A large-scale analysis for significance assessment of frequencies relative to potentially strong sigma 70 promoters - comparison of 32 prokaryotic genomes -

Christine Sinoquet, Sylvain Demey, Frédérique Braun

A large-scale analysis for significance assessment of frequencies relative to potentially strong sigma 70 promoters:

comparison of 32 prokaryotic genomes

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

†Lina - Laboratoire d'Informatique de Nantes-Atlantique, CNRS - FRE 2729, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex, France, ‡INSERM U601, Département de Recherche en Cancérologie, Université de Nantes, 9 quai Moncousu, 44093 Nantes Cedex 01, France

— Computational Biology —

RESEARCH REPORT

N° hal-00153303

June 2007

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

A large-scale analysis for significance assessment of frequencies relative to potentially strong sigma 70 promoters: comparison of 32 prokaryotic genomes

24 p.

A large-scale analysis for significance assessment of frequencies relative to potentially strong sigma 70 promoters:

comparison of 32 prokaryotic genomes

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

christine.sinoquet@univ-nantes.fr

Abstract

This report presents a computational analysis of high ORF expression potentialities in prokaryotic genomes, in relation with medical and economic relevance. Given a bacterial genome and the description of a structured motif, the software BACTRANS^2 implements the search of the occurrence most similar to the motif, in the regulatory region of each gene. In this work, the software BACTRANS^2 was run over 32 prokaryotic genomes, to identify putative strong $\sigma 70$ promoters. We focused in particular on $\sigma 70$ promoters harbouring an UP element, which enhances transcription initiation. We performed four computational analyses per genome, combining two promoter strength levels (CI ¿ CII) with either mandatory or optional UP element presence. We compared the frequencies obtained for 32 bacterial genomes, under these four constraint specifications.

First, we show that an over-representation of putative strong promoters differentiates the AT-rich Firmicutes' genomes from other genomes. Another interesting result is that strong promoters of relatively lesser quality (CII) are more frequently associated with an UP element than strong promoters of better quality (CI).

Then, per each bacterial genome studied, we generated at random 100 artificial genomes. Such genomes are only constrained as to have the same two following characteristics as the bacterial genome: same total number of genes and same proportions of A, C, T and G nucleotides in the 350 nucleotide-long region upstream of start codon. The $\sigma 70$ promoter frequency observed on average over these 100 genomes is compared to the frequency observed for the bacterial genome, under each of the four constraint sets aforementioned. Thus, the statistical significance of the $\sigma 70$ model is discussed for each genome, under each constraint set. For most genomes, and especially for *Firmicutes*, a meaningful difference is statistically ascertained. Besides, the comparison between *Firmicutes* genomes and equally AT-rich *Proteobacteria* genomes also confirm that the *Firmicutes* specificity is not related to genome size bias. Hence, *Firmicutes* would appear as genomes more favoured by nature with respect to high intrinsic transcription potentiality.

Throughout the report, we discuss the influence of AT-richness on promoter frequencies, implementing various correlation analyses. We show that an influence is only observed when the UP element is required. Then we evaluate whether the statistical significance of the $\sigma 70$ model is related or not to AT-richness. Interestingly, we find that the relation is loose except when the UP element is required, and under the more stringent constraint (CI). Thus, we distinguish the AT-bias, whose influence is more or less noticeable for bacterial genomes as well as randomly generated genomes, whatever the species, and the species bias such as the one identified for *Firmicutes*. Finally, we compare the AT-percentages of three sub-regions of the 350 bp-long region upstream of start codon, distinguishing between genes harbouring strong promoters and genes not harbouring any such strong promoters. We can show no evidence that the over-representation characterizing *Firmicutes* is due to a local AT-bias.

To our knowledge, our work is the very first statistical approach thoroughly analysing the presence significance of various potentially strong $\sigma 70$ promoter models, including models harbouring the UP element enhancer, in the context of a genome-comparative study. The presence of the enhanced promoter has been proven significant in all eight large *Firmicutes* genomes studied and between ten and thirteen non *Firmicutes* large genomes studied (depending on the Z-score threshold considered).

1 Foreword

The project BACTRANS^2 was first initiated in an unformal way, in january 2003, after fruitful discussions with Frédérique Braun who was working at that time at the UMR C.N.R.S. 6204 - "Biotechnology, Biocatalysis et Bioregulation" team, under the direction of its head, Professor Vehary Sakanyan, at the Biotechnology Laboratory of the University of Nantes. Initially, the project dealt with identifying putative strong $\sigma 70$ promoters in *Thermotoga maritima* genome, with the objectives of gaining in fundamental knowledge about this thermophilic model and enabling advances in biotechnologies. *Thermotoga maritima* is an hyperthermophilic bacterium (80^oC) encountered in geothermal marine areas. In the last decade, this bacterium was thoroughly studied by Professor Vehary Sakanyan's team.

The very core of the platform was written by Christine Sinoquet. It soon appeared that BACTRANS^2 project aroused the interest from both the bioinformatician and biologist communities. Between june 2003 and june 2004, four students contributed to the platform design, under the direction of Christine Sinoquet. Then Sylvain Demey was assigned the task to improve and extend the platform, integrate all previous components in a software suite, homogenize the interfaces, implement other functionalities. This, he achieved between april 2005 and september 2006.

BACTRANS^2 is a protected platform at the disposal of biologists for the study of putative strong promoters in prokaryotic genomes. It is protected through GNU License. An exhaustive presentation of BACTRANS^2's functionalities is far beyond the scope of the present report. Generic software platform BACTRANS^2 currently provides such putative strong promoters for 45 genomes. Moreover, BACTRANS^2's genericity allows the user to analyse genomes with respect to any other motif consisting of 3 or 4 boxes. BACTRANS^2 is accessible at http://www.sciences.univ-nantes.fr/lina/bioserv/BacTrans2/.

2 Introduction

As one of the simplest known bacterial models, *E. coli* K-12 has been subject to intensive research, especially with regard to transcription [27, 25, 10, 33, 15, 24, 37, 28]. Thus knowledge was gained about the $\sigma 70$ factor, whose two canonical binding sites' consensi are respectively TTGACA and TATAAT. Another feature is the relative conservation of distances: the optimal fixation of RNA polymerase requires that the site with the consensus TTGACA should be located between 35 bp and 30 bp or thereabouts upstream of first transcribed nucleotide. This former site is thus called the -35 box. The Pribnow box, TATAAT, is called -10 box for similar reasons [39]. In the canonical $\sigma 70$ promoter, these boxes are separated by 15 to 21 bp. The more similar to canonical $\sigma 70$ promoter is a given promoter, both in terms of content and structure, the more potentially strong is this promoter with respect to transcription initiation. The RNA polymerase is conserved through evolution in bacteria, which legitimates searches for $\sigma 70$ factor binding sites through other prokaryotic genomes. Research was thus extended to other bacteria [22, 21, 32, 34, 30]. Meanwhile, the number of complete prokaryotic genomes sequenced has increased at a high speed (524 in june 2007). At the same time, various platforms devoted to bacterial genome analysis were made available, with different aims ([36], HOBACGEN [38], GenoExpertBacteria [19], alone or as part of platform Genostar [20], RegulonDB [45], EcoCyc [12], to name but a few). In particular, RegulonDB and EcoCyc are the reference databases for *E. coli* curated knowledge, the former including experimentally validated knowledge about σ promoters. Also various methods and softwares devoted to bacterial promoter prediction were proposed (for an illustration of the former, see [31, 9]; for an example of the latter, see BPROM, accessible through Softberry platform, http://www.softberry.com/berry.phtml). We do not mention here the various softwares made available for inferring a motif common to a set of biological sequences.

Our contribution lies within the scope of computer-assisted identification and study of potentially strong promoters, with the objectives of enabling advances in biotechnologies as well as fundamental

knowledge about prokaryotic genomes. Genetic engineering implements enhancement for the expression of a gene of interest through the association of this gene with a strong promoter. Thus any progress towards speeding the identification of potentially strong promoter candidates in prokaryotic genomes is valuable to institutions involved in biotechnologies. A study was still missing: here, our concern is genome-scale and genome-comparative analysis of high transcription potentiality, with an emphasis on strength reinforcement through the UP element presence. The UP element is an enhancer for transcription and thus for ORF expression [43, 13, 14]. In about 3% of the *E. coli* promoters, the UP element is located approximately 4 bp upstream of the -35 region conferring additional strength to the promoter. The high conservation of the domain of the alpha subunit of the RNA polymerase involved in the interaction with the UP element strongly suggests that the UP element consensus should be valid throughout the bacterial kingdom. Nonetheless, to our knowledge, in addition to *E. coli* genome, the UP element was only experimentally identified in *Bacillus subtilis* [16], *Vibrio natriegens* [1] and *Geobacillus stearothermophilus* [46]. Besides, beyond gaining fundamental knowledge about prokaryotic genomics, we wish to assist in selecting the promoters which should be tested in priority *in vitro*, since their expression potentiality is expected to be reinforced by the presence of an UP element. To our knowledge, the only other work devoted to *in silico* identification of putative strong promoters harbouring an UP element is by M. Dekhtyar, A. Morin and V. Sakanyan (Sakanyan, personal communication).

A scoring function is required to identify the putative promoters with highest potentialities. The fine-tuning of a promoter mainly relies on the following parameters: (i) binding sites, (ii) σ factor, (iii) transcription factors and (iv) three-dimensional conformations of the RNA polymerase and the DNA. All previous parameters may not be at their best each. It has been established that a low similarity between binding sites' sequences and consensi descriptions entails weakness of the promoter. Nevertheless, compensations may operate through regulations performed by specific proteins, the transcription factors [23, 8]. For example, Huerta and Collado-Vides [31] established that more than 50% of experimentally verified promoters are not the promoters with the highest scores when scoring relies on the similarity to the canonical promoter, both in terms of consensi similarity and optimal bp distances between boxes. This statement was checked on 111 promoters constituting a training set designed by Gralla and Collado-Vides [21]. It was confirmed on a test set containing 392 known promoters. Moreover, an *in silico* study grounding the design of a promoter prediction method brought precious biological insight: Huerta and Collado-Vides showed that the major part of regulatory regions in *E. coli* display high densities of potential RNA polymerase-$\sigma 70$ binding sites, forming clusters of overlapping promoter-like signals. In contrast, such signal densities are not detected elsewhere in *E. coli* genome. These authors checked that functional promoters experimentally verified are often identified within these clusters, even when isolated sites with higher potential binding affinity for RNA polymerase exist elsewhere within the region. Thus, the method devoted to promoter prediction in *E. coli* and implemented by Huerta and Collado-Vides partially relies on the distribution of occurrences of promoter-like motifs in regulatory regions. Also did Huerta and Collado-Vides confirm that the regulatory regions in other large bacterial genomes have a high density of $\sigma 70$ promoter occurrences [30].

In this work, we perform a genome-wide analysis of the putative **strongest** promoters' frequencies in 32 bacterial genomes. These genomes belong to different genera across all major bacterial phyla. Our purpose is two-fold: (i) we compare genomes with respect to promoters identified with the highest transcription potentialities; (ii) in particular, we focus on promoters harbouring an UP element. Thus, in our *in silico* analysis, the scoring function used takes into account three parameters: (i) the similarity to consensi contents, (ii) the closeness to optimal bp distances between boxes and (iii) the presence of the UP element. The 32 genomes compared belong to ten *Firmicutes*, thirteen *Proteobacteria*, three *Actinobacteria*, two *Spirochaetales*, one *Chlamydia* and three other taxa outside latter phyla. We distinguish two strength levels, depending on the relaxation allowed with respect to the canonical $\sigma 70$ promoter, and combine them with either mandatory or optional UP element presence. Thus, we actually perform four genome-comparative studies.

3 System and methods

3.1 Genome analysis upon request

Describing the software suite BACTRANS^2 [2] is beyond the scope of this report. Historically, it was first devoted to hyperthermophilic bacterium *Thermotoga maritima* [35, 44, 7]. Subsequently, it was extended into a generic platform. For each genome studied, BACTRANS^2 takes as an input the Fasta genome sequence provided by GenBank (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) together with the corresponding genome annotation. For each gene, we scan up to 350 nucleotides upstream of start codon's first nucleotide. Once the region of interest is obtained, occurrences of the promoter binding sites are searched for under constraints relative to (i) distances between binding sites or distances between binding sites and translation signals playing the role of "anchors" and (ii) the maximal number of mismatches allowed with respect to each binding site consensus. Hereafter the number of mismatches between a sequence and a consensus will also be referred to as the *Hamming distance*. A putative promoter region is identified starting from the start codon and successively looking for the Shine-Dalgarno signal, the -10 box, the -35 box (and the UP element when required). The Shine-Dalgarno signal is a mandatory sequence involved in translation. In the sequel, the Shine-Dalgarno signal will be denoted SD. With the distance ranges observed for *E. coli*, the motif searched for is specified in 3' to 5' direction as: <start codon> [2-10] <SD> [10-200] <-10 box> [15-20] <-35 box> [0-15] <UP element>, where $<box_1>$ $[d_{min}$-$d_{max}]$ $<box_2>$ states the minimal and maximal bp distances allowed between the two boxes concerned. A value of 200 bp was chosen for the maximal distance between start codon and SD; it was selected on the basis of the average 5'UTR region's length (50 or thereabouts, with variations between 0 and 200). Translation and transcription signal contents are the following: <SD> = GGAGG; <-10 box> = TATAAT; <-35 box> = TTGAC; <UP element> = NNAAAWWTWTTTTNNAAAANNN. In addition to these previous bp distance constraints, our study considers two different constraint configurations for Hamming distances. Notation **(**$d_H(UP), d_H(-35\ box), d_H(-10\ box)$**)** specifies the maximal Hamming distances relative to UP element, -35 box and -10 box respectively. Given this notation, the four configurations retained in our study are described as follows: **(**4,2,1**)** and **(**4,3,2**)**, UP element required; the same as previously, UP element optional. From now on, with increasing transcription strength, these two configurations will be respectively denoted CI and CII. Thus the four former configurations will be identified as: CI, UP element required; CII, UP element required; CI, UP element optional; CII, UP element optional. Our choice of a shorter consensus for the -35 box rather than the canonical -35 box itself (TTGACA) is motivated by sixth nucleotide having the lowest conservation level. Our choice simply amounts to taking into account the canonical -35 box either allowing 3 or 4 mismatches at most, depending on the constraint chosen (CI or CII).

3.2 Scoring function used

Considering the 5 nucleotide-long -35 box, there exist $\binom{5}{3}$, *i.e.* 10 possible combinations with exactly 3 mismatches wherever these mismatches occur. A mismatch occurs for any of the 3 nucleotides differing from that of the consensus. Therefore there are 90 possible contents for the -35 box, under these conditions. Besides, when an occurrence with one mismatch is identified, the word starting 2 nucleotide upstream or downstream of the beginning of previous occurrence is yet another occurrence with at most 3 mismatches. Thus a criterion is required to sort the different (possibly overlapping) candidates. A scoring function taking into account bp distances and Hamming distances is designed to identify the putative promoter with the highest potentiality in the region scanned. In the sequel, $d_H(b)$ denotes the Hamming distance observed with respect to the consensus box b; d_1 denotes the bp distance observed between -35 box and -10 box; d_2 denotes the bp distance observed between UP element and -35 box. The score is calculated as follows: $score = 0.60\ d_H(-10\ box) + 0.40\ d_H(-35\ box) + t_1 + d_H(UP) + t_2$,

where $t_1 = 0$ if d_1 belongs to $[17, 19]$ else $t_1 = 5 * d_1$, and $t_2 = 0$ if d_2 ranges in interval $[3, 5]$ else $t_2 = 3 * d_2$. When no UP element can be identified, the score is merely computed as: $score = penalty + 0.60\ d_H(-10\ box) + 0.40\ d_H(-35\ box) + t_1$. The lower the score, the more likely the identified putative promoter is a strong one. The penalty value is set in order to systematically favour a candidate with an UP element within the regulatory region. Depending on the constraints, we obtain a result of 0 or 1 strong transcription promoter per each gene. In either constraint set CI or CII, we denote sp and $upsp$ the numbers of putative strong promoters respectively obtained over a given genome when the presence of the UP element is optional or required. From now on, we will refer to sp_{CI}, sp_{CII}, $upsp_{CI}$ and $upsp_{CII}$.

The difference with the algorithm of Dekhtyar *et al.* and the one presented here lies in six major points (V. Sakanyan, personal communication): (i) the former takes into account genes coding for m-RNAs as well as t-RNAs and r-RNAs; (ii) thus, contrary to ours, the algorithm of Dekhtyar *et al.* does not benefit from the supplementary clue consisting of the Shine-Dalgarno sequence; (iii) the former algorithm is solely devoted to strong promoters harbouring an UP element; (iv) the scoring function is more sophisticated than ours and emphasizes the similarity requirement with regard to the -10 box; (v) the retrieval of the structured motif is performed in 5' to 3' direction in Dekhtyar *et al.*'s approach whereas our method scans the regulatory regions in 3' to 5' direction, which allows relying on the most specific "anchors" in priority; (vi) because a dynamic programming alignment algorithm is run to successively retrieve the -35 and -10 boxes, the minimal similarity thresholds regarding these binding sites are specified by the user as minimal alignment scores. Regarding the latter point, we favoured mismatch error specification as being a more intuitive approach for tuning the algorithm.

3.3 Comparison with randomly generated genomes

For each bacterial genome considered in this study, we compare the sp values (resp. $upsp$ values) observed with respect to the corresponding values expected on average for a similarly AT-rich genome generated at random. This latter artificial genome is only constrained to have the same following characteristics as the prokaryotic genome considered: same total number of genes coding for m-RNAs and same proportions of A, C, T and G nucleotides in the 350 nucleotide-long region upstream of the start codon. Before implementing such a comparison, we need evaluate the values expected on average for a genome generated at random. There are two ways to address the problem of computing such expected values: a probabilistic method would yield a theoretical mean value; an empirical approach computes the appropriate mean value over a sufficiently high number of artificial genomes. In our problem, the start codon is the only box whose location is known for each gene. Thus we have to evaluate the probability to find an occurrence of the *super-motif* <-35 box>[gap3]<-10 box>[gap2]<SD>[gap1] and the probability to find an occurrence of the super-motif <UP element>[gap4]<-35 box>[gap3]<-10 box>[gap2]<SD>[gap1] considering the start codon as a "right" anchor.

3.3.1 Probabilistic approach

In our study, the super-motif embedding the $\sigma 70$ promoter is either described as three or four boxes separated by gaps. Various ways for calculating the probability of occurrence of a motif consisting of a single box, subject to mismatches, have been proposed [6, 17, 42]. A method was also proposed for dealing with two boxes separated by a gap of given length and allowing no mismatch [49]. A still more difficult instance of the problem has been addressed by Robin and Daudin [41]: it deals with two boxes subject to mismatches, separated by a gap whose length is specified to vary in a given interval. In this latter instance, the difficulty arises from the variability of the gap's length: at a given location in a sequence, one has to consider several putative occurrences of the motif, possibly overlapping. Indeed, computing the *exact* probability of occurrence for such a motif is a hard task since possible overlappings have to be taken

into account . In [41], a method was implemented for gaps varying in a $[length, length+2]$ interval, and more specifically, for a maximal number of errors over the two boxes equal to 1. Nevertheless, in addition to the difficulty related to overlapping, the maximal Hamming distances meeting realistic description entail a high complexity of such computations (CI: 1 and 2 mismatches at most respectively for the -10 box and the -35 box; CII: 2 and 3 mismatches at most respectively for the -10 box and the -35 box, combined with at most 4 mismatches for the UP element when the latter is required). Furthermore, to comply with the same realistic requirements, the length for $gap2$ varies in interval $[10, 200]$. An exact method relying on strictly counting non-overlapping words was designed to compute the probability of occurrence for the two motifs we are interested in (unpublished). Through this method, we were able to compute the probability of occurrence of super-motif <-35 box>[15-20]<-10 box>[gap2], supposing that the Shine-Dalgarno location was known. Depending on the maximal Hamming distances stated and the gap2 interval specified, between one and five hours were necessary to process each of the genomes tested, on a 3.40 GHz personal computer fitted out with a 2-MO RAM. The maximal Hamming distances specified were 1 for the -10 box and either 2 or 1 for the -35 box: indeed, increasing gap2's length soon compelled us to restrain $d_H(-35\ box)$ to 1. Therefore, an empirical approach presently remains the only efficient alternative to compute the probabilities expected for constraints CI and CII, not to speak of the addition of a fourth box (UP element).

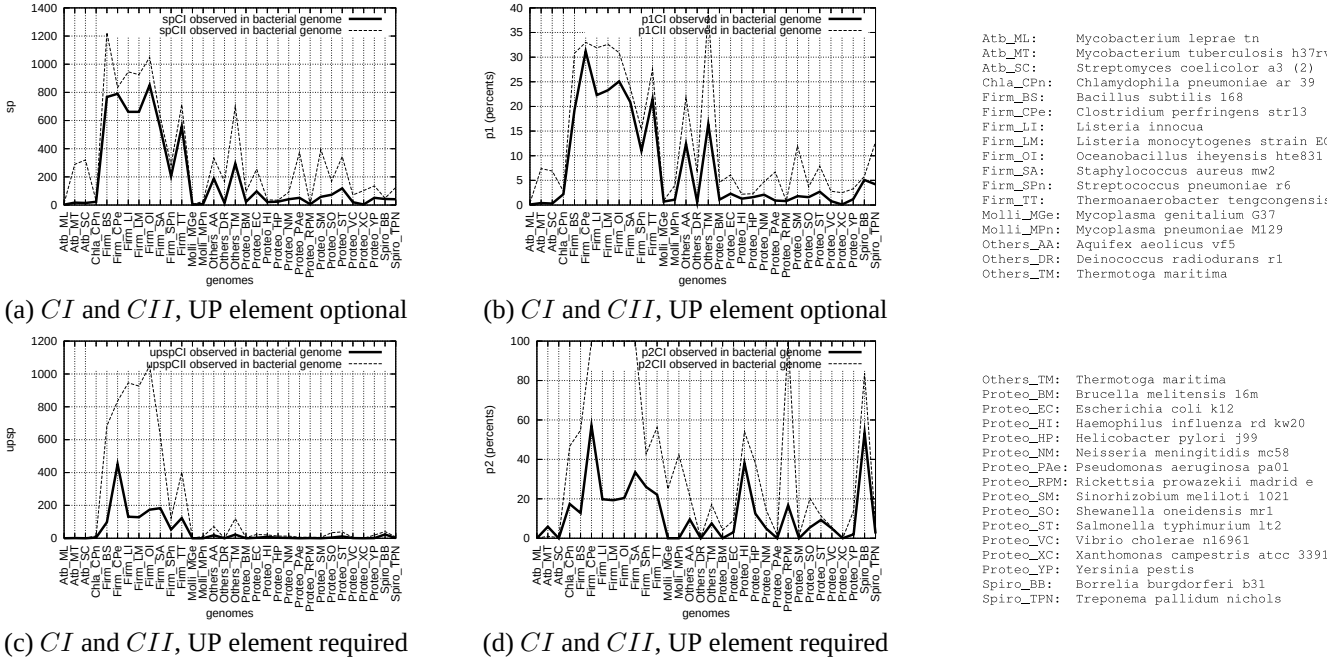
3.3.2 Empirical approach

For each of the 32 bacterial genomes, we systematically compute the minimum, maximum, mean and standard deviation for sp and $upsp$ values, over 100 randomly generated genomes. Such a calculation is performed for each of the four cases studied: CI constraints, CII constraints, the same two with the presence of the UP element required. Thus our software identifying the potentially strongest promoters was run $4 \times 32 \times 100$ times. In the sequel, we examine whether the values observed on bacterial genomes are significantly different from the values observed on "average" genomes generated at random and having the same characteristics as described at the beginning of present Subsection. For this purpose, we rely on the empirical method to yield the mean and standard deviation required to compute Z-scores. For a given bacterial genome, we compute the Z-score as $\frac{|obs - M_{exp}|}{\sigma_{exp}}$, where obs is an $spCI$ value (respectively $spCII$, $upspCI$, $upspCII$ value) and M_{exp} and σ_{exp} respectively denote the mean and standard deviation obtained for $spCI$ (respectively $spCII$, $upspCI$, $upspCII$) over 100 genomes generated at random. Significance will be discussed with respect to a minimal threshold for Z-scores. The choice of this minimal threshold must take into account the fact that the constraint relative to $gap2$'s length is weak (but conforms to reality).

4 Results and discussion

4.1 Are potentially strong promoters frequent?

We compare 32 prokaryotic genomes of different sizes. These genomes divide between the following sets with respect to their total numbers of genes coding for m-RNAs (g): 4 below 1000, 7 between 1000 and 2000, 10 between 2000 and 3000, 3 between 3000 and 4000, 7 between 4000 and 5000, 1 over 5000. The total number of genes coding for m-RNAs (g) and the size of this genome are proven to be correlated over the 32 genomes studied (correlation coefficient: 0.93). To escape the size bias when comparing genomes, we define the percentage $p1$ ($p1 = 100 \times sp/g$). Top section of Figure1 ((a) and (b)) depicts the variations of sp values and $p1$ percentages through genomes (also see Supplementary Data, Appendix 1, Table 1.1). For illustration, the output files regarding *E. coli* genome are provided on line (see Supplementary Data, Appendix 2).



(a) CI and CII, UP element optional

(b) CI and CII, UP element optional

(c) CI and CII, UP element required

(d) CI and CII, UP element required

Figure 1: Frequencies of genes harbouring a putative strong promoter, under four constraint sets, in 32 prokaryotic genomes. See text, Subsection "Genome analysis upon request" for the definition of CI and CII constraints. (a) and (b): UP element optional; (c) and (d): UP element required. Along the x-axis, the following phyla and groups are encountered: *Actinobacteria*, *Chlamydia*, *Firmicutes* (among which *Mollicutes*), "Others" group, *Proteobacteria*, *Spirochaetales*. (a) y-axis: number of genes harbouring a **S**trong **P**romoter (sp); (b) y-axis: ratio $p1$ of genes harbouring a strong promoter (sp) to the total number of genes encoding proteins in the genome (g), $p1 = 100 \times sp/g$; (c) y-axis: number of genes identified with an **UP** element harboured in the **S**trong **P**romoter ($upsp$); (d) y-axis: ratio $p2$ of the number of genes with an UP element in the strong promoter ($upsp$) to the number of genes with a strong promoter (sp), $p2 = 100 \times upsp/sp$).

As a first result, we check that the number of putative strong promoters identified increases when constraints are relaxed from CI to CII. Secondly, we observe that for the AT-rich genomes of *Firmicutes*, putative strong promoters are over-represented under the two constraints specified. This differentiates *Firmicutes* from all other genomes studied. Nonetheless, among *Firmicutes*, the numbers of strong promoters may differ in high proportions (1 to 6 under CI constraints; 1 to 10 under CII constraints); *Streptococcus pneumoniae* has always the lowest value whereas *Bacillus subtilis*, *Oceanobacillus ihenyensis* and *Clostridium perfringens* happen to show peaks depending on the constraint. The differentiation between *Firmicutes* and other genomes holds for $p1$ percentage. The non *Firmicutes* genomes pointed out by the highest $p1$ percentages (over 5%) are *Aquifex aeolicus*, *Thermotoga maritima* and *Borrelia burgdorferi*. Thirdly, a more thorough examination shows that the genomes with the highest numbers of genes (g) are not necessarily those with the highest numbers of putative strong promoters (sp). Percentage $p1$ is variable and no correlation can be shown to exist between sp and g. In the following, $p1_{CI}$ and $p1_{CII}$ will respectively denote the percentages obtained under constraint sets CI and CII. For example, *E. coli* model ($g = 4173$) is characterized by $p1_{CI} = 2.3\%$ and $p1_{CII} = 6.1\%$. With a number of genes not quite so different ($g = 3979$), *B. subtilis* model is described by $p1_{CI} = 19.3\%$ and $p2_{CII} = 30.8\%$. The

highest percentage is observed for *Clostridium perfringens* ($g = 2532$; $p1_{CI} = 31.2\%$; $p1_{CII} = 33.0\%$) whereas low percentages are observed for the genome with the highest number of genes (*Pseudomonas aeruginosa*; $g = 5565$; $p1_{CI} = 0.9\%$; $p1_{CII} = 6.7\%$). Finally, as a fourth result, we show that AT-content does not interfere with $p1$: the linear correlation coefficient between $p1_{CI}$ and AT-content is 0.51, over the 32 genomes; the correlation coefficient between $p1_{CII}$ and AT-content is equal to 0.31. When we consider all bacteria but *Firmicutes*, such coefficients go down to 0.26 (CI) and -0.12 (CII) respectively. When the 10 AT-richest genomes are concerned (*Firmicutes*), such coefficients go down to 0.31 and -0.20 respectively. Anyway, in the latter case, 10 is a borderline value regarding correlation analysis validity. Nevertheless, generally speaking, we retain that AT-content does not interfere with percentage p_1.

4.2 Are potentially strong promoters harbouring an UP element frequent?

We recall that sp is merely the number of genes with putative **S**trong **P**romoters. We defined $upsp$ as the number of genes with an **UP** element in the putative **S**trong **P**romoter. We now define percentage $p2$ ($p2 = 100 \times upsp/sp$). Bottom section of Figure 1 ((c) and (d)) depicts the variations of $upsp$ and $p2$ among the 32 micro-organisms, for a given constraint set; it also allows the comparison of the variation magnitudes over genomes when relaxing the constraint set from CI to CII (also see Supplementary Data, Appendix 1, Table 1.2). For illustration, the output files relative to *E. coli* genome are provided (see Supplementary Data, Appendix 3).

We first show that the differentiation between *Firmicutes* and other genomes holds, but it is more subdued for $p2$ percentage than for $p1$ percentage. Together with *Aquifex aeolicus*, *Thermotoga maritima* and *Borrelia burgdorferi* already pointed out by the highest $p1$ percentages (over 5%), four more non *Firmicutes* genomes, *Helicobacter pylori*, *Rickettsia prowazekii*, *Chlamydophila pneumoniae* and *Haemophilus influenza*, show the highest $p2$ percentages (over 10%). Secondly, we observe that strong promoters of relatively lesser quality (constraint set CII) are more frequently associated with an UP element than strong promoters of better quality (constraint set CI) (Figure 1 ((c) and (d)): the ratio $\frac{p2_{CII}}{p2_{CI}}$ is calculable for 23 genomes and its average is 3.02; the average computed for all *Firmicutes* but *Mollicutes* is 3.55.

Not surprisingly, *Mycobacterium leprae* ($sp_{CI} = 3$; $sp_{CII} = 12$), one of the studied genomes having the lowest number of putative strong promoters together with *Mycoplasma genitalium* ($sp_{CI} = 3$; $sp_{CII} = 4$), shows no UP element under either constraint set (see Figure 1 (c)). No UP element can be identified either under any constraint set for *Deinococcus radiodurans* ($sp_{CI} = 14$; $sp_{CII} = 167$) and *Xanthomonas campestris* ($sp_{CI} = 4$; $sp_{CII} = 103$). More interestingly, another result is that some genomes having few strong promoters show in contrast a high ($p2$) percentage of them harbouring an UP element, whatever the constraint: *Haemophilus influenza* ($sp_{CI} = 21$; $sp_{CII} = 37$; $upsp_{CI} = 8$; $upsp_{CII} = 20$; $p2_{CI} = 38\%$, $p2_{CII} = 54\%$), *Borrelia burgdorferi* ($sp_{CI} = 43$; $sp_{CII} = 49$; $upsp_{CI} = 23$; $upsp_{CII} = 41$; $p2_{CI} = 54\%$, $p2_{CII} = 84\%$). Other such AT-rich genomes with few strong promoters show a high proportion of them harbouring an UP element, only under CII constraints (*Chlamydophila pneumoniae, Helicobacter pylori* and *Mycoplasma pneumoniae*). As an extreme trend, we observe that the very few promoters identified for *Rickettsia prowazekii* under CII relaxed constraints are all associated with an UP element ($sp_{CII} = upsp_{CII} = 6$; $p2_{CII} = 100\%$).

Finally, we show that the correlation coefficient between $p2_{CI}$ and AT-content is 0.70 when the 32 genomes are considered; the correlation between $p2_{CII}$ and AT-content is more pronounced (0.82). These coefficients are quite similar when *Firmicutes* are not taken into account (0.71 and 0.85 respectively). As previously, the 10 *Firmicutes* still actually seem to show no correlation confirmed (0.62 and 0.37 respectively). However, drawing a conclusion is delicate under CII constraints: half of the *Firmicutes* observed show an identical percentage saturation (100%) (Figure 1 (d)); this induces a disputable low correlation coefficient when *Firmicutes* are considered alone; this saturation might instead introduce a

	Proteobacteria		*Firmicutes*							
	HI	HP	LI		LM		SPn		TT	
AT-content	61.9%	60.8%	62.6%		62.0%		60.28%		62.4%	
g	**1673**	**1478**	**2962**	*HI - HP*	**2837**	*HI - HP*	**1861**	*HI - HP*	**2588**	*HI - HP*
sp_{CI}	21	24	662	*37-48*(*)	662	*35-46*	204	*23-30*	557	*33-42*
sp_{CII}	37	34	946	*66-68*	926	*63-65*	285	*41-43*	715	*57-60*
$upsp_{CI}$	8	3	131	*14-6*	128	*14-6*	53	*9-4*	123	*12-5*
$upsp_{CII}$	20	13	946	*35-26*	926	*34-25*	122	*22-16*	402	*31-23*

Table 1: Comparison of the numbers of putative strong promoters for two *Proteobacteria* and four *Firmicutes* characterized by close (high) AT-contents (range $[60.2\%, 62.4\%]$), under conditions CI and CII, and with or without UP element required. g: total number of genes encoding proteins in the genome considered; HI: *Haemophilus influenza*, HP: *Helicobacter pylori*; LI: *Listeria innocua*, LM: *Listeria monocytogenes*, SPn: *Streptococcus pneumoniae*, TT: *Thermoanaerobacter tengcongensis*. (*) With HI then HP taken as a reference, these columns in italics yield the predicted values sp_{CI}, ..., $upsp_{CII}$ on the basis of proportionality to total gene number; the two predicted values in italics have to be compared with the value on their left. Example: the reference being HI, predicted sp_{CI} for LI is $sp_{CI}(LI) = \frac{sp_{CI}(HI)\times 2962}{1673} = 37$, with $sp_{CI}(HI) = 21$; $sp_{CI}(LI)$ has to be compared with the value of 662 observed for LI.

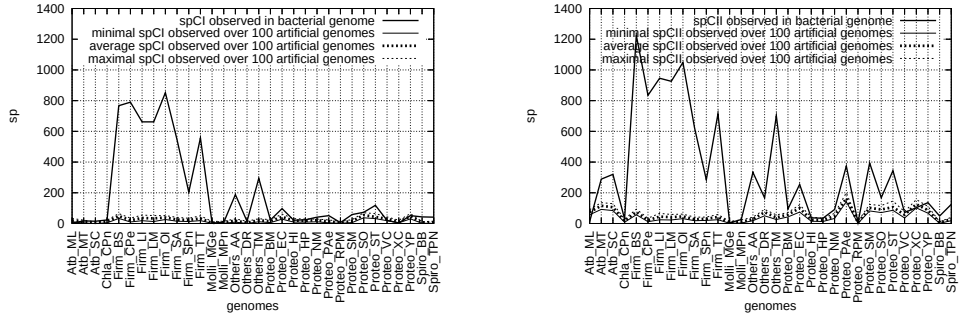
bias regarding the computation of the correlation coeffi cient. We recall that 7 out of the 22 nucleotides of the UP element consensus are nucleotides A, 5 are nucleotides T and 3 are A or T (W). We now recapitulate the results obtained regarding AT-richness influence on $p1$ and $p2$: (i) AT-richness does not interfere so long as the UP element is not considered ($p1$); (ii) on the contrary, AT-content and percentage $p2$ seem to be correlated. A pending question is then: does AT-richness alone entail high $upspCI$ and $upspCII$ values? To answer this question, we will compare *Firmicutes*' genomes with similarly AT-rich *Proteobacteria genomes* as well as similarly AT-rich genomes generated at random.

4.3 Comparing *Firmicutes* with similarly AT-rich *Proteobacteria*

Table 1 compares four *Firmicutes* with two *Proteobacteria*, all characterized by close (high) AT-contents (range $[60.2\%, 62.4\%]$). Table 1 takes into account the bias due to the differences between the total numbers of genes coding for m-RNAs and characterizing the genomes considered. Table 1 predicts what would be the numbers of strong promoters identifi ed for the four *Firmicutes*, when either *Proteobacterium Haemophilus influenza* (HI) or *Helicobacter pylori* (HP) is taken as a reference, which means considering a proportionality rule based on the total number of genes for the reference and the total number of genes for the *Firmicute* genome. In the majority, the thus predicted values are much below a tenth of the values obtained with BACTRANS[2] software. In particular, the case of *Streptococcus pneumoniae* (SPn) is all the more striking as its total number of genes is close to those of the two references HI and HP; nevertheless, there is a signifi cant disproportion between predicted and observed values. The next step just reproduces the experiment of Table 1, now considering genomes generated at random in place of the bacterial genomes (see Supplementary Data, Appendix 4). As expected, we verify that the orders of magnitude for observed and predicted values are identical. Therefore we conclude that the disproportion between *Proteobacteria* and *Firmicutes*, with respect to the total number of genes (g), is not the reason for the high values observed regarding sp_{CI} and sp_{CII} values or $upsp_{CI}$ and $upsp_{CII}$ values.

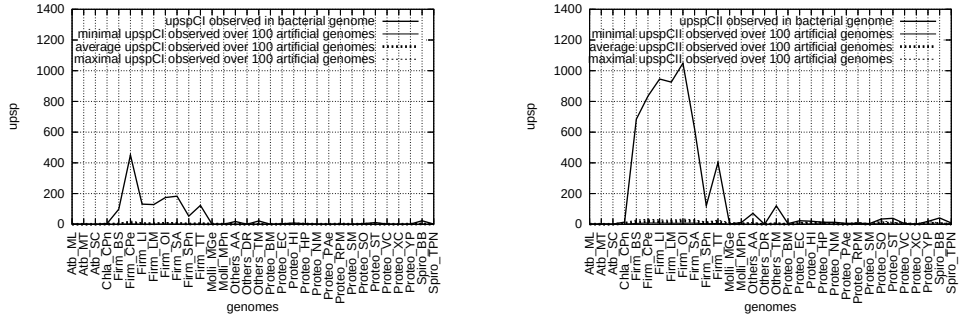
4.4 Comparing observations in bacterial genomes with expectations in randomly generated genomes

The objective is two-fold: (i) we perform a statistical study of strong promoter frequencies to gauge the signifi cance of the motif over all 32 genomes studied and under the four constraints considered; (ii) we wish to check whether *Firmicutes* "average" genomes generated at random still differentiate from other genomes. Figure 2 compares the numbers of genes harbouring a $\sigma 70$ promoter in artifi cial genomes with the numbers of genes associated with such promoters in bacterial genomes. For comparison purposes, a common scale is used in the four pictures of Figure 2 (The reader interested in details is referred to Figure 5.1 (see Supplementary Data, Appendix 5) for a magnifi cation relative to artifi cial genomes' results). Besides, we recall that we rely on Z-scores to measure the difference between a bacterial genome and the corresponding "average" genome generated at random. The difference is ascertained if the Z-score value is greater than a given threshold. To take into account the fact that the gap's length between SD box and -10 box is specifi ed as a wide interval (but conforms to reality), we lead our investigation examining Z-scores with respect to two thresholds (2 and 5).



(a) CI, UP element optional

(b) CII, UP element optional



(c) CI, UP element required

(d) CII, UP element required

Figure 2: Observed bacterial genome values *versus* minimal, average and maximal values observed over 100 similarly AT-rich genomes generated at random, for sp and $upsp$ respectively, under 4 constraint sets. See Figure 1 for defi nition of sp and $upsp$, and for genome abbreviations. See text, Subsection "Genome analysis upon request" for the defi nition of CI and CII constraints.

We start our analysis focusing on the CI case. As already seen for four *Firmicutes* in Table 1, Figure 2 (a) (CI) shows that putative strong promoters are signifi cantly more frequent in *Firmicutes* genomes than in corresponding artifi cial genomes. From now on, we distinguish the 2 *Mollicutes* from the other 8 *Firmicutes*: *Mollicutes* are *Firmicutes* characterized by small genomes. Given as quadruplets (minimum, maximum, **average**, standard deviation), Z-scores (defi ned in paragraph "Empirical approach") are

as follows: *Firmicutes* except *Mollicutes*: (39.5, 165.32, **103.24**, 37.77); *Proteobacteria*: (1.42, 10.9, **4.57**, 3.92). We check that all 8 previous *Firmicutes*' Z-scores are above threshold 5. Regarding the 12 *Proteobacteria* studied (*Rickettsia prowazekii* excluded because of its small genome size), 9 have their Z-scores above threshold 2, among which 4 have their Z-scores above threshold 5. In particular, the Z-score obtained for *E. coli* genome is 7.33. Not surprisingly, low sp_{CI} values are observed for species with small genomes: *Borrelia burgdorferi* (0.91 Mbp), *Chlamydophila pneumoniae* (1.22 Mbp), *Mycoplasma genitalium* (0.58 Mbp), *Mycoplasma pneumoniae* (0.81 Mbp), *Rickettsia prowazekii* (1.11 Mbp) and *Treponema pallidum nichols* (1.13 Mbp). All previous six species are either obligate intracellular pathogens, symbionts or animal commensal parasites.

When examining the 26 species with large genomes under constraints CI, we observe that 23 have their Z-scores over threshold 2 and 14 have their Z-scores over threshold 5. Therefore, this ascertained difference between prokaryotic genomes and artificial genomes shows a biological specificity. For a detailed comparison with sp_{CII}, $upsp_{CI}$ and $upsp_{CII}$ values (Figure 2, (b), (c) and (d)), the reader is referred to Tables 5.1 and 5.2 in Supplementary Appendix 5. We summarize the main results and conclusions in the following two paragraphs.

On the basis of two thresholds τ_1 and τ_2, 2 and 5 respectively, we now recapitulate the numbers of large genomes (among 26) for which the total number of promoters identified is shown to be significantly different from those of genomes generated at random: CI:(τ_1: 23, τ_2: 14); CII:(τ_1: 26, τ_2: 22); CI, UP element required:(τ_1: 15, τ_2: 11), computed over 21 genomes with calculable Z-scores; CII, UP element required:(τ_1: 21, τ_2: 18), computed over 25 genomes with calculable Z-scores (also see Suplementary Data, Appendix 5, for a recapitulation relative to non *Firmicutes* large genomes and for details about *E. coli*). Thus, as a first result, we show that relaxing the constraints from CI to CII entails an increase of the Z-score. We conclude that relaxing the matching constraint is not antagonistic to motif significance. This is not a trivial result, as the opposite was expected instead. Secondly, we show that adding the UP element constraint lowers the significance under CI constraints. On the contrary, the significance reinforcement due to constraint relaxing (CII) is relatively not much affected by the addition of the UP element requirement. This is a remarkable result since the UP element has only been identified experimentally in four genomes (*E. coli*, *Bacillus subtilis*, *Vibrio natriegens*, *Geobacillus stearothermophilus*). Thirdly, we confirm that *Firmicutes* clearly show a specific trend, with an average Z-score above 70 in CII conditions and when the UP element is required, and above 100 in all other three conditions.

Moreover, for randomly generated genomes, when the UP element is optional and whatever the constraint, we check that the orders of magnitude are comparable between *Firmicutes* and other species (see Supplementary Data, Appendix 5). On the contrary, when the UP element is required, *Firmicutes* artificial genomes slightly differentiate from other artificial genomes, thus showing the influence of AT-richness. This result was expected as the -10 box and the UP element are AT-enriched, and *Firmicutes* are the AT-richest genomes. We check that non *Firmicutes* artificial genomes with an AT-richness over 60% also show this AT-bias (*Haemophilus influenza*, *Helicobacter Pylori*, *Rickettsia prowazekii*, *Borrelia burgdorferi*). We would insist on distinguishing the AT-bias, whose influence is more or less noticeable for bacterial genomes as well as randomly generated genomes, whatever the species, and the species bias such as the one identified for *Firmicutes*. To recapitulate, current and previous Subsections definitely prove that the explanation for the *Firmicutes* specificity neither lies in a bias related to the total number of genes in these genomes nor in their high AT-richness.

4.5 About the influence of the global AT-bias

The quality of softwares devoted to promoter prediction is evaluated through the numbers of false positive and false negative occurrences obtained. The AT-bias is known to interfere with the number of false positive and thus some works proposed corrections relative to motif recovering in biased genomes [47, 26]. At genome scale, the number of false positive can only be deduced from confrontation with reposito-

ries gathering knowledge about experimentally verified promoters, such as RegulonDB [45], which is devoted to *E. coli*. The purpose of BACTRANS^2 being to identify the promoters with the highest intrinsic potentialities, there is no motive in this case to consider AT-bias as a prejudicial interference source. This stated, we wish to recapitulate all informations regarding At-bias.

In the first two subsections of Section "Results and discussion", we showed that no linear correlation exists between AT-contents and $p1_{CI}$ or $p1_{CII}$ percentages whereas a correlation exists between AT-contents and $p2_{CI}$ or $p2_{CII}$ percentages. The correlation was shown weaker for $p2_{CI}$ than for $p2_{CII}$. As expected, these results are in accordance with the direct observations on genomes generated at random.

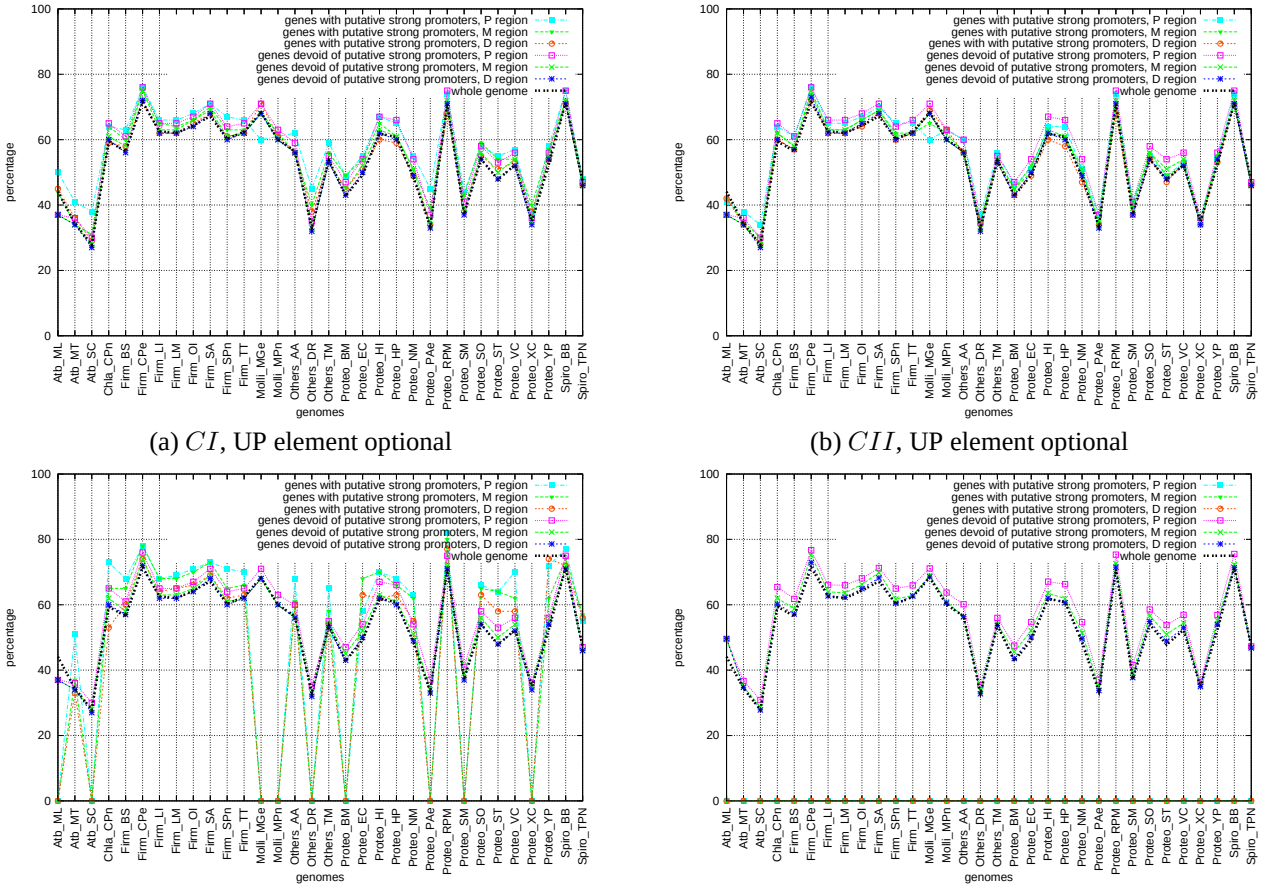
It is interesting to evaluate whether promoter significance is related or not to AT-richness. The linear correlation coefficient between AT-richness and Z-score is calculated over all genomes with calculable Z-scores: this coefficient amounts to 0.47, 0.40, 0.49 and 0.42 for $spCI$, $spCII$, $upspCI$ and $upspCII$ values respectively; escaping small genome bias, the coefficient amounts to 0.67, 0.62, 0.78 and 0.62. Thus the correlation between Z-score and AT-richness is subject to a slight increase when one escapes the bias of species with small genomes. It was not foreseeable that such a loose correlation between Z-score and AT-richness would exist under CII conditions and when the UP element is required. Here, in contrast with conclusion relative to $p2$, the only statistically strong correlation between AT-content and Z-score is observed in the $upspCI$ case, and in this case only.

4.6 Comparison of the AT-contents in three regions upstream of the start codon

A further investigation is required to attempt to explain the *Firmicutes* specificity: we now analyse more thoroughly the AT-content distribution in the 350 nucleotide-long regions upstream of the start codon (SC). We chose the value 350 in this genome-wide analysis, to be sure that the longest region scanned by BACTRANS^2, 278, would be included (see subsection "Genome analysis upon request"). In this 350 bp-long region, we consider three equally 116 bp-long sub-regions. The three regions are denoted P (proximal with respect to SC), M (middle) and D (distal). For each genome, we compute three AT-content averages relative to the P, M and D sub-regions upstream of the genes associated with strong promoters (G_{SP}); symmetrically, we compute three AT-content averages relative to the P, M and D sub-regions upstream of the genes containing no strong promoter (G_{noSP}). Figure 3 shows the six curves relative to the local AT-content averages observed in the three sub-regions, over the two sets of genes (G_{SP} and G_{noSP}). We observe the following: (i) AT-content averages in the proximal sub-regions are the highest ones, for any gene set considered (G_{SP} or G_{noSP}) and the difference with respect to other sub-regions is around 5%; (ii) the highest difference between a curve relative to G_{SP} and the corresponding curve relative to G_{noSP} is less than 5%. Indeed, it is sufficient to replace 6 nucleotides G or C with 6 nucleotides T or A in a 116 nucleotide-long sub-region to increase AT-content by 4%. In conclusion, no significant local AT-bias is shown to distinguish sub-regions upstream of the genes harbouring a strong promoter from sub-regions upstream of the genes devoid of such strong promoters. In particular, this conclusion holds for *Firmicutes* genomes. Regarding the structural *Firmicutes* specificity, our successive investigations lead towards suggesting natural evolution as its cause.

4.7 Putative strong promoters *versus* experimentally verified functional promoters

We remind the reader that BACTRANS^2 is intended to change the scale in genome analysis of potentially high ORF expression. We would insist that BACTRANS^2 was not designed with the purpose of predicting promoters, but indeed with the aim of pointing out potentially strong promoters which should be experimentally checked in priority. Relying on the list of promoters sorted in decreasing score order (UP elements at the top of the list), one can carry out the experimental validation of the top promoters, for a



(a) CI, UP element optional

(b) CII, UP element optional

(c) CI, UP element required

(d) CII, UP element required

Figure 3: Comparison of the AT-contents in three regions upstream of the start codon, for genes with no promoter identified and genes harbouring a promoter (all genes coding for m-RNAs). 32 genomes are considered under constraints CI and CII. A 350 nucleotide-long region upstream of the start codon (SC) is considered. We consider the three 116 nucleotide-long sub-regions: P (proximal with respect to SC), M (middle) and D (distal).

given genome. On the other hand, data relative to experimentally verified functional promoters are available. We will refer to RegulonDB and PromEC databases. Both databases provide annotations relative to experimentally validated promoters in *E. coli* genome. *In vivo*, transcriptional regulations are known to compensate for promoter weakness [23, 8] and it is not known whether some functional promoters might also be intrinsic strong promoters. Therefore, for an investigation as complete as possible, a confrontation between BACTRANS[2] outputs and known *E. coli* functional promoters is interesting. RegulonDB is the reference database for *E. coli* curated knowledge [45, 40]; 601 $\sigma 70$ promoters are listed inside. PromEC is dedicated to functional $\sigma 70$ *E. coli* promoters [29]; it includes 471 entries, among which some are common to RegulonDB. To our knowledge, no such large promoter databases are available for any other genome. *E. coli* genome contains 4173 genes. In 5.6 RegulonDB release (january 2007, http://regulondb.ccg.unam.mx/data/Promoter Set.txt), we listed 1632 genes, among which 601 (36%) are associated with a $\sigma 70$ annotation. The field showing evidence of experimental validation contains at least

one of the following items (or possibly sub-items): transcription initiation mapping, RNA polymerase footprinting, inferred from mutant, inferred from direct assay, inferred from genetic interaction, inferred from experiment, inferred from physical interaction. Some promoters among the earlier listed in RegulonDB had not been assigned an evidence field. Nevertheless, they are indeed experimentally verified promoters and are in the process of being fully annotated (Salgado, personal communication).

Only few genes of *E. coli* harbouring the potentially strongest $\sigma 70$ promoters identified by BACTRANS^2 are also listed in RegulonDB and PromEC databases as harbouring a functional promoter. Not surprisingly, the distance between the transcription start site (TSS) of the putative strong promoter and the TSS of the functional promoter may vary in a large range. Under CI conditions, 97 strong promoters are identified by BACTRANS^2; 14 out the corresponding 97 genes are referred to in at least one of the two databases aforementioned. Under CII conditions, 21 out of the 255 genes identified with strong promoters are cited in at least one database. The reader is referred to Supplementary Appendix 6 for a detailed comparison. We confirm that according to our scoring function, if a functional $\sigma 70$ promoter is known for a gene, it is weaker than the putative strong promoter identified by BACTRANS^2.

On the other hand, we recall that Huerta and Collado-Vides established through two sets of 111 and 392 known promoters that only 50% to 60% of these promoters have the strongest scores (according to their first scoring function) among all candidates located in the same regulatory region [31]. This was an argument for further adding other parameters in score computation (such as knowledge about the distribution of promoter candidates in the regulatory region). Since these known promoters which are also the strongest ones are not identified by BACTRANS^2, an explanation has to be put forward. It might happen that some if not all strongest promoters among known promoters do not satisfy the minimal similarity constraints required by BACTRANS^2. Conversely, strong promoters identified by BACTRANS^2 would not fit well to any of the 288 specific weight matrices of Huerta *et al* (the -10 box length varies between 8 bp and 10 bp, the -35 box varies from 8 bp to 22 bp, see Table 1 in [31]). Besides, Shultzaberger and co-workers recently compiled models from 684 functional promoters listed in RegulonDB and PromEC databases [48]. One model is provided for each possible length of the gap between -35 and -10 boxes (in range $[15, 20]$) (see [48], Figure 2). Such models are consistent with our specification of the $\sigma 70$ strong promoter in terms of bp distance constraints. Nonetheless, the sequence logos provided in [48] compile trends and it is remarkable that on average, the specificities of these models are rather low. The sequence logos of the -35 boxes indicate that the first two nucleotides of the consensus TTGAC are more likely to be encountered simultaneously in the functional promoters than any other pair of nucleotides in the consensus; this description is compatible with constraint CII but not with constraint CI. However, the sequence logos of -10 box suggest that it is very unlikely that more than 3 nucleotides are simultaneously conserved with respect to consensus TATAAT; the least drastic condition, CII, requires that no more than 2 nucleotides differ from the -10 consensus. Hence, we did constrain our strong $\sigma 70$ promoter model in a way consistent with biological reality (or what is known of it at the present time), that is with -35 box less specific than -10 box. Anyway, we constrained our model a degree higher with respect to biological reality, which is the least expected regarding intrinsically strengthened promoters. In Huerta and Collado-Vides' study, there was no contribution from PromEC. But if we suppose that there is no deviation between PromEC promoters (not also belonging to RegulonDB) and RegulonDB promoters, in view of the remarks drawn from Shultzaberger and co-workers' compilations, we can explain how functional promoters (possibly first ranked according to another scoring function than ours) are likely to be rejected by BACTRANS^2.

Furthermore, only 38% and 11% of the genes coding for m-RNAs in *E. coli* are listed in RegulonDB and PromEC, respectively. For each of the "remaining" genes not referred to in RegulonDB or PromEC, we do not know whether the reason lies in promoter experimental validation failure, or in this gene having not yet been studied. Anyway, the remarks in previous paragraph make us inclined to think that further experimentations are not likely to reveal the existence of *strong functional* promoters in the "remaining" gene set: various transcriptional activators are known to correct intrinsic promoter weakness.

Nonetheless, (i) it was interesting to compare BACTRANS^2's promoter sets with those of RegulonDB and PromEC; (ii) we checked that functional $\sigma 70$ promoters are intrinsically weak promoters with respect to our description of strong promoters (even under CII conditions); this difference in strengths is intended for genetic engineering purpose: *in vitro* constructions do require intrinsically strong promoters. Theoretical intrinsic strength is a first requirement. Verifications of functionality as well as high protein synthesis have to be implemented.

Finally, we study the distribution of distances between start codon and +1 transcription, regarding genes harbouring strong promoters. We compare the distributions obtained under the four constraints (CI, CII, UP element optional, UP element optional) (see Supplementary Data, Appendix 7). This distribution is available for 599 $\sigma 70$ known promoters listed in RegulonDB (see [31], Figure 1). Regarding the searches performed when the UP element is optional, we acknowledge the highest percentages around distances comprised between 11 and 30 bp. This result is in accordance with what is observed for the 599 promoters mentioned above. Then, when comparing on a common distance range ([0,220]), the flattening of the distribution corresponding to the highest distances is more pronounced for the 599 promoters than for the strong promoters identified with BACTRANS^2 (see Figure 7.1 (a) and (b)). The 3 strong promoters recovered with an UP element under CI constraints are located in the distance range $[141, 180]$ upstream of the start codon (see Figure 7.1 (c)). Under the more relaxed constraints CII, the 23 distances observed for strong promoters harbouring an UP element range in interval $[0, 210]$ (see Figure 7.1 (d)). No trend can be highlighted, except a percentage peak in interval $[141, 150]$.

4.8 Experimental verification for some strong promoters in *Thermotoga maritima*

A different consideration is that of checking whether a putative strong promoter is really *functional* or is actually a *strong* promoter. In the context of another study devoted to hyperthermophilic bacterium *Thermotoga maritima*, the activity of seven putative strong promoters harbouring an UP element identified by BACTRANS^2 has been measured in *E. coli* cell free extracts [44]. Among these seven promoters, four were identified under the most constrained conditions CI ($TM1016$, $TM0373$, $TM0477$, $TM1667$). The other three were identified under CII conditions ($TM0032$, $TM1429$, $TM1780$). All of them promote protein synthesis, indicating that they are all functional promoters. Moreover, except $TM0032$, all provided a higher protein yield than that of the well-studied pTac promoter. Promoter pTac is a strong hybrid promoter consisting of the -35 region of the trp promoter and the -10 region of the lacUV5 promoter/operator [11]. $TM0477$ has been shown to be twice as strong as others regarding protein yield. Thus these results show that six promoters among the seven tested really favour high expression in *E. coli* cell free extracts.

5 Concluding remarks

Our work contributes to shedding new light on potentially high expressed ORFs in prokaryotic genomes. So far, we focused on potentially high transcription. Studying high translation potentialities in prokaryotic genomes is currently under work. Our computational approach does not merely rely on intrinsic characteristics such as similarity to canonical $\sigma 70$ binding sites and adequacy with optimal bp distances between sites; it takes into account the UP element. In itself, this latter feature introduces originality with respect to comparative studies devoted to bacterial transcription promoters. As an *in silico* approach taking into account the transcription enhancer in the context of a genome-comparative analysis, our work is complementary to that of Huerta and Collado-Vides, regarding transcription potentiality of genomes. In addition to aforementioned reference works, our statistical study discusses the significance of the $\sigma 70$ promoters identified. Under the most relaxed constraint, this significance is rigourously proven for nearly all large genomes. More peculiarly, *Firmicutes* would appear as genomes more favoured by nature with respect to high intrinsic transcription potentiality. Attempting to infer why this would characterize *Fir-*

micutes and whether this would be the result of conservation or speciation is beyond the scope of this article. On the other hand, the conservation of RNA polymerase through evolution in bacteria justifies the search of the canonical *E. coli* $\sigma 70$ promoter in other genomes. Nevertheless, the UP element has been identified by experimentation in four genomes only. Thus our comparative study also brings new knowledge about the potentialities of various genomes regarding enhanced $\sigma 70$ promoters. Indeed, the presence of the enhanced promoter has been proven significant in all eight large *Firmicutes* genomes studied and between ten and thirteen non *Firmicutes* large genomes studied (depending on the Z-score threshold considered). Statistical relevance is supported by the bias existing with respect to genomes generated at random. The possible influence of AT-richness on this former bias could not be proven. Regarding AT-richness, the correlation analyses performed sustain its absence when the UP element is not required. When the UP element is required, the results of various correlation analyses are not consistent with one another. Therefore it cannot be indisputably proven in this way whether a strong AT-bias influences or not the results relative to strong promoters harbouring an UP element. However, the case of *Firmicutes* brings all the more brilliant a refutation of such an existence as these genomes show the highest AT-richnesses: no similarly AT-rich genomes generated at random show the exceptionally high frequencies observed for *Firmicutes*. Finally, generic software platform BACTRANS^2 currently provides such putative strong promoters for 45 genomes. Moreover, BACTRANS^2's genericity allows the user to analyse genomes with respect to any other motif consisting of 3 or 4 boxes.

Acknowledgements

Sylvain Demey was supported by the Pays de la Loire Region ("Technological Innovations and Postgenomics" C.P.E.R. program) and by Ouest-Genopole consortium (National Network of Genopoles, National Genomics Research Consortium). Especially regarding the former fund ressource, the scientific committee was particularly receptive to first author's commitment in the project launching and realisation. The first author is thankful to V. Sakanyan for fruitful discussions.

References

1. Aiyar, S.E., Gaal T. and Gourse, R.L. (2002) rRNA promoter activity in the fast-growing bacterium *Vibrio natriegens*. *J. Bacteriol.*, **184**(5), 1349–58.

2. BacTrans^2. http://www.sciences.univ-nantes.fr/lina/biose rv/BacTrans2/

3. http://www.sciences.univ-nantes.fr/lina/bioserv/BacTrans 2/supplementary_data/illustrations/

4. http://www.sciences.univ-nantes.fr/lina/bioserv/BacTrans 2/supplementary_data/rDBchecking/

5. Softberry. http://www.softberry.com/berry.phtml

6. Blom, G. and Thorburn, D. (1982) How many random digits are required until given sequences are obtained? *J. Appl. Prob.*, **19**, 518–531.

7. Braun, F., Marhuenda, F.B., Morin, A., Guevel, L., Fleury, F., Takahashi, M. and Sakanyan, V. (2006) Similarity and divergence between the RNA polymerase alpha subunits from hyperthermophilic *Thermotoga maritima* and mesophilic *Escherichia coli* bacteria. *Gene*, **380**, 2, 120–126.

8. Browning, D.F. and Busby, S.J. (2004) The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.*, **2**, 57-–65.

9. Bulyk, M.L., McGuire, A.M., Masuda, N. and Church, G.M (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res.*, **14**, 2, 201-208.

10. Collado-Vides, J., Magasanik, B. and Gralla, J.D. (1991) Control site location and transcriptional regulation in *Escherichia coli*. *Microbiol. Rev.*, **55**, 371–394.

11. de Boer, H.A., Comstock, L.J. and Vasser, M. (1983) The tac promoter: a functional hybrid derived from the trp and lac promoters. *Proc. Natl. Acad. Sci. USA.*, **80**, 21-25.

12. EcoCyc. http://www.ecocyc.org/

13. Estrem, S.T, Gaal, T., Ross, W. and Gourse, R.L. (1998) Identifi cation of an UP element consensus sequence for bacterial promoters. *Proc. Natl. Acad. Sci. USA*, **95**, *9761–9766, august.*

14. Estrem, S.T., Ross, W., Gaal, T., Chen, Z.W., Niu, W.,Ebright,R.H. and Gourse,R.L. (1999) Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev.*, **13**, 2134–2147.

15. Fenton, M.S., Lee, S.J. and Gralla, J.D. (2000) *Escherichia coli* promoter opening and -10 recognition: Mutational analysis of sigma70. *EMBO J.*, **19**, 1130–1137.

16. Fredrick, K., Caramori, T., Chen, Y.F., Galizzi, A. and Helmann, J.D. (1995) Promoter architecture in the flagellar regulon of *Bacillus subtilis*: high-level expression of flagellin by the sigma δ RNA polymerase requires an upstream promoter element. *Proc. Natl. Aca. Sci. USA*, **92**, 2582–86.

17. Fu, C. J. (1996) Distribution of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, **6**, 957–974.

18. GenBank. http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi.

19. GenoExpertBacteria. http://www-geb.inrialpes.fr/.

20. Genostar. http://www.genostar.com/

21. Gralla, J. and Collado-Vides, J. (1996) Organization and function of transcription regulatory elements. *Escherichia coli and Salmonella, Cellular and Molecular Biology (Neidhart,F.C., Curtiss,R., Ingraham,J., Lin,E.C.C., Low,K.B., Magasanik, B., et al., eds), American Society for Microbiology, Washington, D.C.*, **57**, 1232–1246

22. Gross, C., Lonetto, M. and Losick, R. (1992) Bacterial sigma factors. *In McKnight, S.L. and Yamamoto, K.R. (Eds.), Transcriptional Regulation, New York Cold Spring Harbor Laboratory Press*, 129-–176.

23. Gross, C.A., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J., Young, B. (1998) The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb. Symp. Quant. Biol.*, **63**, 141–155.

24. Gruber, T.M. and Gross, C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.*, **57**, 441–466

25. Harley, C.B. and Reynolds, R.P. (1987) Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.*, **15**, 2343–2361.

26. Hasan, S. and Schreiber, M. (2006) Recovering motifs from biased genomes: application of signal correction. *Nucleic Acids Res.*, **34**(18), 5124–5132.

27. Hawley, D.K. and and McClure, W.R. (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.*, **25**; 11(8), 2237–2255.

28. Herring, C.D., Raffaelle, M., Allen, T.E., Kanin, E.I., Landick, R., Ansari, A.Z. and Palsson, B.O. (2005) Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *J. Bacteriol.*, **187**, 6166–6174.

29. Hershberg, R., Bejerano, G., Santos-Zavaleta, A. and Margalit, H. (2001) PromEC: An updated database of *Escherichia coli* mRNA promoters with experimentally identifi ed transcriptional start sites *Nucleic Acids Res.*, **29**(1), 277.

30. Huerta, A.M., Francino, M.P., Morett, E. and Collado-Vides, J. (2006) Selection for Unequal Densities of $sigma$70 Promoter-Like Signals in Different Regions of Large Bacterial Genomes. *PLoS Genet.*, **10**; 2(11).

31. Huerta, A.M. and Collado-Vides, J (2003). Sigma70 promoters in Escherichia coli: specifi c transcription in dense regions of overlapping promoter-like signals. *J Mol Biol.*, **17**; 333(2), 261-278.

32. Li, H., Rhodius, V., Gross, C. and Siggia, E.D. (2002) Identifi cation of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. USA*, **99**, 11772-11777.

33. Lisser, S. and Margalit, H. (1993) Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res.*, **21**, 1507–1516.

34. Martinez-Antonio, A. and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.

35. Morin, A., Huysveld, N., Braun, F., Dimova, D., Sakanyan, V., and Charlier, D. (2003) Hyperthermophilic *Thermotoga* arginine repressor binding to full-length cognate and heterologous arginine operators and to half-site targets *J. Mol. Biol.*, **332**(3), 537–53.

36. Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Bhattacharyya, A., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N., Kyrpides, N. (2003) The ERGOTM genome analysis and discovery system. *Nucleic Acids Res.*, **31(1), 164–171.**

37. Pager, M.S. and Helmann, J.D. (2003) The sigma 70 family of sigma factors. *Genome Biol.*, **4**, 203.1-203.6.

38. Perri`ere, G, Duret, L. and Gouy, M. (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385, http://pbil.univ-lyon1.fr/databases/hobacgen.html.

39. Pribnow, D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA*, **72**, 784–788.

40. RegulonDB. http://regulondb.ccg.unam.mx/data/Promoter Set.txt

41. Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2003) Occurrence probability of structured motifs in random sequences. *J. Comp. Biol.*, **9**, 761–773.

42. Robin, S. and Daudin, J.-J. (2001) Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.*, **36**, 179–193.

43. Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A. *et al* (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*, **262**, 1407–1413.

44. Sakanyan, V., Dekhtyar, M., Morin, A., Braun, F. and Modina, L. (2003) Method for the identifi cation and isolation of strong bacterial promoters. *European patent application*, 3290203.3, january 27th.

45. Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A. and Collado-Vides, J. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions *Nucleic Acids Res.*, Jan 1; 34(Database issue): D394-7.

46. Savchenko, A., Weigel, P., Dimova, D., Lecocq, M. and Sakanyan, V. (1998) The *Bacillus stearothermophilus argCJBD* operon harbours a strong promoter as evaluated in *Escherichia coli* cells. *Gene*, **212**(5), 167–177.

47. Schreiber, M. and Brown, C. (2002) Compensation for nucleotide bias in a genome by representation as a discrete channel with noise. *Bioinformatics*, **18**, 507–-512.

48. Shultzaberger, R.K., Chen, Z., Lewis, K.A. and Schneider, T.D. (2007) Anatomy of *Escherichia coli* σ70 promoters. *Nucleic Acids Res.*, **35**(3), 771–788.

49. van Helden, J., Rios, A.F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.*, **28**, 1808–1818.

A large-scale analysis for significance assessment of frequencies relative to potentially strong sigma 70 promoters: comparison of 32 prokaryotic genomes

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

Abstract

This report presents a computational analysis of high ORF expression potentialities in prokaryotic genomes, in relation with medical and economic relevance. Given a bacterial genome and the description of a structured motif, the software BACTRANS[2] implements the search of the occurrence most similar to the motif, in the regulatory region of each gene. In this work, the software BACTRANS[2] was run over 32 prokaryotic genomes, to identify putative strong $\sigma 70$ promoters. We focused in particular on $\sigma 70$ promoters harbouring an UP element, which enhances transcription initiation. We performed four computational analyses per genome, combining two promoter strength levels (CI ¿ CII) with either mandatory or optional UP element presence. We compared the frequencies obtained for 32 bacterial genomes, under these four constraint specifications.

First, we show that an over-representation of putative strong promoters differentiates the AT-rich Firmicutes' genomes from other genomes. Another interesting result is that strong promoters of relatively lesser quality (CII) are more frequently associated with an UP element than strong promoters of better quality (CI).

Then, per each bacterial genome studied, we generated at random 100 artificial genomes. Such genomes are only constrained as to have the same two following characteristics as the bacterial genome: same total number of genes and same proportions of A, C, T and G nucleotides in the 350 nucleotide-long region upstream of start codon. The $\sigma 70$ promoter frequency observed on average over these 100 genomes is compared to the frequency observed for the bacterial genome, under each of the four constraint sets aforementioned. Thus, the statistical significance of the $\sigma 70$ model is discussed for each genome, under each constraint set. For most genomes, and especially for *Firmicutes*, a meaningful difference is statistically ascertained. Besides, the comparison between *Firmicutes* genomes and equally AT-rich *Proteobacteria* genomes also confirm that the *Firmicutes* specificity is not related to genome size bias. Hence, *Firmicutes* would appear as genomes more favoured by nature with respect to high intrinsic transcription potentiality.

Throughout the report, we discuss the influence of AT-richness on promoter frequencies, implementing various correlation analyses. We show that an influence is only observed when the UP element is required. Then we evaluate whether the statistical significance of the $\sigma 70$ model is related or not to AT-richness. Interestingly, we find that the relation is loose except when the UP element is required, and under the more stringent constraint (CI). Thus, we distinguish the AT-bias, whose influence is more or less noticeable for bacterial genomes as well as randomly generated genomes, whatever the species, and the species bias such as the one identified for *Firmicutes*. Finally, we compare the AT-percentages of three sub-regions of the 350 bp-long region upstream of start codon, distinguishing between genes harbouring strong promoters and genes not harbouring any such strong promoters. We can show no evidence that the over-representation characterizing *Firmicutes* is due to a local AT-bias.

To our knowledge, our work is the very first statistical approach thoroughly analysing the presence significance of various potentially strong $\sigma 70$ promoter models, including models harbouring the UP element enhancer, in the context of a genome-comparative study. The presence of the enhanced promoter has been proven significant in all eight large *Firmicutes* genomes studied and between ten and thirteen non *Firmicutes* large genomes studied (depending on the Z-score threshold considered).

LINA, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3