Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models

Pierre Nicolas1,2,*, Laurent Bize3, Florence Muri2, Mark Hoebeka1, François Rodolphe1, S. Dusko Ehrlich3, Bernard Prum2 and Philippe Bessières1

1Laboratoire de Mathématique, Informatique et Génome, INRA, Route de Saint-Cyr, F-78026 Versailles cedex, France, 2Laboratoire de Statistique et Génome, CNRS, Tour Évry2, 523 place des terrasses de l’Agora, F-91034 Évry, France and 3Laboratoire de Génétique Microbienne, INRA, F-78352 Jouy-en-Josas cedex, France

Received August 6, 2001; Revised and Accepted January 24, 2002

ABSTRACT

We present here the use of a new statistical segmentation method on the *Bacillus subtilis* chromosome sequence. Maximum likelihood parameter estimation of a hidden Markov model, based on the expectation-maximization algorithm, enables one to segment the DNA sequence according to its local composition. This approach is not based on sliding windows; it enables different compositional classes to be separated without prior knowledge of their content, size and localization. We compared these compositional classes, obtained from the sequence, with the annotated DNA physical map, sequence homologies and repeat regions. The first heterogeneity revealed discriminates between the two coding strands and the non-coding regions. Other main heterogeneities arise; some are related to horizontal gene transfer, some to f-rich enriched composition of hydrophobic protein coding strands, and others to the codon usage fitness of highly expressed genes. Concerning potential and established gene transfers, we found 9 of the 10 known prophages, plus 14 new regions of atypical composition. Some of them are surrounded by repeats, most of their genes have unknown function or possess homology to genes involved in secondary catabolism, metal and antibiotic resistance. Surprisingly, we notice that all of these detected regions are a + t-richer than the host genome, raising the question of their remote sources.

INTRODUCTION

Numerous factors are known to affect statistical composition of chromosome DNA sequences, such as constraints related to coding properties (1), gene transfers (2) and statistical biases related to replication (3). Horizontal gene transfer between bacteria species (4), often due to mobile elements, is now recognized as playing an important role in the acquisition of adaptive traits, such as pathogenicity (5–7), resistance to antibiotics (8) or heavy metals, such as mercury (9,10) or arsenic (11). Horizontal transfer has been shown to occur in a wide variety of ecosystems (12), raising questions about the consequences of dispersion of genetic constructions from genetically modified organisms. More generally, horizontal transfer is considered as a driving force of bacterial evolution (13–15). Bacteria are known to integrate prophages (16), and to have other ways of integrating foreign DNA sequences (17–19). These transfers can correspond to DNA segments, which have different statistical properties from those of the host. A classical method of horizontal transfer detection, based on codon usage frequencies, has been introduced by Médigue et al. (20), and some other approaches have been recently reviewed in Karlin (21).

Hidden Markov models (HMMs) are good statistical tools for the analysis of this heterogeneity (22–27). We applied them to the *Bacillus subtilis* chromosome (4.2 Mb long). In these models, one assumes that a DNA sequence is made up of successive segments, each one belonging to one of a finite number q of types.

*To whom correspondence should be addressed at: Laboratoire de Mathématique, Informatique et Génome, INRA, Route de Saint-Cyr, F-78026 Versailles cedex, France. Tel: +33 1 30 83 33 52; Fax: +33 1 30 83 33 59; Email: nicolas@versailles.inra.fr
Present address: Laurent Bize, Laboratoire de Biométrie, INRA, chemin de Borde-Rouge, Auzeville, BP 27, F-31326 Castanet-Tolosan cedex, France
MATERIALS AND METHODS

Hidden Markov models

A DNA sequence can be represented by a finite series \(y_1, \ldots, y_n\), each base \(y_i\) being taken from the alphabet \(Y = \{a, c, g, t\}\). HMMs are characterized by two processes [see for instance Rabiner (32)]. The hidden state process \(s = \{s_1, \ldots, s_n\}\), such that \(s_j \in S = \{1, \ldots, q\}\), which in our set-up governs the succession of the segment types along the sequence, and the observed process \(y = \{y_1, \ldots, y_n\}\) which corresponds to the observed DNA sequence. Hidden states are generated according to a homogeneous first order Markov chain (M1) with transition probabilities \(P(s_i = v | s_{i-1} = u), u, v \in S\). On condition that the hidden process \(s = \{s_1, \ldots, s_n\}\), the observed process \(y = \{y_1, \ldots, y_n\}\) is a heterogeneous Markov chain: base \(y_i\) appears in the sequence with a probability distribution that depends on the actual hidden state \(s_i\), as well as on previous bases \(y_{i-k}, \ldots, y_{i-1}\). Higher order Markovian dependencies will not be considered for the hidden chain, as we expect to identify large compositional segments, but the number \(q\) of hidden states will vary; similarly the order and other structural features of the observed chain will also vary. Thus, according to the characteristics of these Markov chains, several models of interest can be constructed.

The M1-M0 model assumes that, conditional on the hidden state \(s_i\), nucleotides \(y_i\) are drawn independently with a probability \(P(y_i = j | s_i = u), j \in Y, u \in S\). Hence, this model takes into account the local base composition of the sequence, and corresponds to the classical HMM described in the literature. More generally, the M1-Mk model assumes, conditional on the actual hidden state, a \(k\) markovian dependence between observations, with a transition probability \(P(y_i = j | y_{i-1} = i, \ldots, y_{i-k} = i_{k-1}, s_i = u)\). This model, introduced by Churchill (22), accounts for the local \(k+1\) nucleotide frequencies of the DNA sequence. We denote the whole set of model parameters by \(\theta\) i.e. transition probabilities between states and between bases.

Given the sequence \(y\), we consider the maximum likelihood estimator of \(\theta\). Consistency and normality results, which justify the maximum likelihood approach, were proved in the M1-M0 model by Baum and Petrie (31), and extended to the M1-Mk model by Muri (23). Several methods exist for estimating the parameters of HMMs, including stochastic likelihood maximization algorithm and Bayesian estimation (23,24,32–36). All these methods do not require any learning set of pre-segmented sequences to estimate \(\theta\). They only require specification of the model structure (number of states, \(q\), and order of the model, \(k\)). In our study we chose not to introduce prior information, and choose the expectation-maximization algorithm (EM) to maximize likelihood \(P(y | \theta)\) (25), which proved to be one of the most effective.

Hidden states are missing data, and the likelihood is a sum over all hidden state paths \(P(y | \theta) = \sum_{s \in \mathcal{S}} P(y, s | \theta)\), which makes it not directly tractable. The EM algorithm is useful in many estimation problems involving missing data, including HMM. It is an iterative procedure that alternates two steps, see Churchill (22), Rabiner (32) and Durbin et al. (35) for detailed description of the HMM case and Dempster et al. (37) for mathematical proof of the convergence toward the maximum likelihood estimator. Given the current value \(\theta^{(m)}\) the expectation E-log likelihood \(Elog(P(s, \theta^{(m)} | y, \theta^{(m)})\) is computed during the E-step and maximized over \(\theta\) during the M-step. In the HMM context, the E-step consists of computing the probability of two consecutive hidden states \(P(s_{i-1} = u, s_i = v | y, \theta^{(m)})\) from which follows \(P(s_i = v | y, \theta^{(m)})\). These probabilities are computed using the Baum–Welch forward–backward recurrence. A new value \(\theta^{(m+1)}\) is obtained in the M-step which increases the likelihood:

\[
P^{(m+1)}(y_{1:t} = y | \theta^{(m+1)}) = \sum_{s_{1:t}} P(s_{1:t} = u, v | y, \theta^{(m+1)})
\]

where \(1 \leq i_1 \leq \ldots \leq i_{v-1} \leq i_v \leq \ldots \leq i_{k-1} \leq i_k \leq t\)

where \(m\) is equal to 1 if the sentence between the brackets is true and 0 otherwise. E and M steps are alternated until we come to an iteration \(M\) for which numerical convergence is reached.

Every limit point of a sequence \((\theta^{(m)})_{m \geq 0}\) generated by EM, satisfies the log-likelihood equations, and \((\theta^{(m)})_{m \geq 0}\) converges towards the maximum likelihood estimator, if the starting point \(\theta^0\) is not too far from the true value of the parameter \(\theta\) (23). This is why we run EM with multiple random initializations, and then select the final result presenting the highest likelihood. Computational cost of the algorithm is proportional to sequence length and to the square of the number of states. Obviously cost also grows proportionally to the number of required iterations which depends on the smoothness of the likelihood landscape. Memory requirement is proportional to sequence length and the number of states, but approximations of the E-step could be done to bypass this problem; they were implemented but not used here.

In order to identify homogeneous segments in \(y\), probabilities of each hidden state were computed at each position \(P(s_i = u | y, \theta^{(M)})\) using the forward–backward recurrence with the maximum likelihood estimator \(\theta^{(M)}\). We did not use the popular Viterbi algorithm (38–41), which consists, given the sequence, of computing the most probable path of the hidden states. In the case of our poorly structured model Viterbi reconstruction is less informative than the one obtained by forward–backward recurrence. Nevertheless, for results interpretation and discussion, we will need to recognize segments. Hence, contiguous positions having \(v\) as the most probable hidden state [i.e. where \(v\) maximizes \(P(s_i = u | y, \theta^{(M)})\)] are identified as a homogeneous segment of class \(v\). In the M1-Mk model, all the homogeneous segments of type \(v\) are characterized by the same \(k+1\) nucleotide composition \(P^{(M)}(y_{1:t} = y_{1:k} = i_k, \ldots, y_{i-1} = i_1, s_i = v)\).

Processing the chromosome

Results were obtained with the software RHOM (Research of HOMogeneous regions in DNA sequences), C++ sources are freely available for UNIX/Linux, at http://www-mig.versailles.inra.fr/ssb/rhom/. RHOM implements the algorithms needed to estimate the parameters of a M1-Mk HMM and to produce a segmentation in the way presented in the previous section. Concerning the model, the user only chooses the number of hidden states and the length of the oligonucleotides taken into account. Different model orders, \(2 \leq k \leq 8\), were used. All models were fitted to the whole sequence of the B.subtilis chromosome (4.2 Mb) through likelihood maximization. To give an idea of the computational cost, processing the chromosome according to the five states
RESULTS

Program behavior

An interesting segmentation of the B. subtilis chromosome is obtained with the M1-M2 model. Segments are long and coincide with genes or groups of them. In contrast, M1-M0 and M1-M1 models give very short segments of a few base pairs, which do not appear related to biological features. Thus, hidden state probabilities plotted along the sequence give intermingled profiles. M1-M2 and higher order Markov models integrate short-range heterogeneities in each segment type, so that the hidden state chain can fit long-range heterogeneities. Therefore the M1-M2 model is a good choice to perform chromosome segmentation. Higher order models do not seem to significantly modify the results, while the number of parameters increases geometrically.

With the two-state M1-M2 model, hidden states fit gene orientation. With the three state M1-M2 model, we typically get two states matching gene orientations (sensitivity, 86.48%; specificity, 90.60% at the nucleotide level), and the third one matching intergenic regions. Such a strong observation may be an indication of the appropriateness of a three-state model for segmenting the chromosome. In terms of oligonucleotide composition, coding strands are a + g rich whereas intergenic regions are a + t rich. Actually, intergenic regions, computed according to GenBank annotation, have an a + t content of 63.2%, and coding strands have an a + g content of 54.0%, in comparison with 56.4% a + t content and 49.9% a + g content for the whole chromosome. In this context, some genes, systematically found associated to the state matching the intergenic regions, appear as atypical. More generally, genes were assigned to the class corresponding to their main hidden state in terms of base pairs.

Searching for gene transfers

In its search for three states RHOM was able to detect a first level of heterogeneity revealing previously identified prophages, and other DNA segments potentially arising from horizontal transfer, both containing genes that we call atypical (for details, see below). Atypical genes belong to the so-called a + t-rich type (a + t content 66.0%) which also contains most of the intergenic regions. These genes have a highly heterogeneous distribution along the chromosome, with a peak at the replication terminus, as shown in Figure 5A.

These a + t-rich regions contain 539 genes, of which 68% have unknown functions, compared with 42% for the complete genome. Here, the term unknown function means genes similar to unknown proteins or without similarities, taken from the functional classification of the bacterium. Genes without any similarity represent 56%, in comparison with 26% for the complete genome. A cross comparison with the codon usage classification reveals that 80% of the genes in the a + t-rich type belong to class III (specificity), while genes of class III represent 13% of the B. subtilis genes. On the other hand, 81% of these genes belong to the atypical state (sensitivity). Thus, there is a strong correspondence between our atypical regions and the codon usage class III genes.

Prophage detection. Literature reports 10 prophages integrated into the B. subtilis chromosome. Seven are putative, or prophage-like sequences P1–P7 (16), since their identification is only based on a + t composition, and all these ‘prophages’ are a + t rich. There is experimental evidence for the three other prophages: PBSX (44), skin (45) and SPβ (46). In this context, the ability of the program to detect experimentally identified prophages and prophage-like sequences provides a biological validation of our approach.

The RHOM software was able to detect all these prophages, except PBSX. The latter is not detectable by RHOM because its
content is too close to the local \textit{B. subtilis} DNA composition, nor is it detectable by simple \(a + t\) content analysis. At least two distinct explanations may be provided for the non-detection of \textit{PBSX}, the first being that it was integrated a long time ago, resulting in the adaptation of its DNA composition to the host context, and the second being that its DNA composition was originally close to that of \textit{B. subtilis}.

Our detection of \textit{SP} \(\beta\) and of the three prophage-like sequences \(P1, P3\) and \(P7\) is in keeping with the literature. For the remaining prophage-like sequences, the detection works, although some boundaries are located differently, or some of their genes remain in the types associated with \textit{B. subtilis}. These differences are probably related to a better segmentation accuracy using HMM than that produced by 10 kb sliding windows with a 5 kb step used in the calculation of the \(a + t\) content (16).

\textbf{New detection of gene transfers.} In a similar manner to prophage detection, we found 14 DNA segments identified as \(a + t\)-rich, and thus potentially arising from horizontal transfer. These segments are presented in Table 1 together with the prophages. In addition to this compositional observation, some other signs exist which strengthen our conclusions.

As reported in Table 1, there is a high correlation between detected segment locations and repeats described in Rocha et al. (19). In particular, only one of these repeats was found not to be associated with the detection of some atypical genes. For example, the 3463–3467 kb segment (Fig. 1) is flanked on both sides by long direct repeats which probably signal a chromosomal rearrangement. These duplicated fragments are larger than the transferred segment and contain five genes, four of which belong to the ABC transporter family. In this case, the putative transfer contains an unknown gene unique to an arsenical resistance operon repressor (\textit{yybA}). Figure 2 shows another kind of repetition, associated with the 4184–4190 kb region of the \(a + t\)-rich type. These sequences are short, \(\sim 100 \text{ bp}\) long, and repeated four times. These repeats are regularly spaced, not correlated with gene borders, and therefore do not present the characteristics of an integrin. Nevertheless, the four repeats surround four resistance related genes, two of them are experimentally proven to confer tetracycline resistance (\textit{tetB, tetL}), and the other two are similar to streptothricin acetyl-transferase (\textit{yyaR}), and to the mercuric resistance operon regulator (\textit{yyaN}), respectively.

The 818–822, 1442–1447 and 4171–4176 kb segments are surrounded by intergenic regions larger than usual, compared with sizes expected in bacterial chromosomes. Another segment located at 3658–3685 kb contains ‘gray holes’. This segment, including teichoic acid metabolism genes, is described in the literature as potentially arising from horizontal gene transfer (17) but the segment detected by the program is larger because it encompasses other genes also coding enzymes involved in cell wall synthesis.

The occurrence of genes having imprecisely known function in these \(a + t\)-rich regions, whose homologies are related to resistance functions, reinforces the hypothesis of gene transfer events. In addition to those previously mentioned in the 3463–3467 and 4184–4190 segments, we found a homology to a multidrug-efflux transporter in 818–822 (\textit{yfmI}), and many other significant homologies in the largest newly detected 570–600 region. Remarkably, this 30 kb segment, adjacent to the \(P2\) ‘prophage’, bears genes with numerous homologies, either related to resistance functions, or to \textit{mocR}, the rhizopine catabolism regulator of \textit{Sinorhizobium melloti}. Rhizopine is a compound found in root nodules resulting from plant–bacteria symbiosis. All homologies of this segment are reported in Table 2.

Finally, a new potential mobile element was found, located between 738 and 747 kb. Genes \textit{yeF} \(\beta\) and \textit{yeF} \(C\), belonging to this segment, are homologous to site-specific recombinases. Moreover, \textit{yeeA} shows similarity with a DNA modification methyltransferase suggesting the presence of a restriction–modification system. These systems are known to be often horizontally transferred (47).

\textbf{Heterogeneities and functional classes}

After identifying the coding strand of genes and atypical segments related to horizontal transfer as being the main heterogeneities, our aim was to find additional ones that can be linked to significant biological features. For example, can we

\begin{table}
\centering
\caption{Coordinates (kb) of potential horizontal transfer regions on the chromosome of \textit{B. subtilis}}
\label{tab:transferredSegments}
\begin{tabular}{lll}
\hline
Functions & HMM & Repeats \\
\hline
\textit{P1} ‘prophage’ & 202–220 & 202–213 \\
\textit{P2} ‘prophage’ & 529–570 & 555–567 \\
\hline
See Table 2 & 570–600 & – \\
\textit{P3} ‘prophage’ & 651–664 & – \\
Site-specific recombinase & 738–747 & – \\
Multidrug-efflux transporter & 818–822 & – \\
& 1124–1130 & – \\
\textit{P4} ‘prophage’ & 1262–1270 & – \\
\hline
\textit{PBSX} prophage (1320–1348) & – & – \\
& 1397–1399 & 1385–1424 \\
& 1442–1447 & – \\
& 1478–1482 & – \\
\textit{P5} ‘prophage’ & 1879–1891 & – \\
& 2038–2041 & – \\
\textit{P6} ‘prophage’ & 2046–2073 & 2050–2060 \\
\hline
\textit{SP} \(\beta\) prophage & 2151–2286 & – \\
\textbf{Skin} prophage & 2652–2701 & 2654–2701 \\
\textit{P7} ‘prophage’ & 2707–2756 & 2725–2735 \\
\hline
\textbf{Competence} & & \\
Arsenic resistance regul. & 3463–3467 & 3462–3469 \\
& – & 3608–3634 \\
Cell wall synthesis & 3658–3685 & 3665–3672 \\
ABC transporter & 4123–4134 & – \\
ABC transporter & 4171–4176 & 4170–4176 \\
Streptothricin, tetracycline, mercury regul. & 4184–4190 & 4189–4190 \\
\hline
\end{tabular}
\end{table}
find other kinds of horizontal transfer that are not \(a + t\) rich? We thus fitted HMMs with more than three states to the chromosome. As a consequence of incrementally adding one more state, the heterogeneity detected is not completely reorganised, but on the contrary, is refined: one of the previously identified states is split. This enables us to present the results in a tree structure as shown in Figure 3. In the following section, we describe the new hidden states in the order in which they appear.

**Hydrophobic proteins.** With the introduction of the fourth and fifth hidden states into the model, we distinguish a minority class of coding sequences, borne by both DNA strands of the chromosome. Compared with the usual \(a + g\)-rich coding strand composition, this new class is \(t\) enriched (31.9 versus 25.7% computed from annotations for the coding strand), and \(a\) depleted (21.8 versus 29.9%), in fact it appears as 56.4% \(t + g\) rich. When compared with the functional classification, these two hidden states are found to be strongly enriched in genes annotated as coding for transport/binding proteins (TBPs). These TBP coding genes represent 9.5% of *B. subtilis* genes, 58% of which belong to the minority class of coding sequences, while TBP genes constitute 36% of the minority class.

The TBP category contains a lot of hydrophobic membrane proteins. Therefore, we looked for amino acid biased composition. Actually, we noticed that proteins belonging to this minority state present differences, compared with the majority state. Their amino acid composition is enriched in: phenylalanine (frequency 0.072 versus 0.039) for which codons, when ordered by *B. subtilis* preference (18), are *tt*, *ttc*; isoleucine (0.096 versus 0.068) *att*, *attc*, *ata*; and leucine (0.13 versus 0.091) *ctg*, *ctt*, *tta*, *ttg*, *ctc*, *cta*. Their composition is simultaneously depleted in: glutamate (0.031 versus 0.080) *gaa*, *gag*; aspartate (0.025 versus 0.057) *gat*, *gac*; and lysine (0.045 versus 0.073) *aun*, *aug*. These amino acid biases of the minority class respectively correspond to a \(t\) nucleotide/hydrophobic amino acid enrichment, and an \(a\) nucleotide/charged amino acid depletion. Thus these two states appear to be associated with genes that code hydrophobic proteins.

A sixth state leads to the separation of atypical \(a + t\)-rich coding sequences according to their transcriptional direction. Thus sensitivity of coding sense detection reaches 98.35% (previously 86.48%) but specificity drops to 85.35% since intergenic regions are counted together with atypical coding sequences. We investigated the trinucleotide composition of the segments obtained with a six-state \(M1-M2\) model, by principal component analysis (Fig. 4). The first three axes explain 77% of the total inertia. The first axis divides the cloud according to coding sense, the second axis distinguishes \(a + t\)-rich segments (atypical coding and intergenic) from the others, while the third axis separates hydrophobic coding sequences. The display of the distribution of the segments according to their associated type along the chromosome (Fig. 5) reports the asymmetrical distribution of coding sequences between leading and lagging strands even for the atypical \(a + t\)-rich segments.

**RNA genes.** Introducing a seventh hidden state requires increasing the model order from \(M1-M2\) to \(M1-M3\). Otherwise, two hidden states remain melted, giving very short segments and intermingled probability profiles, as described in Results.
Table 2. Homologies found in the 30 kb long atypical 570–600 kb region, downstream of the P2 prophage

<table>
<thead>
<tr>
<th>Genes</th>
<th>Homologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ydeL</td>
<td>Transcriptional regulator MocR (GntR family)</td>
</tr>
<tr>
<td>ydeP</td>
<td>Cinnamoyl ester hydrodase</td>
</tr>
<tr>
<td>ydeQ</td>
<td>General stress protein 14 of B. subtilis</td>
</tr>
<tr>
<td>ydeR</td>
<td>Antibiotic resistance translocase</td>
</tr>
<tr>
<td>ydeS</td>
<td>Transcriptional regulator (TetR/AcrR family)</td>
</tr>
<tr>
<td>ydeT</td>
<td>Transcriptional regulator (ArsR family)</td>
</tr>
<tr>
<td>ydfA</td>
<td>Arsenical pump membrane protein</td>
</tr>
<tr>
<td>ydfB</td>
<td>Antibiotic resistance protein</td>
</tr>
<tr>
<td>ydfC</td>
<td>Permease</td>
</tr>
<tr>
<td>ydfD</td>
<td>Transcriptional regulator MocR (GntR family)</td>
</tr>
<tr>
<td>ydfE</td>
<td>Transcriptional regulator (ArsR family)</td>
</tr>
<tr>
<td>ydfH</td>
<td>Nitrate/nitrite sensor protein</td>
</tr>
<tr>
<td>ydfI</td>
<td>Nitrate/nitrite sensor protein</td>
</tr>
<tr>
<td>ydfJ</td>
<td>Antibiotic transport-associated protein</td>
</tr>
<tr>
<td>ydfK</td>
<td>Putative transport protein</td>
</tr>
<tr>
<td>nap</td>
<td>Naproxen carboxylesterase (experimental evidence)</td>
</tr>
<tr>
<td>ydfL</td>
<td>Multidrug-efflux transcriptional regulator</td>
</tr>
<tr>
<td>ydfM</td>
<td>Cation efflux system cobalt-zinc-cadmium</td>
</tr>
<tr>
<td>ydfN</td>
<td>Nitroreductase</td>
</tr>
<tr>
<td>ydfO</td>
<td>ABC transporter</td>
</tr>
<tr>
<td>ydfQ</td>
<td>Thioredoxin</td>
</tr>
</tbody>
</table>

Homologies have been found using BLAST against the non-redundant protein database at the NCBI and only highly significant homologies are reported (with E-value < 1 × 10^-10).

Figure 3. Tree representation of state subdivisions as a function of the number of hidden states. Numeric labels for the six-state HMM correspond to those used in Figures 4 and 5. (+) and (–) indicate the coding strand.

DISCUSSION

General trends of nucleotide compositions

To summarize, general trends are an a + t enrichment for intergenic regions and those we identified as resulting from gene transfer, and a coding strand enrichment in a + g (purine) for the majority class of genes, as opposed to a coding strand enriched in t + g (keto) for a minority class of genes corresponding to those coding hydrophobic proteins. The trends are mixed; for instance, we were able to distinguish transcriptional orientation of a + t-rich horizontally transferred genes, according to their a + g content.

The gc-skew (Nc – Nt)/(Nc + Nt) and the at skew are positive on the replication leading strand in this species, as previously described (3). These skews are linked to a + g enrichment of coding strands, due to a high preference in B. subtilis for encoding of proteins on the leading strand (Fig. 5B and C). These compositional biases violate the second Chargaff rule. This empirical statement assumes that the relations Nc = Nt and Nc = Nt are not only valid on double-stranded DNA but also on each of the strands, if the sequence is long enough. Thus, two violation levels are clearly observed due to asymmetrical evolutionary pressure: the first is related to transcription (kilobase scale), the second to replication (megabase scale). Whereas some work noticed that results of asymmetric pressure related to the replication is observable (50, 51), our study does not notice asymmetric bias that could not be due to the transcription, probably because of its low magnitude. Hypotheses explaining strand asymmetry compositions as results of mutation or selection pressures have been extensively discussed by Frank and Lobjy (52).

Whereas the positive gc skew of the leading strand is a general characteristic of bacterial genomes, a positive at skew is.

Highly expressed genes. An eighth state extracts a subcategory from majority coding sequences from the (+) strand (the one given in the sequence file, in opposition to its reverse complement). This subcategory contains most of the codon usage class II genes, which have been characterized as highly expressed. Genes of class II on the (+) strand cover 2.5% of the chromosome, 92% are found in this new state, but they only represent 44% of the subcategory that might interestingly extend class II. It is well known that such highly expressed proteins exhibit greater codon biases than others (48), producing a different statistical composition from the remaining genes in bacteria [see Makrides (49) for a review on high-level expression strategies in E. coli]. They are especially concentrated within the 118–156 kb region close to the replication origin, where most of them are related to transcription and translation: ribosomal proteins, RNA polymerase α and β subunits, translation initiation and elongation factors.

The highly expressed genes in this region were already detected with a two-state M1-M2 model fitted to the 100–200 kb segment of the chromosome, while we had to go up to eight states when running on the complete chromosome. This exemplifies the interest and complementarity of using both local and complete genome analyses.

For M1-M0 and M1-M1 models. The newly identified composition type corresponds to structural RNA genes. In terms of base pairs, these genes cover 1.28% of the chromosome. We identified structural RNA genes with a sensitivity and a specificity which, at the nucleotide level are 96.6 and 90.8%, respectively. At the gene level, we found all rRNA genes and 78 out of the 88 annotated tRNA genes. Thus the correspondence of this compositional type to structural RNA is highly accurate.
is more exceptional, and depends on the codon base position referred to (3,53). Moreover, the preferred choice of $a$ versus $t$ at the third codon position is not the same for all the encoded amino acids (54).

**Gene transfers**

Concordance between atypical nucleotide composition, occurrence of repetitions, and presence of genes related to resistances or unknown functions, makes the horizontally transferred origin of the detected regions very likely. Over-representation of genes of unknown function within these regions could then be due to adaptive characters of these genes, which are only expressed in natural soil surroundings of the bacteria, and not required by them in the laboratory.

All prophages, plus 14 other segments we identified, are $a + t$ rich in comparison with the chromosome composition. We cannot explain this general property, as we expected nucleotide composition heterogeneity among these horizontal transfers due to their supposedly diverse origins. This is intriguing, particularly since $a + t$ enrichment of horizontal transfers seems common in genomes presenting compositional heterogeneity, including *E.coli*. However, we know of some cases where horizontal transfer is $g + c$ richer than the host genome, see for example the mu-like prophage integrated in the *Haemophilus influenzae* genome (55) or the *E.coli* yagH gene (56).

The widespread hypothesis is to explain the compositional heterogeneity of a genome as a snapshot of transferred DNA fragments progressively adapting to the host composition (2),
these transfers having originally a distinct nucleotide composition related to the source organism. If this assumption is correct, horizontal transfers should generally come from source organisms which are a + t richer than the host.

According to the analysis results on *E.coli* and *Salmonella* sequences of Syvanen (4), an alternative interpretation may be suggested, whereby horizontal transfers essentially occur among strains of the same or close species. In this case, we can imagine a pool of adaptive genes that are shared by recurrent transfers. This may explain the relative homogeneity in a + t richness of the detected fragments, because their shared nucleotide content could then be due to the same evolutionary pressure. Possible origins of the a + t composition bias have been extensively discussed from different angles, not only in relation to gene transfers (4), but also in the interpretation of the interspecific a + t content differences (57), of the isochore existence among vertebrates (58), or even of bacterial intraspecific a + t content heterogeneity (54), the latter relating to the context of our study.

**Protein hydrophobicity**

The t + g richness of hydrophobic proteins appears as one of the main heterogeneity factors at the DNA composition level, coming just after heterogeneities derived from coding properties (coding/non coding, and the transcription direction), and atypical a + t richness related to gene transfers. This is due to the preferential occurrence of t in the second codon position of hydrophobic amino acids.

Influence of amino acid hydrophobicity at the nucleotide composition level has been previously considered (52), because global hydrophobicity has been shown as a main factor for protein variation in amino acid content (59). We were surprised about the importance of this phenomenon on nucleotide composition, even without introducing any frame consideration. It could be interesting to take this heterogeneity into account to improve gene detection based on Markov models, as is done for genes of codon usage class III (41,60).

**Perspectives**

Fitting parameters of a HMM to the oligonucleotide composition of the chromosome, by likelihood maximization through the EM algorithm, leads to a segmentation correlated to biological features of the DNA sequence. It is somewhat remarkable that such structures could be identified without having to specify a window length, or any learning set. From here, the mathematical challenge would be to choose adequate selection criteria of the Markov order and of the number of states that define the model structure.

Initially considered as a tool for detecting horizontal gene transfers, this approach enables one to reveal many more heterogeneities, mainly linked to characteristics of coding sequences. On the basis of these results, it would be interesting to extend the model to phased sequences using a HMM that changes the hidden state periodically according to the codon position. These models allows for a more realistic representation of coding sequences. One promising feature of using such a model is to enable the combination of gene detection and heterogeneity description, in a similar manner to Besemer et al. (61). In addition, focusing the analysis on the heterogeneity of intergenic regions appears to be very promising. To summarize, our general goal is to produce a finer integrated description of the structure of the chromosome, in terms of the statistical composition of biological features.

**ACKNOWLEDGEMENTS**

We are grateful to Annie Bouvier and Fabrice Lepage for their extensive participation in the RHOM program development and we thank Kevin Bryson for his careful reading of the manuscript.

**REFERENCES**


