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

Christine Sinoquet, Sylvain Demey, Frédérique Braun

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

comparison between 32 bacterial 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

— Combinatorics and Bioinformatics —

RESEARCH REPORT

N° hal-00153303

October 2007

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

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

18 p.

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

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

christine.sinoquet@univ-nantes.fr

Abstract

The platform BACTRANS^2 has been designed to help select putative strong $\sigma 70$-like promoter candidates in prokaryotic genomes. It was run to investigate the importance of $\sigma 70$-like potentially high transcription in bacteria other than *Escherichia coli*. We performed a genome-comparative analysis of high ORF expression potentialities over 32 prokaryotic genomes. Besides, we put an emphasis on transcription strength reinforcement through the UP element presence and on translation potentiality enhancement through an optimal Shine-Dalgarno sequence.

We compared frequencies of putative strong promoters between various genomes. We show that in the AT-rich *Firmicutes*' genomes, frequencies of potentially strong $\sigma 70$-like promoters are exceptionally high. Besides, though they contain a low number of strong promoters, some genomes may show a high proportion of promoters harbouring an UP element. Putative strong promoters of lesser quality are more frequently associated with an UP element than putative strong promoters of better quality. A meaningful difference is statistically ascertained when comparing frequencies in bacterial genomes with frequencies in similarly AT-rich genomes generated at random; the difference is the highest for *Firmicutes*. Comparing some *Firmicutes* genomes with similarly AT-rich *Proteobacteria* genomes, we confirm the *Firmicutes* specificity. We show that this specificity is neither explained by AT-bias nor genome size bias but originates in the abundance of optimal Shine-Dalgarno sequences, a typical and significant feature of *Firmicutes* more thoroughly analysed in our study.

The generic software platform BacTrans2 currently provides such putative strong promoters for 45 genomes. These data may be of interest to select a subset of promoters for experimental characterization and possible further use in biotechnological applications. Finally, BacTrans2's genericity allows the user to analyse genomes with respect to any other super-motif consisting of 3 or 4 boxes.

To our knowledge, this work is the very first genome-comparative study thoroughly analysing the significance of various potentially strong sigma 70-like promoter models, including models harbouring the UP element enhancer.

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

This work addresses potentially high ORF expression related to $\sigma 70$-like promoters, in bacterial genomes. In these genomes, a single enzyme, the RNA polymerase, is responsible for the synthesis of all RNA types. The core holoenzyme $\alpha^2\beta\beta'$ is competent for transcribing a specific region of the DNA strand into an RNA molecule. However, transcription can only be initiated (at the so-called +1 transcription site) through a temporary biochemical complex. This complex is composed of the four previous sub-units and of a protein, the σ factor, the primary one being $\sigma 70$. As one of the simplest known bacterial models, *E. coli* K-12 has been subjected to intensive research, especially with regard to transcription (Hawley and McClure, 1983; Harley and Reynolds, 1987; Collado-Vides *et al.*, 1991; Lisser and Margalit, 1993; Fenton *et al.*, 2000; Gruber and Gross, 2003; Pager and Helmann, 2003; Herring *et al.*, 2005). Knowledge was therefore gained about the *E. coli* $\sigma 70$ factor's binding sites. Their consensuses are respectively TTGACA and TATAAT, in the 5' to 3' direction. The optimal fixation of the RNA polymerase requires that the site with the consensus TTGACA should be located between 35 bp and 30 bp or thereabouts upstream of the first transcribed nucleotide. This former site is thus called the -35 box. The Pribnow box, TATAAT, is called - 10 box for similar reasons (Pribnow, 1975). These sites are separated by 15 to 21 bp in the known functional promoters, the canonical $\sigma 70$ promoter being characterized by the optimal distance of 17 bp. Various methods and softwares devoted to the prediction of functional promoters in *Escherichia coli* genome have been developped (Huerta and Collado-Vides, 2003; Eskin *et al.*, 2003; Bulyk *et al.*, 2004; Shultzaberger *et al.*, 2007; to restrain to a few examples). We do not mention here the numerous softwares designed to uncover a motif common to a set of biological sequences.

Not only is the RNA polymerase conserved through evolution in bacteria, there seems to be a *single* $\sigma 70$ factor, responsible for housekeeping gene transcription, across the bacterial kingdom (Wosten, 1998; Mittenhuber, 2002). Both points legitimate searches for $\sigma 70$-like binding sites in other prokaryotic genomes (Morrison and Jaurin, 1990; Gross *et al.*, 1992; Gralla and Collado-Vides, 1996; Li *et al.*, 2002;

Martinez-Antonio and Collado-Vides, 2003). Furthermore, the number of complete prokaryotic genomes sequenced has increased at a high speed (594 in october 2007), which allows genome-wide computational investigations. In the domain of *in silico* analyses related to $\sigma 70$ factor transcription, a reference contribution showed that $\sigma 70$ promoter-like sequences are present throughout the kingdom of prokaryotic organisms (Huerta *et al.*, 2006). This former study demonstrated that the density of promoter-like sequences is high within regulatory regions, in contrast to coding regions and regions located between convergently transcribed genes. For instance, an average of 38 promoter-like sequences was computed for *E. coli*, within each 250 bp sub-region located upstream of the start codon. Density differences between regulatory and non-regulatory regions were detected in most of the large genomes analysed.

In vivo, transcriptional regulations are known to compensate for promoter weakness (Gross *et al.*, 1998; Browning and Busby, 2004). For example, Huerta and Collado-Vides established that more than 50% of experimentally verified promoters are not the promoters with the highest scores when scoring relies on the proximity to the canonical promoter, both in terms of consensus similarity and optimal bp distances between boxes (Huerta and Collado-Vides, 2003). This statement was checked on the 111 promoters constituting a training set designed in a former work (Gralla and Collado-Vides, 1996). On the other hand, in *E. coli* genome, it has been shown that mutations in the -10 box or the -35 box that bring the promoter sequence closer to the $\sigma 70$ consensus tend to increase the strength of the promoter, and conversely, mutations decreasing homology to the $\sigma 70$ consensus tend to lower the promoter strength (Hawley and McClure, 1983). Thus, the more similar to the canonical $\sigma 70$ promoter, the more potentially strong this promoter would be, with the noteworthy exception that the *consensus* promoters may actually be weak because RNA polymerase binds them so strongly that it cannot escape (Ellinger *et al.*, 1994). Therefore, it is attractive to study and compare genomes from the point of view of potentially high transcription, allowing for mismatches, under a minimal similarity constraint. This large-scale comparative analysis is feasible through an *in silico* approach.

No computational method can capture the biological features and environmental conditions involved *in vivo*, to predict functional *strong* promoters. Besides, even for the most intensively studied prokaryotic genome, *E. coli*'s, the available repositories of $\sigma 70$ promoters do not provide annotations about promoter strength. The measurement of promoter activity in cellular or cell-free expression systems cannot be applied on a large scale. ChIP on chip assays allow the identification of transcription factor binding sites, under given environmental conditions, but high-throughput promoter strength measurement cannot be implemented using this technique. Thus, before such large-scale array experimentations may be conducted on the 32 genomes we are interested in, an *in silico* genome-comparative analysis focused on intrinsically high transcription potentiality is worth being performed.

In our work, we intentionally focus on the subset of putative strong $\sigma 70$ promoters already potentially favoured by the presence of an optimal Shine-Dalgarno sequence (GGAGG). The presence of the SD sequence has been ascertained for a large number of bacteria (Osada *et al.*, 1999) and it was established that the extent to which a SD sequence is conserved relates to its translation efficiency (Ma *et al.*, 2002). Besides, our study also puts emphasis on strength transcription reinforcement through the UP element presence. The Upstream Promoter element is an enhancer for transcription and thus for ORF expression (Ross *et al.*, 1993; Estrem *et al.*, 1999). In about 3% of *E. coli* promoters, an UP element has been identified 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 suggests that the UP element consensus should be valid throughout the bacterial kingdom. To our knowledge, in addition to *E. coli* genome, the UP element has been experimentally identified in *B. subtilis* (Fredrick *et al.*, 1995), *V. natriegens* (Aiyar *et al.*, 2002) and *G. stearothermophilus* (Savchenko *et al.*, 1998). At the present time, 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.).

In this report, we perform a comparison of the frequencies observed for the putative **strongest** promot-

ers over 32 bacterial genomes. 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 perform four genome-comparative studies. We discuss the statistical significance of our results through comparisons with randomly generated genomes, highlighting and elucidating the specific case of *Firmicutes*.

3 System and methods

3.1 Genome analysis upon request

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 encoding a protein, the tool first extracts the sub-region spanning to 350 nucleotides upstream of start codon's first nucleotide. Then, occurrences of the $\sigma 70$ promoter binding sites are searched for under constraints relative to (i) bp 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 consensus. In GenBank files, the only location annotation available is that of the start codon. Hence, for each gene, the start codon (SC) is considered a right anchor and each region upstream of SC is scanned to retrieve in priority the structured motif $[UP\ element]$ <3-18> $[-35\ box]$ <15-20> $[-10\ box]$ <10-200> $[SD]$ <2-10> $[SC]$ (described in the 5' to 3' direction), where SD denotes the Shine-Dalgarno sequence and $[box_1]$ <d_{min}-d_{max}> $[box_2]$ states the minimal and maximal bp distances allowed between the two boxes concerned. Actually, the full motif identification is performed in the 3' to 5' direction, successively considering each possible occurrence of the current box as a right anchor. In the absence of any UP element, the structured motif $[-35\ box]$ <15-20> $[-10\ box]$ <10-200> $[SD]$ <2-10> $[SC]$ is looked for.

For each genome, the consensuses used have been adapted from *E. coli* $\sigma 70$ promoter, relying on the work of Huerta and co-workers (Huerta *et al.*, 2006). These authors first identified a pair of Position-Specific Scoring Matrices (PSSMs), corresponding to the -35 and -10 boxes, associated with an interval of minimal and maximal bp distances, best describing *E. coli* $\sigma 70$ functional promoters (see latter reference, Matrix_18_15_13_2_1.5 in Figure 2). Second, for any genome other than *E. coli*, they normalized the frequencies of the pair of *E. coli* PSSMs, using the *a priori* nucleotide probabilities characterizing this genome. Then, they relied on the normalized PSSM pair, to identify a set of promoter-like sequences within each genome. Finally they computed the -10 and -35 consensuses for each genome. In our study, for each genome, the consensuses retained are the subsequences of the consensuses of Huerta and co-workers, corresponding to the locations of the canonical TTGAC and TATAAT *E. coli* consensuses. We were careful to set accordingly the optimal bp distance between the -10 and the -35 boxes. As a result, the two -10 consensus TATAAT and TAAAAT have been used respectively for 20 and 12 genomes; TTGAC, TTGAA and TTTAA were the three -35 consensuses used to scan 5, 19 ad and 8 genomes respectively. 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). The UP consensus used is that of *E. coli*, AAAWWTWTTTTNNAAAA (The genuine UP element has NN and NNN respectively as 5' and 3' termini).

For each binding site, minimal similarity is described through a maximal number of mismatches allowed. Notation **(**$err(UP), err(-35\ box), err(-10\ box)$**)** specifies the maximal numbers of mismatches allowed with regard to the UP element, the -35 box and the -10 box respectively. Given this notation, two mismatch constraints are retained in our study; they are described as follows: **(4,2,1)** and **(4,3,2)**. From now on, the two mismatch constraints **(4,2,1)** and **(4,3,2)** will be respectively denoted CI and CII. CI is more stringent than CII. Finally, four configurations will be considered in our analysis: CI, UP element required; CII, UP element required; CI, UP element optional; CII, UP element optional. The

requirement of a greatest specifi city for the -10 box compared to the -35 box is modeled after observations relative to functional $\sigma 70$ promoters.

Hereafter, we denote sp the number of strong $\sigma 70$ promoter-like sequences obtained from a given genome, when the presence of the UP element is optional. Similarly we defi ne $upsp$ when the UP element is required. From now on, we will refer to sp_{CI}, sp_{CII}, $upsp_{CI}$ and $upsp_{CII}$.

3.2 Scoring function used

In the sequel, $err(b)$ denotes the number of mismatches observed with respect to the consensus box b; d_1 denotes the bp distance observed between the -35 box and the -10 box; d_2 denotes the bp distance observed between the UP element and the -35 box. The score is calculated as follows: $score = 0.60\, err(-10\ box) + 0.40\, err(-35\ box) + t_1 + err(UP) + t_2$, 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 [6-8] else $t_2 = 3 * d_2$. When no UP element can be identifi ed, the score is merely computed as: $score = penalty + 0.60\, err(-10\ box) + 0.40\, err(-35\ box) + t_1$. The penalty value is set in order to systematically favour a candidate with an UP element within the regulatory region. This scoring function takes into account the specifi city increase of the -10 box with respect to the -35 box. The choice of the coeffi cients 0.6 et 0.4 may be debatable. The most important point remains that the ratio between these coeffi cients be consistent with the behaviour of RNA polymerase as observed through functional promoters. Besides, we wished to emphasize the UP element weight, in the case when two promoter candidates harbour an UP-like element. Therefore, we assigned a value of 1 to the coeffi cient of the UP element. Finally, BACTRANS[2] outputs 0 or 1 putative strong promoter per gene encoding a protein.

The scoring function is one of the six major differences with the approach by Dekhtyar *et al.* (V. Sakanyan, personal communication). 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 benefi t 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 specifi c "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 specifi ed by the user as minimal alignment scores. Regarding the latter point, we favoured mismatch error specifi cation 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* value (resp. *upsp* value) observed with respect to the corresponding value expected *on average* for a similarly AT-rich genome generated at random. This latter artifi cial genome is only constrained to have the same following characteristics as the prokaryotic genome considered: same total number of genes coding for proteins and same proportions of A, C, T and G nucleotides in the 350 nucleotide-long region upstream of the start codon. Due to the high bp distance allowed between the -10 box and the SD sequence (200), and the numbers of mismatches allowed, the calculation of the theoretical expected value would not be tractable. Thus, for each genome, and under the four conditions studied, we computed the minimum, maximum, mean and standard deviation for sp and $upsp$ values, over 100 such randomly generated genomes. To evaluate whether two distributions are statistically different when the latter are not of the Gaussian type and when their variances are not in the same order of magnitude, we relied on the Wilcoxon test. The H_0 hypothesis is stated

as follows: the populations from which the two distributions are taken have identical median values. This test first ranks all n_1 + n_2 values from both distributions (n_1 and n_2) combined, then sums the ranks on each distribution, ws being the smallest sum and ws' being computed as $n_1(n_1 + n_2 + 1) - ws$. If either ws or ws' is smaller than the theoretical value mentioned in Wilcoxon tables for n_1 and n_2 and an *a priori* level of significance, then hypothesis H_0 is rejected. We also computed the Z-score as the absolute difference between the number of strong promoters obs observed in the prokaryotic genome and the average number M_{emp} of promoters computed from the 100 artificial genomes, divided by the standard deviation σ_{emp} computed over these 100 latter genomes: Z-score = $\frac{|obs - M_{emp}|}{\sigma_{emp}}$, where obs is an $spCI$ value (respectively $spCII$, $upspCI$, $upspCII$ value). Again, statistical significance will be discussed, this time, with respect to several Z-score thresholds.

4 Results and discussion

4.1 Are potentially strong promoters frequent?

The 32 genomes compared belong to ten *Firmicutes*, thirteen *Proteobacteria*, three *Actinobacteria*, two *Spirochaetales*, one *Chlamydia* and three other taxa outside latter phyla. We draw the reader's attention to the case of small genomes: *B. burgdorferi* (0.91 Mbp), *C. pneumoniae* (1.22 Mbp), *M. genitalium* (0.58 Mbp), *M. pneumoniae* (0.81 Mbp), *R. prowazekii* (1.11 Mbp) and *T. pallidum nichols* (1.13 Mbp). All previous six species are either obligate intracellular pathogens, symbionts or animal commensal parasites and have undergone massive gene decay, as well as numerous genomic rearrangements. The presence of functional $\sigma 70$ promoters is disputable in these genomes. Hereafter the two *Firmicutes M. genitalium* and *M. pneumoniae* will be referred to as *Mollicutes*. Nevertheless, except for *R. prowazekii*, these genomes were investigated in the reference work of Huerta and co-workers (Huerta *et al.*, 2006). We will follow this line, taking great care regarding the discussion. The total number of genes g encoding proteins in a genome and the size of this genome are proven to be correlated over the 32 genomes studied (linear correlation coefficient: 0.93). To escape the size bias when comparing genomes, we define the percentage $p1$ ($p1 = 100 \times sp/g$). The top section of Figure 1 ((a) and (b)) depicts the variations of sp values and $p1$ percentages through genomes (also see Supplementary Data, Appendix 1). For illustration, the output files relative to *E. coli* genome are provided (see Supplementary Data, Appendix 2).

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 CI and CII. This differentiates *Firmicutes* from all other genomes studied. Nonetheless, among *Firmicutes*, the numbers of strong promoters may differ in high proportions (1 to 4 under CI and CII constraints); *S. pneumoniae* is always characterized by the lowest value whereas *B. subtilis*, *O. ihenyensis* and *C. 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 *A. aeolicus*, *T. maritima* and *B. 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). The percentage $p1$ is variable and no linear correlation can be shown to exist between sp and g. More comments are provided in Supplementary Appendix 3, including a brief report about investigating the nature of genes associated with putative strong promoters.

The high AT-richness of *Firmicutes* could justifiably be suspected to yield these high numbers of $\sigma 70$ promoter-like sequences. Indeed, we show that AT-content does not interfere much with $p1$: over the 32 genomes, the linear correlation coefficient between $p1_{CI}$ and AT-content is **0.52**; the correlation coefficient between $p1_{CII}$ and AT-content is equal to **0.30**, which was expected indeed under relaxed constraints allowing more blurred occurrences of the $\sigma 70$ promoter model. When we take into account

all bacteria but *Firmicutes*, such coeffi cients go down to 0.26 (CI) and -0.14 (CII) respectively. When the 10 AT-richest genomes are considered (*Firmicutes*), the coeffi cients are 0.27 and 0.20 respectively. Anyway, in the latter case, 10 is a borderline value regarding correlation analysis validity.

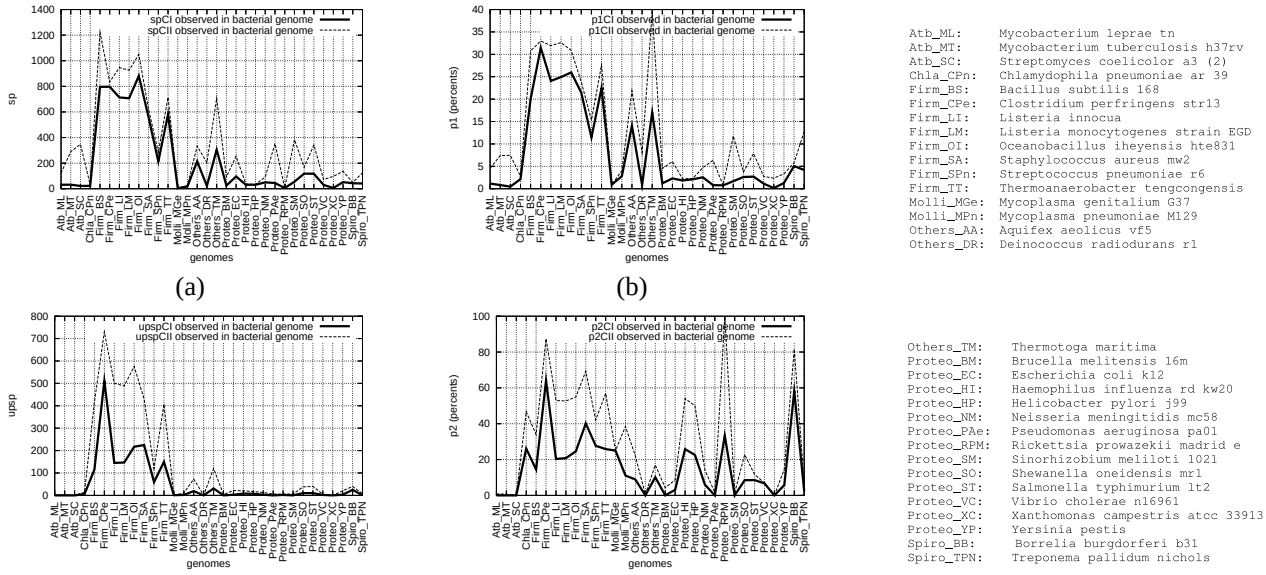
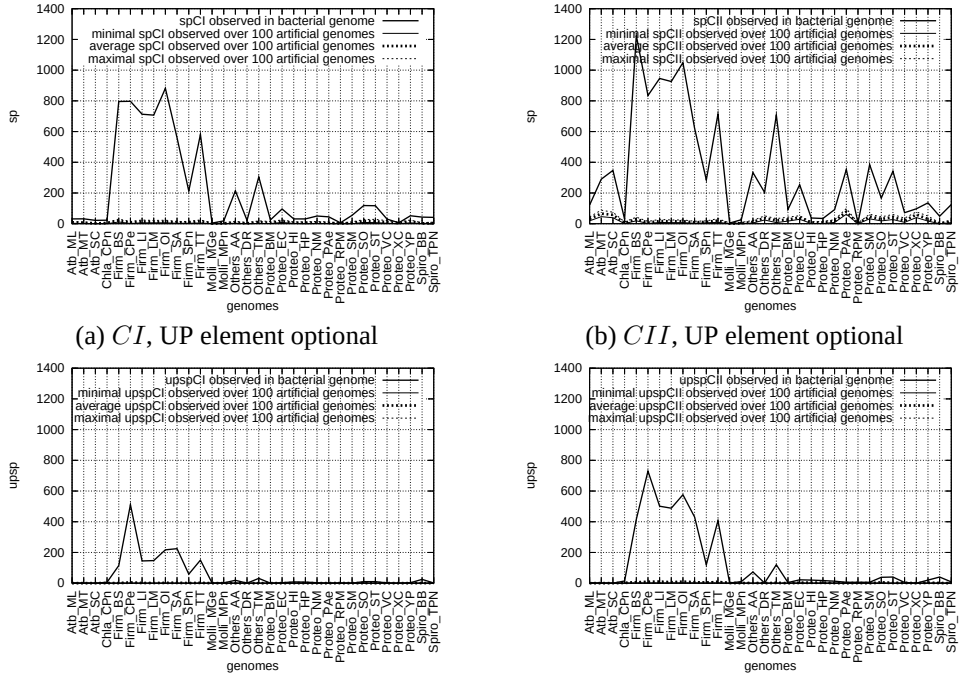


Figure 1: Frequencies of genes harbouring a putative strong promoter, under four constraint sets, in 32 prokaryotic genomes. See text, Subsection "Genome analysis upon request" for the defi nition of CI and CII constraints. (a) and (b): UP element optional; (c) and (d): UP element required. Along the 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 identifi ed 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), $p2 = 100 \times upsp/sp$).

4.2 Are potentially strong promoters harbouring an UP element frequent?

We now defi ne percentage $p2$ as follows: $p2 = 100 \times upsp/sp$. The bottom section of Figure 1 ((c) and (d)) depicts the variations of $upsp$ and $p2$ among the 32 micro-organisms, under CI and CII constraints (also see Supplementary Data, Appendix 1). The output fi les relative to *E. coli* genome are provided (see Supplementary Data, Appendix 4).

Again, detailed complements to the present paragraph may be found in Supplementary Appendix 3. We fi rst show that the differentiation between *Firmicutes* and other genomes holds, but it is more subdued for $p2$ percentage than for $p1$ percentage. Secondly, we observe that $\sigma 70$ promoter-like sequences of relatively "lesser quality" (constraint CII) are more frequently associated with an UP-like element than sequences of "better quality" (constraint set CI) (Figure 1 ((c) and (d)): the ratio $\frac{p2_{CII}}{p2_{CI}}$ is calculable for 24 genomes and its average is 2.13; the average computed for all *Firmicutes* but *Mollicutes* is 2.07.



(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 definition of sp and $upsp$, and for genome abbreviations. See text, Subsection "Genome analysis upon request" for the definition of CI and CII constraints.

Thirdly, we show that some genomes characterized by a low number of strong promoters show in contrast a high ($p2$) percentage of them harbouring an UP element, whatever the constraint (see Supplementary Appendix 3 for more details).

We calculate a correlation coefficient between $p2_{CI}$ and AT-content of **0.84** when all 32 genomes are considered; the correlation between $p2_{CII}$ and AT-content is similarly high (**0.87**). A high correlation is still observed when *Firmicutes* are not taken into account (0.82 and 0.86 respectively). In contrast with the case when no UP element was required, the 10 *Firmicutes* clearly show a correlation between $p2$ and AT-content (0.87 and 0.65 respectively). As expected, a stronger correlation is observed for $p2$ with respect to $p1$, since 7 out of the 17 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) depending on the species considered, AT-richness interferes but moderately so long as the UP element is not considered ($p1$); (ii) on the contrary, AT-content and percentage $p2$ are highly correlated. A pending question is then: does AT-richness alone entail high $upspCI$ and $upspCII$ values? To answer this question, we will in particular compare *Firmicutes*' genomes with similarly AT-rich genomes generated at random.

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

For each genome, we compare the frequency of putative strong promoters with that obtained for a similarly AT-rich "average" genome generated at random (Figure 2). For comparison purposes, a common scale is used in the four pictures of Figure 2 (The reader interested in details is referred to Supplementary Data, Appendix 5, for a magnification relative to artificial genomes' results).

We start our analysis focusing on the CI case. Figure 2 (a) (CI) shows that strong $\sigma 70$ promoter-like sequences are significantly more frequent in *Firmicutes* genomes than in corresponding artificial genomes. From now on, we distinguish the 2 *Mollicutes* from the other 8 *Firmicutes*. Given as quadruplets (minimum, maximum, **average**, standard deviation), Z-scores are as follows: *Firmicutes* except *Mollicutes*: $(81.3, 308.5, \mathbf{193.0}, 66.1)$; *Proteobacteria*: $(1.0, 32.4, \mathbf{16.0}, 9.5)$. We check that the 8 *Firmicutes*' Z-scores are above threshold 140, except for *L. monocytogenes* (81.3). Concerning the 12 large *Proteobacteria* genomes studied, 10 have their Z-scores above threshold 7, among which 6 have their Z-scores above threshold 15. In particular, the Z-score obtained for *E. coli* genome is 21.7.

When restraining our examination to the 26 species with large genomes, under condition CI, we observe that 24 genomes have their Z-scores over threshold 7, among which 15 have their Z-scores over threshold 15 and finally 10 Z-scores exceed threshold 80. For a detailed description relative to sp_{CII}, $upsp_{CI}$ and $upsp_{CII}$ values (Figure 2, (b), (c) and (d)), the reader is referred to Tables 5.1 through 5.4 in Supplementary Appendix 5. Table 5.3 focuses on *E. coli*. We recapitulate the main results and conclusions in the following paragraph.

First, we confirm that, except for the slightly more subdued case of *L. monocytogenes*, *Firmicutes* clearly show a specific trend, with Z-scores above thresholds 160, 100 and 150 respectively under CII condition (UP optional), and CI and CII conditions (UP required). Yet, under all four conditions, the Z-scores calculated for *L. monocytogenes* stay rather high (they range in interval $[69, 93]$). Secondly, relaxing the constraint from CI to CII entails no decrease of the Z-score (see Supplementary Appendix 5, Table 5.1). We conclude that relaxing the stringency is not antagonistic to motif significance. This is not a trivial result, as the opposite was expected instead. Besides, the number of putative strong promoters harbouring an UP element, observed in the average random genome under CI condition, drastically decreases down to 0 for 26 species out of 32. Under this latter condition, it is obvious that both observed and expected $upsp_{CI}$ distributions strongly differ from one another. More rigorously, and more generally, the Wilcoxon test successively performed on $p1_{CI}$, $p1_{CII}$, $p2_{CI}$ and $p2_{CII}$ allows us to conclude that the difference between observed values and values expected by chance is statistically significant under all four conditions, for the 0.05 threshold. Thus, the $\sigma 70$ promoter-like sequences retrieved in bacterial genomes are not due to mere chance. Additionally, Table 5.2 in Supplementary Appendix 5 enables evaluation of the statistical significance for each non *Firmicute* genome with respect to the Z-score thresholds 7, 15 and 80. Table 5.4 recapitulates the number of large genomes for which statistical significance is ascertained with regard to these thresholds: at least half of them under CI and CII conditions, for threshold 15, which we consider a high threshold; nearly all of them for threshold 7. Finally, since similarly AT-rich average genomes generated at random are far from yielding such high frequencies as those observed for the 8 corresponding *Firmicutes* genomes, AT-richness is clearly not the reason for the *Firmicutes* specificity.

Another lead is thoroughly examined to attempt to explain the *Firmicutes* difference. Due to the lack of space, we refer the reader to Tables 5.5 and 5.6 in Supplementary Appendix 5. We demonstrate therein that the *Firmicutes* difference is neither explained by genome size bias. Summarizing, in this section, we have characterized the statistical significances for all genomes, under four conditions of stringency, and with respect to three Z-score thresholds. We have proven the existence of a specificity for *Firmicutes* (large) genomes with regard to our definition of potentially high transcription. Moreover, this specificity is neither an artifact due to high AT-richness nor to differences in gene numbers between genomes.

4.4 Discussing the *Firmicutes* case

To explain the fact that putative strong $\sigma 70$ promoters appear much more frequently in *Firmicutes* than in other bacteria, including - paradoxically -*E. coli*, we recall that we adopted the consensus GGAGG. In *E. coli*, GGAGG is a very strong SD sequence; more frequent SDs are the submotifs GGAA, GGAG, GAGG, AGGA and AAGG (Gold, 1988; Ma *et al.*, 2002). On the other hand, ribosomes from many Gram-positive bacteria depend much more stringently upon a strong SD interaction for initiation (Roberts and Rabinowitz, 1989). For instance, in *B. subtilis* genome, most SD sequences are close to the consensus sequence AAAGGAGG (Rocha *et al.*, 1999). This, we suggest, could be the reason for the abundance of putative strong promoters in *Firmicutes* genomes. This point has been investigated further. We show that the percentage p_{bact} of genes associated with an optimal SD sequence ranges between 2.21% and 39.8% for the 26 large genomes. Immediately behind *T. maritima*, which shows the highest ratio, the 8 large *Firmicutes* genomes rank first with respect to this p_{bact} ratio ([15.3%, 32.6%]). The percentages p_{rand} expected for similarly AT-rich genomes generated at random have been calculated. The calculus is described in Supplementary Appendix 6. The p_{bact} and p_{rand} distributions are proven statistically different through a Wilcoxon test (threshold 0.05). Furthermore, the correlation coefficient between p_{rand} and AT-richness is -0.97, over the 32 artificial genomes. This high negative value was expected since the optimal SD sequence is enriched with four G nucleotides. In contrast, the correlation coefficient between p_{bact} and AT-richness is low when computed over the 32 bacterial genomes (0.22). This point argues in favour of the biological significance of such GGAGG sequences in the close neighbourhood of start codons. Moreover, regarding this criterion, the Wilcoxon test also ascertains the statistical significance of the difference between the 8 *Firmicutes* and the 18 other species with large genomes. This difference is reflected by the Z-scores. Z-scores range in interval [3.2, 363.9] when all genomes are considered (mean: 86.9, standard deviation: 103.1). The Z-scores calculated for the 8 large *Firmicutes* genomes range between 86.8 (*S. pneumoniae*) and 363.9 (*C. perfringens*). When all large genomes but *Firmicutes*' are considered, the mean and standard deviation are respectively equal to 41.8 and 40.0. Outside the *Firmicutes* taxon, *T. maritima* and *A. aeolicus* are the only two bacteria showing as outstanding Z-scores as *Firmicutes* (respectively 168.7 and 106.2). Again, we emphasize that both previous genomes are also characterized with high AT percentages (54.6% and 57.6%), which confirms a bias for the presence of optimal SD sequences in some genomes.

Anyway, such bias exists for all genomes. For example, in the light of the previous explanation, we now explain the scarcity of putative strong promoters associated with optimal SD sequences, in *E. coli*, through the low p_{bact} percentage of 6.2% observed. Though, the percentage expected is 0.9%. The bias measured through the Z-score is 37.9. Therefore, this point suggests that even in *E. coli*, hazard would only contribute for 15% ($\frac{0.9}{6.2}$) to yield false positive optimal SD sequences. Finally, considering the criteria retained in our analysis (high intrinsic transcription potentiality combined with strong SD interaction), we conclude that *Firmicutes* would appear as genomes more favoured by nature, especially with respect to other similarly AT-rich genomes.

4.5 Putative strong promoters *versus* experimentally verified functional promoters in *Escherichia coli* genome

In vivo, activation by various factors is ascertained to compensate for promoter weakness. However, it is not known whether some functional promoters might also be intrinsic strong promoters. So far, data compilations relative to experimentally verified functional promoters are only available for *E. coli* genome, through two repositories, RegulonDB and PromEC (Salgado *et al.*, 2006; Hershberg *et al.*, 2001). Therefore, we could compare the putative strong promoters identified by BACTRANS[2] in *E. coli* genome with known *E. coli* functional promoters.

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. 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 condition, 96 putative strong promoters are identified by BACTRANS^2; 12 out of the corresponding 96 genes are referred to by at least one of the two databases aforementioned. Under CII condition, 20 out of the 254 genes identified with strong promoters are cited in at least one database. For more details and an explanation of the results obtained, the reader is referred to Supplementary Appendix 7. We confirm that in *E. coli* genome, according to BACTRANS^2 scoring function, when a functional $\sigma 70$ promoter is known for a gene, it is intrinsically weaker than the putative strong promoter identified by BACTRANS^2 if the latter exists under CI or CII conditions.

4.6 Experimental verification of putative strong promoters identified in *Thermotoga maritima* genome

The hyperthermophilic model *Thermotoga maritima* has been intensively studied (Morin *et al.*, 2003; Braun *et al.*, 2006). In the context of a former study, the activity of thirteen putative strong promoters harbouring an UP element has been measured in *E. coli* cell free extracts (Sakanyan *et al.*, 2003). The present work thereby benefits from these experimentations. The protocol used is described in Supplementary Appendix 8. Seven putative strong promoters harbouring an UP element identified by BACTRANS^2 were thus tested. Four were identified under the most constrained condition CI ($TM1016$, $TM0373$, $TM0477$, $TM1667$). The other three were identified under CII condition ($TM0032$, $TM1429$, $TM1780$). 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. $TM0477$ has been shown to be twice as strong as others regarding protein yield. Therefore six promoters among the seven tested really favour high expression in *E. coli* cell free extracts.

Conclusion

Our work contributes to shedding new light on potentially high ORF expression in prokaryotic genomes, focusing on potentially high transcription combined with the presence of an optimal Shine-Dalgarno sequence. Our approach also puts emphasis on transcription initiation potentially enhanced through UP-like elements. In itself, this latter feature introduces originality with respect to other genome-comparative studies devoted to bacterial promoters. Our analysis clearly departs from other works, since it considers four different conditions of stringency and discusses in each framework the statistical significance of the presence of $\sigma 70$ promoter-like sequences. Under all four conditions, we identified the species showing statistically significant differences between the bacterial genome and an average similarly AT-rich genome generated at random. In particular, *Firmicutes* would appear as genomes more favoured by nature with respect to other genomes, including the cases when an UP-like element is required. A rigorous discussion allowed us to dismiss AT-richness and genome size bias as determining factors to explain the *Firmicutes* specificity. To explain this specificity, the hypothesis of the abundance of optimal SD sequences in *Firmicutes*' large genomes has been checked with success. Besides, so far, the UP element has been identified by experimentation in four genomes. Thus our comparative study also brings novel knowledge about the statistical significance of the presence of putative $\sigma 70$ promoters enhanced with an UP-like element, in various genomes.

The generic software platform BACTRANS^2 currently provides such putative strong promoters for 45 genomes. These data may be of interest to select a subset of promoters for experimental characterization and possible further use in biotechnological applications. In this latter field, inserting in cellular or cell-free expression systems regulatory regions including promoters enhanced with an UP element and an optimal SD sequence may be advocated, instead of inserting artificial binding sites in a synthetic se-

quence. A more thorough study of high translation potentiality related to high transcription potentiality in prokaryotic genomes is attractive and is currently under work. Finally, BACTRANS^2's genericity allows the user to analyse genomes with respect to any other super-motif consisting of 3 or 4 boxes.

Acknowledgements

S. 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). The first author is thankful to V. Sakanyan for valuable comments and critically reading the manuscript. Thanks are also due to J. Bourdon for insightful discussions.

References

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.

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.

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

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.

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

Ellinger, T., Behnke, D., Bujard, H. and Gralla, J.D. (1994) Stalling of *Escherichia coli* RNA polymerase in the +6 to +12 region *in vivo* is associated with tight binding to consensus promoter elements. *J. Mol. Biol.*, **239**(4), 455–465.

Eskin, E., Gelfand, M. and Pevzner, P. (2003) Genome-wide analysis of bacterial promoter regions. *Pacific symposium on Biocomputing*, **8**, 29-40.

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.

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.

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.

Gold, L. (1988) Posttranscriptional regulatory mechanisms in *Escherichia coli*. *Ann. Rev. Biochem.*, **57**, 199-233.

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.

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

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.

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

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

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

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.

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

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).

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

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

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

Ma, J., Campbell, A. and Karlin, S. (2002) Correlation between Shine-Dalgarno sequence and gene features such as predicted expression levels and operon structure. *J. Bacteriol.*, **184**, 5733-5745.

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

Mittenhuber, G. (2002) An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. *J. Mol. Microbiol. Biotechnol.*, **4**, 77-91.

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.

Morrison, D.A. and Jaurin, B. (1990) *Streptococcus pneumoniae* possesses canonical *Escherichia coli* (sigma 70) promoters. *Mol. Microbiol.*, Jul; **4**(7):1143-52.

Osada, Y., Saito, R. and Tomita, M. (1999) Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics*, **15**, 578-581.

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

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

Roberts, M.W. and Rabinowitz, J.C. (1989) The effect of *Escherichia coli* ribosomal protein S1 on the translational specificity of bacterial ribosomes. *J. Biol. Chem.*, **264**, (4), Feb, 2228-2235.

Rocha, E.P.C, Danchin, A. and Viari, A. (1999) Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis, *Nucleic Acids Res.*, **27**(3), 3567-3576.

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.

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

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.

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.

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.

Wosten, M. M. (1998) Eubacterial sigma-factors. *FEMS Microbiol. Rev.*, **22**, 127-150.

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

Christine Sinoquet†, Sylvain Demey†, Frédérique Braun‡

Abstract

The platform BACTRANS^2 has been designed to help select putative strong σ70-like promoter candidates in prokaryotic genomes. It was run to investigate the importance of σ70-like potentially high transcription in bacteria other than *Escherichia coli*. We performed a genome-comparative analysis of high ORF expression potentialities over 32 prokaryotic genomes. Besides, we put an emphasis on transcription strength reinforcement through the UP element presence and on translation potentiality enhancement through an optimal Shine-Dalgarno sequence.

We compared frequencies of putative strong promoters between various genomes. We show that in the AT-rich *Firmicutes*' genomes, frequencies of potentially strong σ70-like promoters are exceptionally high. Besides, though they contain a low number of strong promoters, some genomes may show a high proportion of promoters harbouring an UP element. Putative strong promoters of lesser quality are more frequently associated with an UP element than putative strong promoters of better quality. A meaningful difference is statistically ascertained when comparing frequencies in bacterial genomes with frequencies in similarly AT-rich genomes generated at random; the difference is the highest for *Firmicutes*. Comparing some *Firmicutes* genomes with similarly AT-rich *Proteobacteria* genomes, we confirm the *Firmicutes* specificity. We show that this specificity is neither explained by AT-bias nor genome size bias but originates in the abundance of optimal Shine-Dalgarno sequences, a typical and significant feature of *Firmicutes* more thoroughly analysed in our study.

The generic software platform BacTrans2 currently provides such putative strong promoters for 45 genomes. These data may be of interest to select a subset of promoters for experimental characterization and possible further use in biotechnological applications. Finally, BacTrans2's genericity allows the user to analyse genomes with respect to any other super-motif consisting of 3 or 4 boxes.

To our knowledge, this work is the very first genome-comparative study thoroughly analysing the significance of various potentially strong sigma 70-like promoter models, including models harbouring the UP element enhancer.

LINA, Universit´e de Nantes
2, rue de la Houssini`ere
B.P. 92208 — F-44322 NANTES CEDEX 3