Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes

Michael Dekhtyar* 1, Amelie Morin2 and Vehary Sakanyan* 2,3

Address: 1University of Tver, 33 Jelyabova, 170000 Tver, Russia, 2Laboratoire de Biotechnologie, UMR CNRS 6204, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France and 3ProtNeteomix, 2 rue de la Houssinière, 44322 Nantes, France

Email: Michael Dekhtyar* - Michael.Dekhtyar@tversu.ru; Amelie Morin - Amelie.Morin@cancer.org.uk; Vehary Sakanyan* - Vehary.Sakanyan@univ-nantes.fr

* Corresponding authors

Abstract

Background: Bacterial promoters, which increase the efficiency of gene expression, differ from other promoters by several characteristics. This difference, not yet widely exploited in bioinformatics, looks promising for the development of relevant computational tools to search for strong promoters in bacterial genomes.

Results: We describe a new triad pattern algorithm that predicts strong promoter candidates in annotated bacterial genomes by matching specific patterns for the group I σ70 factors of Escherichia coli RNA polymerase. It detects promoter-specific motifs by consecutively matching three patterns, consisting of an UP-element, required for interaction with the σ subunit, and then optimally-separated patterns of -35 and -10 boxes, required for interaction with the σ70 subunit of RNA polymerase. Analysis of 43 bacterial genomes revealed that the frequency of candidate sequences depends on the A+T content of the DNA under examination. The accuracy of in silico prediction was experimentally validated for the genome of a hyperthermophilic bacterium, Thermotoga maritima, by applying a cell-free expression assay using the predicted strong promoters. In this organism, the strong promoters govern genes for translation, energy metabolism, transport, cell movement, and other as-yet unidentified functions.

Conclusion: The triad pattern algorithm developed for predicting strong bacterial promoters is well suited for analyzing bacterial genomes with an A+T content of less than 62%. This computational tool opens new prospects for investigating global gene expression, and individual strong promoters in bacteria of medical and/or economic significance.

Background

Efficient promoter recognition is crucial in the synthesis of the gene-encoded products required by bacteria to allow them to grow rapidly and to adapt to different environmental conditions. The general architecture and protein-DNA interaction interfaces appear to be conserved in RNA polymerases of different bacteria, to judge by a comparison of the resolved structures of the multi-subunit protein or its subunits [1]. This structural information suggests that the principles of DNA recognition by RNA polymerases are universal, and this constitutes a basis for in silico prediction of promoters that are recognized by families of sigma factors. Research in bioinformatics has developed approximate matching methods to detect con-
served sequences in nucleic acids [2-5], including promoter-specific sequences that are invaluable in helping to elucidate the overall organization of transcriptional signals and regulatory circuits in evolutionarily distant bacteria [6-14]. Most promoter prediction programs so far proposed use statistical or motif-based methods, and take into consideration what is known about experimentally defined promoter architectures, and extract conserved sequences from the genomes under analysis. Attempts have been made to improve promoter prediction by introducing statistical mechanical methods to measure the stress-induced destabilization or bendability of the duplex DNA region located upstream of the transcription initiation site required for the local dissociation of strands to start mRNA synthesis [15-17]. The steady increase in the number of sequenced bacterial genomes of medical and economic significance means that there is an increasing need for computational tools to predict promoters, especially those responsible for high-level gene expression in organisms, of which there has been little experimental investigation.

Many housekeeping genes in Escherichia coli are transcribed from promoters possessing the recognition elements referred to as -35 and -10 sites (boxes), for which the TTGACA and TATAAT consensi, respectively, have been identified by compiling characterized RNA polymerase-binding regions in the DNA [18,19]. The -35 and -10 sites, which are separated from each other by a 15–20-bp spacer [20], are specifically recognized by Eσ70 RNA polymerase, an RNA polymerase holoenzyme bearing the group I σ70 factor [21]. Experimental data have shown that high transcription rates of genes correlate with the level of conservation of three promoter parameters, with the consensus -35 and -10 hexanucleotide boxes, and with the 17 ± 1-bp spacer separating them [22]. This fact has been widely used to construct vectors for protein overexpression in bacterial cells [23].

However, the strength of strong promoters also depends on a fourth parameter, an AT-rich UP element 17–20 bp length, which is located just upstream of the -35 site, and which is recognized by the α subunit of Eσ70 RNA polymerase, which was first discovered for ribosomal RNA promoters [24]. The C-terminal domain of this subunit binds both to the UP element and to transcription regulation proteins, whereas the N-terminal domain makes contact with other subunits during the assembly of RNA polymerase [25]. A 17-bp consensus 5′AAAWWT-WTTITNNNAAA (where W is A or T, and N can be any of the four bases) has been identified for the UP element by analyzing the patterns, selected by the SELEX method, which mediate increases of between 10- and 300-fold in gene expression in E. coli cells [26-28]. Two preferred sub-sites have been identified within the UP element. They are centered approximately at the -42 and the -52 positions respectively, and appear to be specifically recognized by one or two monomers of a dimeric α subunit in the RNA polymerase.

It is noteworthy, that a virtual analysis of patterns located upstream from the consensus -35 had since long time suggested their functional significance [3]. The sequences reminiscent of UP element have been detected in the E. coli genome by the algorithms PWM [29] and PlatPram [30]. A software-based (GCG, version 9.0) dissection of regions located upstream of the E. coli promoters had made it possible to detect putative promoters with ≤ 4 mismatches in the full UP element consensus [28]. Several UP elements have also been visually identified, and characterized by their ability to direct high level gene expression in vivo or in vitro in Bacillus subtilis [31], Geobacillus (formerly Bacillus) stearothermophilus [32] and Vibrio natriegens [33]. Recently, a comparative analysis of Eσ70-specific promoter and non-promoter regions indicated that upstream regions of E. coli ribosomal and T4 phage early promoters possess electrostatic elements that could be responsible for modulating promoter activities due to ADP-ribosylation of RNA polymerase α subunit [34]. However, no specific algorithms have yet been proposed to detect strong promoters in bacterial genomes, and so this remains an important task for genomic and proteomic research in microbiology.

In this study, we have developed a triad pattern algorithm that detects strong promoter candidates composed of a UP-element, and two consensi, -35 and -10 boxes, which are optimally distanced from each other. All four parameters are required for efficient DNA recognition, and the initiation of mRNA synthesis by an Eσ70-like RNA polymerase. The data presented indicate that the frequency of strong promoters is a function of the A+T content of the corresponding genomes. The proposed prediction program is flexible, and can be modified by users to modulate the search stringency criteria depending on the promoter features of the genome under analysis. The accuracy of detection has been experimentally validated for putative strong promoters predicted in a hyperthermophilic bacterium Thermotoga maritima.

Implementation

Overview of the approach

The promoter activity in cells is determined by regulatory proteins (activators and repressors) that can recognize overlapping sequences specific for Eσ70 RNA polymerase sites, and thereby mask the true promoter strength. In addition, almost 20% of E. coli RNA polymerase Eσ70-specific promoters possess an extended -10 sequence that might compensate for the absence of a clear -35 site [35]. Different prediction programs based on statistical and
motif-searching approaches have been developed to detect a variety of binding sites in DNA, and both position-specific weight matrices [36] and hidden Markov models [37] have been used to improve the accuracy of the prediction of promoter sequences in bacterial genomes [38-40]. These programs usually detect hexanucleotide dyad patterns of RNA polymerase-promoter binding sites, such as -35 and -10 boxes, and none of them is free of false-positives, which correspond to similar, non-promoter sequences in bacterial genomes [for a review, see [41]].

In this study, we exploited the strengths of the "triad pattern" approach to develop an algorithm able to detect strong promoters by matching three nucleotide sequences recognized by the σ70 and α subunits of bacterial RNA polymerase. Theoretically, the presence of a UP element may not be essential for relatively strong promoter activity if two -35 and -10 boxes are well conserved and optimally distant. Similarly, the presence of a well conserved UP element may compensate for a poor -35 box in some promoters. However, it seems very likely that the strongest promoters probably possess all three essential sequences. The specific interaction between the UP element and the α subunit significantly amplifies the association of RNA polymerase with promoter DNA [27,28]. Therefore, to improve the filter to exclude possible false-positive due to short hexanucleotide similar sequences scattered throughout the genome, our algorithm starts by first matching the UP element, and only then identifying the -35 and -10 boxes located further downstream.

**Design of the triad pattern algorithm**

We designed an algorithm able to detect the triad nucleotide patterns in bacterial genomes. The core of the algorithm is the FIND_TRIAD procedure, which given an input nucleotide string, $s$, returns the substring $s'$ of $s$, which is the best approximation of a given triad pattern of the form $(pat(1), L1)-(l1, l2)-(pat(2), L2)-(d1, d2)-(pat(3), L3)$, where each $pat(i)$, $i = 1, 2, 3$, is a nucleotide string, $L_i$ is its length, $l1$ and $l2$ are the minimum and maximum distances respectively between the first and the second patterns, and $d1$ and $d2$ are the minimum and maximum distances respectively between the second and the third patterns. To avoid making a "bad" approximation, three scores $S_{c1}$, $S_{c2}$ and $S_{c3}$ are used as input parameters for the procedure. The resulting substring, $s'$, can then be represented as $(spat(1), Ls1)-(ls1)-(spat(2), Ls2)-(ls2)-(spat(3), Ls3)$, where each $spat(i)$, $i = 1, 2, 3$, is a substring of $s$ aligned to $pat(i)$. $Ls1$ is its length, $ls1$ is the distance between $spat(1)$ and $pat(1)$, $Ls2$ and $ls2$ is the distance between $pat(2)$ and $spat(2)$, and $Ls3$ is not less than $Sci$ (or the number of "mismatches" does not exceed $(Li - Sci)$):

(1) for each $i = 1, 2, 3$ the similarity score (weight) $Wi$ of the match or alignment of $pat(i)$ and $spat(i)$ is not less than $Sci$ (or the number of "mismatches" does not exceed $(Li - Sci)$):

(2) $(l1 \leq ls1 \leq l2)$ and $(d1 \leq ls2 \leq d2)$.

For each of the three patterns, one can either forbid insertions/deletions or allow them. In the former case, $Lsi = Li$ and the weight = $Wi$ are computed as the sum of matching pairwise symbols, whereas in the latter case, the difference $[Lsi - Li]$ between $spat(i)$ and $pat(i)$ is bounded by a value $Ri$ for the permissible deletions/insertions (gaps), an optimum alignment, and its weight, $Wi$, are computed by the standard dynamic programming method for global string alignment [42]. In both cases, a symbol scoring matrix $Mi(x,j)$ is used to define the weight of the symbol $x$ in the position $j$, $1 \leq j \leq Lsi$, of $spat(i)$. If symbol $x$ occurs in position $j$ of $pat(i)$, then $Mi(x,j) = 1$, otherwise $Mi(x,j) \leq 1$. To choose the best approximation of the triad pattern from substrings satisfying conditions (i) and (ii), FIND_TRIAD uses a total score function with the form:

$$
tot_{sc} = C1*nsc1(L1,W1)+D12*nsc_{dist12}(11,l2,ls1) + C2*nsc2(L2,W2) + D23*nsc_{dist23}(d1,d2,ls2) + C3*nsc3(L3,W3),$$

where $nsci(Li,Wi)$, $i = 1, 2, 3$, are normalized scores of matching (alignments) of $pat(i)$ and $spat(i)$, $0 \leq nsci(Li,Wi) \leq 1$, and $nsc_{dist12}(11,l2,ls1)$ and $nsc_{dist23}(d1,d2,ls2)$ are the normalized scores of the distances between the first and the second, and the second and the third patterns, respectively, and $0 \leq nsc_{dist12}(11,l2,ls1), nsc_{dist23}(d1,d2,ls2) \leq 1$. The linear coefficients $C1$, $C2$, $C3$, $D12$, and $D13$ are chosen so that their sum is equal to 1. They indicate the relative importance of the corresponding sub-patterns of the triads; and the distances between them. So, the best value of $tot_{sc}$ is 1.

**Application of the algorithm to searching for strong promoter candidates**

Here we describe the main parameters of the FIND_TRIAD procedure used to detect strong promoter candidates in bacterial genomes. In this study, a bacterial promoter is assumed to be a nucleotide sequence, located upstream from genes encoding proteins, tRNAs or rRNAs that could be recognized by an RNA polymerase holoenzyme containing a major σ factor (using E. coli Ecσ70 RNAP as the reference). The triad patterns defined for strong promoter candidates include three specific nucleotide subregions: (i), a UP element, which is a 17-nt prefix of the strong promoter, and has the following consensus pattern: $pat(1) = P_{up} = aaaaWWTWtttNNNaa$; (ii) the -35 site, which is located downstream of the UP element, and has
the pattern \( \text{pat}(2) = P_{35} = tcttgaca \) (underlining indicates a commonly used consensus for group 1 \( \sigma^{70} \) factors; however, the \( \sigma_4 \) domain of these factors appears to be in contact with 9 nucleotides in the region extending from -30 to -38 [43,44]); (iii) the -10 site, which is located downstream of the -35 site, and has the pattern \( \text{pat}(3) = P_{10} = tataat \) (this site is highly conserved). We used the following boundaries for the distances between the sub-regions: \( l_1 = 0, l_2 = 5 \) (these boundaries are standard for the distance between the -35 site and the -10 site). To search for the first pattern \( \text{pat}(1) \) of the UP-element, the simple matching algorithm was chosen with an \( a \) and \( t \) mismatch score of 0.5. The reason is that in the full UP-element consensus and the consensuses of two of its subsides – distal and proximal – in some places do not distinguish between \( a \) and \( t \). We assumed that the consensus for the -35 site of length 9 is less conserved than that of the -10 site, and so in order to detect the second pattern \( \text{pat}(2) \) of the -35 site we used a dynamic programming algorithm to search for optimal alignment, with boundaries for the number of permissible deletions/insertions of \( R_2 = 2 \). For the most of -35 sites, which were detected by algorithm, no insertions/deletions were applied. However, this scoring system allowed us to identify some stronger promoter candidates. Thus, the insertion of C between two AA in the sequence TCTTGAAT of TM1016, increases the score of a putative site.

To define the total score function, \( \text{tot}_s c \) (formula 1), we chose the following normalized scores for the three patterns and for the distance between the -35 site and -10 site (no information was available about the best values for the distance between the UP element and the -35 site):

\[
\text{nsc}_1(17,W1) = \text{nsc}_{up} = 1 - (17 - W1)/20, \quad \text{(2)}
\]
\[
\text{nsc}_2(9,W2) = \text{nsc}_{35} = 1 - (9 - W2)/10, \quad \text{(3)}
\]
\[
\text{nsc}_3(6,W3) = \text{nsc}_{10} = 1 - (6 - W3)/2, \quad \text{(4)}
\]

and the values of the normalized distance score, \( \text{nsc} \_\text{dist}23(14,20,ls2) = \text{nsc} \_\text{dist} \), are defined as follows:

\[\text{nsc} \_\text{dist} = \frac{\text{bs2} \_\text{distance between the -35 and -10 sites}, \text{nt} \_1 \_1 \_1 17 \_1 16 \_1 18 \_1 15 \_1 19 \_1 14 \_1 20}{1 \_1 0.95 \_1 0.85 \_1 0.7}\]

We also chose linear coefficients \( C_1 = 0.3, C_2 = C_3 = 0.25, D_{12} = 0, \) and \( D_{23} = 0.2 \). These coefficients indicate the relative importance of corresponding sub-regions for evaluating the total score of a candidate sequence. They were chosen empirically, after preliminary tests with several annotated genomes, assuming a higher significance of the UP element, equal significance of the -10 and -35 boxes, and lower significance of the distance between them. In this application, the value \( D_{12} = 0 \) means that we ignore the variations of the distance between a putative UP element and -35 box because \( a \_\text{priori} \) it is not known what value is the best in the interval 0–5 nt.

Formulas 2, 3 and 4 reflect the lack of exact matching for the different sub-regions. If the -10 box is highly conserved and is essential for initiation of transcription [22], then the penalty for its mismatches is higher than for those of the other parameters. For example, for 2 mismatches, the penalty is \( (6 - 4)^2/10 = 0.4 \) for the -10 site, whereas it is \( (9 - 7)/10 = 0.2 \) for the -35 site, and \( (17 - 15)/20 = 0.1 \) for the UP element. The choice of the normalized score functions in equations 2, 3 and 4 is based on empirical observations, and on common sense, and may seem to be arbitrary. We want to stress that, in fact, the total score function \( \text{tot}_s c \) also has a further role: it does not significantly change the set of the best candidates identified by the algorithm. This set is defined by the three score bounds \( \text{sc1} = \text{sc} \_up \) for UP element, \( \text{sc2} = \text{sc} \_35 \) for -35 site, and \( \text{sc3} = \text{sc} \_10 \) for -10 site. The total score affects only the ordering of these candidates amongst themselves.

The general scheme of the algorithm is as follows. It has the following input: (i) the name of a genome file in GenBank format; (ii) three parameters of scores: \( \text{sc} \_up, \text{sc} \_35 \) and \( \text{sc} \_10 \), determining the minimum acceptable similarity between candidate sequences of the UP element, the -35 box, and the -10 box, respectively, and the \( E. \_\text{coli} \) consensus patterns. For each gene in the genome input file that is not inside an operon, the algorithm runs in two steps:

(i) it extracts a 300-bp DNA region, \( s \), upstream of the annotated coding sequences for tRNA, tRNA or proteins (we limited the search to 300 bp, since most \( E. \_\text{coli} \) promoters fall within this length inter-gene space [41,45]);

(ii) then it uses the FIND_TRIAD procedure to identify the best strong promoter candidate within \( s \) that satisfies conditions (1) and (2) above. If such a candidate is found, it is added to the output list of strong promoters.

We recommend to read attentively the "ReadMe" information [see Additional file 1] before to start proceeding the "strong_promoters.doc" software [see Additional file 2]. The algorithm is implemented by a program that produces the results in two forms: (i) a Text-format table which lists all strong promoter candidates in a genome, and provides additional information about the operon organization of genes located downstream (for example, see Fig. 1); (ii) a Word-format table which lists strong promoter candidate sequences. A 20-nt sequence preceding a
LOCUS AE000512 1860725 bp DNA circular BCT 28-DEC-2005
DEFINITION Thermotoga maritima MSB8, complete genome.

Total number of genes on the 'direct' strand: 1056
Total number of genes on the 'complement' strand: 870

UP-ELEMENTS for genes on the 'direct' strand

 unos_eventually_used_to_create_the_algorithm_for_example_a_weakly_conserved_-35_or_-10_box

Figure 1
Text-format presentation of strong promoter candidates.

possible initiation codon of each ORF is also included in the annotation, as this could be useful for the visual examination of the translation signals of the corresponding genes. Lastly, the user can select a convenient score for each sequence-specific motif taking into consideration the promoter features of the annotated genome if they differ from the E. coli-specific patterns used to create the algorithm (for example, a weakly conserved -35 or -10 box).

Methods
Construction of recombinant linear DNAs
Putative promoter regions in the T. maritima genome, identified by the algorithm described above, were amplified by PCR using appropriate oligonucleotide primers connected to the previously-described G. stearothermophilus argC gene [46]. This reporter gene encodes N-acetyl glutamylphosphate reductase, a thermostable and soluble protein that is easily detectable after exposing E. coli cleared lysates to 65°C. In order to increase protein yield, the ribosome-binding site of G. stearothermophilus argC was modified to the sequence GAGGAGGGAACAT-AIG (the modified Shine-Dalgarno site and the initiation codon are underlined), and the distance between the -10 promoter site and the Shine-Dalgarno site was shortened to 15 bp (Fig. 2). The DNA fragment carrying the argC gene was connected to T. maritima or control promoters by two consecutive PCR steps, as described previously [47]. The quantity and quality of the amplified DNAs were determined with a 2100 Bioanalyzer (Agilent Technologies).

Two well-characterized strong promoters, Ptac and PargC, were used as references to compare the strength of the putative promoters of T. maritima. The strong promoter Ptac contains an AT-rich nucleotide sequence upstream of a -35 site [48], which has no defined UP element; it was obtained from the vector pBTac2 (purchased from Boehringer Mannheim). PargC, a strong promoter of G. stearothermophilus, contains the UP element, as demonstrated
both in vivo and in vitro, and was amplified from the plasmid pHAV2 [32].

**Cell-free protein synthesis**

PCR-generated linear DNA fragments carrying a promoter region fused to the argC reporter gene were used to evaluate the promoter strength in a coupled transcription-translation system, as described previously [49]. The cell-free extracts were prepared from the E. coli strain BL21 (DE3) Star recBCD (our laboratory construction) as described by Pratt [50]. Protein synthesis was carried in the presence of pyruvate oxidase to generate ATP [51]. Typically, 50 ng of PCR-amplified DNA was added to a pre-mix containing all necessary compounds and 10 μCi of [35S]-L-methionine (specific activity 1000 Ci/mmol, 37 TBq/mmol, Amersham-Pharmacia Biotech), and E. coli S30 cell-free extracts. The reaction mixture was incubated at 37°C for 90 min, and heated to 65°C for 10 min. After centrifuging, the supernatant was precipitated with acetone, and then protein samples were separated by SDS-PAGE and bound to 3 MM paper. The ArgC protein synthesized in vitro was quantified by counting the radioactivity of the corresponding band with a PhosphorImager 445 SI (Molecular Dynamics).

The bacterial genome sequences were extracted from available data banks. The logo of *T. maritima* promoter consensus sequences was generated at the WebLogo site as described in [52,53].

**Results**

**The number of strong promoters reflects the A+T content of bacterial genomes**

In our algorithm, 26 of the 32 symbols used to evaluate matches in the three promoter-specific patterns, namely in the UP element and the -35 and -10 boxes, are a and t. One could expect the number of genes transcribed from potential strong promoters to depend on the A+T content of a given genome. To find out whether this is indeed the case, we compared the frequency of candidates in 300-bp regions located upstream of genes of annotated bacterial genomes and in random sequences of the same regions generated by computing. First, we calculated the (A+T)% in all 300-bp regions preceding each gene or operon in the annotated genomes (Table 1). The A+T content in these DNA regions was found to be slightly higher than that of the entire genomes of almost all bacteria that have been analyzed. Next, we generated 10,000 random sequences with the same A+T content for all the 300-bp regions of each genome. The algorithm was applied to detect strong promoter candidates in the 300-bp real genomic and random-generated regions of 43 bacterial genomes.

We tested different matching stringencies and empirically found that the score parameters $sIP = 13$, $s35 = 5.5$ and $s10 = 4.5$ satisfied the criteria required for scaled comparative analysis without grossly exaggerating the number of candidate sequences identified in the various genomes. This analysis revealed that the real genomes with an A+T content of less than 50% contained many more potential strong promoters than their simulated counterparts (see

---

**Figure 2**

Diagram of the fusion DNA constructs used to express the *G. stearothermophilus* argC-reporter gene from putative strong promoters of *T. maritima* in a cell-free system. The argC gene was amplified with forward 5’-GGAGGGGGAACATATGATGAA and reverse 5’-GGACCACCGCGCTACTGCCG primers from pHAV2 [32] by conserving a 112-bp downstream region carrying transcriptional terminators of the vector DNA.
Table 1: A+T content of bacterial genomes and 300-bp regions located upstream of genes and the percentage of strong promoter candidates predicted in 300-bp real genomic and random-generated regions of the same content.

| N° | Genome                                    | (A+T)% of                       | % of candidates in |
|----|-------------------------------------------|--------------------------------|-------------------|
|    |                                           | Bacterial genomes*            | 300-bp genomic regions | 300-bp random sequences |
| 1  | Deinococcus radiodurans R1 (AE000513)     | 32.99                         | 34.19              | 0.19 | 0 |
| 2  | Caulobacter crescentus (AE005673)         | 32.77                         | 34.40              | 0.05 | 0 |
| 3  | Ralstonia solanacearum GMII1000 (AL646052)| 34.51                         | 34.50              | 0.20 | 0 |
| 4  | Pseudomonas aeruginosa PA01 (AE004091)    | 33.44                         | 35.38              | 0.27 | 0 |
| 5  | Xanthomonas campestris pv. campestris     | 33.15                         | 35.64              | 0.05 | 0 |
| 6  | Xanthomonas axonopodis pv. citri 306      | 35.23                         | 36.02              | 0.05 | 0 |
| 7  | Mesorhizobium loti (NC002678)             | 37.25                         | 39.09              | 0.13 | 0 |
| 8  | Sinorhizobium meliloti (AL591688)         | 37.27                         | 39.66              | 0.24 | 0 |
| 9  | Mycobacterium leprae TN (AE008922)        | 34.93                         | 35.64              | 0.05 | 0 |
| 10 | Mycobacterium tuberculosis (AL123456)     | 34.39                         | 35.69              | 0.20 | 0 |
| 11 | Xanthomonas axonopodis pv. citri 306      | 35.23                         | 36.02              | 0.05 | 0 |
| 12 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 13 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 14 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 15 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 16 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 17 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 18 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 19 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 20 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 21 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 22 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 23 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 24 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 25 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 26 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 27 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 28 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 29 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 30 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 31 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 32 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 33 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 34 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 35 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 36 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 37 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 38 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 39 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 40 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 41 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 42 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |
| 43 | Mycobacterium tuberculosis (AL123456)     | 34.93                         | 35.64              | 0.05 | 0 |

* A+T content of bacterial genomes was calculated from corresponding genomic DNA sequences available in gene banks.
** Similar values were found for the E. coli K12 genome.
The percentage of candidate sequences was very low in the bacterial genomes with an A+T content of between 33% and 47%, and these sequences were completely absent in the corresponding 300-bp, random-generated sequences. When the A+T content increased from 47% to 78%, the percentage of strong promoter candidates increased dramatically, whereas the difference between the real and random sequences decreased, and virtually disappeared when the A+T content exceeded 62%. There were two exceptions where the genomes analyzed did not display this pattern at an A+T content of less than 62%. One was M. pneumoniae, the genome of which had an A+T content of about 60%, and in which the promoters had no -35 consensus [54]. The other example is the hyperthermophilic species A. aeolicus (~58% AT-rich genome). This species is very close to the Archaea, and occupies a unique position in the bacterial kingdom [55].

Our data show that the number N(A+T) of strong promoter candidates in 300-bp random-generated sequences corresponding to upstream regions of bacterial genes satisfies the "exponential low" of the form $N(A+T) = \exp [c_1 (A+T) + c_2]$. The distribution of strong promoter candidates in real genomes indicates that the critical point of the A+T content is close to 62% (Fig. 3). Above this level, the number of random sequences reminiscent of strong promoter patterns increases markedly.

**Strong promoter candidate sequences are located upstream of gene-coding regions**

Another important aspect of the quality of detection is the location of candidate sequences with regard to coding regions in the genome analyzed. We compared the frequencies of strong promoter-like patterns identified upstream and downstream of the initiation codon in all the genomes. The frequency of candidate sequences was clearly greater in the upstream region of ORFs in most of the genomes with an A+T content of less than 62% (Table 2). No difference was detected in T. pallidum (~47% AT-rich genome), which belongs to a distinct phylum of Spirochetes that appear to use different DNA patterns for the promotion and regulation of transcription [56].

The fact that more candidate sequences were identified upstream of ORFs highlights the fact that they are not randomly distributed in bacterial genomes, which suggests that the detection of strong promoter candidates in genomes with an A+T content of less than 62% is fairly reliable.

**Experimental validation of virtual prediction: analysis of putative strong promoters of T. maritima**

Taking our cue from the results of the virtual prediction, we sought to find out whether, and if so, to what extent the putative promoters are functional in a biological context. To do this we used reporter-gene technology, which relies on the fusion of an assayable sequence with a promoter being investigated, and the subsequent evaluation of promoter strength in a cell-free system (see Fig. 2). The genome of the hyperthermophilic bacterium T. maritima [57] was used to evaluate the feasibility of the algorithm experimentally.

63 candidate sequences were detected in the T. maritima genome using the matching scores described above. We increased the penalty for mismatching of -35 and -10 boxes by raising the scores of $s_{35}$ and $s_{10}$ to 6 and 5, respectively. This reduced the number of candidate sequences to 34 (Table 3). In this shorter list, 28 T. maritima strong promoter candidates possessed a total score higher than the 0.8475 calculated for the reference strong promoter, Ppac, that does not have a typical UP element [48]. 15 of these candidates had a total score higher than 0.8775, as estimated for PargC, another reference strong promoter that has a well defined UP element [32,49]. It is worth mentioning that 6 candidate DNA regions in T. maritima had a total score higher than 0.91, a value estimated for E. coli promoters that govern the transcription of 16S ribosomal RNA, and which were used as models for studying the stimulating effect of the UP element on gene expression [58].
We selected 13 candidate promoter sequences for further analysis by evaluation of the ArgC thermostable protein production in a coupled transcription-translation system. These sequences all exhibited a total score $t = 0.8475$, apart from TM1490 (see Table 3). The amplified DNA regions were connected to the reporter gene argC, and used directly to assess promoter activity in vitro (see Fig. 2). All putative promoters of *T. maritima* were found to be active; the protein yield ranged from 0.3 to 2.7-times that of the reference *Ptac* promoter (Fig. 4). The gene expression from the promoter PTM1272 was similar to that of *Ptac*, whereas PTM0032 was reduced almost threefold. However, higher expression was detected from the other 11 promoters; the greatest expression level was observed for PTM0477, PTM1016, PTM1429 and PTM45. Reporter gene expression was also higher for the strong promoter argC, which carries the UP element.

We next aligned experimentally analyzed promoters of *T. maritima* (Fig. 5). The most conserved sequence was the -10 box, which was identical to the E. coli consensus. The -35 box was also highly conserved, except that cytosine...
Table 3: Strong promoter candidates identified in T. maritima MSB8*.

| Downstream located gene(s)** | Strong promoter candidate sequence*** | Total score**** |
|------------------------------|--------------------------------------|-----------------|
| TM_0013 conserved hypothetical protein Operon: 2 genes | ACAATTTTATCTGATATTTTTTACACccaccatgtagTATAAC | 0.8475 |
| TM_0110 transcriptional regulator, XylR-related Operon: 6 genes | ACCTTGATTTAATTTTCTCGCAATaaattgtagaAATAAT | 0.805 |
| TM_0280 hypothetical protein Operon: 6 genes | GCAATATTGTCAGAAAATGATTGAGAATaggatgtTATAAT | 0.88 |
| TM_0339 hypothetical protein Operon: 3 genes | AGAAAATTTTTTGAGACCTTGCAAttaattgtagaTAAAT | 0.8975 |
| TM_0373 dnaK protein Operon: 2 genes | TTTTAAATTCTCTACGACCCCTTGCAATagagatgtTATAAT | 0.94 |
| TM_0657 rubrerythrin Operon: 3 genes | TAATGTAACATTCACAAATCTTACCAgttagaTAAAT | 0.8125 |
| TM_0682 hypothetical protein Operon: 3 genes | GAATCTCTTGCAAGGATTCTGTGATCAtaattgtagaTAAAT | 0.915 |
| TM_1016 hypothetical protein Operon: 2 genes | TAAAAATTTCTGAAATTTTCTGTAATctgtagaAATAAT | 0.9175 |
| TM_1167 hypothetical protein Operon: 3 genes | GAAAATGTTACAGAAAAGTACCTGCTTGAATagagatgtTAAAT | 0.865 |
| TM_t27 rRNA-Aas Operon: 3 genes | TCATTGTGATTTTTTGTAATACCAACTGCAATgtagaTAAAT | 0.8675 |
| TM_1205 conserved hypothetical protein Operon: 13 genes | GTTTTTTATCTTCTACTAATTAGGTTGCAAtttagaTAAAT | 0.88 |
| TM_1318 Operon: 2 genes | AGAAACAATTGTGAAATTGATCCATGGACTCattgtagaTAAAT | 0.8325 |
| TM_t34 tRNA-Leu Operon: 3 genes | GCAAATTTAGTTGAAATTGATCCATGGACTCattgtagaTAAAT | 0.855 |
| TM_1429 glycerol uptake facilitator protein Operon: 3 genes | GCATTGTGATTTTTTGTAATACCAACTGCAATgtagaTAAAT | 0.9175 |
| TM_t39 tRNA-Ala Operon: 3 genes | AAAAAATTTTTGGAGACCTTGCAAttaattgtagaTAAAT | 0.8725 |
| TM_1667 xylose isomerase Operon: 2 genes | AAGTATATCCTAAAAATATTTTCTGTAATctgtagaAATAAT | 0.85 |
| TM_1780 arginosuccinate synthase Operon: 6 genes | GAAAATAACAGTGAATTTTCTGTAATctgtagaAATAAT | 0.875 |
| TM_0150 (complem.) ribosomal protein L32 Operon: 5 genes | AAAAAATTTTTGGAGACCTTGCAAttaattgtagaTAAAT | 0.855 |
| TM_0477 (complem.) outer membrane protein alpha Operon: 3 genes | ACACAAAAATTTTGAAATTGATCCATGGACTCattgtagaTAAAT | 0.9425 |
| TM_0625 (complem.) hypothetical protein Operon: 2 genes | ATATCTCGTTTGGAATGAGGCTACTTTGAATTCTGCTTAGT | 0.805 |
| TM_0656 (complem.) conserved hypothetical protein Operon: 2 genes | AACTTAAGTACAAATTACTGCTGCAATgtagaTAAAT | 0.8925 |
| TM_0755 (complem.) conserved hypothetical protein Operon: 2 genes | AGAAATTCTTTGAAAATCTTACGAAActtacgctgtaatgtagaTAAAT | 0.85 |
| TM_0971 (complem.) hypothetical protein Operon: 3 genes | AAATATAAATCTGAATTTACTAAATTCACAtttagaTAAAT | 0.895 |
| TM_1015 (complem.) glutamate dehydrogenase Operon: 2 genes | ATATTTTTGGGAAAAATCTTACGACAtttagaTAAAT | 0.855 |
Table 3: Strong promoter candidates identified in *T. maritima* MSB8*. (Continued)

| Operon: 22 genes | TM_1490 (complem.) | glutamyl tRNA-Gln amidotransferase... |
|------------------|--------------------|--------------------------------------|
| **Operon: 3 genes** | TAGTATCTACCCCTAATCTTCCTAAttcggttcgaatatTGAAT | 0.835 |
| **Operon: 2 genes** | TTTCACTTTTGCAATTAACCTTGACCGAagctgtaatgttgtatgTTAAT | 0.8975 |
| **Operon: 3 genes** | AATATTGAATTTGCAACTATAATTCGAAAatctcctctctctATATAT | 0.86 |
| **Operon: 1 genes** | GGTGAAAATAGCCCGAGGAAATTTGGAATTATTTTAA | 0.845 |

*The genome annotation of *T. maritima* AE000512 used for analysis was dated 28th December 2005.

**The gene order for the first 34 candidate sequences is shown on both strands as described in the annotated genome [49]. The complementary strand is noted as (complem).**

***The spacer between -35 and -10 sites and the region located downstream of the -10 site are shown in lowercase; the initiation codons of the ORFs are shown in capital letters at the end of the corresponding sequences.***

****The first 34 candidate sequences were detected with the score parameters $s_{UP} = 13$, $s_{S35} = 6$, $s_{I0} = 5$; TMt11, TM1272, TM0032 and TM1490 were detected with $s_{UP} = 12$, $s_{S35} = 6$, $s_{I0} = 5$ and used for analysis in a cell-free system (see Fig. 3).****

preceded the -35 site in 9 promoters, and no significant preference was detected for the nucleotides at the 5th and 6th positions. An 18-bp spacer appeared to be more representative than a 17-bp distance between the -35 and -10 boxes. Although all candidates possessed an AT-rich region upstream of the -35 site, some of them had only one A-rich tract, suggesting that they harbor only a single sub-site of a putative UP element. In any case, the high score attributed to 11 identified promoters was corroborated by elevated activity in vitro. Taken together, the alignment data and the expression data from the cell-free system, suggest that *E. coli* RNA polymerase efficiently recognizes putative strong promoters of *T. maritima*, and that the presence of an UP-like element might contribute to the strength of the promoter.

Two regions, (2.4 and 4.2) of the four domains of $\sigma^{70}$ are involved in the recognition of the -10 and -35 boxes of *E. coli* promoters, respectively [59]. Several amino acids involved in contact with DNA have been also identified in the $\alpha$ subunit [60]. These DNA-binding regions in both $\sigma^{70}$ and $\alpha$ subunits of *E. coli* and *T. maritima* RNA polymerases share high similarity (data not shown), which highlights the fact that -35 and -10 boxes and UP-like element all contribute to the high promoter activity in the thermophilic host.

**Discussion**

Bacterial promoters can be arbitrarily classified as weak, moderate and strong promoters, depending on the level of expression of mRNAs or of the corresponding proteins. We have developed an algorithm that can predict strong promoters in bacterial genomes by matching the triad pat-
tern specific for the group I σ70 factor of *E. coli* RNA polymerase. The first step in the proposed triad pattern approach involves matching the UP element located 300 bp upstream of a gene-coding sequence, and then matching two optimally separated -35 and -10 boxes.

The accuracy of the computational prediction of bacterial promoters depends on the A+T content of the genomes, which means that the matrix has to be adjusted to account for this factor in the DNA under analysis [29]. The data presented highlight the fact that the detection accuracy is lower in genomes with a high A+T content. The number of potential strong promoters identified in 43 bacterial genomes, is a direct function of their A+T content; this implies that the accuracy of the prediction is lower for genomes with A+T content higher than 62%.

(A)

| Name    | UP-element | -35               | -10               |
|---------|------------|-------------------|-------------------|
| TM0373  | TTACAAATTCATCGACCCCTTGA    | $<18 \text{ bp}>$ TATAAT |
| TM1016  | TAAAAAAAAAATGAAAAATTCTTTGA | $<16 \text{ bp}>$ TTTAAT |
| TM1272  | TCCACATTGGTATTATACACTTGA | $<17 \text{ bp}>$ TTTAAT |
| TM1429  | CATTTGATTTTTGTAACTTATTTGA | $<17 \text{ bp}>$ TATAAT |
| TM1667  | CAAGTATATCCCTAAAAATTTTTGA | $<18 \text{ bp}>$ TATAAT |
| TM1780  | GAAATTACAGTTGAAGGACTTCTTA | $<20 \text{ bp}>$ TATAAT |
| TMt11   | AAAAGGTTATCAGAAATACTTTGA | $<17 \text{ bp}>$ TAAAAT |
| TM032   | ATATACTGATTTTGTAAATCTTTGA | $<18 \text{ bp}>$ CATTTA |
| TM0477  | AAAAAAAACTTTAGAAAAACTTCTGA | $<18 \text{ bp}>$ TATAAT |
| TM1067  | GATTATTTTATACGAAAGCCCTTGAC | $<18 \text{ bp}>$ TATTAT |
| TM1271  | GTGATATTTCAACATTTAATTTTTGA | $<18 \text{ bp}>$ TATAAT |
| TMt45   | AGAAGGAAAGGAAATGAAAGACTTGA | $<17 \text{ bp}>$ TATAAT |
| TM1490  | TGAAATATATGCCAGGAAAGCTTGACT | $<17 \text{ bp}>$ TAAAAT |

(B)

**T. maritima**

| Name    | UP-element | -35               |
|---------|------------|-------------------|
| TM0373  | TTACAAATTCATCGACCCCTTGA    | $<18 \text{ bp}>$ |

**E. coli**

| Name    | UP-element | -35               |
|---------|------------|-------------------|
| TM0373  | TTACAAATTCATCGACCCCTTGA    | $<17 \text{ bp}>$ |

(C)

- **P_tac**
  
  GCAATATTTCTGAAATGAGCTGTGGAGACATTTATCTCGGCTCGTATAAT

- **PargC**
  
  TCGAAATATTTAAATACATTTGATTTTATTTTTATACATTTATAAT

**Figure 5**

**Organization of strong bacterial promoters.** (A), Alignment of 13 promoter candidates of *T. maritima*; (B) consensus sequences of *T. maritima* and *E. coli* strong promoters; consensus of the *E. coli* UP element is described in [26, 27]; (C) the strong promoters P_tac and PargC were used as references in this study.
The choice of the matching score is yet another difficulty in identifying DNA-binding sites including promoters, as the highest score may not be the one most biologically relevant for genome-scale predictions [61,62]. It is therefore helpful to use additional criteria to eliminate false-positives. It looks as if the total score of 0.8475, calculated for the reference promoter Ptac, could be used as a reasonable criterion for identifying real strong promoters recognized by an Eσ⁷⁰-like RNA polymerase. In particular, using the scores applied to genomes analysis (see Tables 1 and 2), the algorithm detects 7 potential strong promoters in *M. tuberculosis* (~34% AT-rich genome) that encodes a variety of *σ* factors, including *σ^A* that recognizes the promoters possessing typical -10 and -35 boxes [63]. However, none of the predicted strong promoters had a total score in excess of 0.8475, and visual inspection indicated that none of these promoters possesses an UP-like sequence, suggesting that this gene expression-stimulating element is absent in *M. tuberculosis*.

The possibility of applying linear PCR-generated molecules for cell-free protein synthesis, without needing to perform DNA cloning in bacteria, is a prerequisite for assessing gene expression on a genome-wide scale. As a first step in this direction, we tested reporter-gene fusions to evaluate the strenght of the promoters identified in the genome of *T. maritima*. Though this approach does not exclude possible masking effects of *E. coli* repressors or activators in the extracts, it is relatively simple, timesaving and informative, all of which are major advantages for evaluating computational predictions. Using the two well-characterized strong promoters (Ptac and PargC) as references, high activity has been demonstrated for 11 out of 13 candidate sequences of *T. maritima*. This is quite a low proportion; however, it suggests that the detection accuracy by the triad pattern algorithm might be close to 85%. The limitations of the algorithm in terms of specificity and sensitivity of the virtual prediction of putative strong promoters might be further experimentally evaluated by analysis of bacterial genomes with high-throughput methods.

This study offers the first insight into the organization and distribution of strong promoters in hyperthermophilic organisms, which probably constitute the longest lineage in the microbial world [64]. Overall, strong promoters of hyperthermophiles are similar to those of mesophilic origin. We have recently shown that the *T. maritima* RNA polymerase α subunit binds to the PargG promoter described here under PTM1780 [65]. It has been found that the substitution of arginine in the hyperthermophilic α subunit, corresponding to the position Arg265 in the *E. coli* subunit and crucial for DNA recognition [60,66], or the deletion of an AT-rich sequence located upstream of the -35 site, decreases the binding affinity for DNA [65].

The *PargG* promoter harbors a UP-like element, and is able to direct high gene expression in vitro. Moreover, this element appears to compensate for a poor -35 box or non-optimal 20-bp spacer of this promoter (see Table 3 and Fig. 5). Hence, these observations, along with the data obtained using other *T. maritima* promoters, allow us to assume that the presence of a UP-like element with less than 5 mismatches out of 17 nucleotides is essential for the strenght of most strong promoters. This is consistent with the conservation of DNA interaction amino acids in the α subunit of the hyperthermoiphilic RNA polymerase. However, sequence-independent upstream DNA interactions within the C-terminal domain of the α subunit could often be required to initiate transcription in *E. coli* cells [67]. Therefore, the functional significance of the UP-like element in gene expression remains to be proven experimentally in hyperthermophilic organisms.

The strong promoters of *T. maritima* direct the transcription of genes involved in tRNA, ribosome synthesis, energy metabolism, transport, and cell movement (see Table 3). However, to our surprise, we found that 15 of the 38 best candidates promote the transcription of hypothetical proteins. The previously uncharacterized hypothetical protein TM1016 (total score 0.9175) turns out to share 28% identity with a biopolymer transport protein of *Vibrio vulnificus* [68]. In this context, recent studies of the *T. maritima* transcriptome have indicated that ABC transporters could play a major role in its ecology [69]. Further characterization of highly expressed hypothetical genes identified in our study might help to elucidate their role in the biology of this hyperthermophilic organism.

The strong promoter candidates prediction could contribute to the wide-scale genome expression analysis of evolutionarily distant bacteria, especially of those that possess an A+T DNA content lower than 62%. As a complement to DNA microarrays, it could help to elucidate the overall response of bacterial genomes to various environmental stresses. Moreover, the triad pattern algorithm can be used to extract the DNA region that carries translational signals; this is useful for investigating ORFs located downstream from the corresponding strong promoters (see Table 3). Thus, almost half of the *T. maritima* ORFs transcribed from putative strong promoters are preceded by a highly conserved Shine-Dalgarno site located 7–9 nucleotides from the ATG initiation codon, which is a characteristic feature of elevated protein synthesis in gram-negative and gram-positive bacteria [70]. This information will be useful for comparing highly synthesized mRNAs with the production of the corresponding proteins using high-throughput transcriptomic and proteomic methods, which is an important challenge in the fields of basic and applied microbiology [71]. Furthermore, the characterization of proteins whose expression is governed by strong
promoters looks like a promising approach to selecting candidate vaccines against microbial diseases and/or to identifying potential new antibacterial targets in the fight against nosocomial infections.

Further quantitative assessment of a dynamic and complex mechanism of protein-DNA and protein-protein interactions involved in transcription might help to develop a more advantageous multi-pattern tool using both DNA and protein parameters to provide a comprehensive prediction of the strength of promoter activity in bacterial cells.

Conclusion
The triad pattern algorithm developed predicts strong promoter candidates by matching UP-like elements and identifying the presence of -35 and -10 boxes optimally distanced from each other in the annotated bacterial genomes. The presence of strong promoters is a function of the A+T content of the bacterial genome, and the number of false-positives is greater for genomes that have an A+T content higher than 62%. The prediction algorithm has been validated by cell-free experimental dissection of putative *T. maritima* promoters. The data indicate that strong promoters govern the transcription of genes coding vital functions, and of genes coding as-yet unknown functions in this hyperthermophilic bacterium. This algorithm is simple to use and flexible, and it could be further adapted to meet the requirements of a genome of interest if its promoter-specific motifs differ from consens recognition by Eσ70-like RNA polymerase.

Availability and requirements
The algorithm is freely accessible for non-commercial use at the web-site [http://www.protneteomix.com](http://www.protneteomix.com). It takes several seconds to analyze the annotated genome sequence available from databases.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
MD developed the algorithm and performed the computational analysis; AM conducted cell-free experiments; VS designed the project, contributed to the development of the algorithm and data analysis, and wrote the manuscript.

Additional material

Additional file 1
ReadMe. Contains information to use the algorithm.
Click here for file.
[http://www.biomedcentral.com/content-supplementary/1471-2105-9-233-S1.pdf](http://www.biomedcentral.com/content-supplementary/1471-2105-9-233-S1.pdf)

Acknowledgements
At the time of the study AM was a research fellow supported by Pays de la Loire. MD acknowledges support from the Conseil Régional des Pays de la Loire and ProtNeteomix for his visit to Nantes University. We should like to thank anonymous reviewers whose suggestions allowed us to improve the manuscript. This study was supported by the "Post-Gène programme des Pays de la Loire", by the EU project EUR-INTAFAr (n°LSHM-CT-2004-512138), and by the R&D program of ProtNeteomix.

References
1. Darst SA: Bacterial RNA polymerase. *Curr Opin Struct Biol* 2001, 11:155-162.
2. Queen C, Wegman MN, Korn LJ: Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. *Nucleic Acids Res* 1982, 10:449-456.
3. Galas DJ, Eggert M, Waterman MS: Rigorous pattern-recognition methods for DNA sequences. *J Mol Biol* 1985, 186:117-128.
4. Soden R: Methods for discovering novel motifs in nucleic acid sequences. *Comput Appl Biosci* 1989, 5:293-298.
5. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994, 2:28-36.
6. Alexandrov N, Mironov A: Application of a new method of pattern recognition in DNA sequence analysis: a study of *E. coli* promoters. *Nucleic Acids Res* 1990, 18:1847-1852.
7. Demeler B, Zhou G: Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Res* 1991, 19:1593-1599.
8. Cardon LR, Stormo GD: Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* 1992, 223:159-170.
9. Horton PB, Kanehisa M: An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucleic Acids Res* 1992, 20:4331-4338.
10. Thieffry D, Salgado H, Huerta AM, Collado-Vides J: Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics* 1998, 14:391-400.
11. Vanet A, Marsan L, Labigne A, Sagot M-F: Inferring regulatory elements from a whole genome. *An analysis of Helicobacter pylori* α family of promoter signals. *J Mol Biol* 2000, 297:335-353.
12. Leung SW, Melish C, Robertson D: Basic gene grammars and DNA-chartparser for language processing of *Escherichia coli* promoter DNA sequence. *Bioinformatics* 2001, 17:226-236.
13. Gordon L, Chervonenkin AY, Gammerman AJ, Shakhmuradov IA, Solovyev VV: Sequence alignment kernel for recognition of promoter regions. *Bioinformatics* 2003, 19:1964-1971.
14. Jacques P-E, Rodrigue S, Gaudreau L, Goulet J, Brzezinski R: Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs. *BMC Bioinformatics* 2006, 7:423. (doi:10.1186/1471-2105-7-423)
15. Benham CJ: Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proc Natl Acad Sci USA* 1993, 90:2999-3003.
16. Kanhere A, Bansal M: A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics* 2005, 6:1-10.
17. Wang H, Benham CJ: Promoter prediction and annotation of microbial genomes based on DNA sequence and structural
responses to superhelical stress. BMC Bioinformatics 2006, 7:248.

18. Ayer SE, Gourley RL: RNA-polymerase-promoter interactions: the comings and goings of RNA polymerase. J Bacteriol 1998, 180:3019-3025.

19. Malrèdes SC: Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol Rev 1996, 60:512-338.

20. Weigel P, Dimova D, Lecocq M, Sakanyan V: Analysis of “extended” promoters in Bacillus stearothermophilus argCJBD by WebLogo: A graphical representation of key evolutionary conservation across multiple sequences. Nucleic Acids Res 2006, 34:5533-5543.

21. Helmann JD: The extracytoplasmic function (ECF) sigma factors. Adv Microb Physiol 2002, 46:87-110.

22. deHastie PL, Zupančič ML, Record MT: Use of the “Per-sonal computer” to analyze and predict the structural models of biological primary sequence information. Genes & Dev 1999, 13:2134-2147.

23. readership. BMC Bioinformatics 2006, 7:248. doi:10.1186/1471-2105-7-248

24. Harley C, Reynolds R: Analysis of E. coli promoter sequences. Nucleic Acids Res 1983, 11:6237-11:6247.

25. Fredrick K, Caramori T, Chen Y, Galizzi A, Helmann JD: Identification of a UP element consensus sequence for bacterial promoters. Proc Natl Acad Sci USA 1998, 95:9761-9766.

26. Autumn SE, Gaal RL: Fine structure of E. coli RNA polymerase promoters: the σ element subunit binding to the UP element minor groove. Genes & Dev 2001, 15:491-506.

27. Aiyar SE, Gourse RL, Ross W: Identification of a UP element consensus sequence for bacterial promoters. Nucleic Acids Res 1995, 23:4652-4657.

28. Estrem ST, Ross W, Gourse RL: Identification of an UP element consensus sequence for bacterial promoters. Proc Natl Acad Sci USA 1998, 95:9761-9766.

29. Ross W, Ernst A, Gourse RL: Fine structure of E. coli RNA polymerase promoters: the σ element subunit binding to the UP element minor groove. Genes & Dev 2001, 15:491-506.

30. Tutukina MN, Shakunov KS, Masulis IS, Ozoline ON: Intragenic promoter-like sites in the genome of Escherichia coli discovery and functional implication. J Bioinform Comput Biol 2007, 5:497-506.

31. Fredrick K, Caramori T, Chen Y, Galizzi A, Helmann JD: Promoter architecture in the flagellar region of Bacillus subtilis: high-level expression of flagellin by the E. coli RNA polymerase requires an upstream promoter element. Proc Natl Acad Sci U S A 1995, 92:2583-2586.

32. Savchenko A, Weigel P, Dimova D, Lecocq M, Sakanyan V: The Bacillus steareothermophilus argCJBD operon harbours a strong promoter as evaluated in Escherichia coli cells. Gene 1998, 212:167-177.

33. Ayer SE, Gaal RL, Gourse RL: rRNA promoter activity in the fast-growing bacterium Vibrio natriegens. J Bacteriol 2002, 184:1349-1358.

34. Sorokin AA, Ospyov AA, Dzhelilyan TR, Beskaravainy PM, Kamzolov SG: Electrostatic properties of promoter recognized by E. coli RNA polymerase Esigma70. J Bioinform Comput Biol 2004, 2:455-467.

35. Mitchell JE, Zheng D, Busby SJ, Mclninch SD: Identification and analysis of “extended” promoters in Escherichia coli. Nucleic Acids Res 2003, 31:4689-4695.

36. Sneath GD, Sneath TD, Gold L: Ehrenfeucht A: Use of the “Perception” algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 1982, 10:2997-3011.

37. Baldi P, Chauvin Y, Hunkapiller T, McClure MA: Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci USA 1994, 91:1059-1063.

38. Jarmer H, Larsen TS, Kragh A, Saxild HH, Brunak S, Knudsen S: Sigma A recognition sites in the Bacillus subtilis genome. Microbiology 2001, 147:2417-2424.

39. Petersen L, Larsen TS, Utserry DW, On SL, Krogh A: RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box. J Mol Biol 2003, 326:1361-1372.

40. Munch RL, Hiller K, Grote A, Scheer M, Klein J, Schobert M, John D: Virtual Footprint and PROFORDIC: an integrative frame-work for regulon prediction in prokaryotes. Bioinformatics 2005, 21:i487-i489.

41. Vatsa A, Marsan L, Sagot M-F: Promoter sequences and algorithmical methods for identifying them. Res Microbiol 1999, 150:779-799.
involved in contact with the DNA enhancer element. EMBO J 1996, 15(16):4358-4367.

61. Eskin E, Keich U, Gelfand MS, Pevzner PA: Genome-wide analysis of bacterial promoter regions. Pac Symp Bioinform 2003:29-40.

62. Huerta AM, Collado-Vides J: Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol 2003, 333:261-278.

63. Manganelli R, Proveddi R, Rodrigue S, Beaucher J, Gaudreau L, Smith I: α factors and global regulation in Mycobacterium tuberculosis. J Bacteriol 2004, 186:895-902.

64. Woese CR, Kandler O, Wheelies ML: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990, 87:4576-4579.

65. Braun F, Marhuenda FB, Morin A, Guevel L, Fleury F, Takahashi M, Sakanyan V: Similarity and divergence between the RNA polymerase α subunits from hyperthermophilic Thermotoga maritima and mesophilic Escherichia coli bacteria. Gene 2006, 380:120-126.

66. Gaal T, Ross W, Blatter EE, Tang H, Jia X, Krishnan VV, Assa-Munt N, Ebright RH, Gourse RL: DNA-binding determinants of the alpha subunit of RNA polymerase: novel DNA-binding domain architecture. Genes & Dev 1996, 10:16-26.

67. Ross W, Gourse RL: Sequence-independent upstream DNA-αCTD interactions strongly stimulate Escherichia coli RNA polymerase-lacUV5 promoter association. Proc Natl Acad Sci USA 2005, 102:291-296.

68. Kim YR, Lee SE, Kim CH, Kim SY, Shin EK, Shin DH, Chung SS, Choy HE, Progulske-Fox A, Hillman JD, Handfield M, Rhee JH: Characterization and pathogenic significance of Vibrio vulnificus antigens preferentially expressed in septicemic patients. Infect Immun 2003, 71:5461-5471.

69. Johnson MR, Conners SB, Montera CI, Chou CJ, Shockley KR, Kelly RM: The Thermotoga maritima phenotype is impacted by syntrophic interaction with Methanococcus jannaschii in hyperthermophilic coculture. Appl Environ Microbiol 2006, 72:811-818.

70. Vellanoweth RL, Rabinowitz JC: The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. Mol Microbiol 1992, 6:1105-1114.

71. Boyce JD, Cullen PA, Adler B: Genomic-scale analysis of bacterial genes and protein expression in the host. Emerg Infect Dis 2004, 10:1357-1362.