Journal of Biotechnology Research

ISSN(e): 2413-3256, ISSN(p): 2413-8878

Vol. 6, Issue. 7, pp: 90-97, 2020

URL: https://arpgweb.com/journal/journal/16

DOI: https://doi.org/10.32861/jbr.67.90.97

Academic Research Publishing Group

Original Research Open Access

Prediction and Recognition of Gram-Negative Bacterial Promoter Sequences: An Analysis of Available Web Tools

Hugo André Klauck
Instituto Federal do Rio Grande do Sul – Campus Bento Gonçalves – RS, Brazil

Gabriel Dall'Alba
Biotechnology Institute – University of Caxias do Sul - RS, Brazil

Scheila de Avila e Silva (Corresponding Author)
Biotechnology Institute – University of Caxias do Sul - RS, Brazil
Email: sasilva6@ucs.br

Ana Paula Longaray Delamare
Biotechnology Institute – University of Caxias do Sul - RS, Brazil

Article History
Received: June 8, 2020
Revised: July 6, 2020
Accepted: July 14, 2020
Published: July 16, 2020

Abstract

Many computational methods aim to improve the prediction and recognition of transcription elements in prokaryotes. Despite this, the natural features of those elements make their prediction and recognition remain as an open field of research. In this paper, we compared the open-access tools BacPP, BPROM, bTSSfinder, CNNPromoter_b, iPro70-PseZNC, NNPP2, PePPer, and PromPredict. First, we listed the overall functionalities of each tool and the resources available on their web pages. Later, we carried out a comparison of prediction results using 206 intergenic regions. When evaluating the prediction using intergenic regions containing a single promoter within each, NNPP2 and BacPP obtained >90% correct predictions, with NNPP2 obtaining the highest values of match between predicted promoter location and location indicated by RegulonDB. Overall, many discrepancies were observed among the results. They may be explained by the differences in the methodologies that each tool applies for promoter prediction, not excluding the natural features of promoters as a factor as well. In any case, the results highlight the necessity to continue the efforts to improve promoter prediction, perhaps combining multiple approaches. Through said efforts, some of the challenges of the postgenomic era may be tackled as well.

Keywords: Gram-negative; Web tools; E. coli; Bioinformatics; Computational methods; Promoter prediction tools.

1. Introduction

Transcription plays a major role in gene expression. It occurs through the interaction between the enzyme RNA polymerase (RNAp) and promoters: DNA sequences located ~70 base pairs (bp) upstream from the transcription starting site (TSS). This interaction between the RNAp and the promoter sequence is mediated by a small subunit protein, known as sigma (σ) factor. Its role in the transcription process is to bind itself to the RNAp and guide the now holoenzyme towards a specific promoter sequence [1].

There are multiple σ factors found within bacteria. In *E. coli*, for instance, there are seven known σ factors: the σ^{19}, σ^{24}, σ^{28}, σ^{32}, σ^{38}, σ^{54}, and the σ^{70}. Each σ factor interacts with specific DNA promoter sequences. As example, the housekeeping factor (σ^{70}) is known to regulate most of the genes involved in the vital processes of the bacteria, also maintaining the bulk of transcription during its growth phase [2, 3]. It also serves as the model for the canonical promoter, composed by two main motifs: one located at a 10 bp distance from the TSS, and another located 35 bp distant. Furthermore, the -10 region is composed by a canonical 5' – TATAAT – 3' sequence, while the -35 region is composed by a 5' – TTGACA – 3' [4]. Each σ factor has its own set of features that distinct themselves from each other, including structural variations in the motifs (e.g., expected sequence, and degree of conservation), sometimes in the expected position of the motifs (such as the σ^{54}, with its motifs located around -12 and -24 bp distant from the TSS), among others [1, 3, 4].

Overall, these distinct features give specificity to the transcription process, and the existence of these multiple σ factors allows the bacteria to survive adverse situations – from pH fluctuations to the stress provoked by heat shocks [3]. Hence the importance of thoroughly studying them, allowing opportunities to expand our comprehension in topics such as: (*i*) the mechanisms of gene expression regulation; (*ii*) comprehension of the mechanisms involved in diseases; (*iii*) development of novel drugs in the combat of bacterial infections, to mention a few [4-6].

The present post-genomic era and the development of high-throughput sequencing methods is highlighted by the genome annotation efforts lagging behind the growing capacity to generate more data [5, 7]. In this context, a still relevant challenge is the prediction of promoter sequences and of other genomic structures. The intrinsic features of promoter sequences (e.g., short sequences, lack of conservation in its nucleotide content, among others) provide a computational challenge to the automated prediction of those elements [8]. As a result, elevated false positive rates (that is, non-promoter DNA sequences incorrectly classified as promoter sequences) is a problem yet to be completely solved [9]. Obtaining higher-quality datasets, that is, larger and more reliable (generated through larger amounts of experimentally verified sequences), proper understanding of the biological features of those sequences, the development of novel methods, and the employment of novel approaches are listed as a few possibilities to solve this problem [5, 9, 10].

In silico approaches are welcomed in that regard. Computational biology and bioinformatics approaches have often been employed in efforts to improve the accuracy of promoter prediction tools. Among the tools designed to predict and/or classify promoter sequences, a few stand out, such as: BacPP [11], BPROM [12], bTSSfinder [13], CNNPromoter_b [10], iPro54-PseKNC [9], iPro70-PseZNC [14], NNPP2 [15], PePPer [16], PromPredict [17].

Throughout the history of *in silico* approaches to promoter prediction, several strategies were deployed. Initially, the prediction and recognition of these regions was made mostly by sequence alignment. One of the most common methods was the Position Weight Matrix (PWM), using (and requiring beforehand, for that matter) the conservation of the -35 and -10 elements of known σ^{70} sequences. Later on, the application of machine learning approaches, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM), contributed to expand the accuracy and thus enable a higher efficacy on promoter prediction. Information on curvature and stability of the DNA molecule has also played a significant role in enhancing the aforementioned overall efficacy when applied as discriminatory features between promoter and non-promoter sequences, going beyond the often-used nucleotide composition [13, 15, 17].

In light of this, the goal of this paper is to describe open-access computational tools focused on gram-negative promoter recognition, carrying out a comparison between their results, and assisting in laying out an overall outlook on the efforts to tackle the promoter prediction challenges.

2. Methodology

2.1. Data Description

We extracted Intergenic sequences from *Escherichia coli* str. K-12 substr. MG1655 (see supplemental material) from IntergenicDB [18]. The K-12 strain choice was made considering that it is one of the considered model genomes for *in silico* approaches, containing well documented and experimentally verified information about its genes and their regulation [19]. From the 695 sequences available on IntergenicDB, we selected only those with a length equal or higher than 81 nucleotides for comparison and standardization purposes – the 81 nucleotides length is the standard sequence size found at RegulonDB [20]. The resulting number of selected sequences was 206. Then, we searched the exact location of promoter regions by comparing the 206 sequences with available data from RegulonDB v. 9.4 [20]. We also excluded all sequences found in RegulonDB that were not experimentally verified nor related to the σ^{70}. In total, 959 σ^{70}-dependent promoter sequences experimentally identified remained.

2.2. Prediction Tools

Three criteria were used to select the promoter recognition tools: (*i*) the tools should be Open Access, free of charge on its usage, designed for academic purposes, and should also have an associated published paper with it; (*ii*) online availability and (*iii*) the tools should focus at or at least show promising results when dealing with gram-negative bacterial genomes. At least eight tools fulfilled these criteria: BacPP [11], BPROM [12], bTSSfinder [13], CNNPromoter_b [10], iPro70-PseZNC [14], NNPP2 [15], PePPer [16], e PromPredict [17]. In table 1, each tool is presented accordingly to each targeted data, type of input data, performance score, and the related paper.

Table-1. Selected promoter prediction tools, Information regarding input format and performance score were extracted from both websites and related references

Name	URL	Training Dataset	Performance score (informed by the authors)	Reference
BacPP – Bacterial Promoter Prediction	http://bacpp.bioinfoucs.com/home	RegulonDB v. 2009	83,6% (σ^{70})	de Avila, *et al.* [11]
bTSSfinder - bacterial Transcription Start Site finder	http://www.cbrc.kaust.edu.sa/btssfinder/	RegulonDB v.2013	89,5% (σ^{70})	Shahmuradov, *et al.* [13]
BPROM - Predicts bacterial promoters	http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb	Not mentioned	80% (Accuracy)	Solovyev and Salamov [12]
CNNPromoter_b - Convolutional Neural Networks Promoter Bacterial	http://www.softberry.com/berry.phtml?topic=cnnpromoter_b&group=programs&subgroup=deeplearn	RegulonDB v. 2016	90% (Sensibility)	Umarov and Solovyev [10]
iPro70-PseZNC - Identifying sigma70 promoters – pseudo-multi-window Z-curve nucleotide	http://lin-group.cn/server/iPro70-PseZNC.html	RegulonDB v. 2016	90% (Accuracy)	Lin, *et al.* [14]

composition				
PePPer - Prediction of Prokaryote Promoter Elements and Regulons	http://genome2d.molgenrug.nl/index.php/prokaryote-promoters	RegulonDB v. 2010	-	De Jong, *et al.* [16]
NNPP2 - Neural Network Promoter Prediction	http://www.fruitfly.org/seq_tools/promoter.html	Not mentioned	75% (Accuracy)	Reese [15]
PromPredict – Promoter Prediction	http://nucleix.mbu.iisc.ernet.in/prompredict/prompredict.html	RegulonDB v. 2004	90% (sensibility)	Bansal and Kanhere [17]

CNNPromoter_b [10] is able to analyze promoter sequences from both prokaryotes and eukaryotes genomes. The classification is made via a co-evolutionary ANN method and a deep learning approach. The tool can predict promoter sequences from five different organisms: humans, rats, *Arabidopsis thaliana*, and both *E. coli* and *Bacillus subtilis*. The authors found promising results in identifying important motifs in biological sequences using a deep learning approach on co-evolutionary ANNs. They highlight a sensibility, specificity, and correlation coefficient of 90%, 96%, and 84% respectively when predicting *E. coli* promoters.

BacPP tool [11] focus is on recognizing and predicting *E. coli* promoters in accordance to each promoter's associated σ factor. The tool's approach involves applying extracted rule values of the trained ANN in a separate way for each σ factor. Overall, the accuracy of the tool for each σ factor is: 86.9% for σ^{24}, 92.8% for σ^{28}, 91.5% for σ^{32}, 89.3% for σ^{38}, 97.0% for σ^{54} and 83.6% for σ^{70}.

BTSSfinder [13] tool uses two different elements on its ANN approach: (*i*) a window of 251 nucleotides, and (*ii*) the classification of a possible TSS at the 201st nucleotide, based on data extracted during their ANN training. The approach is based on prediction models of promoters associated with the factors σ^{70}, σ^{38}, σ^{32}, σ^{28} and σ^{24} from *E. coli* and the σ^{A}, σ^{C}, σ^{H}, σ^{G} e σ^{F} factors from cyanobacteria. According to the authors, sensibility for *E. coli* σ factors ranges from 86% to 92%, depending on the σ factor. For cyanobacteria, it varies from 72% to 92%.

BPROM [12] is a σ^{70} promoter recognition tool that achieves ~80% of accuracy and specificity values. The tool applies a Linear Discriminant Function (LDF) to combine information of functional motifs and oligonucleotide composition – features of promoter sequences. To achieve this, the authors used a PWM of five conserved regions of a promoter sequence: the sequences located on the -10 and -35 regions regulated by the σ^{70}; sequences with a length of 7 nucleotides at the following positions: -60 to -40; -11 to +10, and sequences with a length of 5 nucleotides located on -31 to -22.

iPro70-PseZNC [14] uses the Z-curve method of analysis, where genomic information and characteristics of nucleotides (calculated by the frequency of each nucleotide) are mapped on a tridimensional model. The tool is built upon a model named "multi-window Z-curve", representing the tridimensional characteristics of a given promoter sequence. Regarding promoter prediction, the authors report an accuracy of ~90%. They applied an SVM and a reference data set composed of 741 σ^{70}-related sequences obtained on the RegulonDB [20]. The negative sample was composed of 1400 sequences were extracted randomly from coding and intergenic regions of *E. coli*.

PePPer [16] is a prediction, mining, and visualization of prokaryotic Transcription Factor Binding Sites (TFBSs) tool. It includes an all-in-one method for transcription factors, TFBSs, promoters, and regulons. The promoter prediction is based on PWM models and Hidden Markov Models (HMM) of motifs (-35 and -10 regions). The authors do not present any information regarding the tool's performance in its paper.

NNPP2 [15] uses an ANN model known as Time-Delay Neural Network to incorporate promoter's elements that presents variable gaps between them. It consists mainly of two layers: one to recognize the TATA-box (5' – TATAAT – 3') and one to recognize the transcription initiator. When tested on the alcohol dehydrogenase gene from the *Drosophila melanogaster* genome, the tool achieved a recognition score of 75%, with a false positive rate on the basis of 1/547. According to the authors, the tool was built upon the *D. melanogaster* example. However, the authors mention that the tool can be applied to any sequence from both eukaryote and prokaryote genomes. Burden et al. (2005) mentions that the training of the tool [15] was also made using 272 *E. coli* promoters, although this data was not published. Due to this last information, the tool is eligible for our analysis.

PromPredict [17] is based on the stability difference of coding and promoter regions. The algorithm calculates the stability difference (ΔGº) between these regions through the division of a sequence in overlapping windows of 15 nucleotides. PromPredict [17] is the only tool that we studied that does not use machine learning techniques. According to the authors, the tool, when applied to *E. coli* sequences, shows an overall sensibility of 90%, and accuracy of 35%.

2.3. Tool Analysis

We divided the tool analysis into three segments: tool's features, available resources, and prediction comparison.

Regarding features, the criteria applied to create an empirical comparison of the tools took into consideration four aspects: (*i*) the content of the help section: if it includes examples and explanations of the tool's resources; (*ii*) any detailed explanation about how the results should be interpreted; (*iii*) tool's capacity of reporting errors on data sets or during its execution, informing why the tool was not able to continue its execution, and (*iv*) the tool's design regarding modern resources, such as having a responsive screen and technologies to meet usage demands.

About the available resources, the analysis was based on the following aspects: (*i*) availability to perform multiple executions per day; (*ii*) analysis of multiple sequences per execution; (*iii*) maximum nucleotide size of sequences that the tool accepts; (*iv*) input data upload, and (*v*) download of results. The intention of such an analysis is not to arbitrarily argue towards one tool in detriment of another, but to summarize their features in a technical way.

The analysis of the results was divided into: (*i*) the accuracy of each tool's prediction, followed by a comparison between each tool's results (i.e., if a promoter was found by one or more tools); and (*ii*) validation of the promoter sequences from step *i* by pairing the prediction results with the data extracted from RegulonDB for both location and sequence integrity. Figure 1 presents a workflow of the described methodology.

No time complexity analysis of the tools was considered due to the main focus of them being on classification and recognition of promoter sequences, not complete genome scans – where the said analysis would be best suited.

3. Results and Discussion

Initially, an evaluation of each tool's features and resources will be presented, followed by an analysis of the prediction capacity of each tool.

3.1. Tool's Features

The evaluation of the tool's features is summarized in Table 2. It can be observed that PePPer presents no web page with a help section included. With exception of PePPer, assistance material or tutorial with examples can be found.

Table-1. The features found on the analyzed tools.

Tool	Tool's Features				Tool's Resources				
	Usage help	Results help	Support for errors	Page project	Searches per day	Amount of sequences	Size of sequences	Input by file	Save results
BacPP	✔	✔		✔	✔	> 10	< 2.000	✔	✔
BPROM	✔	✔		✔		1	> 25.000	✔	
bTSSfinder		✔		✔	✔	> 10	< 500	✔	✔
CNNPromoter_b	✔	✔		✔		> 10	> 25.000	✔	
iPro70-PseZNC	✔		✔		✔	> 10	< 500		
NNPP2	✔				✔	> 10	> 25.000		
PePPer				✔	✔	> 10	> 25.000		✔
PromPredict	✔	✔			✔	> 10	< 10.000		✔

Regarding the approach of each tool to sequence input, iPro70-PseZNC verifies if the input sequences present a valid format, while the other tools execute even if the sequence format is not valid. All the tools allow the input of sequences on its web page and, additionally, the tools BPROM, BacPP, bTSSfinder, and CNNPromoter_b offer the possibility of uploading files. The tools that offer the possibility of downloading the results are: bTSSfinder, PePPer, PromPredict, and BacPP.

With exception of BPROM and CNNPromoter, unlimited daily access is allowed, some of the requiring the user to register. BPROM and CNNPromoter has a limitation of 15 searches per academic domain per day.

3.2. Prediction Analysis

We opted to exclude BPROM and CNNPromoter_b from this section of the analysis due to the limited executions availability (Table 2). In order to adequate our dataset of intergenic regions to the maximum number of nucleotides allowed by each tool, we fragmented the dataset accordingly. After the analysis, the intergenic regions were reattached.

Beforehand, RegulonDB data reveals that 70 promoter sequences can be found within the 206 promoter regions. The distribution of promoters among the intergenic regions follow: most of the regions contain a single promoter (62/70 promoters), one intergenic region contained three promoters, while four regions contained two promoters.

We fed the six selected tools with the 62 intergenic regions containing a single promoter (Figure 2). BacPP predicted 60 promoters (96.77%), while iPro70-PseZNC predicted 14 out of 62 (22.58%). Intermediary values were obtained for the remaining tools, with good performance values demonstrated by PromPredict and bTSSfinder, and good results by NNPP2.

Intriguingly, NNPP2 is the only tool that reportedly was not trained using RegulonDB data. Burden et al. (2005) only mention a cross-validated dataset of 272 *E. coli* promoters. This avoids biased results in our analysis using RegulonDB data as well, since no clear advantage is given to one or more tools due to similarities between their training datasets and our analysis dataset. Yet, NNPP2 achieved ~92% correctly predicted promoters.

Following the performance analysis, we verified if a given promoter sequence was predicted equally among the tools (Table 3). To clarify this, take as example the 60 predicted promoters by BacPP. From them, 47 were also predicted by bTSSfinder, and only 12 by iPro70-PseZNC. Although the three tools were built on ANN approaches and trained using RegulonDB datasets, no uniformity was found in their prediction. It is possible that the different dataset versions may explain the discrepancies (Table 3), however that isn't self-explanatory per se. It would be wise to search for further explanations on the data's own heterogeneity, as well as on the approaches that each tool employed.

iPro70-PseZNC performance values are also intriguing. Although projected to identify σ^{70}-related promoters, its prediction fell below the expected when comparing it with tools that were not built upon specific σ factors datasets. Pairing iPro70-PseZNC with BacPP, they share 12 predicted promoters, and zero between iPro70-PseZNC and bTSSfinder.

Table-3. Prediction similarity among tools by pairing their promoter prediction results

	BacPP	bTSSfinder	iPro70 PseZNC	NNPP2	PePPer	PromPredict
BacPP	60 (100%)	47/60 (78%)	12/60 (20%)	56/60 (93%)	24/60 (40%)	38/60 (63%)
bTSSfinder	47/47 100%	47 (100%)	0	46/47 (97%)	24/47 (51%)	38/47 (80%)
iPro70-PseZNC	12/14 85%	0	14 (100%)	10/14 (70%)	0	0
NNPP2	56/57 98%	46/57 80%	10/57 17%	57 (100%)	24/57 (42%)	38/57 (66%)
PePPer	24/24 100%	24/24 100%	0	24/24 100%	24 (100%)	21/24 (87.5%)
PromPredict	38/39 100/%	38/39 97%	0	38/39 97%	21/39 53%	39 (100%)

On the other hand, NNPP2 and BacPP shares 56 predicted promoters (Table 3), even though the training of the NNPP2 was not made using RegulonDB data. It also shares 46 promoters with bTSSfinder, ten with iPro70-PseZNC, 21 with PromPredict, and 24 with PePPer. PromPredict, the only tool that does not built under a machine learning-related approach, identified 39 promoters.

Additionally, a closer look at the region indicated as containing a promoter by each tool is not shared among them. For instance, Figure 3 shows one intergenic region that precedes the *ampG* gene with a size of 459 nucleotides. For comparison purposes, we used the promoter's location indicated by RegulonDB: from position 356 to 437 (highlighted in Figure 3). Each tool identified different fragments of the intergenic region as the promoter itself, with significant mismatches from the intended promoter. Surprisingly, iPro70-PseZNC did not identify any segment of the 356-437 range as a promoter, even as it is known to be a classic σ^{70} promoter.

This comparison was carried out for all the 62 single promoters found within intergenic regions by matching predicted sequences' start and end nucleotide positions given by a tool with the location indicated by RegulonDB (Figure 4).

NNPP2 and BacPP obtained matches between prediction and model RegulonDB data above 70%, with NNPP2 achieving the highest values (85.96%). On the other hand, the remaining tools were unable to surpass an overall match of ~40%. The natural features of promoter sequences can explain these results. They are often perceived as short-sized and AT-rich sequences, presenting low degrees of conservation (i.e. the motifs located on the -10 and -35 regions often are not found to be in accordance with their canonical motif - both in structure and position) [3]. For instance, promoter sequences related to the σ^{54}, instead of presenting its motifs located on the -10 and -35, are commonly found in the -12 and -24 regions. Therefore, the various features found in promoter sequences poses one of the many challenges that promoter prediction and recognition faces [3].

Additionally, terminator sequences – which also are relevant elements found within the intergenic regions – are similarly AT-rich sequences. Therefore, it is possible that promoter prediction tools may incorrectly classify them as promoters, resulting in a false-positive. It has also been related that some promoter prediction tools may accept non-promoter sequences that carries the classical TATA sequence [21]. Mishra, *et al.* [8] also mentions novel, not yet fully comprehended difficulties arising for promoter prediction efforts, such as overlapping coding regions, short intergenic regions between genes, and multiple TSSs. Alongside this, the knowledge of transcription initiation from locations other than the TSS (known as pervasive transcription) [22], further expands the list of factors that should be taken into account when the objective is to predict promoters.

4. Conclusions

This paper had the objective of analyzing open-access and online tools focused on the prediction of promoter sequences. The majority of the tools present friendly and straightforward web pages, which contribute to a more agile execution by its user. However, there is a lack of user support in case of errors, since many tools accept invalid characters (which results in poorly formatted or invalid outputs and, sometimes, without delivering any form of feedback about the errors). As described in the methodology, 62 intergenic regions were identified as containing at least one promoter. However, none of the tools were able to predict this scenario correctly in all the 62 regions. Although many of the tools have been built using RegulonDB [20] data for training or validation of results – and also being described in the literature as achieving high-performance values (table 1), it was expected that the tools would identify at least the 62 regions.

The discrepancies found between the results of each tool stimulates the continuation of researches that focuses on enhancing the performance of prediction tools. With such a diversity of results, it would be interesting to use all tools together to obtain close to the ideal results, as the location of a promoter sequence should be consensual when different prediction tools are employed the task. Moreover, the efforts in promoter prediction are still short on being able to efficiently predict promoters of a broad range of microorganisms. An increase in the complexity of data that is fed to prediction tools (e.g., multiple features of a promoter sequence, besides nucleotide composition, curvature, and stability) could possibly tackle this problem. With this concept in mind, a web service is under development, focusing on the integration of regulatory elements tools. Improving promoter prediction tools is beneficial to further bridge the gap between genomic data generation and analysis in the postgenomic era. More automation in a process that is currently behind data generation may speed our capacity to gather and study meaningful data in the many fields that compose biological and medical research. Overall, an expansion in the capacity to discover novel genes and comprehend transcriptional mechanisms in an organism may also be seen. Promoter identification is one of the most important steps in genome annotation (and one of the most difficult tasks), making every effort towards improvements in this field not only a necessity, but a very encouraged endeavor.

Acknowledgements

We thank the University of Caxias do Sul for the support throughout the development of this research.

References

[1] He, W., Jia, C., Duan, Y., and Zou, Q., 2018. "70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features." *BMC Systems Biology*, vol. 12, pp. 4-9.

[2] Barrios, H., Valderrama, B., and Morett, E., 1999. "Compilation and analysis of sigma(54)-dependent promoter sequences." *Nucleic Acids Res.,* vol. 27, pp. 4305–13. Available: https://doi.org/10.1093/nar/27.22.4305

[3] Dall'Alba, G., Casa, P. L., Notari, D. L., Adami, A. G., Echeverrigaray, S., and de Avila, e. S., S., 2019. "Analysis of the nucleotide content of Escherichia coli promoter sequences related to the alternative sigma factors." *Journal of Molecular Recognition*, vol. 32, p. e2770.

[4] Krebs, J. E., Goldstein, E. S., and Kilpatrick, S. T., 2017. *Genes XII*. 12Th Ed. ed. Jones and Bartlett Publishers.

[5] Coelho, R. F., Dall'Alba, G., de Avila, e. S. S., Echeverrigaray, S., and Delamare, A. P. L., 2020. "Toward algorithms for automation of postgenomic data analyses: Bacillus subtilis promoter prediction with artificial neural network." *Omics: A Journal of Integrative Biology*, vol. 24,

[6] Payne, S. R., Pau, D. I., Whiting, A. L., Kim, Y. J., Pharoah, B. M., Moi, C., Boddy, C. N., and Bernal, F., 2018. "Inhibition of bacterial gene transcription with an rpon-based stapled peptide." *Cell Chemical Biology*, vol. 25, pp. 1-8.

[7] Bervoets, I. and Charlier, D., 2019. "Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology." *FEMS Microbiology Reviews*, vol. 43, pp. 304-339.

[8] Mishra, A., Dhanda, S., Siwach, P., Aggarwal, S., and Jayaram, B., 2020. "A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics." *Bioinformatics*, vol. 36, pp. 2375-2384.

[9] Lin, H., Deng, E., Ding, H., Chen, W., and Chou, K., 2014. "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition." *Nucleic Acids Research*, vol. 42, pp. 12961–12972.

[10] Umarov, R. K. and Solovyev, V. V., 2017. "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks." *PLoS ONE.*, vol. 12, p. e0171410.

[11] de Avila, e. S. S., Echeverrigaray, S., and Gerhardt, G. J., 2011. "BacPP: Bacterial promoter prediction: a tool for accurate sigma-factor specific assignment in enterobacteria." *Journal of Theoretical Biology*, vol. 287, pp. 92–99.

[12] Solovyev, V. and Salamov, A., 2011. *Automatic annotation of microbial genomes and metagenomic sequences. In: Li RW (ed), Metagenomics and its applications in agriculture, biomedicine and environmental studies*. New York: Nova Science Publishers, Hauppauge. pp. 61–78.

[13] Shahmuradov, I. A., Razali, R. M., Bougouffa, S., Radovanovic, A., and Bajic, V. B., 2017. "bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia Coli." *Bioinformatics*, vol. 33, pp. 334–40.

[14] Lin, H., Deng, E., Ding, H., Chen, W., and Chou, K., 2017. "Identifying sigma70 promoters with novel pseudo nucleotide composition." *IEEE/ACM Trans Comput Biol Bioinform*, vol. 16, pp. 1316-1321.

[15] Reese, M. G., 2001. "Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome." *Comput. Chem.*, vol. 26, pp. 51-6.

[16] De Jong, A., Pietersma, H., Cordes, M., Kuipers, O. P., and Kok, J., 2012. "PePPer: a webserver for prediction of prokaryote promoter elements and regulons." *BMC Genomics*, vol. 13, p. 299.

[17] Bansal, M. and Kanhere, A., 2005. "A novel method for prokaryotic promoter prediction based on DNA stability." *BMC Bioinformatics. 6.*, Available: https://doi.org/10.1186/1471-2105-6-1

[18] Notari, D. L., Molin, A., Davanzo, V., Picolotto, D., Ribeiro, H. G., and de Avila, e. S. S., 2014. "IntergenicDB: a database for intergenic sequences." *Bioinformation*, vol. 10, pp. 381-383.

[19] Rangel-Chávez, C., Galan-Vasquez, E., and Martinez-Antonio, A., 2017. "Consensus architecture of promoters and transcription units in Escherichia coli: design principles for synthetic biology." *Molecular Biosystems*, vol. 13, pp. 665-676.

[20] Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muniz-Rascado, L., Garcia-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., *et al.*, 2016. "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond." *Nucleic Acids Res.*, vol. 44, pp. D133–43.

[21] Oubounyt, M., Louadi, Z., Tayara, H., and Chong, K. T., 2019. "Deepromoter: Robust promoter predictor using deep learning." *Frontiers in Genetics*, vol. 10, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6460014/

[22] Wade, J. and Grainger, D., 2014. "Pervarsive Transcription: illuminating the dark matter of bacterial transcriptome." *Nature Reviews Microbiology*, vol. 12, pp. 647-653.

Figure-1. Simplified workflow of the described methodology

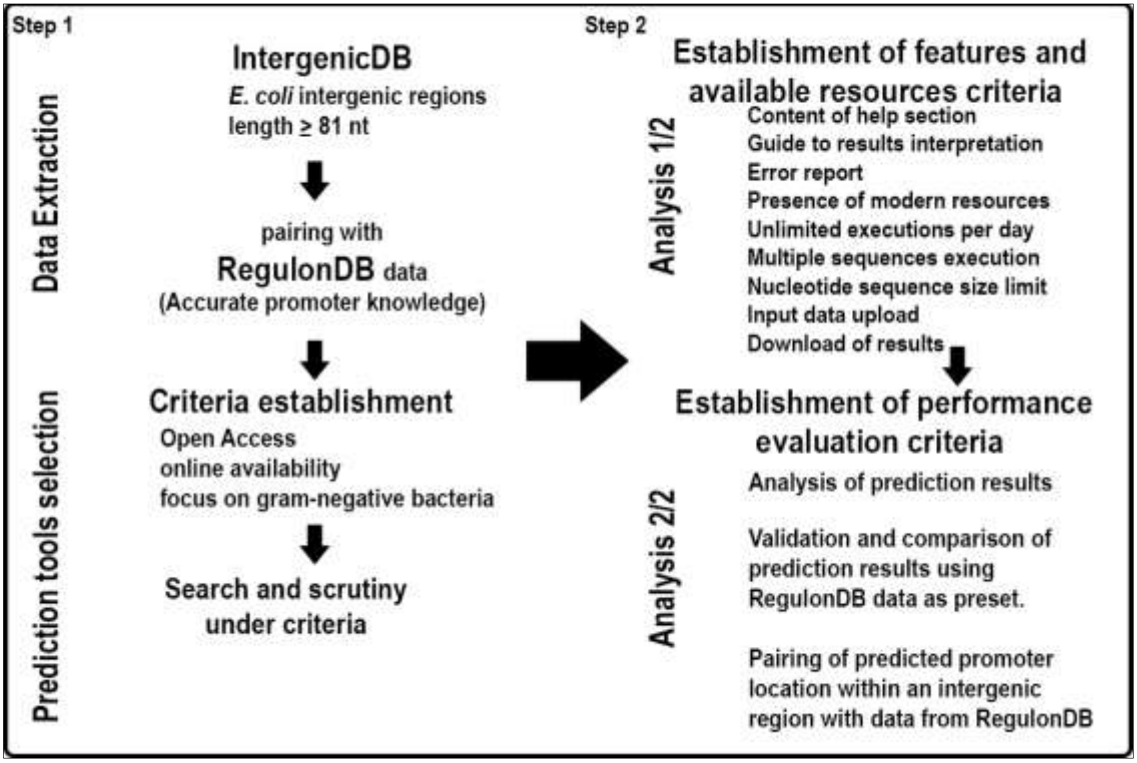

Step 1
Data Extraction
IntergenicDB
E. coli intergenic regions
length ≥ 81 nt
pairing with
RegulonDB data
(Accurate promoter knowledge)
Prediction tools selection
Criteria establishment
Open Access
online availability
focus on gram-negative bacteria
Search and scrutiny
under criteria
Step 2
Establishment of features and
available resources criteria
Analysis 1/2
Content of help section
Guide to results interpretation
Error report
Presence of modern resources
Unlimited executions per day
Multiple sequences execution
Nucleotide sequence size limit
Input data upload
Download of results
Establishment of performance
evaluation criteria
Analysis 2/2
Analysis of prediction results
Validation and comparison of
prediction results using
RegulonDB data as preset.
Pairing of predicted promoter
location within an intergenic
region with data from RegulonDB

Figure-2. Prediction of promoters using 62 intergenic regions–each containing a single promoter

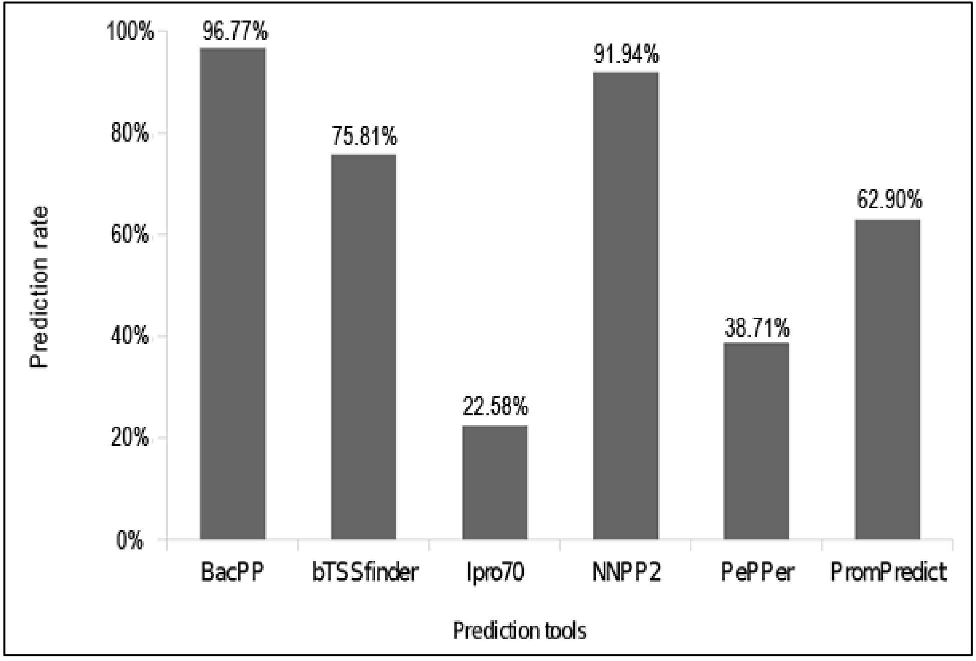


Figure-3. Promoter identification within the intergenic region preceding the ampG gene. The Figure shows the RegulonDB promoter (marked in gray), and the location that each prediction tool pointed out as a possible promoter. It is also shown the score and the exact position inside the intergenic region of each tool

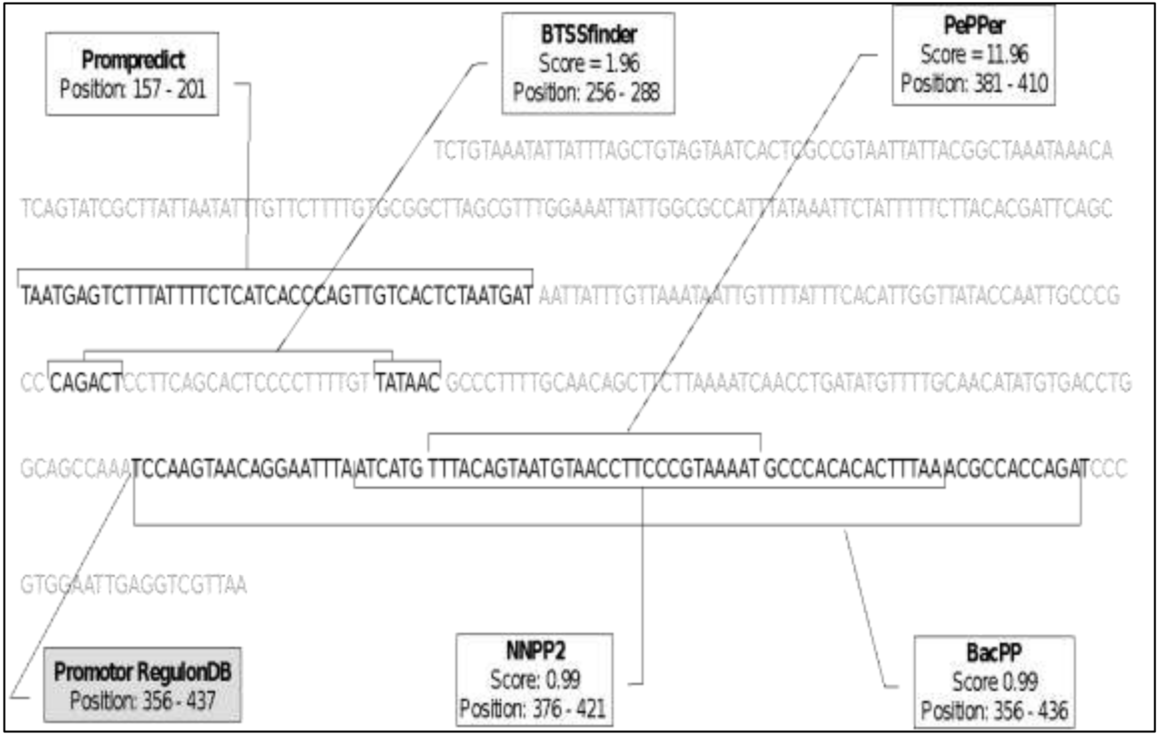
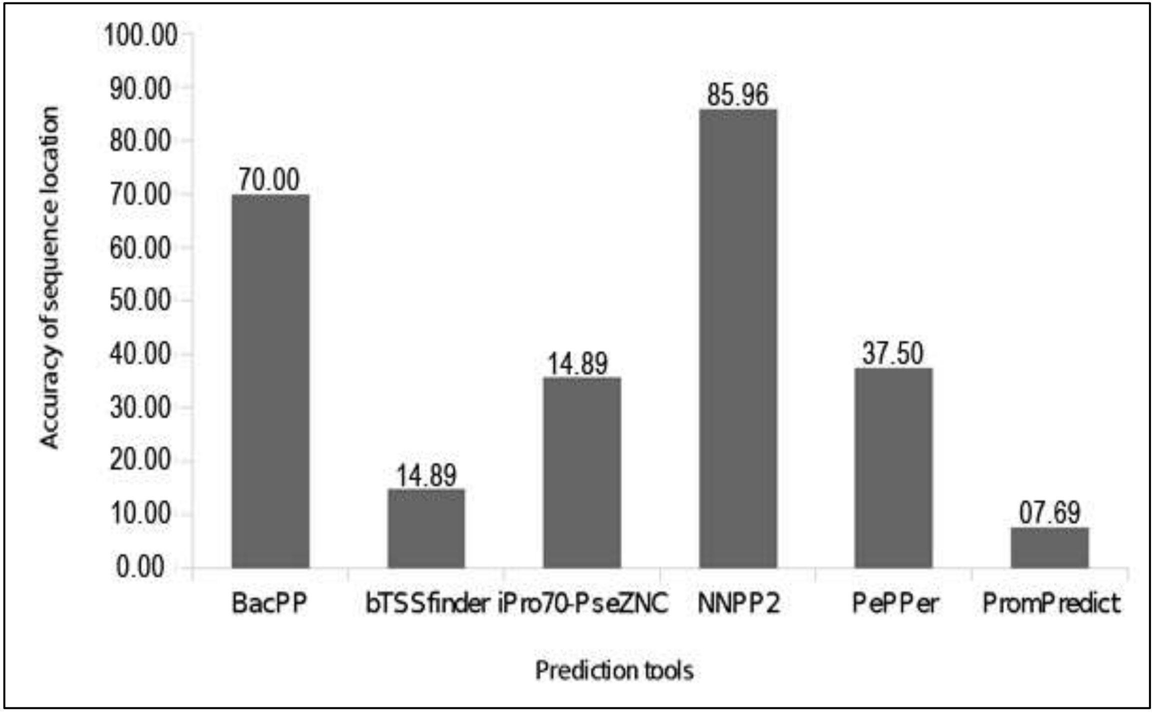

Prompredict
Position: 157 - 201
BTSSfinder
Score = 1.96
Position: 256 - 288
PePPer
Score = 11.96
Position: 381 - 410
TCTGTAAATATTATTTAGCTGTAGTAATCACTCGCCGTAATTATTACGGCTAAATAAACA
TCAGTATCGCTTATTAATATTTGTTCTTTTGTGCGGCTTAGCGTTTGGAAATTATTGGCGCCATTATAAATTCTATTTTTCTTACACGATTCAGC
TAATGAGTCTTTATTTTCTCATCACCCAGTTGTCACTCTAATGAT AATTATTTGTTAAATAATTGTTTTATTTCACATTGGTTATACCAATTGCCCG
CC CAGACTCCTTCAGCACTCCCCTTTTGT TATAAC GCCCTTTTGCAACAGCTTCTTAAAATCAACCTGATATGTTTTGCAACATATGTGACCTG
GCAGCCAAATCCAAGTAACAGGAATTTAATCATG TTTACAGTAATGTAACCTTCCCGTAAAAT GCCCACACACTTTAAACGCCACCAGATCCC
GTGGAATTGAGGTCGTTAA
Promotor RegulonDB
Position: 356 - 437
NNPP2
Score: 0.99
Position: 376 - 421
BacPP
Score 0.99
Position: 356 - 436

Figure-4. Accuracy of sequence location inside intergenic regions. The figure shows the match between the predicted location given by a tool and promoter's original location, indicated by RegulonDB. The results are relative to the number of regions identified by each tool (Figure 2).


Accuracy of sequence location
BacPP
bTSSfinder
iPro70-PseZNC
NNPP2
PePPer
PromPredict
Prediction tools