,B.A., Nussinov,R., Owens,J. and Jernigan,R.L. (1988) Helix stability in prokaryotic promoter regions. *Biochemistry*, **27**, 5179–5188.
37. Lisser,S. and Margalit,H. (1994) Determination of common structural features in Escherichia coli promoters by computer analysis. *Eur. J. Biochem.*, **223**, 823–830.
38. Vollenweider,H.J., Fiandt,M. and Szybalski,W. (1979) A relationship between DNA helix stability and recognition sites for RNA polymerase. *Science*, **205**, 508–511.
39. Olivares-Zavaleta,N., Jauregui,R. and Merino,E. (2006) Genome analysis of Escherichia coli promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. *Genomics*, **87**, 329–337.
40. Meysman,P., Collado-Vides,J., Morett,E., Viola,R., Engelen,K. and Laukens,K. (2014) Structural properties of prokaryotic promoter regions correlate with functional features. *PLoS One*, **9**, e88717.
41. Rangannan,V. and Bansal,M. (2009) Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. *Mol. Biosyst.*, **5**, 1758–1769.
42. Rangannan,V. and Bansal,M. (2007) Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J. Biosci.*, **32**, 851–862.
43. Kanhere,A. and Bansal,M. (2005) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, **6**, 1.
44. Rangannan,V. and Bansal,M. (2010) High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics*, **26**, 3043–3050.
45. Wang,H., Noordewier,M. and Benham,C.J. (2004) Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. *Genome Res.*, **14**, 1575–1584.
46. Mallios,R.R., Ojcius,D.M. and Ardell,D.H. (2009) An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis sigma66 promoters. *BMC Bioinformatics*, **10**, 271.
47. Bland,C., Newsome,A.S. and Markovets,A.A. (2010) Promoter prediction in E. coli based on SIDD profiles and Artificial Neural Networks. *BMC Bioinformatics*, **11**(Suppl. 6), S17.
48. Askary,A., Masoudi-Nejad,A., Sharafi,R., Mizbani,A., Parizi,S.N. and Purmasjedi,M. (2009) N4: a precise and highly sensitive promoter predictor using neural network fed by nearest neighbors. *Genes Genet. Syst.*, **84**, 425–430.
49. Bustamante,C., Smith,S.B., Liphardt,J. and Smith,D. (2000) Single-molecule studies of DNA mechanics. *Curr. Opin. Struct. Biol.*, **10**, 279–285.
50. Rief,M., Clausen-Schaumann,H. and Gaub,H.E. (1999) Sequence-dependent mechanics of single DNA molecules. *Nat. Struct. Biol.*, **6**, 346–349.
51. Ohler,U., Niemann,H., Liao,G. and Rubin,G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17**(Suppl. 1), S199–206.
52. Wang,H. and Benham,C.J. (2006) Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*, **7**, 248.
53. Yus,E., Guell,M., Vivancos,A.P., Chen,W.H., Lluch-Senar,M., Delgado,J., Gavin,A.C., Bork,P. and Serrano,L. (2012) Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.*, **8**, 585.
54. Breiman,L. (2001) Random Forest. *Mach. Learn.*, **45**, 5–32.
55. Dietterich,T.G. (2000) *Multiple Classifier Systems – First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings*. Springer, Berlin, Heidelberg, pp. 1–15.
56. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res.*, **24**, 4420–4449.
57. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
58. Ishii,T., Yoshida,K., Terai,G., Fujita,Y. and Nakai,K. (2001) DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Res.*, **29**, 278–280.
59. SantaLucia,J. Jr and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
60. Graves,M.C. and Rabinowitz,J.C. (1986) In vivo and in vitro transcription of the Clostridium pasteurianum ferredoxin gene. Evidence for 'extended' promoter elements in gram-positive organisms. *J. Biol. Chem.*, **261**, 11409–11415.
61. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
62. Thais Mayumi Oshiro,P.S.P. and José Augusto Baranausakas. (2012) In: Perner,P (ed). *Machine Learning and Data Mining in Pattern Recognition*. Springer-Verlag, pp. 154–168.
63. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
64. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

65. Wodke,J.A., Alibes,A., Cozzuto,L., Hermoso,A., Yus,E., Lluch-Senar,M., Serrano,L. and Roma,G. (2014) MyMpn: a database for the systems biology model organism Mycoplasma pneumoniae. *Nucleic Acids Res.*, **43**, D618–D623.
66. Mendoza-Vargas,A., Olvera,L., Olvera,M., Grande,R., Vega-Alvarado,L., Taboada,B., Jimenez-Jacinto,V., Salgado,H., Juarez,K., Contreras-Moreira,B. *et al.* (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One*, **4**, e7526.
67. Raghavan,R., Sloan,D.B. and Ochman,H. (2012) Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio*, **3**, doi:10.1128/mBio.00156-12.