Computational and Structural Biotechnology Journal 20 (2022) 4969–4974

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

SAPPHIRE.CNN: Implementation of dRNA-seq-driven, species-specific promoter prediction using convolutional neural networks

Lucas Coppens [a,b], Laura Wicke [b,c], Rob Lavigne [b,*]

[a] Department of Bioengineering and Imperial College Centre for Synthetic Biology, Imperial College London, London, UK
[b] Laboratory of Gene Technology, Department of Biosystems, KU Leuven, Kasteelpark Arenberg 21, Box 2462, 3001 Leuven, Belgium
[c] Institute for Molecular Infection Biology (IMIB), Medical Faculty, University of Würzburg, Josef-Schneider-Straße 2, 97080 Würzburg, Germany

ARTICLE INFO

Article history:
Received 6 June 2022
Received in revised form 3 September 2022
Accepted 5 September 2022
Available online 9 September 2022

ABSTRACT

Data availability is a consistent bottleneck for the development of bacterial species-specific promoter prediction software. In this work we leverage genome-wide promoter datasets generated with dRNA-seq in the Gram-negative bacteria *Pseudomonas aeruginosa* and *Salmonella enterica* for promoter prediction. Convolutional neural networks are presented as an optimal architecture for model training and are further modified and tailored for promoter prediction. The resulting predictors reach high binary accuracies (95% and 94.9%) on test sets and outperform each other when predicting promoters in their associated species. SAPPHIRE.CNN is available online and can also be downloaded to run locally. Our results indicate a dependency of binary promoter classification on an organism's GC content and a decreased performance of our classifiers on genera they were not trained for, further supporting the need for dedicated, species-specific promoter classification tools.

1. Introduction

As key drivers of the process of transcription, promoter sequences represent fundamental genetic features across all domains of life. Consequently, computational tools to predict promoter features from raw DNA sequences of sequenced genomes have become a key area of study within the field of bioinformatics. Over the past two decades, efforts have been made to develop effective software for the task of promoter prediction in prokaryotes. For the development of the most cited prokaryotic promoter prediction tool, BPROM, Solovyev et al. [13] applied a linear discriminant analysis, relying on conserved features in promoter sequences, most notably the well characterised −35 and −10 sequence motifs of σ70 promoters. These conserved motifs were encoded in position weight matrices, representing the prevalence of each nucleotide at each position in a set of *Escherichia coli* promoters. BacPP, another popular prokaryotic promoter classifier, aimed to leverage neural networks for the task of promoter prediction [3]. Neural network models provide improved complexity compared to linear models of position weight matrices, but do require substantially more data to be trained. Interestingly, despite limited data, BacPP also provided tools for the prediction of σ24, σ28, σ32, σ38 and σ54 in addition to σ70 promoters. Many other creative approaches have been developed for prokaryotic promoter prediction. For instance, propensity for stress-induced DNA duplex destabilisation was found to be a good predictor of specific promoter regions [15]. Furthermore, a variety of machine learning models other than traditional neural networks like the ones in BacPP have been leveraged for prokaryotic promoter prediction, such as Random Forests [8], Support Vector Machines [5], Convolutional Neural Networks [14] and a Capsule Network [9].

Besides a few works which also covered promoter prediction in *Bacillus Subtilis* [14,15], to our knowledge, no promoter classification research exists for the majority of prokaryotic genera. Shahmuradov et al. [10] had previously highlighted this limitations for other prokaryotic genera and developed bTSSfinder, enabling promoter prediction in cyanobacteria [10]. Similarly, for the prediction of promoter sequences within the genus *Pseudomonas* we have previously developed SAPPHIRE, a neural network which was trained on a limited set of −35 and −10 promoter motifs [2]. Nevertheless, a shortage of qualitative training data has consistently proven a major bottleneck for the development of tools for species-specific promoter prediction.

The emergence of high-throughput sequencing approaches provides the potential to force a breakthrough in this regard. Genome-wide and experimentally verified promoter data can be generated through by the dRNA-seq method, which relies on enriching

* Corresponding author.
E-mail address: rob.lavigne@kuleuven.be (R. Lavigne).

https://doi.org/10.1016/j.csbj.2022.09.006

primary transcripts prior to sequencing [11,12]. We here propose that curated dRNA-seq data serve as valuable training data to develop promoter prediction tools for prokaryotic genera which currently lack predictive software. In this manuscript, we leverage dRNA-seq datasets for the development of SAPPHIRE.CNN, implementing species-specific σ70 promoter prediction models for *Pseudomonas aeruginosa* and *Salmonella enterica*, respectively. Our results indicate that models trained on the data of one bacterial species lack accuracy in the prediction of promoters in other species, confirming the need for species-specific promoter classifiers. In addition to the development and publication of two new promoter predictors, our methods are publicly available so that they can be implemented for the development of classifiers using custom data.

2. Model development

2.1. Data

Datasets of 3,066 manually curated transcription start sites (TSS) for *P. aeruginosa* and 3,583 TSS for *S. enterica* were retrieved from works in which the dRNA-seq technique was applied to obtain genome-wide TSS data [7,16]. Per base average sequence content plots of the regions upstream of the TSSs showed strong deviations from average nucleotide contents around the −35 and −10 regions, hinting the presence of promoter motifs (Fig. 1A). Motifs elucidated from these regions were found to match known consensus sequences at the −35 and −10 positions of prokaryotic promoters in the σ70 family. These −35 and −10 motifs were used as Position-Specific Scoring Matrices (PSSMs, Supplementary Tables 1 and 2) to evaluate all regions upstream of TSSs in *P. aeruginosa* and *S. enterica* for their similarity to the σ70 consensus sequence. A scatter plot of these motif scores revealed a large, distinct cluster of promoter regions with high sequence similarity to the −35 and −10 consensus sequences (Fig. 1B and 1C). DBSCAN clustering with an epsilon value that resulted in an optimal DBCV score was used to assign promoter sequences to this cluster, yielding a set of 2113 putative σ70 family promoters for *P. aeruginosa* and 2,928 for *S. enterica*. These sequences were used as positive examples for training of the promoter predictions models. For each species, 10,000 background sequences were obtained by randomly selecting 5,000 coding and 5,000 non-coding sequences from their respective genomes, avoiding overlap with any of the experimentally determined promoter sequences. Training sequences were one-hot encoded to allow feeding into neural networks. One-hot encoding represents the nucleotides A, C, G and T as binary codes of zeros and ones, which is a necessary conversion when using DNA sequences as input for neural network models.

2.2. Model design and training

Five neural network architectures were hand-designed and tested to identify a suitable architecture for our model. The Python3 keras package was used to construct these networks [1]. The five networks included tree conventional architecture types: a traditional fully connected neural network, a convolutional neural network (CNN) with one layer of convolutional kernels and a recurrent neural network (RNN) using one layer of LSTMs. Furthermore, two combinations of the latter two were examined: a network with a convolutional layer followed by a layer of LSTMs (CNN-LSTM) and a network with an LSTM layer, a convolutional layer, and another LSTM layer (LSTM-CNN-LSTM). In each of the evaluated networks, the number of nodes per layer varies between 10 and 30, layer sizes appropriate for the length of the input sequences which was tentatively set to 45. The Rectified Linear Unit (ReLU) activation function was chosen for all network nodes except the final node in each network. The ReLU activation function works particularly well for deep networks, especially for supervised tasks with large labeled datasets [4]. The final node in each network has a sigmoid activation function for the binary classification of sequences as non-promoter or promoter. For convolutional layers, a kernel size of 6 was chosen to match the lengths of the known six-base "TTGACA" −35 and "TATAAT" −10 promoter motifs. The Python3 code for these five architectures and their parameters can be found online at https://github.com/LoGT-KULeuven/SAPPHIRE_CNN_model_development. The potential of each architecture to classify *Pseudomonas aeruginosa* and *Salmonella enterica* promoter sequences was evaluated using fivefold cross-validation, in each iteration retaining the values for sensitivity and specificity on the validation set corresponding to the lowest loss encountered during training. The results of these model evaluations on the *P. aeruginosa* and *S. enterica* datasets are shown in Fig. 2A and 2B. The fully connected neural network was significantly outperformed by all other architectures. The standard CNN

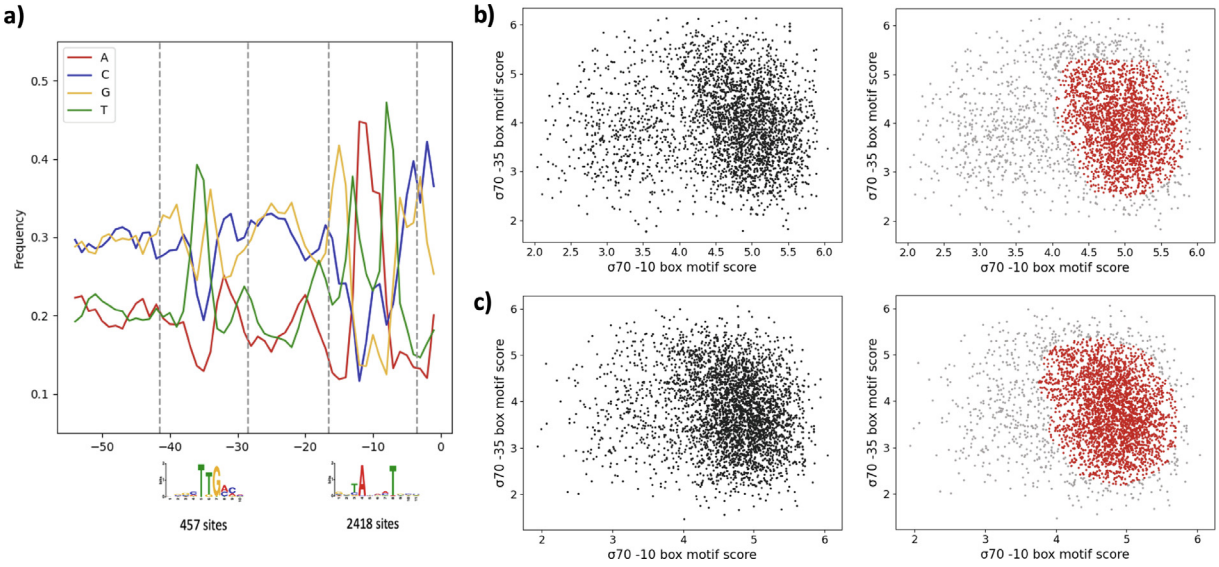

Fig. 1. a) Per base average sequence content of regions upstream of TSSs and σ70 promoter motifs found in these regions. b) σ70 motif scores and clustering for all TSSs obtained by dRNA-seq for P. aeruginosa. c) σ70 motif scores and clustering for all TSSs obtained by dRNA-seq for *S. enterica*.

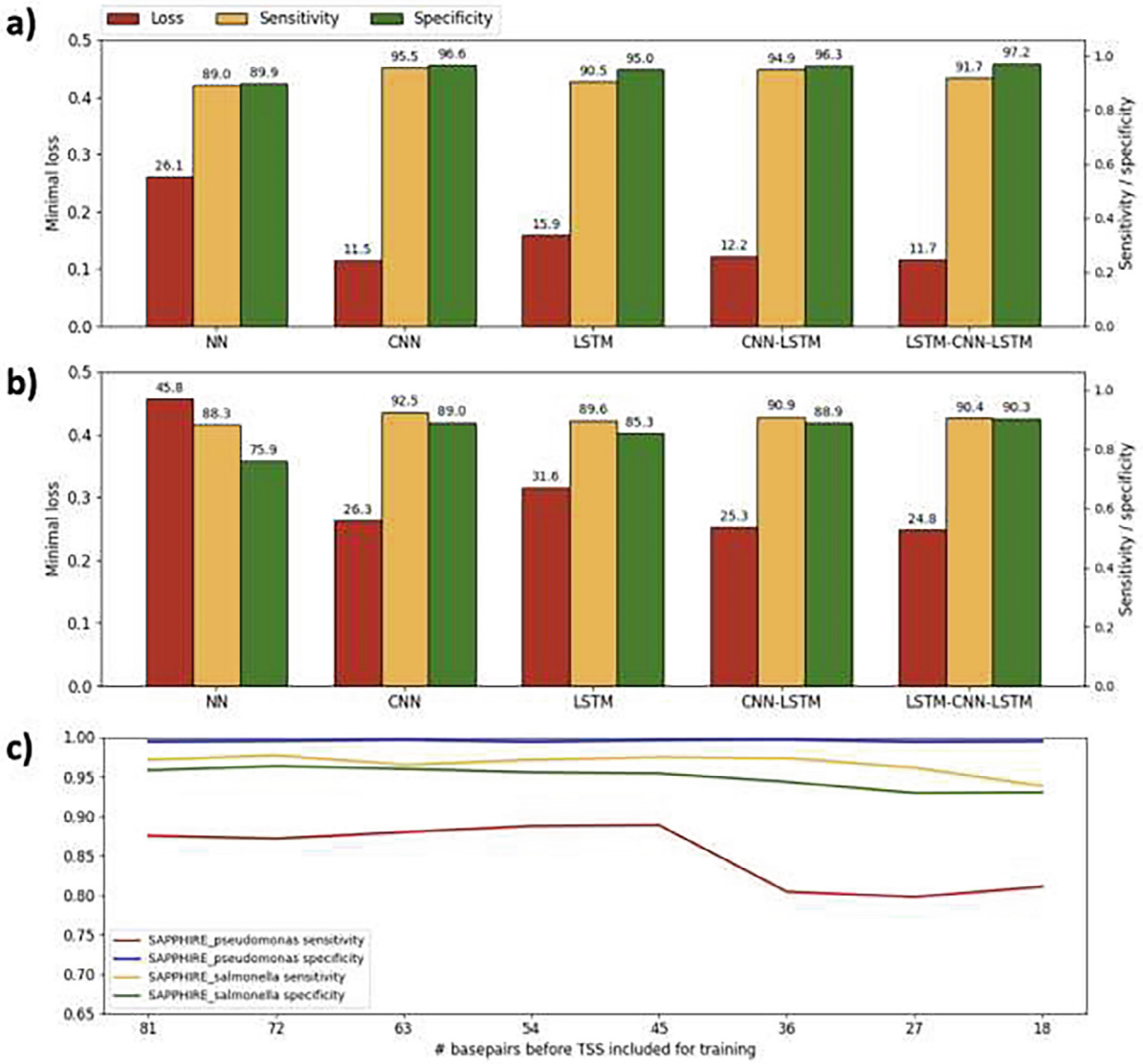

Fig. 2. a) Lowest loss and corresponding sensitivity and specificity achieved on the validation set encountered during training for five different types of neural networks on the *P. aeruginosa* dataset. b) Lowest loss and corresponding sensitivity and specificity achieved on the validation set encountered during training for five different types of neural networks on the *S. enterica* dataset. c) Average sensitivity and specificity of multiple iterations of training of CNNs for both species on promoter sequences with various lengths of basepairs included before the TSSs.

showed the optimal overall performance, indicating that further increasing the complexity beyond a CNN with a single convolutional layer did not improve model performance for our datasets. The CNN was therefore the architecture of choice for subsequent development of the predictors SAPPHIRE.CNN.pseudomonas and SAPPHIRE.CNN.salmonella.

The Adam optimiser, a computationally efficient stochastic optimisation algorithm, was used to train the CNNs. The default learning rate of 0.001 of the Adam optimiser as implemented in the keras library was retained. Binary cross entropy was used as loss function, as it is well suited for binary classification problems. A validation set of 10% of the training data was separated to assist in training, to retain the model from the epoch with highest sensitivity and specificity on the validation set. A maximum of 250 epochs was used, which is comfortably larger than the number of epochs that was required for the sensitivity and specificity of the models on training and validation set to reach a plateau during training (Supplementary figure S3).

Predictive performance of the CNNs did not improve by increasing the length of the promoter sequences for training beyond −45 basepairs with respect to the TSS for training (Fig. 2C). The length of 45 was therefore kept for the training sequences for the models. This length appears to match our current understanding of prokaryotic σ70 promoters, which is centered around DNA motifs in the −35 and −10 locations with respect to the TSS.

3. Evaluation of SAPPHIRE.CNN

We evaluated the performances of SAPPHIRE.CNN.pseudomonas and SAPPHIRE.CNN.salmonella on independent test sets of promoters that were separated from the dRNA-seq sequences before training, as well as a genome-wide *E. coli* dRNA-seq dataset retrieved from the EcoCyc database [6]. The results of this evaluation are shown in Tables 1A and 1B Each of the models performs best on their respective test sets, reaching about 95% binary accuracy. Accuracy decreases for the test sets of species they were not trained for. Interestingly, the sensitivity of SAPPHIRE.CNN.pseudomonas is higher on the.

Table 1A
Performance of SAPPHIRE.CNN.pseudomonas on the different test sets.

	Pseudomonas test set	*Salmonella* test set	*E. coli* test set
Sensitivity	94.5	98.6	78.2
Specificity	95.5	85.7	88.3
Binary accuracy	95.0	92.2	83.3

Table 1B
Performance of SAPPHIRE.CNN.salmonella on the different test sets.

	Pseudomonas test set	*Salmonella* test set	*E. coli* test set
Sensitivity	81.5	95.2	59.0
Specificity	99.3	94.7	95.6
Binary accuracy	90.4	94.9	77.3

Salmonella test set than the *Pseudomonas* test set. Similarly, the SAPPHIRE.CNN.salmonella specificity is higher on the *Pseudomonas* test set than the *Salmonella* test set. This can be explained by looking at how these species-specific models adapted during training to different GC contents of these organisms. Promoter sequences generally have lower GC content than the average GC content of the host organism. Trained promoter classifiers will therefore be more prone to classifying sequences with low GC content as promoters. However, *Salmonella's* GC content (~52%) is about 15% lower than *Pseudomonas'* GC content (~67%) (Fig. 3). Consequently, the SAPPHIRE.CNN.pseudomonas model, trained for higher GC promoters and background sequences, will be prone to classify GC-low *Salmonella* sequences as promoters, resulting in a higher sensitivity yet lower specificity on the *Salmonella* test set. The inverse reasoning explains the low sensitivity and high specificity of the SAPPHIRE.CNN.salmonella classifier on the GC-high *Pseudomonas* test set. These observations further justify the need for species-specific promoter classification software.

To further validate the quality and species specificity of our classifiers, as well as compare them to other promoter classification software, we retrieved all the annotated promoters from four genera of Gram-negative bacteria from the first 100 results that came up after querying for the genus of interest combined with keyword "promoter" on NCBI nucleotide (see accession numbers Supplementary Table S4). We subjected the retrieved sequences to promoter classification by BPROM [13] and BacPP [3], two highly cited predictors for which the online tools are still available and straightforward to use. For BacPP, a cut-off probability of 0.5 for σ70 promoters was used. In addition, the sequences were subjected to the previous version of SAPPHIRE [2], SAPPHIRE.CNN.pseudomonas and SAPPHIRE.CNN.salmonella. The results are shown in Table 2. SAPPHIRE.CNN.pseudomonas and SAPPHIRE.CNN.salmonella outperform the other classifiers across all tested

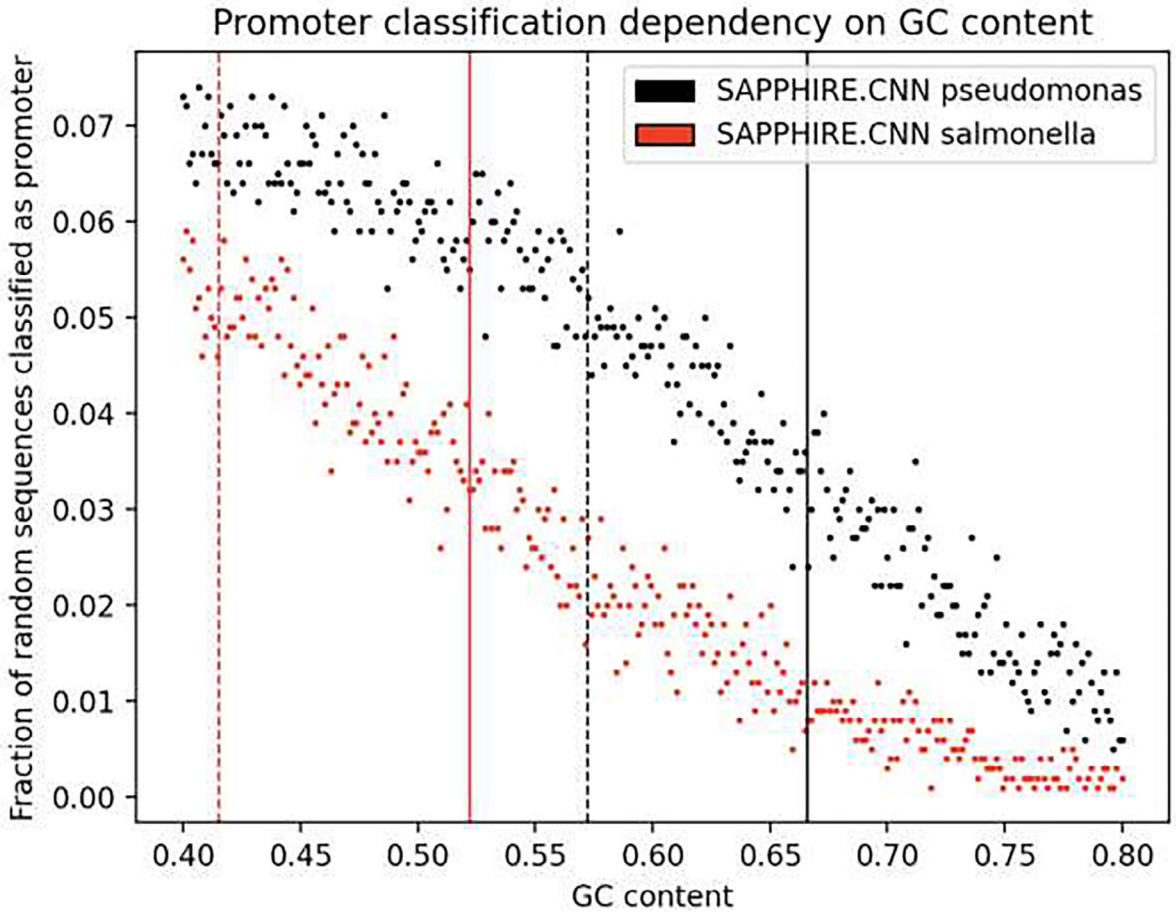

Fig. 3. Promoter classification dependency on GC content. Dots represent how many of groups of 100 randomly generated sequences with a certain GC content are classified as promoters by the respective predictors. Full black line: average GC content of the *P. aeruginosa* genome (PA01, accession: NC_002516). Dashed black line: average GC content of the *P. aeruginosa* promoter sequences used to train SAPPHIRE.CNN.pseudomonas. Full red line: average GC content of the *S. enterica* genome (subsp. enterica serovar Typhimurium str. ST4/74, accession: CP002487). Dashed red line: average GC content of the *S. enterica* promoter sequences used to train SAPPHIRE.CNN.salmonella. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Number of promoters identified by various promoter classifiers in promoter sequences retrieved from NCBI Nucleotide for various Gram-negative genera. For each genus/species, the best performing classifier is highlighted in green.

	BPROM	BacPP	SAPPHIRE	SAPPHIRE.CNN pseudomonas	SAPPHIRE.CNN salmonella
P.aeruginosa (250)	51 (20%)	57 (23%)	124 (50%)	211 (84%)	206 (82%)
P. putida (34)	7 (21%)	3 (9%)	18 (53%)	19 (56%)	19 (56%)
P. syringae (184)	14 (8%)	24 (13%)	6 (3%)	50 (27%)	21 (11%)
Pseudomonas (other) (123)	12 (10%)	16 (13%)	15 (12%)	32 (26%)	24 (20%)
Salmonella enterica (57)	12 (21%)	6 (11%)	36 (63%)	39 (68%)	40 (70%)
Escherichia coli (143)	31 (22%)	47 (33%)	73 (51%)	110 (77%)	100 (70%)
Vibrio (various species) (54)	12 (22%)	38 (70%)	19 (35%)	33 (61%)	33 (61%)

genera except for *Vibrio*, for which BacPP remains the superior tool. Furthermore, SAPPHIRE.CNN.pseudomonas was the best classifier to detect *Pseudomonas* sequences while SAPPHIRE.CNN.salmonella was the best for *Salmonella* sequences. The predictive performance for both new predictors is lower when it comes to predicting sequences for genera and species they were not trained for. This again supports our principal that species-specific classifiers are needed for bacterial species which do not currently have them.

4. Application

The SAPPHIRE.CNN software was written in Python 3.7. A user-friendly browser interface is available (https://sapphire.biw.kuleuven.be/). Input DNA sequences should be at least 45 nucleotides long and should be provided in FASTA file format. Sequences can either be uploaded as a file or pasted directly into the interface. After submission, SAPPHIRE.CNN scans the full length of each sequence for promoters, subsequently re-turning a list of hits and providing the corresponding estimated transcription start site and p-value. Alternatively, the SAPPHIRE.CNN software can be downloaded from the same website, permitting users to run it locally from a command line interface.

5. Conclusion

We presented SAPPHIRE.CNN, comprising the models SAPPHIRE.CNN.pseudomonas and SAPPHIRE.CNN.salmonella, which unlock promoter prediction for two bacterial species currently lacking such tools. We illustrate that genome-wide TSS datasets generated by the dRNA-seq method provide a suitable starting point for the development of such models. CNNs trained on σ70 promoter sequences of 45 basepairs performed well and reached test set accuracies of about 95%. The dependence of promoter prediction on GC content of promoters and background sequences is discussed, suggesting that promoter prediction tools are biased by the GC content of the dataset and therefore organism for which they are trained. Finally, evaluating the models using data sets of different genera showed decreased performance in the genera for which the models were not trained. This observation corroborates the need for species-specific promoter prediction beyond the many tools based on promoter data in *E. coli*. However, the concept of leveraging dRNAseq data for promoter prediction will enable a straightforward scaling towards other species, as well as other promoter motifs beyond σ70. To help researchers create custom promoter prediction models based on their own datasets, the pipeline for the training of neural networks on genomic promoters and background sequences has been made available https://github.com/LoGT-KULeuven/SAPPHIRE_CNN_model_development.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.006.

References

[1] Chollet F. Keras. Github 2015.
[2] Coppens L, Lavigne R. SAPPHIRE: A neural network based classifier for σ70 promoter prediction in Pseudomonas. BMC Bioinf 2020;21(1):1–7. https://doi.org/10.1186/s12859-020-03730-z.

[3] de Avila e Silva S, Echeverrigaray S, Gerhardt GJL. BacPP: Bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria. J Theor Biol 2011;287(1):92–9. https://doi.org/10.1016/j.jtbi.2011.07.017.
[4] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. Proc Fourteenth Internat Conf Artif Intell Stat 2011;15:315–23.
[5] He W, Jia C, Duan Y, Zou Q. 70ProPred: A predictor for discovering sigma70 promoters based on combining multiple features. BMC Syst Biol 2018;12 (Suppl 4). https://doi.org/10.1186/s12918-018-0570-1.
[6] Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res 2017;45(D1):D543–50. https://doi.org/10.1093/nar/gkw1003.
[7] Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, et al. An infection-relevant transcriptomic compendium for salmonella enterica serovar typhimurium. Cell Host Microbe 2013;14(6):683–95. https://doi.org/10.1016/j.chom.2013.11.010.
[8] Liu B, Yang F, Huang DS, Chou KC. IPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 2018;34(1):33–40. https://doi.org/10.1093/bioinformatics/btx579.
[9] Moraes L, Silva P, Luz E, Moreira G. CapsProm: a capsule network for promoter prediction. Comput Biol Med 2022;147(December 2021):105627. https://doi.org/10.1016/j.compbiomed.2022.105627.
[10] Shahmuradov IA, Mohamad Razali R, Bougouffa S, Radovanovic A, Bajic VB. BTSSfinder: A novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. Bioinformatics 2017;33(3):334–40. https://doi.org/10.1093/bioinformatics/btw629.
[11] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 2010;464(7286):250–5. https://doi.org/10.1038/nature08756.
[12] Sharma CM, Vogel J. Differential RNA-seq: The approach behind and the biological insight gained. Curr Opin Microbiol 2014;19(1):97–105. https://doi.org/10.1016/j.mib.2014.06.010.
[13] Solovyev V, Salamov A, Seledtsov I, Vorobyev D, Bachinsky A. Automatic annotation of bacterial community sequences and application to infections diagnostic. In: BIOINFORMATICS 2011 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. p. 346–53. https://doi.org/10.5220/0003333703460353.
[14] Umarov RK, Solovyev VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS ONE 2017;12 (2):1–12. https://doi.org/10.1371/journal.pone.0171410.
[15] Wang H, Benham CJ. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC Bioinf 2006;7:1–15. https://doi.org/10.1186/1471-2105-7-248.
[16] Wicke L, Ponath F, Coppens L, Gerovac M, Lavigne R, Vogel J. Introducing differential RNA-seq mapping to track the early infection phase for Pseudomonas phage ϕKZ. RNA Biol 2021;18(8):1099–110. https://doi.org/10.1080/15476286.2020.1827785.