P&TIT: A computer tool for predicting prototypical transcription promoter and terminator elements by conserved motifs

Marco Di Salvo[a], Eva Pinatel[b], Gianluca De Bellis[b], Adelfia Talà[a], Clelia Peano[b], Pietro Alifano[a]

[a] *Department of Biological and Environmental Sciences and Technologies, University of Salento, Lecce, Italy*
[b] *Institute of Biomedical Technologies National Research Council, Segrate, Milan, Italy.*

Abstract

Over the last few decades, computational genomics has tremendously contributed to decipher biology from genome sequences and related data. Considerable effort has been devoted to prediction of transcription promoter and terminator sites that represent the essential "punctuation marks" for DNA transcription. In this study we have investigated the possibility to identify putative promoters and Rho-dependent terminators in prokaryotes on the basis of evolutionarily conserved motifs. The final aim of this work is to develop computer software to predict location of bacterial promoters and terminators based on nucleic acid motifs.

Introduction

The evolutionary relatedness and/or functional analogies between transcription initiation and termination mechanisms in all three domains of life prompted us to explore the possibility to recognize promoter and terminator elements in a prokaryotic genome based on conserved structured motifs [1]. For promoter prediction we focused on possible G-quadruplex structures upstream of AT-rich elements. G-quadruplexes, non-canonical nucleic acid topologies with little established biological roles, are increasingly considered for conserved regulatory element discovery [2]. For Rho-dependent terminator prediction, based on current models for Rho-dependent termination we used an algorithm based on C>G ratio, cytosine spacing, and hairpin structures (possible RNA polymerase pausing sites) downstream from a "C>G content" motif. Results of genome mining by were then validated by RNAseq data in the actinobacterium *Streptomyces ambofaciens*.

Materials and methods

Algorithms to identify the motifs of interest were generated by using the programming language "Python" v.3.5 taking as input data the annotated genome sequence of *S. ambofaciens* ATCC 23877 for promoter prediction, and RNAseq for validation. Methods to identify putative promoter and terminator elements are described afterwards.

Results

Method to identify putative promoters and results

A three-step procedure was used to detect putative promoters (Figure 1): i.) The first step consisted in the identification of the putative promoter "AT-rich element". To this purpose, within each intergenic sequence (≥50 nt) associated with transcribed genes, the 25-bp long region with the highest AT% (herein referred to as "AT-max" motif) was selected; ii.) The second step was the identification of putative G-quadruplex motifs extended up to 50 bp upstream from the 5'-end of the selected "AT-max" motif. Motif $G_xN_yG_xN_yG_xN_yG_x$, with $2 \leq x \leq 4$ and $1 \leq y \leq 10$ is commonly used to predict the presence of G-quadruplexes [3]; iii.) The third step consisted in validation of promoter prediction by RNAseq data. We considered the 75-bp long elements (including the G-quadruplex and "AT-max" motifs) as possible promoters if there was an increase of read values by a factor of at least 1.5 between the read value at the "AT-max" motif 5'-end point and the read value 10 bp downstream from the "AT-max" motif 3'-end point, the latter with a read value ≥5. The results are shown in Table 1. In Figure 2 are represented some examples of putative promoter signals mapped by the RNAseq graphical user interface.

Methods to identify putative Rho-dependent terminators and results

The procedure used to detect Rho-dependent terminators was the following: i.) The first step consisted in the identification of putative RUT site. To this purpose, within each deduced intercistronic (≥78 nt) or intracistronic RNA sequence of transcribed genes, 78-nt long regions with C/G ratio ≥1.3 and regularly spaced cytosine residues (every 11-13 nt) were selected (the so called RUT sites) [4]; ii.) The second step was the identification of hairpin structure (often serving as RNAP pausing sites) in a region extended up to 150 nt downstream from the 3'-end point of the "C>G content" motif; iii.) The third step consisted in validation of Rho-dependent transcription terminator prediction by RNAseq data. We considered these regions as putative Rho-dependent terminators if there was a decrease of read value by a factor of at least 1.5 between the read value at the "C>G content" motif 5'-end point and the read value 150 nt downstream from the "C>G content" motif 3'-end point, the first with a read value ≥5. The results are shown in Tables 2 and 3. In Figure 3 are represented some examples of putative Rho-dependent terminator signals mapped by the RNAseq graphical user interface.

Methods identify putative intrinsic terminators and results

In order to identify intrinsic (Rho-independent) terminators, in all the intergenic sequences of the genome we looked for RNA hairpins followed by a run of U residues (min 3, max 8). We considered these regions as possible intrinsic terminators if there was a decrease of RNAseq read value by a factor of at least 1.5 between a point located 50 nt upstream from the U residues and a point located 5 nt downstream from the U residues, the first with the read value ≥5. The results are shown in Table 4.

Conclusions

In this study we investigated the possibility to identify promoter and terminator elements by detecting evolutionarily conserved motifs. The algorithm predicted putative promoters in about 42.2-42.8% of intergenic sequences associated with transcribed genes, and prediction was validated by RNAseq data in a percentage of cases ranging from 38.3% to 42.4% (Table 1). Interestingly, these high percentages very similar to those found in other evolutionarily distant promoters including yeast [5] and human gene

promoters more than 40% of which were shown to contain one or more G-quadruplexex [6]. Our results further support the idea that G-quadruplex is a prototypical motif involved in general promoter function/regulation.

For Rho-dependent terminator prediction, the algorithm predicted putative Rho-dependent terminators in about 30.2-30.4% of intercistronic sequences associated with transcribed genes, and prediction was validated by RNAseq data in a percentage of cases ranging from 31.1% to 36.4% (Table 2). In contrast, an algorithm predicted the presence of putative Rho-independent (intrinsic) terminators in a very small percentage of transcribed intercistronic sequences ranging from 6.3% to 6.7% with percentages of validation by RNAseq data ranging from 27.8% to 36.2% (Table 4). This very limited number of putative Rho-independent terminators in *S. ambofaciens* suggests that Rho (or factor)-dependent termination mechanisms may be predominant in this microorganism. More interestingly, the presence of putative Rho-dependent terminators was predicted in about 92.4-92.8% of transcribed intracistronic sequences, and results were validated in a percentage of cases ranging from 71.4% to 73.4% (Table 3). Future efforts will be dedicated to extending our analysis to other organisms with the final aim to better characterize what appear to be prototypical elements of the "punctuation marks" for DNA transcription.

References

[1] Burton SP, Burton ZF. The σ enigma: bacterial σ factors, archaeal TFB and eukaryotic TFIIB are homologs. Transcription 5:e967599. (2014).

[2] Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res 43:8627-37. (2015).

[3] Kikin O, D'Antonio L and Bagga PS. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequenze. Nucleic Acids Res 34.

[4] Alifano P, Rivellini F, Limauro D, Bruni CB, Carlomagno MS. A consensus motif common to all Rho-dependent prokaryotic transcription terminators. Cell 64:553–563. (1991).

[5] Capra JA, Paeschke K, Singh M, Zakian VA. G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in *Saccharomyces cerevisiae*. PLoS Comput Biol 6:e1000861. (2010).

[6] Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 35:406-413. (2007).

Table 1. Statistics of putative promoters that were identified by G-quadruplex and "AT-max" motif-based algorithm.

Sampling time (h)	Intergenic sequences with putative promoters (%)	Validated promoters (%)	Validated promoters with "TANNNT" consensus (%)	Validated promoters with "TTGAC" consensus (%)
48	42.8	38.3	43.8	4,6
72	42.3	39.9	45.6	4,5
96	42.5	38.5	43.4	4,1
120	42.2	42.4	42.5	4,8

Table 2. Statistics of putative extracistronic Rho-dependent terminators that were identified by C>G content RNA motif-based algorithm.

Sampling time (h)	Extracistronic sequences with Rho-dependent terminators (%)[a]	Validated extracistronic Rho-dependent terminators (%)[b]
48	30.2	31.1
72	30.2	32.6
96	30.3	33.4
120	30.4	36.4

[a]Percentages of extracistronic sequences that possess at least one possible RUT sequence.
[b]Percentage of extracistronic sequences with at least one possible RUT sequence, which we consider as possible Rho-dependent terminators, according to of RNAseq data.

Table 3. Statistics of putative intracistronic Rho-dependent terminators that were identified by C>G content RNA motif-based algorithm.

Sampling time (h)	Intracistronic sequences with Rho-dependent terminators (%)[a]	Validated intracistronic Rho-dependent terminators (%)[b]
48	92.4	71.4
72	92.7	72.1
96	92.6	71.5
120	92.8	73.4

[a]Percentages of intracistronic sequences that possess at least one possible RUT sequence.
[b]Percentage of intracistronic sequences with at least one possible RUT sequence, which we consider as possible Rho-dependent terminators, according to of RNAseq data.

Table 4. Statistics of putative Rho-independent terminators that were identified by the algorithm[a]

Sampling time (h)	Intercistronic sequences with Rho-independent terminators (%)[a]	Validated intercistronic Rho-independent terminators (%)
48	6.7	27.8
72	6.7	30.7
96	6.5	29.8
120	6.3	36.2

[a] RNA hairpins followed by a run of U residues (min 3, max 8)

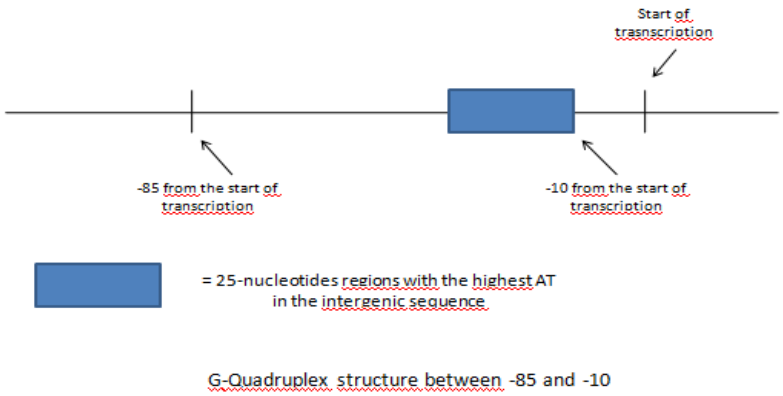

Fig.1. Graphical description of the procedure used to detect putative promoters

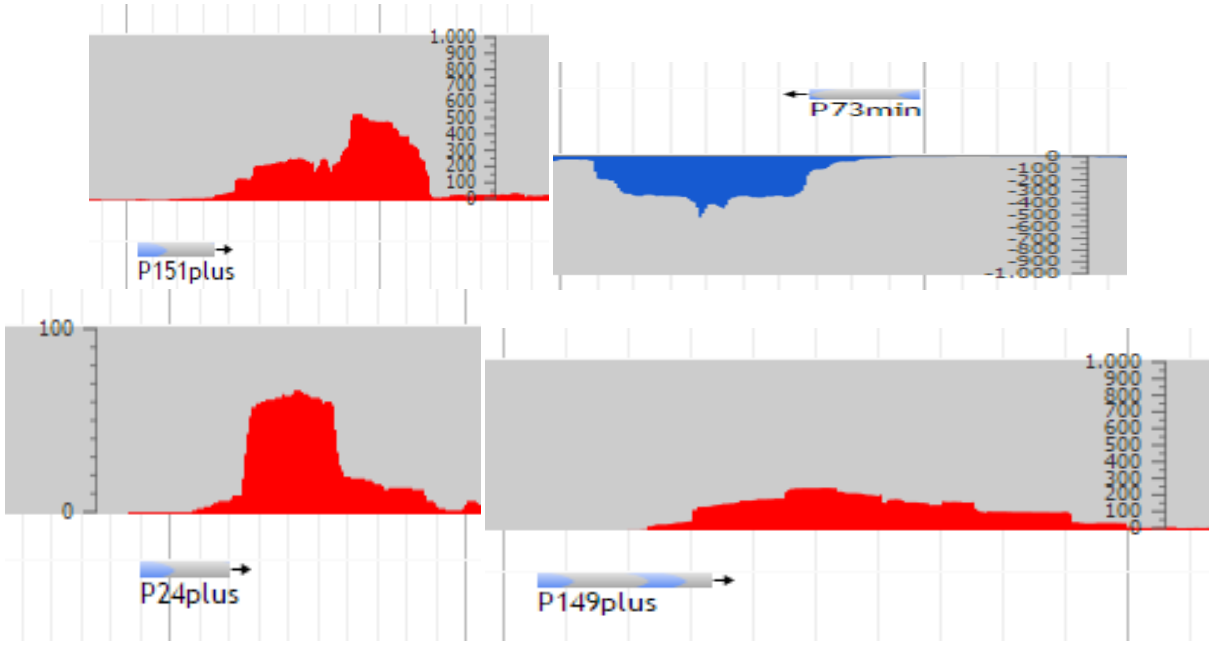

Fig. 2. Examples of putative promoter signals mapped by the RNAseq graphical user interface.

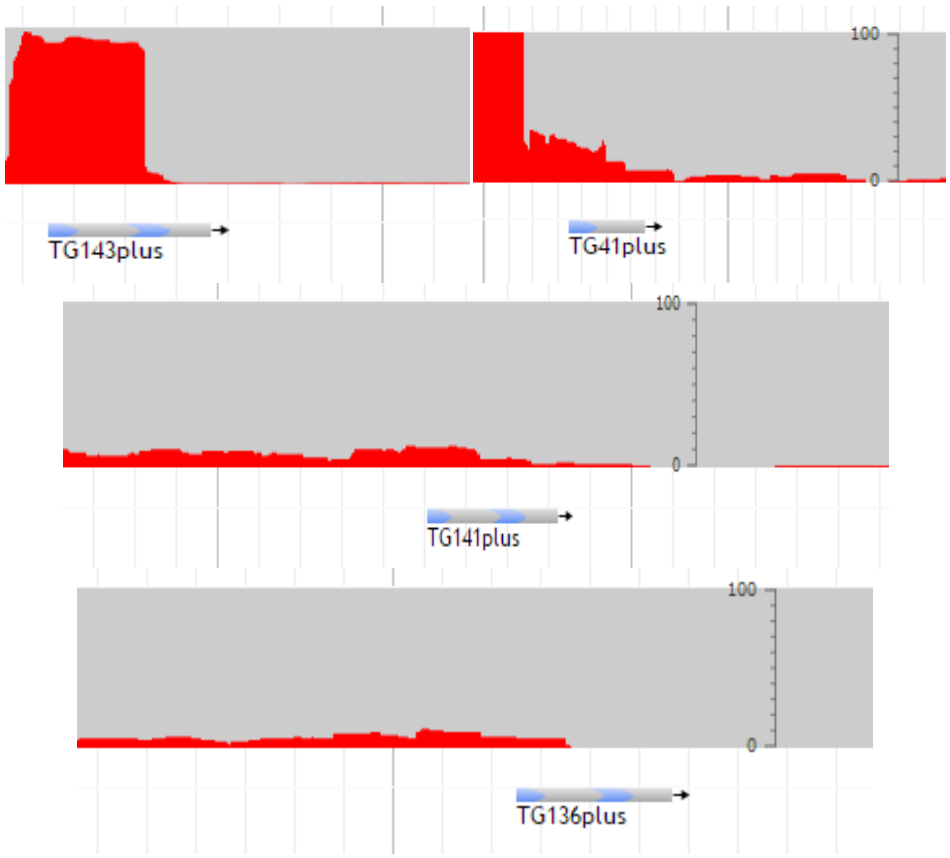

Fig. 3. Examples of putative Rho-dependent terminator signals mapped by the RNAseq graphical user interface