doi:10.1016/S0022-2836(03)00034-2 *J. Mol. Biol.* (2003) **326**, 1361–1372

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

*Rpo*D Promoters in *Campylobacter jejuni* Exhibit a Strong Periodic Signal Instead of a −35 Box

**Lise Petersen[1,2]*, Thomas S. Larsen[1], David W. Ussery[1]
Stephen L. W. On[2] and Anders Krogh[3]**

[1]*Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby Denmark*

[2]*Department of Bacteriology Danish Veterinary Institute DK-1790 Copenhagen Denmark*

[3]*Bioinformatics Center University of Copenhagen DK-2100 Copenhagen Denmark*

**Corresponding author*

We have used a hidden Markov model (HMM) to identify the consensus sequence of the *Rpo*D promoters in the genome of *Campylobacter jejuni*. The identified promoter consensus sequence is unusual compared to other bacteria, in that the region upstream of the TATA-box does not contain a conserved −35 region, but shows a very strong periodic variation in the AT-content and semi-conserved T-stretches, with a period of 10–11 nucleotides. The TATA-box is in some, but not all cases, preceded by a TGx, similar to an extended −10 promoter.

We predicted a total of 764 presumed *Rpo*D promoters in the *C. jejuni* genome, of which 654 were located upstream of annotated genes. A similar promoter was identified in *Helicobacter pylori*, a close phylogenetic relative of *Campylobacter*, but not in *Escherichia coli*, *Vibrio cholerae*, or six other Proteobacterial genomes, or in *Staphylococcus aureus*. We used upstream regions of high confidence genes as training data ($n = 529$, for the *C. jejuni* genome). We found it necessary to limit the training set to genes that are preceded by an intergenic region of >100 bp or by a gene oriented in the opposite direction to be able to identify a conserved sequence motif, and ended up with a training set of 175 genes. This leads to the conclusion that the remaining genes (354) are more rarely preceded by a (*Rpo*D) promoter, and consequently that operon structure may be more widespread in *C. jejuni* than has been assumed by others.

Structural predictions of the regions upstream of the TATA-box indicates a region of highly curved DNA, and we assume that this facilitates the wrapping of the DNA around the RNA polymerase holoenzyme, and offsets the absence of a conserved −35 binding motif.

Keywords: Campylobacter jejuni; promoter; hidden Markov model; DNA periodicity; *Rpo*D

Introduction

Campylobacter jejuni is a frequently reported human gastrointestinal pathogen, with the number of reported cases currently exceeding 80 per 100,000 inhabitants in several developed countries.[1] Moreover, the incidence of infection has been increasing for over a decade, for reasons unknown. The main route of transmission is assumed to be food-borne, but pets and contaminated water may also serve as sources of infection. Molecular epidemiological studies have implied that certain clones may be more pathogenic than others,[2,3] but such hypotheses are difficult to test, since the mechanisms by which *C. jejuni* causes disease remain largely unclear. A better insight into gene organization, function, and regulation in *C. jejuni* is clearly desirable to provide an understanding of its fundamental biology, and for possible exploitation in novel rational control strategies.

The complete genome sequence of one *C. jejuni* isolate from human diarrhoea has been determined.[4] To initiate transcription of a gene or operon, the RNA polymerase holoenzyme, to which a sigma factor contributes specificity, has to recognize and bind to the upstream promoter

Present address: L. Petersen, Bioinformatics Center, University of Copenhagen, DK-2100 Copenhagen, Denmark.

Abbreviations used: HMM, hidden Markov model; RBS, ribosomal binding site; TP, true positive; FP, false positive.

E-mail address of the corresponding author: lpe@vetinst.dk

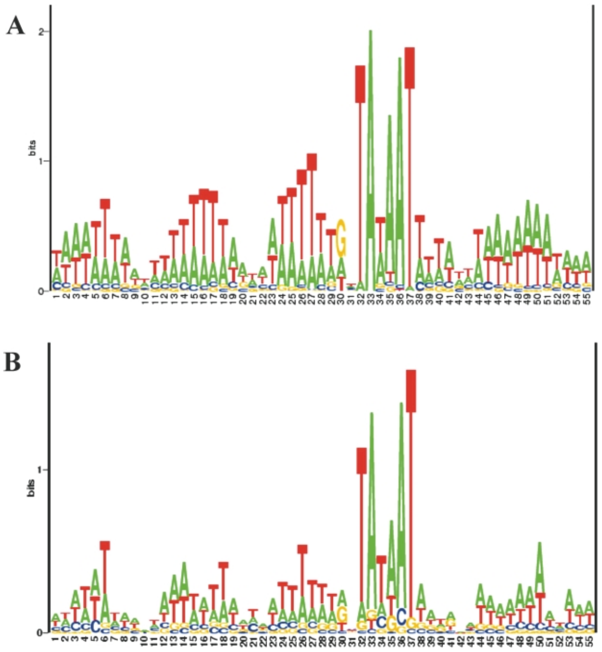

Figure 1. Sequence logo plot of predicted promoter sequences (aligned at first T in TATA-box) from (A) *C. jejuni* (184 sequences), and (B) *H. pylori* (65 sequences). The vertical axis shows information content in bits. The height of a nucleotide is proportional to its frequency in the sequence at that position.[20]

region. Only three sigma factor genes have been identified in the *C. jejuni* genome,[5,6] the housekeeping sigma-factor *Rpo* D, and sigma-factors *Fli*A and *Rpo*N that, among others, regulate genes related to flagellar motility.[7,8] However, *C. jejuni* does not appear to have heat shock or stationary phase sigma factors, unlike the enteric pathogens belonging to the Enterobacteriaceae.[5,6] On the basis of these observations, the question arises of how differential gene regulation in the *C. jejuni* genome is achieved? Wösten and colleagues[9] attempted to characterize the *Rpo* D promoter of *C. jejuni* and concluded that the promoter is somewhat unusual and poorly conserved. Recent work in our group has shown that the structural characteristics of the *C. jejuni* genome are distinct in a number of ways.[10]

In this study, we use a hidden Markov model (HMM) to identify conserved motifs upstream of functional genes, within presumed promoter regions. HMMs are probabilistic models that can be used to describe classes of symbol sequences as sets of states with transitions between them. In the case of a DNA-sequence, each state has specific probabilities for the four nucleotides, and one can say that they "emit" nucleotides according to this probability distribution. Transition and emission probabilities are estimated using a set of sequences of which the majority contain the target motif(s) ("training" of the model). Subsequently, other sequences can be run through the model ("decoding"), and a score for the extent of similarity to the conserved motif in the training sequences can be obtained.

An HMM can be constructed from several modules that are trained separately, and then combined. The modelling of complex sequence motifs with partly unknown structures, like the promoter described here, is a highly suitable task for HMMs due to the flexibility inherent in the HMM framework. The principles of HMMs have been described in detail elsewhere.[11,12]

We chose to use "posterior decoding",[11,13]which in this application essentially calculates for every nucleotide the probability that it was emitted by one of the states modelling the TATA box. The advantage of posterior decoding over the commonly used Viterbi decoding is primarily that one gains access to alternative parses of the sequence given the model, in the sense that alternative or overlapping promoter locations are derivable from the output. This is particularly relevant when two promoters with very similar scores are predicted close to each other. They may both be functional, or the prediction that scores slightly lower may be the right one. In both cases, both predictions are relevant output. The non-looped architecture is distinct from the otherwise rather similar model architecture described by Jarmer and colleagues.[14] When this architecture is used, the score of one promoter becomes independent of the surrounding sequence, or of other nearby or overlapping promoter motifs.

HMMs have been used to model promoter regions in bacterial genomes[14,15] as well as the human genome,[16] and a sensitivity of genome-wide promoter prediction of 70% has been achieved.[14] Other approaches include the expectation maximization algorithm described by Cardon & Stormo,[17] and the motif-based approach taken by Vanet and colleagues;[18] however, a comparable performance has not been achieved.

DNA structures were predicted within the predicted promoter regions, in terms of curvature, melting, and flexibility[19] to elucidate the mechanism of the identified *Rpo* D promoters.

Results

We trained an HMM using a training set of 175 promoters containing sequences of length 121 bp to estimate the HMM parameters (see Methods). The initial simple model contained only the motif TATA and the ribosomal binding site/spacer region (RBS-model[13]), with no restrictions on location, or distance between the two motifs. The model was then expanded and improved gradually in successive training rounds, by the incorporation of tendencies inferred from the trained model file or visualized by sequence logos,[20] in combination with prior knowledge of promoter structure. The region between the TATA-box and RBS was modelled first, by using the assumption

Table 1. Bacterial genomes

Species	Strain	Length	% AT	Accession no.	Reference
Aquifex aeolicus	VF5	1551335	56	AE000657	48
Borrelia burgdorferi	B31	910724	71	AE000783	49
Campylobacter jejuni	NCTC11168	1641481	69	AL111168	4
Escherichia coli	K-12, MG1655	4639221	49	U00096	50
Haemophilus influenza	Rd	1830138	61	L42023	51
Helicobacter pylori	J99	1643831	60	AE001439	52
Pasteurella multocida	PM70	2257487	59	AE004439	53
Pseudomonas aeruginosa	PA01	6264403	33	AE004091	54
Rickettsia prowazeekii	Madrid E	1111523	70	AJ235269	55
Staphylococcus aureus	N315	2813641	67	BA000018	56
Vibrio cholerae (chromosome I)	N16961	2961149	52	AE003852	57

that a shift in dinucleotide composition will occur around the transcription start-site.[19] When the resulting model was used for decoding, a logo plot of the aligned sequences showed a periodic signal upstream of the TATA-box, and this periodicity was subsequently included in the model.

The final model included states modelling everything in a genome that is not a promoter (null-model). These states were trained separately before being incorporated into the model. The final model can be used to predict promoters in sequences of any length, including whole genomes (see Methods).

A sequence logo[20] of the consensus sequence of the presumed *Rpo* D promoters in *C. jejuni* is depicted in Figure 1(A). Upstream of the TATA-box, a distinct AT-rich periodic signal is seen that extends beyond the TATA-box. The transition probabilities in the trained model (transitions from periodic signal state 10 to state 1, and from periodic signal state 11 to state 1) show that the average period is 10.56 nucleotides. The TATA-box is preceded by a semi-conserved TGx. When the model was trained on *Helicobacter pylori* sequences, a similar promoter structure was found (Figure 1(B)). In the remaining bacterial genomes tested (Table 1), a TATA-box of varying intensity but no periodic signal could be seen in sequence logos of predicted promoters aligned by the model (data not shown).

Predictions of DNA structural parameters of the region surrounding the TATA-box showed distinct periodicity in stacking energy, position preference, and DNaseI sensitivity and extreme (high as well as low) values around the TATA-box of those three parameters, whereas DNA curvature showed a distinct peak immediately upstream of the TATA-box, and lower value downstream (Figure 2).

The performance of the whole genome model was tested by five-fold cross-validation, where the training set was divided into five sets of 35 sequences. The model was repeatedly trained on four subsets and tested on the fifth, until all subsets had been tested once. The training set consisted of sequences of which a high percentage was likely to contain the predominant promoter. It should be noted that we do not actually know the level of true positives, false positives or false negatives produced by the model. However, we assume that promoters predicted in the test sequences in the cross-validation experiment are true positives (TP). We also assume that predicted promoters in the 500,000 bp random DNA sequence are false positives (FP). Establishing the threshold is a compromise between wanting to predict all true positives, and wanting to avoid false positive predictions. The performance of the model in the cross-validation experiment is shown in Figure 3, where the TP-rate at a given log-odds score is plotted against number of hits in the random sequence, and number of hits in the genome, respectively. In both graphs, the points are marked where the plots depart from linearity, and the number of hits in the test sequence starts to grow rapidly. The log-odds threshold is set at 2.6, corresponding to the marked points. At this threshold the model predicts 119 sequences to have a promoter in the cross-validation (sensitivity 68%), and predicts ten promoters in 500,000 bp of random sequence and 764 in the *C. jejuni* genome. On the basis of the results from the random sequence, our best estimate of the false positive rate is 66 predictions in a genome of 1.641 Mb (1,641,481/500,000 × 2 × 10).

When a set of sequences containing 27 experimentally mapped *C. jejuni* promoters were decoded with the model, 19 were predicted correctly by the model, one was predicted but an additional prediction 10 bp downstream scored slightly higher, two had a log-odds score below the established cut-off, and five were missed (Table 2). This indicates a sensitivity of the model of around 74%.

Figure 4 shows that the majority of TATA-boxes predicted in the genome sequence start 35-40 bp upstream of the annotated start codon, whereas a smaller fraction actually start downstream of start codons. The Figure indicates that in the *C. jejuni* genome the first T of the TATA-box is located predominantly in the area 150 bp upstream of annotated start codons to 25 bp downstream. A total of 654 promoters were predicted in such regions, upstream of 541 genes (two or more promoters were predicted in the upstream regions of some genes). In addition, 110 predicted promoters were found outside the upstream regions

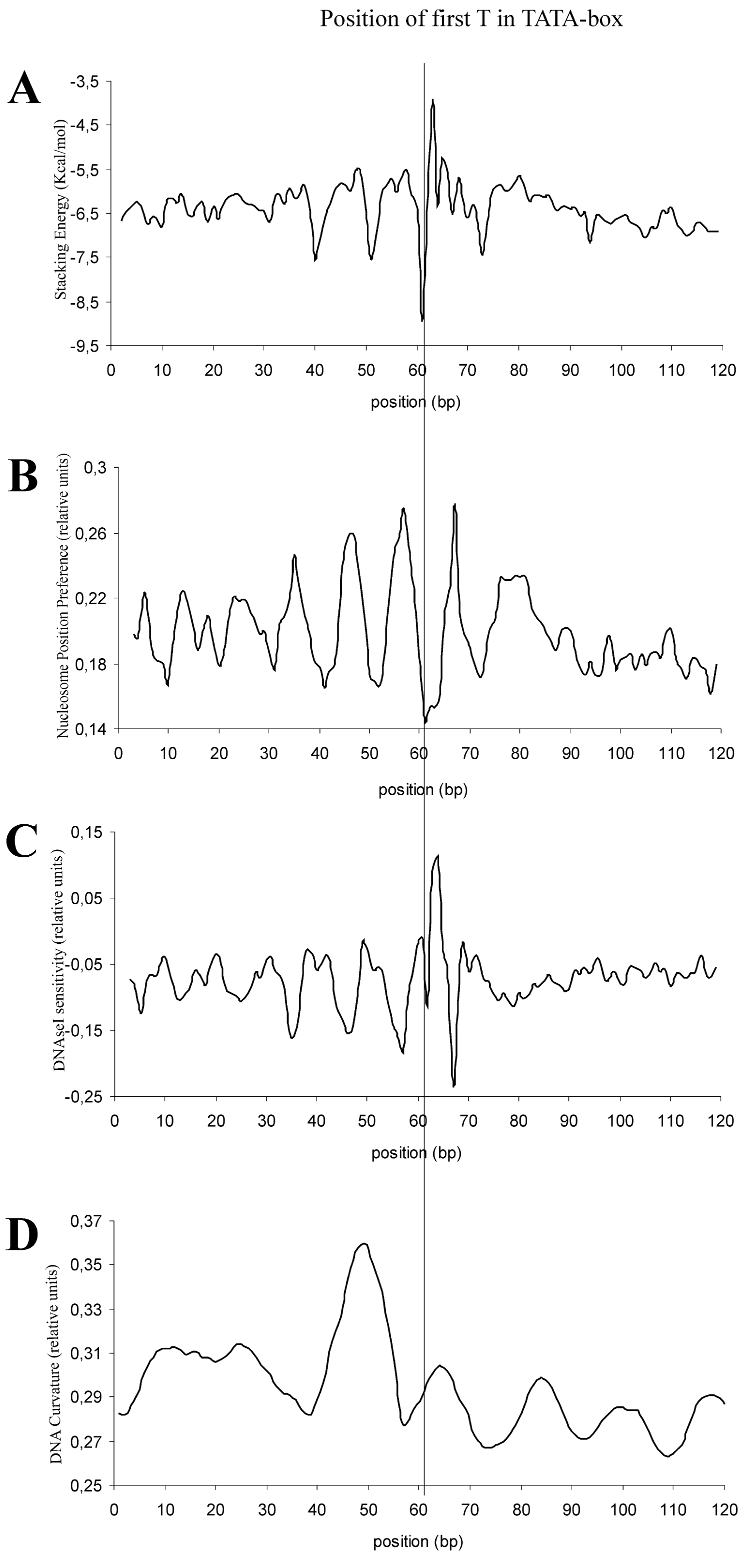

Figure 2. DNA structural predictions in 184 aligned promoters. (A) Stacking energy/DNA meltability;[45] (B) flexibility (position preference; lower values correspond to greater flexibility);[16,46] (C) flexibility (DNaseI sensitivity; higher values correspond to greater flexibility);[47] (D) DNA Curvature (higher values correspond to greater curvature).[44]

(specificity: $((764 - 110)/764) \times 100 = 86\%$). Of the 541 genes with a predicted promoter, 278 are followed by a series of genes (two or more) separated by intergenic regions of less than 20 bp, but with no predicted promoter. Several well-characterized genes are contained in such "operon candidates" consisting of genes with related function (Table 3).

Discussion

Earlier attempts to characterize the *Rpo* D promoter structure in *C. jejuni* have concluded that the promoter is unusual, and the −35 box is weakly conserved,[9,21] whereas a T-rich domain has been identified upstream of the TATA-box.[9] We report here the identification of the structure of the

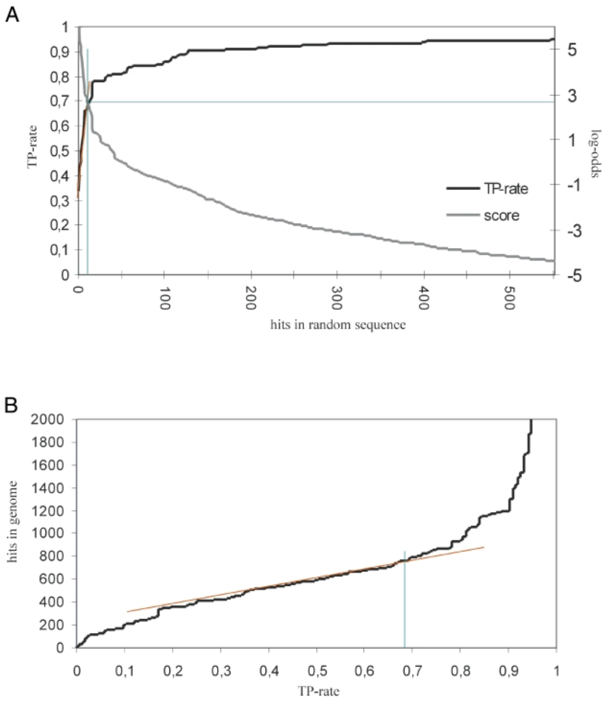

Figure 3. Performance of the whole genome model in the cross-validation experiment. (A) Number of predicted promoters in cross-validation experiment in 175 sequences (TP-rate) *versus* number of predicted promoters in random sequence at corresponding log-odds scores. At log-odds 2.6 and below and at aTP-rate of 0.68 and above (indicated with red and blue lines) the number of hits in random sequence grows rapidly compared to the TP-rate. (B) Number of predicted promoters in the *C. jejuni* genome *versus* TP-rate. As indicated by the red line, the number of predicted promoters in the genome starts to grow rapidly at or above a TP-rate of 0.68.

Table 2. Decoding results of 22 sequences containing 27 experimentally mapped *C. jejuni* promoters

Accession	Log-odds	Gene/origin of sequence	Reference
AF044271	18.9831	*ahp*C[a]	58
AF044271	21.6445	*fdx*A	59
AJ002027	5.29105	Glu-tRNA	9
AJ002415	−0.82076	*met* K	9
AJ002416	12.5444	Clone 1b7	9
AJ002417	8.06119	Clone 1g9[b]	9
AJ002418	2.07037	*icd*	9
AJ002419	13.6167	Clone 3D8	9
AJ002420	5.81369	Clone 14b7[b]	9
AJ002421	6.20396	Clone 2a12[b]	9
AJ002422	5.48556	Clone 11b4[b]	9
L25627	4.96666	*hup* B	60
M63448	Not predicted	*lys*S	61
M74579	9.63052	*pro*A	62
U06951	11.7635	Orf3	63
U06951	9.67093	*rps*O	63
U08132	9.9314	*sod* B[a,b]	64
U15295	9.33809	*ile*S	65
U38524	7.88633	*psp*A[a]	9
X53816	Not predicted	*gly*A	66
X85954	13.8935	*tig*[c]	67
X95910	Not predicted	*fts*A	68
Y13333	15.5419	*clp* B	69
Z36940	18.0575	*hip*O[a]	9

When a promoter is predicted at or close to the location of the experimentally mapped promoter (±10 bp) the log-odds score is shown.

[a] Additional promoters were predicted in the upstream region of that gene.

[b] Positions of experimentally mapped and predicted promoters diverge by <11 bp.

[c] Two additional experimentally mapped promoters (one upstream and one downstream) were not predicted.

predominant promoters in *C. jejuni* (the presumed *Rpo* D promoters), a structure that is unique among other bacterial promoters that have been described to date. We used a hidden Markov model that proved very powerful for this purpose.

Our model did not identify similar motifs in most of the other bacteria investigated. When the model is trained on a set of sequences from a given bacterium, it will try to estimate emission and transition probabilities that allow for the target structure (a periodic signal, a TATA-box, a shift in dinucleotide composition after less than 9 bp, a stretch of unspecified length and a ribosomal binding site). During the decoding, motifs that are dissimilar to the target structure will receive a low score. The fact that motifs from different bacteria aligned by the trained models do not show upstream periodicity shows that the predominant (*Rpo* D) promoters in those bacteria do not share the structure of the *C. jejuni* promoters. The finding of a promoter structure in *H. pylori* that resemble the *C. jejuni* promoter is in concordance with the close relationship that is well established between *Helicobacter* and *Campylobacter* spp.,[22] and the distinctness of the Epsilobacteria.[23] The absence of the unique *Rpo* D promoters in *Aquifex aeolicus* is noteworthy, since this free-living organism was recently classified as an epsilobacterium on the basis of a comprehensive consideration of bacterial cell structure and processes.[23] Although certain structural features of the genome sequence of *Aq. aeolicus* are shared by those of *C. jejuni* and *H. pylori*,[24] the hyperthermophilic nature, and difference in *Rpo* D promoter sequence (this study) of *Aq. aeolicus* suggest that the classification of the latter as an Epsilobacterium requires further study.

The *C. jejuni Rpo* D promoters to some degree resemble the “extended −10” promoter that has been described in *E. coli*[25,26] and is widespread in Gram-positive bacteria.[27,28] Extended −10 promoters are characterized by the presence of a conserved TGx upstream of the TATA-box, and the absence of a conserved −35 box. The sigma-80 promote of *H. pylori* described by Vanet and colleagues[18] shares some similarity with the consensus sequence that we have identified, and would most likely be recognized by our model. The approach used by Vanet and colleagues[18] is

Table 3. Operon candidates: clusters of genes with a promoter in front of the first gene, intergenic distances <20 bp (unless otherwise stated below), and in most cases related function

Genes	Function	Log-odds score	Position of first T in the TATA-box
*cdt*A, *cdt* B, *cdt*C	Cytolethal distending toxin	6.03236	91,109
atp F′, *atp* F, *atp* H, *atp*A, *atp*G	ATP-syntase subunits	7.51324	109,965
ceu B, *ceu*C, *ceu* D, *ceu* E	Uptake of ferric siderophore[70,71]	3.91651	1,283,965
*arg*C, Cj0225, *arg* B, *arg* D	Arginine biosynthesis[72]	8.64344	209,246
*gat*C, Cj0399,[a] *fur, lys*S, *gly*A, Cj0403, Cj0404, *aro* E	Mixed function[41]	13.2638 (0.52538)	364,284 (364,637)
hyp B, *hyp*C, *hyp* D, *hyp* E, *hyp*A	Hydrogenase isoenzyme formation	9.56402	584,061
gro ES, *gro* EL[b]	Stress response chaperonins[73]	9.47424	1,149,133
*kps*M, *kps*T, *kps* E, *kps* D, *kps* F, Cj1442c, *kfi* D, Cj1440c	Capsule formation	3.02765	1,387,744
*Chu*A, *chu* B, *chu*C, *chu* D	Haemin uptake system	17.2276	1,540,763
*sda*C, *sda*A	Serine uptake	14.8996	1,554,422
*leu*A, *leu* B, *leu*C, *leu*C	3-Isopropyl malate modification	4.68085	1,631,966

[a] Intergenic distance = 68 bp. Promoters are located in front of first and second gene.
[b] Intergenic distance = 21 bp.

based on the assumption that the −10 and −35 regions are conserved and most regions between them non-conserved, and therefore would not capture the promoter structure identified in this study. It is not known whether two or more different sigma-80 promoter types exist in *H. pylori*. A strong periodic signal similar to what we found in *C. jejuni* and *H. pylori* promoters has not been reported in bacterial promoters to our knowledge, though a certain periodicity in human promoters has been described.[16]

It is well known that the estimated period of 10.56 nucleotides corresponds approximately to one helix turn and that curvature of DNA requires a number of in-phase curved regions,[29] in our case stretches of T bases (or A bases on the other strand). The predictions of intrinsic curvature indicate the presence of a highly curved region upstream of the TATA-box (Figure 2(D)). Position preference and DNaseI sensitivity are both measures of DNA flexibility (Figure 2(B) and (C)).[19] Such periodic flexibility is thought to enhance curvature of DNA.[16] The TATA-box is in-phase with the structural oscillations; therefore, it is likely to have a constant angle to the bending direction. Earlier investigations of the *Escherichia coli* sigma-70 promoter have shown that approximately 90 bp of the DNA strand surrounding the promoter is wrapped around the RNAP (RNA polymerase) holoenzyme before transcription

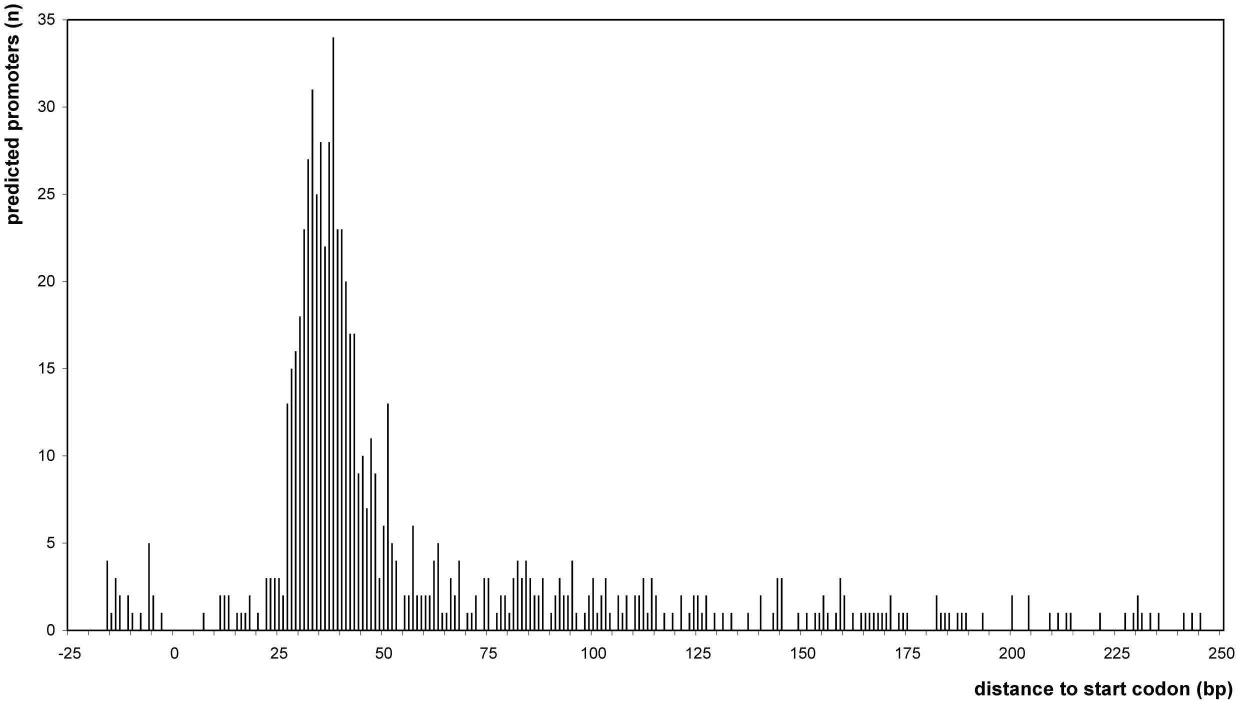

Figure 4. Position of promoters predicted in the genome sequence (first T in TATA-box) in relation to the annotated start codon (from 250 bp upstream to 25 bp downstream of the start codon).

initiation.[30] The curvature and flexibility of this region may play a significant role here, so that a conserved −35 box is no longer needed. Several investigations have tried to establish the role of DNA structure upstream of the −35 region in promoter function, and often found that conserved sequence motifs play a significant role. Investigations of the region upstream of the −35 box in the *E. coli* rRNA promoter have shown that the presence of either the UP-element, a conserved motif upstream of the promoter,[31,32] a curved region,[33] or one or more A-tracts of length five or six nucleotides arranged with a period of ten nucleotides[34] greatly enhance promoter strength. Multiple in-phase A-tracts result in macroscopic curvature of DNA.[34,35] However, it is unclear whether curvature in itself can enhance transcription, or a conserved sequence motif is required.[34] Our attempts to model conserved motifs upstream of the TATA-box were unsuccessful, but it can be seen from the sequence logo (Figure 1) that thymines bases are predominant in regions −18 to −24, and −29 to −35. We speculate whether in this particular case, DNA structure in combination with the TATA-box and the semi-conserved TG at position −17 are the only determinants of sigma factor recognition and binding, or whether the (semi-conserved) phased T-stretches observed in the upstream regions play a role as specific binding sites. The level of curvature in regions with phased A-stretches (or T-stretches on the other strand) is influenced by temperature, salt etc.,[36,37] and we hypothesize that the observed periodicity plays a role in environmental regulation of expression levels, and may explain the absence of a stress response sigma factor. However, this needs to be further investigated experimentally. The predictions of stacking energy (DNA meltability) show a distinct peak at the position of the TATA-box, consistent with the melting of this region during open complex formation,[6] prior to initiation of transcription.

The sensitivity and specificity of our promoter predictor is difficult to establish in the absence of a gold standard; we simply do not know which genes do in fact have an *Rpo* D promoter. There must be at least two other promoter types in the genome (corresponding to sigma-factors *Fli*A and *Rpo*N), which function as regulatory elements of genes related to flagellar motility, among others.[7,38] Two different sigma-70/*Rpo* D promoter types have been identified in other bacteria[14,25–27] and we do not know whether that is the case in *C. jejuni*, or to what degree variant promoters are recognized by the model. Furthermore, second or later genes in operons do not, in most cases, have a promoter. Finally, a considerable fraction of genes in the *C. jejuni* genome have unknown or hypothetical functions assigned to them in the GenBank file, a certain amount of over-annotation has been assumed,[39] and there may be genes that are not yet annotated. We do not know the maximum distance to a start codon for a functional promoter but, on the basis of the distribution presented in Figure 4, we chose to base further calculation on a maximum distance of 150 bp, as our best estimate. As shown in Figure 4, our model sometimes predicts promoters downstream of translation −25, where they should be in order to be able to initiate translation at the annotated start codons. Such a result does not make biological sense; however, it is very common that genes are annotated with a wrong start codon upstream of the actual start, and this may partly explain our results.[13] It can be speculated whether the combination of gene finders with promoter models similar to that described here will improve the performance of both.

We predicted *Rpo* D promoters in 119 of 175 upstream regions in the cross-validation (68%), and in front of 184 of 529 high-confidence genes (35%), and 541 of 1708 annotated genes in the *C. jejuni* genome (32%). The model successfully identified 20 of 27 experimentally mapped promoters (74%). However, we are not certain which of those experimentally mapped promoters are actually regulated by *Rpo* D. Furthermore, the level of predictions in the random sequence indicate a false positive (FP) level of approximately 66 in the *C. jejuni* genome, but 110 FPs were found in the genome under the assumptions described above. On the one hand, the FP-level based on the random sequence is clearly an approximation, on the other hand the FP-level based on the genome is influenced by the validity of the GenBank annotation as well as the assumptions that we made on cut-off and maximum distance to start codons. The actual FP-level of our model is likely to lie somewhere between the two. The performance of our model is therefore comparable to the *Bacillus subtilis* promoter model described by Jarmer and colleagues.[14] We used posterior decoding, whereas Viterbi decoding was used in the *B. subtilis* model.[14] A training set of 236 experimentally verified promoters was used in this study; in contrast, we had an insufficient amount of experimentally verified promoters to use as training data, and knew little about the structure of the promoter. Nonetheless, it was possible to identify this unique promoter structure, and do genome-wide promoter prediction with a reasonable performance, which establishes the validity of our approach. The expectation maximization approach described by Cardon & Stormo[17] used a training set of 231 sequences that were known to contain promoters, and predicted over 90% of *E. coli* promoters in the training set. However, this study differs considerably from ours in that the actual fractions of *Rpo* D-promoter containing sequences in our training set or test sets are unknown, and our sensitivity is therefore likely to be higher than the estimated minimum of 68%. Furthermore, our model has the advantage of being able to predict promoters in sequences of any length. Vanet and colleagues[18] used a motif-based approach to identify consensus sequences of −10 and −35 regions in the *H. pylori* genome. A total of 56 putative promoters were identified in the *H. pylori* genome using this approach, yielding a sensitivity that is clearly inferior to that of our method.

When we started developing the model, we tried to train a simple model that could identify the TATA-box and ribosomal binding site on the entire set of upstream regions from 529 genes with known function and high confidence starts. The trained model did not reveal conserved sequences. However, when we reduced the training set as described in Methods, including only upstream regions from genes that are unlikely to be second or later gene in an operon, we quickly managed to identify the pattern as presented here. This observation alone indicates strongly that this pattern is absent from a considerable fraction of the 529 sequences, and applying the model to the 529 upstream regions supported this. Consequently, we suggest that operon structure may be more common in *C. jejuni*, than hitherto proposed.[4] It should be noted that when we reduced the training set to contain only genes that are unlikely to be part of an operon and that are not related to flagellar motility, some legitimate promoter-containing sequences were most likely excluded, such as those that are located in an upstream coding region. However, as the resulting training set of 175 sequences is expected to be sufficiently large to reliably train an HMM, it is a minor concern that some sequences are not used for training. The trained model should ideally be able to predict those. Indeed, a total of 57 predicted promoters overlap the upstream gene, and promoters were predicted in front of a subset of genes related to flagellar motility (data not shown).

Salgado and colleagues[40] have investigated intergenic distances in *E. coli*, within and outside operons, and found that operons in this organism are characterized by intergenic distances of around −20 (overlapping genes) to 20 bp, whereas intergenic distances at the borders of transcription units tend to be longer. In contrast, operon structure in *C. jejuni* is not well described. We found 278 examples of operon candidates, based on the criteria of predicted promoter in front of the first gene and short intergenic distances (<20 bp). (A complete list of operon candidates identified in this study is available†.) This very simplified approach to identify operons is clearly preliminary and insufficient. An "operon finder" should take transcription termination, other promoter types etc. into account. However, we managed to identify two previously described Campylobacter operons that are listed in Table 3, together with examples of operon candidates that contain only genes with related function. Another two described operons had longer intergenic distances, but we predicted promoters in front of the first gene, and in the case of the *fur* operon, a predicted promoter (with a lower cut-off) was identified in front of the second gene, as described by Van Vliet and colleagues.[41] This agreement with experimental results further establishes the validity of the promoter model.

In conclusion, we have identified a likely consensus *Rpo* D promoter sequence in *C. jejuni*, assessed its distribution in the genome sequence of NCTC11168 and inferred that genes may be organized in operons more commonly than was realized hitherto. Our results call for experimental verification of our hypotheses, as well as additional work on the identification, distribution, and significance of other promoters in *C. jejuni*, and investigations of operon structure.

Methods

HMM architecture

The general architecture of the promoter model is depicted in Figure 5. In addition to the promoter model itself, it consists of begin-states and end-states, a second-order null-model trained on coding and shadow (reverse-complement of coding) regions, respectively; and a first-order null-model trained on intergenic regions. The promoter model consists of a branch state, followed by 11 first-order states modelling the region upstream of the TATA-box. A loop is included to model a periodic signal of length 10–11 nucleotides in this region. Then follows the TATA-box model, a number of first-order "background" states modelling the dinucleotide distributions around transcription +1, and an RBS model, modelling the ribosomal binding site as well as the nucleotides up to the start codon. The RBS model is identical with that described by Larsen & Krogh,[13] except that the maximum distance from the ribosomal binding site to start codon is reduced, to reflect the distances found in this particular bacterium. Transitions are allowed from the branch state and directly to the RBS model, to take into account occurrence of promoter types that are specific for other sigma-factors. Due to the way training sequences are selected, they are not likely to be co-translated with the previous gene, and therefore must have a Shine–Dalgarno sequence. During decoding, the model may encounter genes that have a full promoter and Shine–Dalgarno motif, a Shine–Dalgarno alone (if they have a different promoter type or are located within an operon) or neither (genes that are located within an operon and co-translated with the previous gene). However, given the way that we defined the scoring system, only full promoters will be recognized by the model and result in a score. Genes that are preceded only by, or lack a Shine–Dalgarno, will not influence the promoter prediction.

Pseudo-counts were used on the six states of the TATA-box model and the second to sixth states of the RBS model to reflect the expected base composition, as described by Durbin and colleagues.[11] Similarly, pseudo-counts were used on transition probabilities in the branch state, and in the two spacer regions. Training was carried out by means of the Baum–Welch algorithm, which is a maximum likelihood approach guaranteed to find the parameters that (locally) maximize the conditional probability of the training set given the parameters.[11]

† www.binf.ku.dk/krogh/CampyPromoters/

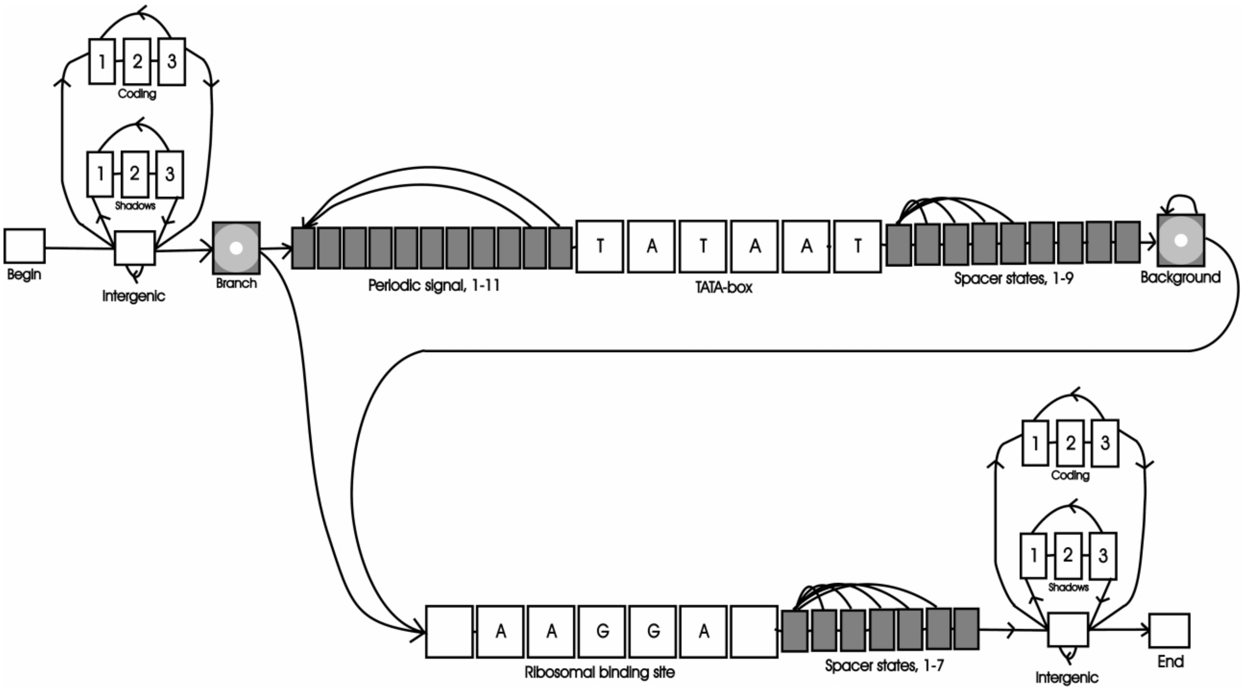

Figure 5. HMM architecture: states and transitions of the whole genome model used for predicting promoters in the *C. jejuni* genome. For other bacterial genomes, the states intergenic, coding and shadow were replaced by simple background states. In order to constrain the training process and make use of existing knowledge, selected emission and transition probabilities were biased manually towards their consensus value. In particular, this regularization was done by adding pseudo-counts to the most conserved nucleotides, to transitions from the branch state, and to different transitions in the spacer regions.[11] After a few tests we added a pseudo-count of 70 to the most conserved nucleotides in the TATAAT region, and a pseudo-count of 100 to the most conserved nucleotides in the AAGGA region of the Shine–Dalgarno sequence, whereas smaller pseudo-counts were added to remaining nucleotides in those regions. Similarly, we added pseudo-counts of 5–60 to the transitions in the two spacer regions.

For non-*C. jejuni* genomes, a reduced model, consisting of a second-order background model, the periodic upstream region, TATA-box, first-order background model and RBS regions, was used.

Training sets

Functional genes with high-confidence start codons were identified for each genome as described.[13,42] A subset of genes that were preceded by an annotated open reading frame (ORF) transcribed in the opposite direction, based on the existing GenBank annotation (Table 1) was assembled. Upstream regions of 120 bp + the first base of the start codon (a total of 121 bp) was included in the training sets. For the *C. jejuni* genome, upstream regions from genes that were preceded by an intergenic region of >100 bp were included in the training set, and genes that in the GenBank description contain the word flagel were excluded (six genes), because such genes are not expected to be regulated by *Rpo* D.[7,8]

For the *C. jejuni* genome, three training sets were used to train the non-promoter regions in the whole genome model. The set of functional genes referred to above ($n = 529$) was used to train the coding model, their reversed complements were used to train the shadow model (reverse strand of coding DNA), and all annotated intergenic regions (what is left when coding and shadow regions are removed) longer than 30 bp were used to train the intergenic region model.

Prediction of promoters

The trained HMM was subsequently used to find all *Rpo* D promoter-like sequences in the genome, by means of posterior decoding.[11] By adding null-models that model all non-promoter sequence, the probability of a given nucleotide being emitted by one of the states modelling the promoter can be divided by the probability of the same nucleotide being emitted by the null-model, and the log-odds score is the logarithm of this ratio. When dividing with the probability of the sequence given the null-model the probability of the sequences flanking the promoter cancels out, and the odds ratio will depend only on the promoter region. Now, all predicted promoters, regardless of whether they are located in the vicinity of other promoters, can be seen as alternative parses through the model for which a log-odds score is produced. The resulting model can be used to decode sequences of any length. The output consists of a rough curve showing the log-odds score as a function of nucleotide positions. This output was further processed by extracting maxima within an 80 bp window, so that “predicted promoters” could occur with a minimum distance of 40 bp.

The model was tested on the following test sets.

- A random sequence of 500,000 bp generated by a third-order Markov chain trained on the *C. jejuni* genome.
- The *C. jejuni* genome sequence.

- Upstream regions (from 120 bp upstream to 50 bp downstream of the annotated start codon) of 529 functional genes with certain starts, of which a subset was used to train the model, as described above.
- A set of 22 sequences containing 27 experimentally mapped promoters (Table 2).

DNA structural predictions

Of the 529 functional genes with high-confidence starts, a set of 184 sequences with strong promoter predictions (log-odds score > 2.6) was selected. The sequences were aligned at the predicted TATA-boxes, and subjected to structural predictions as described.[19,43] Briefly, DNA curvature was calculated using a modified version of the Curvature programme.[44] Stacking energy or "DNA meltability" was derived from the dinucleotide values described by Ornstein and colleagues.[45] Position preference was derived from the trinucleotide values described by Satchwell and colleagues[46] as described by Pedersen and colleagues.[16] DNaseI sensitivity was derived from the trinucleotide values described by Brukner and colleagues.[47]

Acknowledgements

The authors thank the people at the Center for Biological Sequence Analysis for their helpful discussions, particularly Peter Wad Hansen and Lars Juhl Jensen. This work was sponsored by a grant from the Danish Agricultural and Veterinary Research Council (grant SJVF-23-00-0285), and a grant to the Center for Biological Sequence Analysis (Søren Brunak) from the Danish National Research Foundation.

References

1. Friedman, C. R., Neimann, J., Wegener, H. C. & Tauxe, R. V. (2000). Epidemiology of *Campylobacter jejuni* infection in the United States and other industrialized nations. In *Campylobacter* 2nd Edition (Nachamkin, I. & Blaser, M. J., eds), pp. 121–139, ASM Press, Washington, DC.
2. On, S. L. W., Nielsen, E. M., Engberg, J. & Madsen, M. (1998). Validity of SmaI-defined genotypes of *Campylobacter jejuni* examined by SalI, KpnI, and BamHI polymorphisms: evidence of identical clones infecting humans, poultry, and cattle. *Epidemiol. Infect.* **120**, 231–237.
3. Petersen, L., Nielsen, E. M., Engberg, J., On, S. L. & Dietz, H. H. (2001). Comparison of genotypes and serotypes of *Campylobacter jejuni* isolated from Danish wild mammals and birds and from broiler flocks and humans. *Appl. Environ. Microbiol.* **67**, 3115–3121.
4. Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D. *et al.* (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.
5. Mittenhuber, G. (2002). An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. *J. Mol. Microbiol. Biotechnol.* **4**, 77–91.
6. Wosten, M. M. (1998). Eubacterial sigma-factors. *FEMS Microbiol. Rev.* **22**, 127–150.
7. Jagannathan, A., Constantinidou, C. & Penn, C. W. (2001). Roles of *rpo*N, *fli*A, and *flg*R in expression of flagella in *Campylobacter jejuni*. *J. Bacteriol.* **183**, 2937–2942.
8. Lüneberg, E., Glenn-Calvo, E., Hartmann, M., Bär, W. & Frosch, M. (1998). The central, surface-exposed region of the flagellar hook protein FlgE of *Campylobacter jejuni* shows hypervariability among strains. *J. Bacteriol.* **180**, 3711–3714.
9. Wosten, M. M., Boeve, M., Koot, M. G., van Nuene, A. C. & van der Zeijst, B. A. (1998). Identification of *Campylobacter jejuni* promoter sequences. *J. Bacteriol.* **180**, 594–599.
10. Petersen, L., On, S. L. W. & Ussery, D. W. (2002). Visualization and significance of DNA structural motifs in the *Campylobacter jejuni* genome. *Genome Letters*, **1**, 16–25.
11. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). Markov chains and hidden Markom models. In *Biological Sequence Analysis* (Durbin, R., Eddy, S., Krogh, A. & Mitchison, G., eds), pp. 46, Cambridge University Press, Cambridge.
12. Larsen, T.S. (2002). Gene and motif finding in prokaryotic DNA using hidden Markov models. PhD thesis, Technical University of Denmark.
13. Larsen, T.S., Krogh, A. (2003). EasyGene: a prokaryotic gene finder that ranks ORFs by statistical significance. Submitted for publication.
14. Jarmer, H., Larsen, T. S., Krogh, A., Saxild, H. H., Brunak, S. & Knudsen, S. (2001). Sigma A recognition sites in the *Bacillus subtilis* genome. *Microbiology*, **147**, 2417–2424.
15. Fouts, D. E., Abramovitch, R. B., Alfano, J. R., Baldo, A. M., Buell, C. R., Cartinhour, S. *et al.* (2002). Genomewide identification of *Pseudomonas syringae* pv. tomato DC3000 promoters controlled by the HrpL alternative sigma factor. *Proc. Natl Acad. Sci. USA*, **99**, 2275–2280.
16. Pedersen, A. G., Baldi, P., Chauvin, Y. & Brunak, S. (1998). DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.* **281**, 663–673.
17. Cardon, L. R. & Stormo, G. D. (1992). Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J. Mol. Biol.* **223**, 159–170.
18. Vanet, A., Marsan, L., Labigne, A. & Sagot, M. F. (2000). Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals. *J. Mol. Biol.* **297**, 335–353.
19. Pedersen, A. G., Jensen, L. J., Brunak, S., Staerfeldt, H. H. & Ussery, D. W. (2000). A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.* **299**, 907–930.
20. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.
21. Wassenaar, T. M. & Meinersman, R. J. (2001). Promoter sequences in *Campylobacter jejuni*. *Int. J. Med. Microbiol.* **291**, 78.
22. On, S. L. W. (2001). Taxonomy of *Campylobacter*, *Arcobacter*, *Helicobacter*, and related bacteria: current status, future prospects, and immediate concerns. *Symp. ser. J. Appl. Microbiol.* **30**, 1S–15S.

23. Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76.
24. Ussery, D., Soumpasis, D. M., Brunak, S., Staerfeldt, H. H., Worning, P. & Krogh, A. (2002). Bias of purine stretches in sequenced chromosomes. *Comput. Chem.* **26**, 531–541.
25. Belyaeva, T., Griffiths, L., Minchin, S., Cole, J. & Busby, S. (1993). The *Escherichia coli* cysG promoter belongs to the "extended −10" class of bacterial promoters. *Biochem. J.* **296**, 851–857.
26. Prost, J. F. & Cozzone, A. J. (1999). Detection of an extended-10 element in the promoter region of the *pck*A gene encoding phosphoenolpyruvate carboxykinase in *Escherichia coli*. *Biochimie*, **81**, 197–200.
27. Sabelnikov, A. G., Greenberg, B. & 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. Voskuil, M. I., Voepel, I. & 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.
29. Sinden, R. R. (1994). *Editor of DNA Structure and Function*, Academic Press, San Diego.
30. Rivetti, C., Guthold, M. & Bustamante, C. (1999). Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex. *EMBO J.* **18**, 4464–4475.
31. Estrem, S. T., Gaal, T., Ross, W. & Gourse, R. L. (1998). Identification of an UP element consensus sequence for bacterial promoters. *Proc. Natl Acad. Sci. USA*, **95**, 9761–9766.
32. Rao, L., Ross, W., Appleman, J. A., Gaal, T., Leirmo, S., Schlax, P. J. *et al.* (1994). Factor independent activation of rrnB P1. An "extended" promoter with an upstream element that dramatically increases promoter strength. *J. Mol. Biol.* **235**, 1421–1435.
33. Nickerson, C. A. & Achberger, E. A. (1995). Role of curved DNA in binding of *Escherichia coli* RNA polymerase to promoters. *J. Bacteriol.* **177**, 5756–5761.
34. Aiyar, S. E., Gourse, R. L. & Ross, W. (1998). Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase α subunit. *Proc. Natl Acad. Sci. USA*, **95**, 14652–14657.
35. Potaman, V. N., Ussery, D. W. & Sinden, R. R. (1996). Formation of a combined H-DNA/open TATA box structure in the promoter sequence of the human Na,K-ATPase alpha2 gene. *J. Biol. Chem.* **271**, 13441–13447.
36. Chan, S. S., Breslauer, K. J., Austin, R. H. & Hogan, M. E. (1993). Thermodynamics and premelting conformational changes of phased (dA)5 tracts. *Biochemistry,* **32**, 11776–11784.
37. Ussery, D. W., Higgins, C. F. & Bolshoy, A. (1999). Environmental influences on DNA curvature. *J. Biomol. Struct. Dynam.* **16**, 811–823.
38. Kinsella, N., Guerry, P., Cooney, J. & Trust, T. J. (1997). The *flg*E gene of *Campylobacter coli* is under the control of the alternative sigma factor sigma54. *J. Bacteriol.* **179**, 4647–4653.
39. Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D. & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17**, 425–428.
40. Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. & Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
41. van Vliet, A. H., Rock, J. D., Madeleine, L. N. & Ketley, J. M. (2000). The iron-responsive regulator Fur of *Campylobacter jejuni* is expressed from two separate promoters. *FEMS Microbiol. Letters*, **188**, 115–118.
42. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucl. Acids Res.* **26**, 2941–2947.
43. Skovgaard, M., Jensen, L. J., Friis, C., Staerfeldt, H. H., Worning, P., Brunak, S. & Ussery, D. (2002). The Atlas visualisation of genome-wide information. In *Methods in Microbiology V33* (Wren, B. & Dorrell, N., eds), pp. 49–63, Academic Press, London.
44. Shpigelman, E. S., Trifonov, E. N. & Bolshoy, A. (1993). CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci.* **9**, 435–440.
45. Ornstein, R. L. & Rein, R. (1978). An optimized potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers*, **17**, 2343–2360.
46. Satchwell, S. C., Drew, H. R. & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome come DNA. *J. Mol. Biol.* **191**, 659–675.
47. Brukner, I., Sanchez, R., Suck, D. & Pongor, S. (1995). Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dynam.* **13**, 309–317.
48. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E. *et al.* (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature*, **392**, 353–358.
49. Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R. *et al.* (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.
50. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V. & Riley, M. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
51. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
52. Alm, R. A., Ling, L.-S. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C. *et al.* (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–181.
53. May, B. J., Zhang, Q., Li, L. L., Paustian, M. L., Whittam, T. S. & Kapur, V. (2001). Complete genomic sequence of *Pasteurella multocida*, Pm70. *Proc. Natl Acad. Sci. USA*, **98**, 3460–3465.
54. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J. *et al.* (2000). Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.
55. Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C., Podowski, R. M. *et al.* (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.
56. Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I. *et al.* (2001). Whole genome

sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–1240.

57. Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J. *et al.* (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.
58. Baillon, M. L., van Vliet, A. H., Ketley, J. M., Constantinidou, C. & Penn, C. W. (1999). An iron-regulated alkyl hydroperoxide reductase (*Ahp*C) confers aerotolerance and oxidative stress resistance to the microaerophilic pathogen *Campylobacter jejuni*. *J. Bacteriol.* **181**, 4798–4804.
59. van Vliet, A. H. M., Baillon, M. L. A., Penn, C. W. & Ketley, J. M. (2001). The iron-induced ferredoxin FdxA of *Campylobacter jejuni* is involved in aerotolerance. *FEMS Microbiol. Letters*, **196**, 189–193.
60. Konkel, M. J., Marconi, R. T., Mead, D. J. & Cieplak, W. (1994). Cloning and expression of the hup gene encoding a histone-like protein of *Campylobacter jejuni*. *Gene*, **146**, 83–86.
61. Chan, V. L. & Bingham, H. L. (1992). Lysyl-tRNA synthetase gene of *Campylobacter jejuni*. *J. Bacteriol.* **174**, 695–701.
62. Louie, H. & Chan, V. L. (1993). Cloning and characterization of the gamma-glutamyl phosphate reductase gene of *Campylobacter jejuni*. *Mol. Gen. Genet.* **240**, 29–35.
63. Miller, S., Pesci, E. C. & Pickett, C. (1994). Genetic organization of the region upstream from the *Campylobacter jejuni* flagellar gene flhA. *Gene*, **146**, 31–38.
64. Pesci, E. C., Cottle, D. L. & Pickett, C. L. (1994). Genetic, enzymatic, and pathogenic studies of the iron superoxide dismutase of *Campylobacter jejuni*. *Infect. Immun.* **62**, 2687–2694.
65. Hong, Y., Wong, T., Bourke, B. & Chan, V. L. (1995). An isoleucyl-tRNA synthetase gene from *Campylobacter jejuni*. *Microbiology*, **141**, 2561–2567.
66. Chan, V. L. & Bingham, H. L. (1991). Complete sequence of the *Campylobacter jejuni* glyA gene encoding serine hydroxymethyltransferase. *Gene*, **101**, 51–58.
67. Griffiths, P. L., Park, R. W. & Connerton, I. F. (1995). The gene for *Campylobacter* trigger factor: evidence for multiple transcription start sites and protein products. *Microbiology*, **141**, 1359–1367.
68. Griffiths, P. L., Dougan, G. & Connerton, I. F. (1996). Transcription of the *Campylobacter jejuni* cell division gene ftsA. *FEMS Microbiol. Letters*, **143**, 83–87.
69. Thies, F. L., Karch, H., Hartung, H. P. & Giegerich, G. (1999). The ClpB protein from *Campylobacter jejuni*: molecular characterization of the encoding gene and antigenicity of the recombinant protein. *Gene*, **230**, 61–67.
70. Richardson, P. T. & Park, S. F. (1995). Enterochelin acquisition in *Campylobacter coli*: characterization of components of a binding-protein-dependent transport system. *Microbiology*, **141**, 3181–3191.
71. Galindo, M. A., Day, W. A., Raphael, B. H. & Joens, L. A. (2001). Cloning and characterization of a *Campylobacter jejuni* iron-uptake operon. *Curr. Microbiol.* **42**, 139–143.
72. Hani, E. K., Ng, D. & Chan, V. L. (1999). Arginine biosynthesis in *Campylobacter jejuni* TGH9011: determination of the argCOBD cluster. *Can. J. Microbiol.* **45**, 959–969.
73. Thies, F. L., Weishaupt, A., Karch, H., Hartung, H. P. & Giegerich, G. (1999). Cloning, sequencing and molecular analysis of the *Campylobacter jejuni gro* ESL bicistronic operon. *Microbiology*, **145**, 89–98.

Edited by G. von Heijne

(Received 3 September 2002; received in revised form 17 December 2002; accepted 20 December 2002)