Microbiology (2001), **147**, 2417–2424 Printed in Great Britain

Sigma A recognition sites in the *Bacillus subtilis* genome

Hanne Jarmer,[1] Thomas S. Larsen,[1] Anders Krogh,[1] Hans Henrik Saxild,[2] Søren Brunak[1] and Steen Knudsen[1]

Author for correspondence: Steen Knudsen. Tel: +45 45252480. Fax: +45 45931585. e-mail: steen@cbs.dtu.dk

Center for Biological Sequence Analysis, Building 208[1], and Section for Molecular Microbiology, Building 301[2], BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark

A hidden Markov model of σ^A RNA polymerase cofactor recognition sites in *Bacillus subtilis*, containing either the common or the extended −10 motifs, has been constructed based on experimentally verified σ^A recognition sites. This work suggests that more information exists at the initiation site of transcription in both types of promoters than previously thought. When tested on the entire *B. subtilis* genome, the model predicts that approximately half of the σ^A recognition sites are of the extended type. Some of the response-regulator aspartate phosphatases were among the predictions of promoters containing extended sites. The expression of *rapA* and *rapB* was confirmed by site-directed mutagenesis to depend on the extended −10 region.

Keywords: sigma factor, HMM, response regulator aspartate phosphatase, extended −10 region

INTRODUCTION

To initiate transcription, RNA polymerase (RNAP) has to recognize and bind to the promoter region. In prokaryotic cells this ability resides in the 'specificity' (in Greek s = σ) factor of the RNAP complex. The genome of *Bacillus subtilis* encodes at least 17 different σ factors (Huang & Helmann, 1998). Growing cells utilize at least six different σ factors: the housekeeping σ^A, and σ^B, σ^C, σ^D, σ^H and σ^L, and *B. subtilis* uses yet another four during endospore formation: σ^E, σ^F, σ^G and σ^K. The remaining seven σ factors were identified after sequencing of the complete genome and are all of the extracytoplasmic function (ECF) subfamily (Huang *et al.*, 1998).

The σ factor in the RNAP complex recognizes and binds a specific conserved DNA pattern upstream of the transcription start site, thereby allowing the RNAP to associate with the DNA strand, first loosely in a 'closed promoter–polymerase complex', and then tightly, melting a local region of the promoter to form an 'open promoter–polymerase complex', resulting in initiation of transcription. When the transcription is initiated, the σ factor is released from the complex.

Every σ factor facilitates binding of the RNAP complex by recognition of a specific binding site, usually located 10 and 35 bp upstream of the transcription start site. σ^A in *B. subtilis* participates in the initiation of transcription of most of the housekeeping genes. The consensus sequence recognized by σ^A, 5′-TTGACA-17 nt-TA-TAAT-3′, is identical to the consensus that σ^{70} of *Escherichia coli* recognizes. σ^A-dependent promoters from *B. subtilis* are easily transcribed by the σ^{70} of *E. coli*, but poorly the other way around (Camacho & Salas, 1999), which suggests that the RNAP of *B. subtilis* has a stricter requirement for binding than the RNAP of *E. coli* (Voskuil & Chambliss, 1998; Camacho & Salas, 1999). This corresponds with the fact that earlier studies have shown that many Gram-positives including *B. subtilis* utilize an extended −10 region in a large number of their σ^A-dependent promoters. This region is located 1 bp upstream of the −10 region and is hence referred to as the −16 region. The consensus of this region is 5′-TRTG-3′, where R = G/A (Helmann, 1995; Voskuil & Chambliss, 1998; Camacho & Salas, 1999), and it is therefore larger than the corresponding 5′-TG-3′ motif found in *E. coli* (Ponnambalam *et. al.*, 1986; Keilty & Rosenberg, 1987). This extension is estimated to exist in less than 10% of the promoters in *E. coli* (Chan *et al.*, 1990) and approximately 45% in *B. subtilis*. Especially in promoters containing this extended signal a series of A- and T-rich regions upstream of the −35 region has been observed. And both σ^A-dependent promoter types have an overrepresentation of A residues downstream of the −10 region (Voskuil & Chambliss, 1998). By

Abbreviations: FP, false positive; HMM, hidden Markov model; RNAP, RNA polymerase; TP, true positive.

extracting this information it is possible to create a model for prediction of new sites.

The complete genome sequence (4·2 Mb) of *B. subtilis* was published in November 1997 (Kunst *et al*., 1997), and at present 4228 genes are annotated (SubtiList, 1999). Approximately one-third of these genes have experimentally identified functions. The function of the second third can be predicted by homology to other known gene products. Among the last third of the genes there is most likely an unknown number of misclassified ORFs, and therefore the exact number of genes in *B. subtilis* remains unknown. It is therefore also not possible to estimate how many promoters exist in the genome of *B. subtilis*. If the annotated genes are correct and if only regions upstream of a gene and downstream of a terminator, or regions between genes arranged head-to-head, are defined as promoter regions, a conservative estimation will be that *B. subtilis* has 1800 promoter regions. The fraction of these that is dependent on σ^A is not known.

There are currently no publicly accessible tools for the prediction of σ^A-binding sites in *B. subtilis*. Nobody has attempted to estimate the number of such sites. We have used hidden Markov models (HMMs) and trained them to recognize σ^A-binding sites in *B. subtilis* from existing experimentally generated data. The goal of this work is to create a tool to predict the number of true signals within the genome. This work has the further aim of recovering possible hidden information in the surrounding sequence of the two types of σ^A-binding sites as they are known today. This will clarify the differences between the sites, and make it easier to distinguish between them.

METHODS

Hidden Markov models. The central idea of an HMM is to embed the statistics of a motif in a set of states with transitions between them. Each HMM state has a specific probability distribution over the four nucleotides and hence one may say that it 'emits' nucleotides according to specific emission probabilities. There is a state for each position in the motif and the emission probabilities essentially end up being equal to the nucleotide frequencies at these positions. Hence, an HMM may be viewed either as a generative model which 'emits' nucleotides according to specific statistics or as a scoring model which may be used to answer questions such as: 'To what extent is a given sequence compatible with/similar to the sequences used to train the HMM?'. These two HMM interpretations are equally valid and the choice between them depends on the application in question (for further introduction to HMMs we recommend Durbin *et al*., 1998).

HMMs are generally well suited for searching for motifs like σ^A-binding sites since they facilitate an easy and intuitive incorporation of prior knowledge about signals associated with the motif in question. Another advantage, compared to techniques such as neural networks, is the ease of relating trained model parameters to sequence information; for instance, it is possible to use the trained emission probabilities to directly read off any consensus signals found and to get a good idea of the information present in these signals.

Prior knowledge may be included in the HMM architecture by addition or deletion of states, by biasing their nucleotide emission probabilities and/or biasing the probabilities of transitions between them.

When a model architecture has been set up, the optimal parameters are estimated by the Baum–Welch algorithm, which maximizes the likelihood of the training sequences given the model – i.e. it finds the HMM parameters which best capture the statistics of the training sequences.

The trained model is then used to analyse sequences not included in the training set. To get an idea of the model's ability to generalize, one may split the initial training set into 10 parts and then repeatedly train on nine parts and test on the remaining part, until all parts have been tested once. This is a common technique known as a 10-fold cross-validation. It provides a way of estimating the extent of expected false positives and false negatives for a given threshold, when using the model to decode new sequences.

'Decoding' is the term applied to the process of evaluating how well a sequence or sub-sequence fits a given HMM model. There are several ways to perform decoding, and we have used *posterior* decoding, where one calculates, for the ith nucleotide x_i in the query sequence x of length L, the total probability that the state π_i emitting it is state k, $P(\pi_i = k \mid x)$. Note that in general there are many paths through the model that could have emitted nucleotide x_i while in state k (i.e. for which $\pi_i = k$), so one must add the probabilities of all these parses to get the total probability. Formally, we have:

$$P(\pi_i = k \mid x) = \frac{P(x, \pi_i = k)}{P(x)} \tag{1}$$

The numerator may be written

$$P(x, \pi_i = k) = P(x_1 \ldots x_i,\ \pi_i = k)\ P(x_{i+1} \ldots x_L \mid x_1 \ldots x_i,\ \pi_i = k)$$
$$= P(x_1 \ldots x_i, \pi_i = k)\ P(x_{i+1} \ldots x_L \mid \pi_i = k) \tag{2}$$

since all observations after x_i depend only on π_i. The first and second term in this product may be calculated recursively by the forward and backward algorithm respectively (Durbin *et al*., 1998). The remaining unknown on the right-hand side of equation (1), $P(x)$, may also be obtained from the forward/backward algorithms.

HMM prediction. For predicting sites in the genome we calculate, for each nucleotide, the posterior probability that it was emitted by the first state of the −10 region in a σ^A-binding motif. Once we have the posterior probabilities of the −10 start-states at all nucleotide positions, we simply regard all probabilities above a certain threshold (determined by the cross-validation procedure described above) as statistically significant. Hence, whenever the posterior probability of the desired motif exceeds the threshold, the model is said to have 'found' a motif at that nucleotide. The better the motif and its contextual signals fit the model, the higher the probability score, and the more confidence will be placed in the prediction.

Fig. 1 is a schematic view of the HMM used to predict σ^A promoters in *B. subtilis*. The model incorporates known information about conserved positions in σ^A binding (Helmann, 1995; Voskuil & Chambliss, 1998), and was trained to pick up additional unknown signals.

'BACKGROUND' is a state whose emission probabilities are obtained from a first-order HMM trained on the entire *B. subtilis* genome (direct strand). This represents a null model.

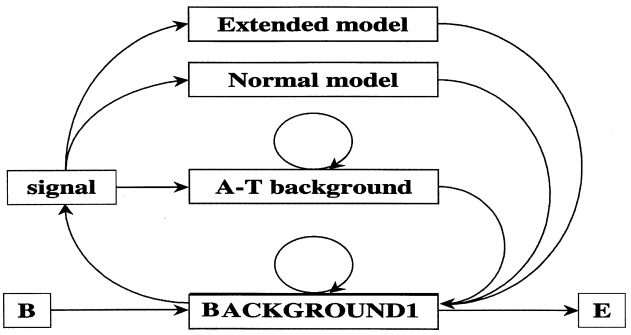

Fig. 1. A schematic drawing of the HMM used to predict σ^A-binding sites in *B. subtilis*. Each box indicates a submodel (see text). The arrows show the possible transitions between submodels. A circular arrow indicates that the model is allowed to loop (stay in the same state for more than one base in the sequence) at the given position (Durbin *et al.*, 1998).

The reason for using a low-order Markov chain for the background is to avoid modelling the genome too explicitly, since most of the genome is presumably coding whereas the promoters generally reside in non-coding regions. If the promoter finder were combined with a gene finder, one could train the background state on supposed non-coding regions and conceivably improve the signal-to-noise ratio. However, it is unclear how the promoter-finding performance would be affected in coding regions. This is certainly a possible path of further investigation.

The model shown in Fig. 1 is used exclusively for decoding (testing). For training purposes a loop model should be avoided, since in the absence of fully labelled training sequences it may end up using several motifs in its maximum-likelihood estimation even though only one is actually present; this will then distort the statistics of the motif states and hence impair decoding performance. Thus, during training a second background (identical to 'BACKGROUND') is included on the right-hand side of Fig. 1 in such a way that all three signal states must pass on to this second background and from here to the end state 'E'.

Promoter regions as well as other intergenic regions are known to be comparatively A and T rich. Hence, in order to prevent prediction of σ^A signals merely on the basis of A and T richness, a state has been added to the model ('A-T background' in Fig. 1). The 'signal' state is included purely for technical reasons in order to switch from the trained left–right model to the looping prediction model shown in Fig. 1.

Note the presence of two alternative σ^A-binding site models in Fig. 1. This is motivated by the finding of two different submodels of binding sites – one with an extended -10 region (extended) and one without (normal) (Helmann, 1995; Voskuil & Chambliss, 1998). As dictated by the data in the training set each model allows a separation of the -10 and -35 region of 16–21 bp and 4–10 bp between the -10 region and the start site of transcription (Helmann, 1995).

Fig. 2 shows a more detailed view of the extended and normal submodels. The consensus sequences of the -10 and -35 regions are clearly marked, as are the A-T-rich states. Dotted lines indicate the presence of more states than could be comfortably shown. The presence of the 9 and 5 extra explicitly modelled states in the extended model reflects the expectation that there is more information in the binding sites. The rationale behind the self-looped states is to model length distributions between e.g. the signal state (Fig. 1) and the start of the -35 region. There are two more looped states in the normal model in order to compensate for the 9 explicit states in the extended model. If the length modelling differed markedly between the extended and normal model, one would risk situations where a submodel was preferred merely on the basis of length. The emission probabilities of the looped states are identical to the background state.

Datasets. Only sequences from experimentally verified σ^A promoters were used for training and testing. We obtained these sequences from a list on John D. Helmann's worldwide web page (http://www.bio.cornell.edu/microbio/helmann/helmann.html) (Helmann, 1995) containing 236 σ^A-dependent promoters. Using a subset of these, which have supporting

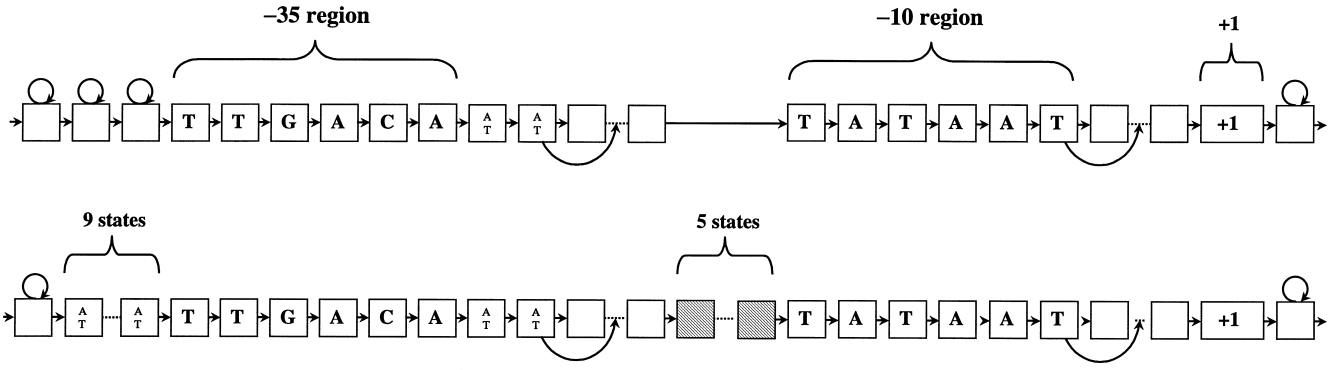

Fig. 2. A schematic drawing of the two σ^A-binding site submodels, the normal (upper) and the extended (lower). This figure uses the same symbols as Fig. 1. Background states are white without letters, the -10, -35 and $+1$ regions are indicated, the TG motif is hatched and other explicitly modelled states are indicated by letters. A dotted line between two boxes indicates that the number of states in this region is greater than two. An arrow pointing at the dotted line between two boxes symbolizes that the model is allowing a bypass of states. This allows the number of states between the -10 and the -35 regions to vary between 16 to 21 and likewise 4 to 10 between the -10 region and the $+1$ state.

Table 1. Bacterial strains and plasmids used in this study

Strain/plasmid	Genotype/description	Ref./source
B. subtilis		
HH263	*trp*C2 (168 wild-type)	C. Anagnostopoulos*
HØJ1	*trp*C2 *amyE*::pB1	HH263/pB1, Neo^R
HØJ4	*trp*C2 *amyE*::pB4	HH263/pB4, Neo^R
HØJ5	*trp*C2 *amyE*::pA1	HH263/pA1, Neo^R
HØJ8	*trp*C2 *amyE*::pA4	HH263/pA4, Neo^R
E. coli		
MC1061	F$^-$ *araD139* Δ(*ara–leu*)7696 *galK16* Δ(*lac*)X74 *rpsL* (Str^R) *hsdR2* (r$^-$ m$^-$) *mcrA mcrB*	Stratagene
Plasmids		
pDG268neo	Ap^R (*E. coli*), Neo^R (*B. subtilis*); pBR322 derivative; vector used for integration of transcriptional *lacZ* fusions into the *amyE* gene	Saxild *et al.* (1996)
pB1	Ap^R (*E. coli*), Neo^R (*B. subtilis*); *Eco*RI–*Bam*HI PCR fragment containing the wild-type promoter of *rapB* (260 bp, 23 bp upstream of the ORF) cloned into pDG268	This work
pB4	As pB1 with a base substitution G to C (in the extended −10 region 42 bp upstream of the ORF)	This work
pA1	Ap^R (*E. coli*), Neo^R (*B. subtilis*); *Eco*RI–*Bam*HI PCR fragment containing the wild-type promoter of *rapA* (93 bp, 27 bp upstream of the ORF) cloned into pDG268	This work
pA4	As pA1 with a base substitution G to A (in the extended −10 region, 43 bp upstream of the ORF)	This work

* C. Anagnostopoulos, INRA, Jouy en Josas, France.

experimental data, and which are labelled at the transcription start sites (109), combined with some (11) determined in our laboratory (H. H. Saxild, unpublished results) and some (10) found in existing literature (Huang *et al.*, 1998; Huang & Helmann, 1998; Lewis *et al.*, 1998; SubtiList, 1999; Zhang & Begley, 1991), a list of 130 σ^A-dependent promoters was constructed. The 130 sequences are 100 bp long and range from approximately −85 to +15 relative to the transcription start site.

Bacterial strains, plasmids and growth conditions. The bacterial strains and plasmids used in this study are listed in Table 1. Strains HØJ1, HØJ4, HØJ5 and HØJ8 have a single-copy *rap–lacZ* transcriptional fusion inserted by a double-crossover recombination event at the *amyE* locus, with and without a base substitution in the extended −10 region. Cells were grown at 37 °C as described previously (Saxild *et al.*, 1995). Spizizen salt-buffered minimal medium supplemented with 100 μg L-tryptophan ml^{-1} was used in the enzyme assay and Luria–Bertani (LB) broth was used as rich medium. The relevant antibiotics were used at the following concentrations: neomycin, 5 μg ml^{-1}; ampicillin, 50 μg ml^{-1}.

DNA manipulations and genetic techniques. Chromosomal and plasmid DNA was isolated as described previously by Saxild *et al.* (1996). Treatment of DNA with restriction enzymes and T4 DNA ligase was performed as recommended by the supplier. Transformations of *E. coli* and *B. subtilis* were performed as described previously by Saxild *et al.* (1996). DNA sequencing was performed by the chain-termination reaction method using dideoxyribonucleotides as described by Sanger *et al.* (1977) using the Amersham Pharmacia Biotech Thermo Sequenase radio-labelled termination cycle sequencing kit. All sequencing was done with double-stranded plasmid, and was performed as described by the supplier. All PCRs were performed as described previously (Zeng & Saxild, 1999). PCR product DNAs were isolated by the use of GFX PCR DNA and gel band purification tubes from Amersham Pharmacia Biotech.

Construction of clones. The promoter regions from *rapA* and *rapB* were obtained by a PCR on chromosomal DNA from the wild-type *B. subtilis* strain 168. In addition to the wild-type promoter region, site-directed mutations were incorporated in the extended −10 region by using PCR primers with mismatches. The amplified promoter fragments with a 5′ *Eco*RI linker and a 3′ *Bam*HI linker were cloned in a transcriptional fusion with the reporter gene *lacZ* using the vector pDG268neo. The plasmid was amplified in *E. coli* MC1061, linearized with *Kpn*I and transformed into *B. subtilis* HH263. The transcriptional fusion was integrated into the *amyE* gene by a double-crossover event (Saxild *et al.*, 1996). All strains containing a transcriptional fusion were confirmed by colony PCR with relevant primers, sequencing of the cloned promoter region and verification of the AmyE$^-$ phenotype, by screening for inability to produce clearing zones on LB plates containing 1% starch. Primers complementary to regions on each side of the cloned fusion were used in PCRs to confirm that no double insertion had occurred.

Primer extension. RNA was isolated from HØJ1 and HØJ5 as described by Saxild *et al.* (1995). The single-stranded DNA primer (annealing just downstream of the *Bam*HI cloning site in pDG268) was radiolabelled at the 5′ terminus using T4 polynucleotide kinase and [γ-^{33}P]ATP. The primer extension was performed by using the displayTHERMO-RT Reverse Transcriptase kit from Display Systems Biotech. The radio-labelled cDNA probes were separated on a 6% poly-

acrylamide sequencing gel next to a sequencing of pB1/pA1 with the same primer, and visualized by autoradiography.

β*-Galactosidase activity assay.** Growing cells were harvested by pouring 25–30 ml culture into a 50 ml centrifuge tube 1/3 full of ice, centrifuging at 7000 ***g for 5 min, washing with 10 ml of a 0·9% NaCl solution, centrifuging at 7000 ***g*** for 5 min, washing with 2 ml of a 0·9% NaCl solution, centrifuging at 15000 ***g*** for 2 min, discarding the liquid phase, gently adding 0·5 ml 30 mM phosphate buffer (pH 7·5), 1 mM EDTA and 1 mM DTT (sonication buffer) without resolving the pellet, and stored at −20 °C. The total amount of protein was determined by the Lowry method. The *β*-galactosidase activity assay was performed using the method of Miller (1972).

RESULTS

Fig. 3 shows the average performance of the trained model on the test sets in the cross-validation experiment. The true positive (TP) rate is the fraction of test sequences which are predicted by the model to be σ^A sites. Ideally, TP should be 1, which would correspond to a sensitivity of 100%, but this is hardly ever feasible without paying a price in terms of a high false positive (FP) rate. The FP rate is the fraction of non-σ^A sites which are nevertheless identified by the model to be σ^A sites. Hence, ideally one wants FP equal to zero, in which case the specificity of the model is 100%. Note that in order to calculate the FP rate, one really needs a set of sequences to which one is sure that σ^A does not bind. Such a set is currently difficult if not impossible to obtain, so in the absence of a better alternative we used, for each of the 10 trained models, 1000 randomly generated sequences of length 100 with a statistic equal to 'BACKGROUND' statistics.

In most classification scenarios there is a trade-off between sensitivity and specificity, and one has to find a balance (by choosing a threshold) which is sensible for the application in question. From Fig. 3 it is clear that one can achieve a TP rate of about 0·7 with a very low FP rate. Thus the probability threshold corresponding to this sensitivity was our chosen threshold. Note that the false negative rate at this threshold is 1 − TP = 0·3, meaning that 30% of all true sites are not reported.

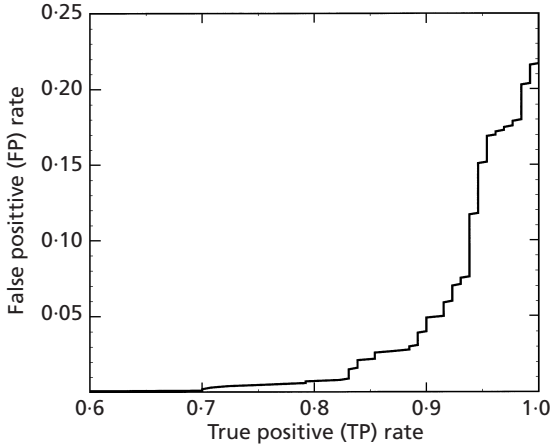

Fig. 3. The rate of false positives versus the rate of true positives.

The predicted signals always contain the whole σ^A-binding motif, and the expected transcription start site, but are reported based on the score from the −10, rather than the −35, sequence. As observed in *E. coli*, the −10 region of *B. subtilis* tends to occur unaccompanied by a −35 signal (or accompanied by an extremely poor one) in σ^A-binding promoters (Camacho & Salas, 1999), though usually dependent on an activator (Lewis *et al.*, 2000). The converse is relatively rare, and presumably such single −35 sites are not sufficient to bind σ^A and should therefore not be counted as hits.

In order to estimate the number of FPs made on the entire genome we generated 100000 random sequences of length 100 with 'BACKGROUND' statistics and counted the number of sequences scoring higher than the chosen cutoff. We did this three times and got 185, 199 and 225 sequences respectively. We then simply assume that the mean of these numbers, 203, is the expected number of FPs made on 100000 candidates. In addition to the first-order Markov statistics used in 'BACKGROUND', we also tried generating the random sequences from kth-order Markov chains for $k = 0$, 2 and 3. The number of sequences found on average in these cases was 791, 265 and 270 respectively – i.e. they all performed worse than the chosen $k = 1$.

Note that it is conceivable that some of the high-scoring random sequences would in fact bind σ^A in an experimental setup and hence are not really FPs. Nevertheless, this is our best estimate. In a genome of length 4·2 Mb the model is therefore expected to find roughly $(4{\cdot}2 \times 10^6/100)(203/100000) = 85$ FPs on both the positive and the negative strand, making a total of 170.

Using the HMM we predict that the entire genome of *B. subtilis* contains 2538 σ^A-binding sites. When examining the list containing the reported results (1927 high-confidence predictions) we were able to locate 1127 of these within the 400 bp upstream regions of the 4228 predicted genes in *B. subtilis* (SubtiList, 1999). Both these lists are available from the authors upon request.

The model further predicts that approximately 50% of the predicted sites are of the 'extended' type, which is a little more than previous findings on smaller samples (45%) (Helmann, 1995).

The sequence logos in Fig. 4 show the profiles of the HMM predictions. From this it is clear that more information exists in both types of σ^A-binding sites than previously thought. It is especially clear that our model has found that the transcription start site in *B. subtilis* promoters dependent on σ^A is highly conserved. The consensus sequence of this signal is 5′-Y**R**TA-3′ (+1 in bold) in the normal type, and 5′-Y**R**NA-3′, where Y = C/T, R = A/G and N = nt, in the extended type. The most frequent observed +1 signal in σ^A-binding

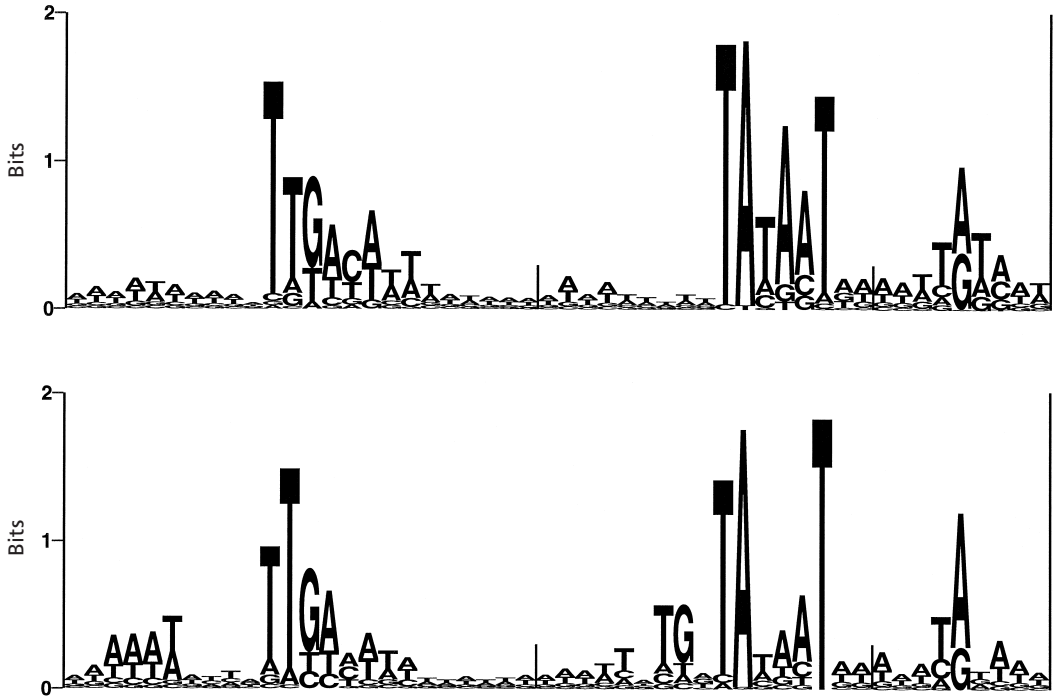

Fig. 4. Logos of the predicted σ^A-binding sites in *B. subtilis*. The logos shown are merged from six individual logos (the merging positions are shown in the figure by horizontal bars), each containing either the −10 or the −35 signal from either the normal (top logo) and the extended (bottom logo) type of σ^A recognition sites. Each of the −10 and the −35 signals represents approximately 500 signals predicted by the HMM. Each of the +1 logos is generated from 350 predicted signals. The logos are constructed by aligning the six types of signals on the first base of the reported signal. For the +1 signal, this base is represented by the highest peak in that area, with an A on the top in both types of binding site. The Shannon information content is shown on the *y* axis; Shannon's unit of non-randomness is the bit (short for 'binary digit') (Shannon, 1948).

promoters is, according to these results, 5′-TATA-3′. The signal in the +1 position is strongest in the extended type, and here an A is much more frequent than a G. This may be to compensate for the fact that the +2 position in this type of binding site is less conserved.

The model has, as expected, found that the extended type of σ^A-binding sites has an A- and T-rich area approximately 4 bp upstream of the −35 region, and that the 3′ end of the consensus of the −35 region in this type of promoters is poorly conserved. The middle section of the −10 region is likewise less conserved in the promoters of the extended type when compared to the normal type. The −16 motif is found to be 5′-TNTG-3′, which almost corresponds with the findings in other Gram-positives and previous findings in *B. subtilis* (Helmann, 1995; Voskuil & Chambliss, 1998; Camacho & Salas, 1999). Both types of promoters seem to have a slightly conserved tail with a length of 2 nt of Ts and As following the −35 region, and likewise it is found that the level of As is above average downstream of the −10 region.

The model identified σ^A-binding sites in the expected promoter regions of eight of the response-regulator aspartate phosphatase encoding genes (the *rap* genes). Our model predicts that *rapB*, *-D*, *-E*, *-I* and *-K* are transcribed from a σ^A-dependent promoter using the extended σ^A-binding site, and that *rapF*, *-G* and *-H* have the normal σ^A-binding site.

rapB	ATACATTA**TG**AT**A**AA**AT**ATAACCAA
rapA	TGTAAATA**TG**AT**A**AA**AT**ATGACATA
rapC	TATAAACA**TG**AT**A**AA**AT**ATGACATA
rapF	AATGTTGA**TG**AT**A**AA**AT**ATGACATA
rapJ	AACAGCTA**TG**AT**A**AA**AT**ATAACATA
rapD	AAAAGTTA**TG**AT**A**TG**AT**AATTATAG
rapH	TTTGGGAT**TG**AT**A**GA**AT**ATGACATA
rapI	GGTTATTC**TG**AC**A**TA**AT**ACAATTAA
rapK	AATGACTA**TG**TT**A**TG**AT**TGTTTTCG
rapE	CGAAAACT**TG**TT**A**AT**AT**TTACAGTA
rapG	GAAAGAGG**TG**TT**A**CT**AT**CAGAATAA

Fig. 5. A multiple alignment of the expected −10 region of the *rap* genes. Fully conserved residues are shown in bold. The TG motif is observed in all the *rap* genes. The positions of transcription start are shown by underlining the first base of the transcript. The transcription starts indicated for *rapA* and *rapD* are experimentally verified (Mueller *et al.*, 1992; Huang & Helmann, 1998), the rest are solely predicted by the HMM, having a score higher than the cutoff used.

In Fig. 5 the expected −10 regions of the *rap* genes are aligned. There appears to be a highly conserved con-

Table 2. *β*-Galactosidase activity

Strain	Relevant genotype	*β*-Galactosidase activity (±SD)*
HØJ1	Wild-type	45 (±18)
HØJ4	G_{42}→C	4·1 (±2·3)
HØJ5	Wild-type	710 (±290)
HØJ8	G_{43}→A	170 (±71)

* The *β*-galactosidase activities are reported as the mean (±standard deviation) of eight independent measurements.

sensus containing a TG motif immediately upstream of a −10 motif in all the *rap* genes, which at least implies that this group of genes are being transcribed from a promoter containing an extended −10 region.

The aligned putative extended −10 region for *rapF* in Fig. 5 is not the one predicted by the model. The model predicts a −10 region 10 bp further downstream. The sequence shown in Fig. 5, however, aligns with the experimentally verified −10 region in the promoter region of *rapA* and *-D*. We suggest that *rapF* might utilize both putative σ^A-binding sites.

Experimental verification of predicted extended sites

We tested these predictions by site-directed mutagenesis of the extended region within the predicted σ^A-binding site. The site-directed mutagenesis (see Table 2) indeed showed a decrease in transcription for both *rapA* and *rapB* throughout a sporulation experiment (*rapA*, *-B* and *-E* are known to play a role in the phosphorelay signal-transduction system of sporulation: Mueller *et al*., 1992; Jiang *et al*., 2000), confirming that this region is necessary for transcription. When the G in the TG motif in the promoter region of *rapB* is substituted with a C, the amount of transcript drops on average 10-fold in the sporulation experiment. Likewise, when the corresponding G upstream of *rapA* is substituted with an A, the level of transcription drops approximately fourfold.

Primer extension of *rapA* confirmed two previously mapped transcription start sites (1 bp apart) (see Fig. 5) (Mueller *et al*., 1992). For *rapB*, we were unable to detect any clear signal in repeated experiments, presumably due to the lower expression level of this gene, to instability of the messenger, or to both.

DISCUSSION

By using the HMM-based prediction tool we have constructed, we are able to predict that the genome of *B. subtilis* contains roughly 2538 σ^A-binding sites. We have generated a list containing 1127 binding sites, which are located within the 400 bp sequences upstream of predicted genes. By examining Fig. 3 it is clear that the constructed model can predict almost 70% of the true sites, virtually without predicting sites that do not actually bind the σ^A factor. It is also clear that the model can be used to predict an even larger percentile of the true sites with a low level of false positive (FPs). The model would predict only 1% FPs when predicting 83% of all true binding sites, or 7% when predicting 94%. In cases where a rate of FPs of almost 22% is acceptable, all true binding sites would theoretically be predicted.

When using the chosen cutoff, we are unable to locate approximately 30% of the true σ^A-binding sites. These false negatives are binding sites that in a variety of ways differ from the average σ^A-binding sites. One example of true σ^A-binding sites that this prediction tool has difficulties in finding are the Spo0A-activated promoters. These promoters are known to have one or several 0A boxes at or near the −35 region, where Spo0A∼P binds and activates transcription. This binding abolishes the negative effect of not only the poorly conserved −35 regions, but also the exceptionally large separation between the −10 and the −35 region (more than 21 bp), which promoters of this type are known to have (Lewis *et al*., 2000). These sites do not fit the model due to the fact that the model only allows a spacing of 16–21 bp. Despite this drawback, we chose to accept this restriction because it gave rise to the model with the best overall performance.

The large spacing and poorly conserved −35 regions, which are often observed in activator-dependent promoters, could explain why the model does not find any true σ^A-binding sites in either *rapA*, *rapC*, *rapE*, or the putative second site in *rapF*, though there apparently exists a strong signal for an extended −10 region in this group of genes. *rapA*, *rapC* and *rapE* are known to be activated by the binding of ComA∼P to a ComA box upstream of the −35 region (Mueller *et al*., 1992; Lazazzera *et al*., 1999; Jiang *et al*., 2000). When the expected promoter regions of the *rap* genes are aligned, it appears that *rapF* has a ComA box consensus site at the same position as *rapA*, *rapC* and *rapE*, which strongly suggests that expression of *rapF* is also dependent on ComA∼P (alignment not shown).

In Fig. 4 it is observed that the HMM has found a highly conserved signal at the +1 position. It appears that the site of initiation of transcription in σ^A-dependent promoters is separated from the −10 Pribnow box by on average 7 bp and has the consensus 5′-pyrimidine-purine-T-A/C-3′ (most frequent: 5′-T**A**TA-3′), and starting transcription at the purine. This corresponds with findings in *E. coli*, where the initiation site in σ^{70}-dependent promoters is -purine-pyrimidine- (Rosenberg & Court, 1979; Pedersen & Engelbrecht, 1995).

From this work, it is suggested that the σ^A-binding sites classified as extended are significantly different from normal σ^A-binding sites in two areas of the promoter sequence. These differences are the −16-motif and the four bases approximately 40 nt upstream of the initiation site, which seem to be rich in A and T. The extended type has likewise been found to be less conserved at position 5 of the −35 region (the C in TTGACA) and at the +2 position in the +1 motif.

In conclusion, we have constructed an HMM that has identified σ^A-binding sites in *B. subtilis* with known sensitivity and specificity. We have estimated the total number of σ^A-binding sites to be around 2538, and found the ratio between extended and normal −10 regions of σ^A-binding sites to be around 1:1. To support these findings we have experimentally verified that two of the predicted promoters indeed depend on an extended type of σ^A-binding site.

The trained HMM is available from the authors upon request. The list of predictions from the trained HMM is available as supplementary data with the online version of this paper at http://mic.sgmjournals.org.

ACKNOWLEDGEMENTS

We thank John D. Helmann for allowing us easy access to his list of σ^A-dependent promoters. We would also like to thank Kristine Bøje Dahlin for her technical assistance in the laboratory. This work was supported by Novo Nordisk A/S and the Danish National Research foundation.

REFERENCES

Camacho, A. & Salas, M. (1999). Effect of mutations in the 'extended −10' motif of three *Bacillus subtilis* sigmaA–RNA polymerase-dependent promoters. *J Mol Biol* **286**, 683–693.

Chan, B., Spassky, A. & Busby, S. (1990). The organization of open complexes between *Escherichia coli* RNA polymerase either with or without consensus −35 sequences. *Biochem J* **270**, 141–148.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). Markov chains and hidden Markov models. In *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*, pp. 46–80. Cambridge, MA: Cambridge University Press.

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res* **23**, 2351–2360.

Huang, X. & Helmann, J. D. (1998). Identification of target promoters for the *Bacillus subtilis* sigma(X) factor using a consensus-directed search. *J Mol Biol* **279**, 165–173.

Huang, X., Fredrick, K. L. & Helmann, J. D. (1998). Promoter recognition by *Bacillus subtilis* sigma W: autoregulation and partial overlap with the sigma X regulon. *J Bacteriol* **180**, 3765–3770.

Jiang, M., Grau, R. & Perego, M. (2000). Differential processing of propeptide inhibitors of Rap phosphatases in *Bacillus subtilis*. *J Bacteriol* **182**, 303–310.

Keilty, S. & Rosenberg, M. (1987). Constitutive function of a prositively regulated promoter reveals new sequences essential for activity. *J Biol Chem* **262**, 6389–6395.

Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Lazazzera, B. A., Kurtser, I. G., McQuade, R. S. & Grossman, A. D. (1999). An autoregulatory circuit affecting peptide signaling in *Bacillus subtilis*. *J Bacteriol* **181**, 5193–5200.

Lewis, R. J., Brannigan, J. A., Offen, W. A., Smith, I. & Wilkinson, A. J. (1998). An evolutionary link between sporulation and prophage induction in the structure of a repressor:anti-repressor complex. *J Mol Biol* **283**, 907–912.

Lewis, R. J., Krzywda, S., Brannigan, J. A., Turkenburg, J. P., Muchová, K., Dodson, E. J., Barák, I. & Wilkinson, A. J. (2000). The *trans*-activation domain of the sporulation response regulator Spo0A revealed by X-ray crystallography. *Mol Microbiol* **38**, 198–212.

Miller, J. H. (1972). Assay of β-galactosidase. In *Experiments in Molecular Genetics*, pp. 352–355. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Mueller, J. P., Bukusoglu, G. & Sonenshein, A. L. (1992). Transcriptional regulation of *Bacillus subtilis* glucose starvation-inducible genes: control of *gsiA* by the ComP–ComA signal transduction system. *J Bacteriol* **174**, 4361–4373.

Pedersen, A. G. & Engelbrecht, J. (1995). Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Proc Int Conf Intell Syst Mol Biol* **3**, 292–299.

Ponnambalam, S., Webster, C., Bingham, A. & Busby, S. (1986). Transcription initiation at the *Escherichia coli* galactose operon promoters in the absence of the normal −35 region sequences. *J Biol Chem* **261**, 16043–16048.

Rosenberg, M. & Court, D. (1979). Regulation sequences involved in the promotion and termination of RNA transcription. *Annu Rev Genet* **13**, 319–373.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467.

Saxild, H. H., Jacobsen, J. H. & Nygaard, P. (1995). Functional analysis of the *Bacillus subtilis purT* gene encoding formate-dependent glycinamide ribonucleotide transformylase. *Microbiology* **141**, 2211–2218.

Saxild, H. H., Andersen, L. N. & Hammer, K. (1996). *dra–nupC–pdp* operon of *Bacillus subtilis*: nucleotide sequence, induction by deoxyribonucleosides, and transcriptional regulation by the *deoR*-encoded DeoR repressor protein. *J Bacteriol* **178**, 424–434.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Tech J* **27**, 379–423, 623–656.

SubtiList (1999). Release R15.1. Current URL http://bioweb.pasteur.fr/GenoList/SubtiList.

Voskuil, M. I. & Chambliss, G. H. (1998). The −16 region of *Bacillus subtilis* and other gram-positive bacterial promoters. *Nucleic Acids Res* **26**, 3584–3590.

Zeng, X. & Saxild, H. H. (1999). Identification and characterization of a DeoR-specific operator sequence essential for induction of *dra–nupC–pdp* operon expression in *Bacillus subtilis*. *J Bacteriol* **181**, 1719–1727.

Zhang, Y. & Begley, T. P. (1991). Cloning, sequencing and regulation of *thiA*, a thiamin biosynthesis gene from *Bacillus subtilis*. *Gene* **198**, 73–82.

Received 31 January 2001; revised 22 May 2001; accepted 6 June 2001.