Prediction of Bacterial Promoter Sequences using Machine Learning

Seongwon Kim[1], Joshua Booth[2], Niharika Vattikonda[3], David W. Aha[4], Leslie N. Smith[4], Trey J. Morris[5], Amina Jackson[1], Dagma H. Leary[1]

[1]Center for Bio/molecular Science and Engineering, Naval Research Laboratory [2]McConnellsburg High School [3]Thomas Jefferson High School
[4]Information Technology Division, Naval Research Laboratory [5]Space Systems Development Department, Naval Research Laboratory

Motivation/Grand Challenge

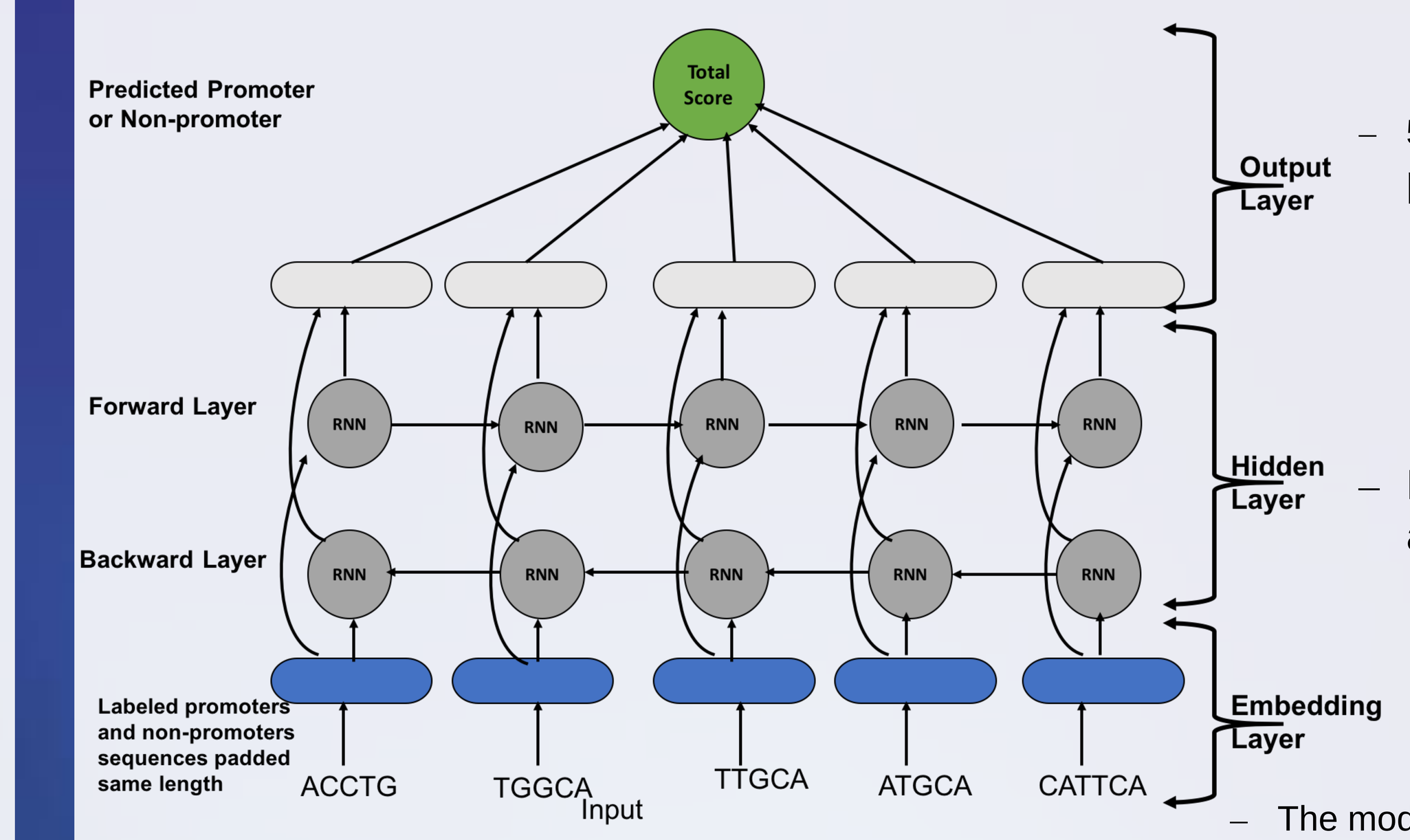
- Knowledge of bacterial promoter sequences can be a valuable resource for synthetic biology effort.
  - Promoters play crucial roles in gene regulation and hence in the production of proteins and metabolites.
  - Known promoters bind a variety of sigma factors that respond to various environmental cues. It is challenging to identify and classify them in non-model organisms.
- Promoters are known to have certain sequence signatures. Their distribution in the genome exhibit more complicated landscape than preciously expected.
  - It is desirable to utilize strong promoters (that enhance corresponding gene expression) in synthetic biology.
- Computational methods for promoter prediction have been mainly focused on human-derived feature extraction.
  - Data exist for several model organisms, but it is challenging to extend them to non-traditional organism. Validation is desired before commitment to experiment.

Project Goals

- In this work, we explore the utility of recent advancements in deep neural network for promoter prediction.
- – They have been shown to be effective in computer vision, speech recognition and natural language processing.
- – Features are learned by the algorithm rather than input by humans (automatic feature extraction)
- – A host of architecture will be examined, with priority on algorithms known to work for text processing.
- – We want to see if deep learning algorithms can extract features that are common in multiple organisms, overcoming the challenge of traditional approaches focused on specific organisms.

Approach

Bi-Directional RNN/LSTM Deep Learning Model



Model Prediction Procedure

- 5 TensorFlow trained models each output their prediction
  - Score weight assignment constitutes confidence in the prediction
  - A mean from the sum of all the 5 scores for a single score per predicted sequence
  - Cluster sequences based on the scores.
- Predicted promoter sequences must score above 80%.

Training and Testing data for model

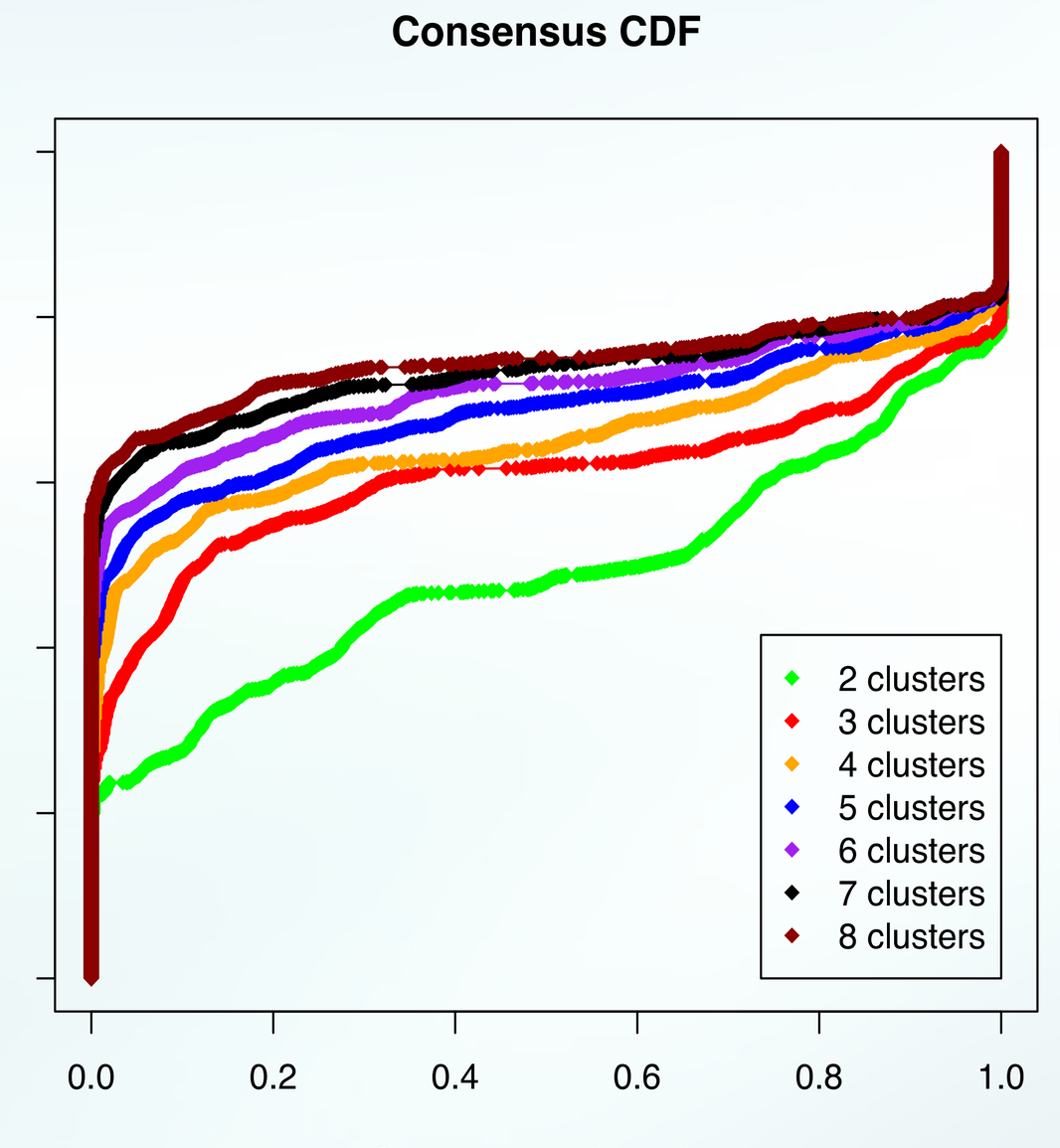
- The model is trained on genome data set of one organism is tested on genome data set of closely related organism.
  - Example: Trained on E. coli K-12 genome data set, tested on E. coli Nissle data set.
- Robustness of the model is trained not more than 50% of the entire dataset on genome of an organism where promoters have been experimentally identified and published.
  - Used 30% of E. coli K-12 genome data set in training, tested on the entire genome for promoter and non-promoter predictions.

Training data pre-processing

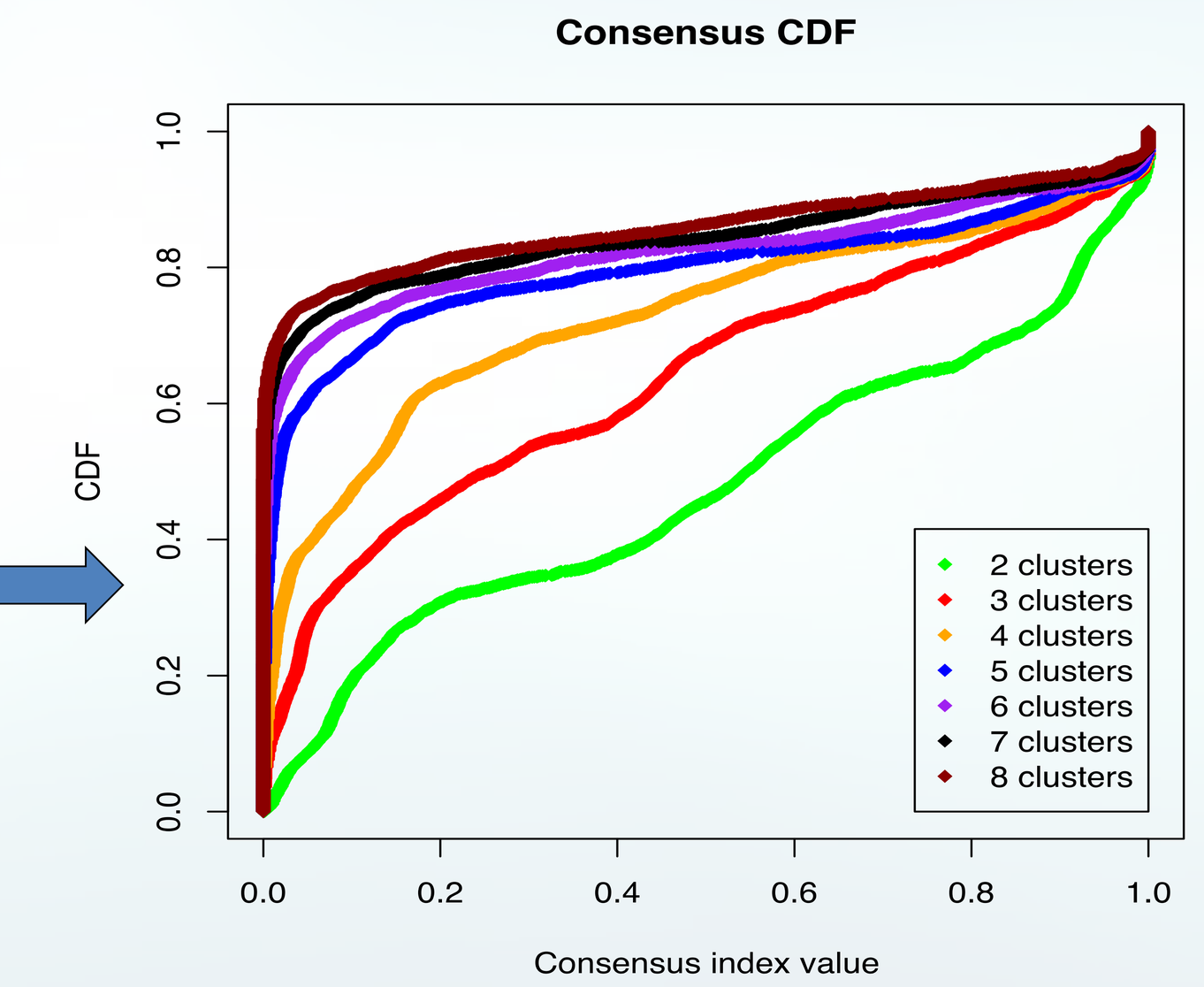
Elements	Parameters
Max sequence length	60 nucleobases
Upstream ORF	>= 200 nucleobases
ORF length	<= 30 nucleobases
Min sequence length	10 nucleobases

Data Normalization for clustering

Before Normalization



After Normalization



Quantile Normalization

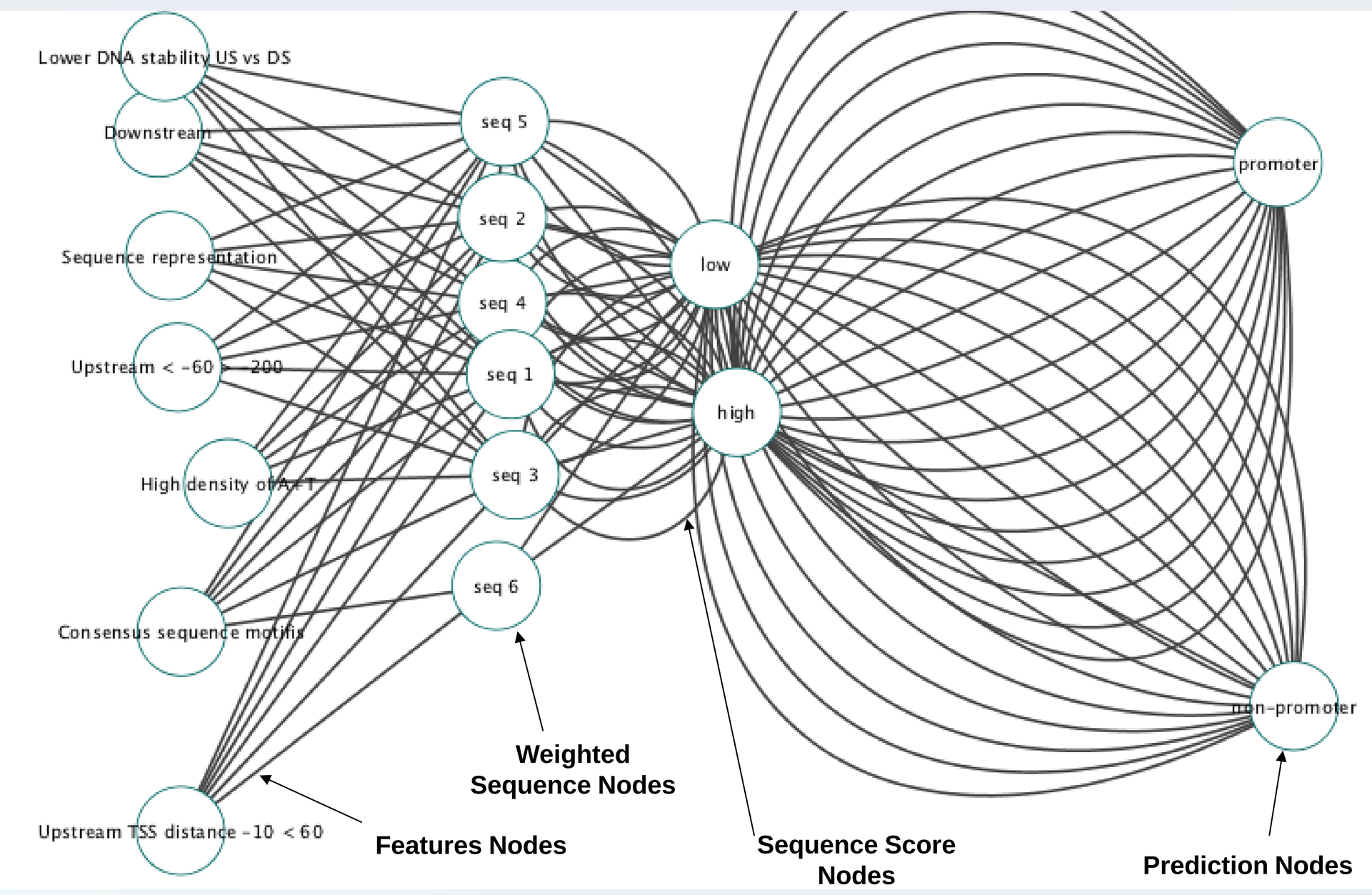
- Used to control biases in the scores
- Gives data identical statistical representation
- Minimizes systematic errors and correlation differences

Conclusion

- Deep learning methodology was applied to the problem of promoter prediction.
- Architectures exhibited different degree of success in terms of transferability of E coli sequences to other organisms.
- Among existing architectures, Bi-directional neural network produced most promising prediction applied to Vibrio strains.
- Recurrent neural networks and Bayesian neural networks ensemble method has demonstrated ability to handle transferability issues between organisms when tested on E.coli and B subtilis
- The findings will be useful in utilizing knowledge of well-studied organisms to learn more about unknown organisms.

Next Steps/Transitions

Bayesian Neural Network Integration



Each sequence starts with the predicted score from RNN neuron. The score is incremented or decremented based on how it maps to the features. The final score determines whether the sequences stays predicted as a promoter or non-promoter.

Feature mapping and scoring system

Feature	Scoring Influence
Upstream transcription start site (TSS) distance < - 60	+
Representation	+
Downstream	0
High density A+T upstream	+
Upstream < 200	-
Inside TSS	-
Lower DNA stability upstream Vs. downstream	-
Consensus sequences	+

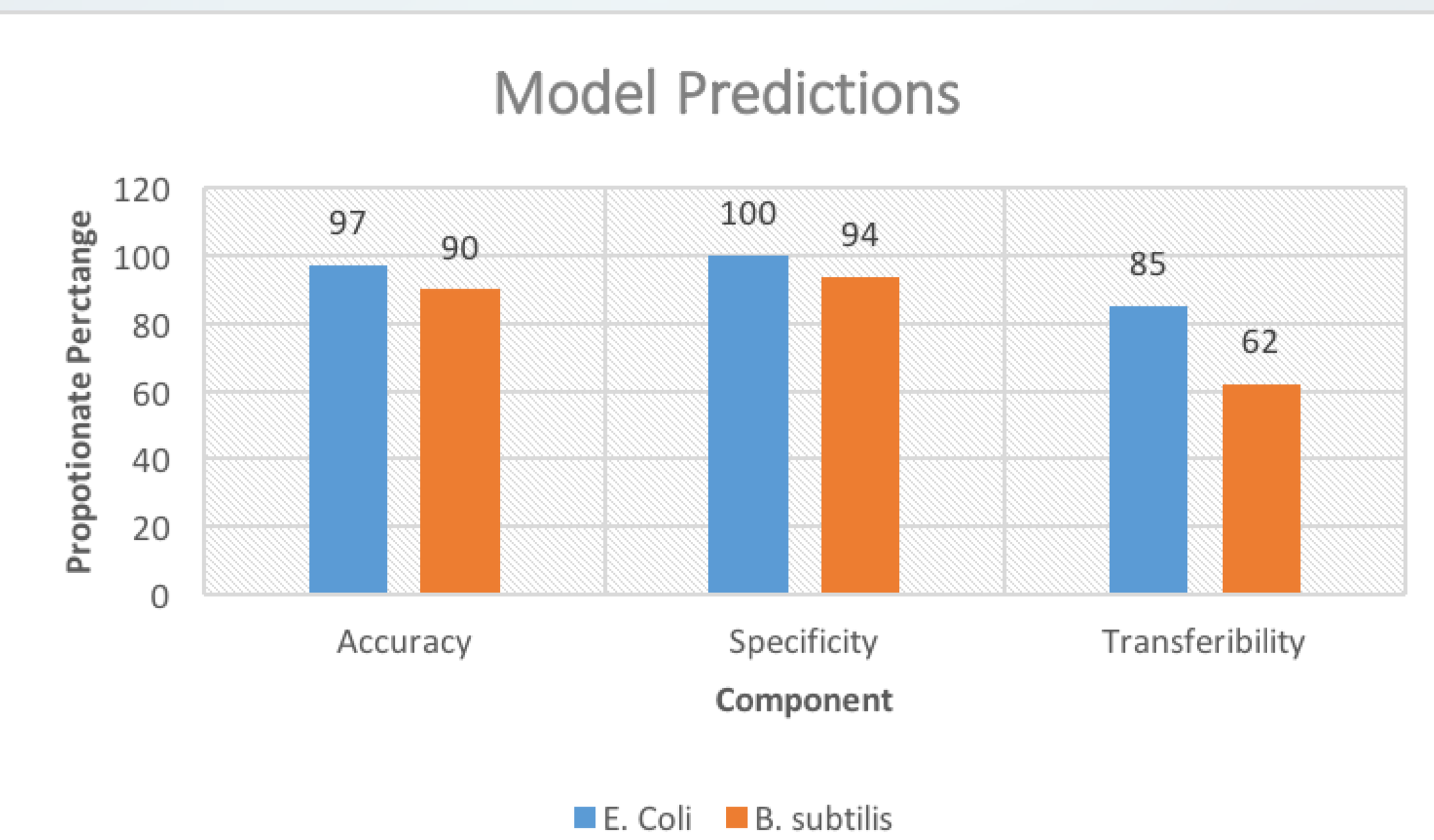
- Predicting on more non-traditional organisms
  - Marinobacter & Vibrio strains genomes
- Additional features for accuracy
- More tests on model transferability between organisms

Results

Output format of predicted promoter sequences

Predicted Promotor Sequence	Start locations of the sequence	Number of appearance
TTCGCTGTCTCGACGATGCTGTAAAGCATCGCGCTCGCATCAGCACCATTTGGTGTGTTTGAGAACAGCCAATTTTTTCT	[12420, 55007, 86577]	3
TGCTCTTTTAAGGCGGCATAGAGGGGTTTGAACTTCTCACTGACTTGGATAACCCACCTTGCCATGGTAGTTCGCGATAG	[55709, 88378]	2
TACAAAATGCAAGCGTTGACAATGAACTCGATTCAGACTGCTCGATGAGCGTTTGCCATTCTTGATTAGTGTGTCGTTTT	[749, 14496, 57091]	3
TTGGTACTACCCTGTTTTATCCTCAATATCTCATCTTCTCCGCGTAGAACCGCGTTCTGAGATGCATCAAACAAAAGTCC	[45596]	1
AACTCAAAAACAAGCTGAAGACGTTGAACTTGTAGGTTGGTTTCATAATCTAGCGAGTCAAGCCGTTGGTCGTGGATTGC	[72452]	1
TAAAACCTAGGAAATCAAAACCTCTGCTGATGTGGGTAATGTGCGTTTTCTCTTCGGAGAGTGTTAAGCCTCTTTCTGTT	[69496]	1
ATACCTTCGATGGAGCAATGCCACCAACCTTTATGATTGGTTTCTTGATAACGAATATTCCCTTTAAAATTTTTGGTAGC	[16254]	1
GACACGTTTACGATTAACGTTATGGCCTTTTTTAGCCATCTCATTTCGAATGCGTCTGCTCCCCATGAACGGATACCGAA	[15109, 17376]	2
TTGGCCAAAAAATCATTTTCCATGGTCAATTGACTAATCTTGGCGTGCAGTTTGTCCACTTCTTCAGAGTTGTCCTTGCC	[17631]	1
TCTTAGAGAAGTTAGGTTTTGAAAAATATGTACGTCCAGAAGCGTATGAGTCAACAGATTCAGGCTCAAGTTGTGGTACT	[28028]	1
AGCACGATTATTATCGATAGATAAATGACCGTCATCGATATAGCGGATGAGCTTGGGCCATTGGCCGAGCGTATACTTAA	[12530, 55117, 86687]	3
GAGTTGCTGGCGCTTAGCATAAAACGTCGAGGGATTGAGTTCATTAAGCTTACAAAATGCAAGCGTTGACAATGAACTCG	[699, 14446, 57041]	3
CTCAGCTTTCGCGAGCTCTTTCATGCACTTGACCATGTAGTCGTAGAGGATAAGGCCGTTGGCTTTCGCTGTCTCGACGA	[12356, 54943, 86513]	3
TATCGATAGATAAATGACCGTCATCGATATAGCGGATGAGCTTGGGCCATTGGCCGAGCGTATACTTAATCGCTTTTCCT	[12541, 55128, 86698]	3
GGTGTGTTTGAGAACAGCCAATTTTTTCTGCCAATGACCAGCGGTTTAATGGCGCGTTCAGCACGATTATTATCGATAGA	[12471, 55058, 86628]	3
GTTATTTTATAAAGTCTATAATTTTCAGATTAAATCAAGAAAACGTTGAGCAGCTGTTAGATACACTAACTCGAGACCTT	[75484]	1
TTGAAATTCAACGTCAATGGAAAAATCTGCGCCATCGGTAGCGAATGTTGATAGGTTTTTTGCTACTCCAAAACCACCAG	[46132]	1
TGCTGATGACGGCTTTGGAGTCTCGATATTCTTCGATGAGTCAGCATTCAATAAAGTCTCCCTGAGTCAGTTGCCAGAGA	[25732]	1
GTCACTTTCGATAATCGCCGCGAGGCCATTGATGGACTTTCTAAAATCGACACTCTCACGATACAGATAAATCTCTGGTG	[427, 14174, 56769]	3
AGAGGATAAGGCCGTTGGCTTTCGCTGTCTCGACGATGCTGTAAAGCATCGCGCTCGCATCAGCACCATTTGGTGTGTTT	[12400, 54987, 86557]	3
TGGTTACGTGATGAAGACCATCATAGGCACCATAGCCATCGGTTTGCAGATAACCACTGTAATTCCCCAAGAAGGCAACA	[55452, 88121]	2
ACTAAGCAACCAGAACCGACCAAGTTCTCGAACTTGGAGTCGAAGGCTGCCTCTAAAGATGAGGTTTTTAAGGTGGCAAT	[4619]	1
TCGACAGTCGCACCTTTTAAGCGCGATTCTATCCCATAGAGCTTCTGGATTTTGGCCAGCGCTTTATCTGCTTTGCCTGA	[12717, 55304, 87973]	3
GCGATTCTATCCCATAGAGCTTCTGGATTTTGGCCAGCGCTTTATCTGCTTTGCCTGACTTACCTTTACCTTGAAGCTTC	[12739, 55326, 87995]	3
TTAAGTCATTGTAAAATCGTTTATTTGAACGGGTTTATGGCCGATAATCGTGAAGCCAGAAAGCAGTCTATCAAGCTCGA	[172, 13919, 56514]	3
GGTGTCACTTTCGATAATCGCCGCGAGGCCATTGATGGACTTTCTAAAATCGACACTCTCACGATACAGATAAATCTCTG	[424, 14171, 56766]	3
CGTCTGCTCCCCATGAACGGATACCGAAGGTGAATTTCGTCAATCATGCGCCGTAGTACAATCTCTTCAGCAGAGAGTCC	[15161, 17428]	2
GAATGCTCAACTCCACATCATTGATGAGAAGTGTCATGTTAGCGGTGGCTATTTGAGCTTGATATTTCGTTGTCTTTTCG	[575, 14322, 56917]	3
TTTTCGACGACCTCTGCTCGTACAAAGCCTTGAGAGACATCTGCTTTTTTGAGTTGCTGGCGCTTAGCATAAAACGTCGA	[649, 14396, 56991]	3
AATCGTTTATTTGAACGGGTTTATGGCCGATAATCGTGAAGCCAGAAAGCAGTCTATCAAGCTCGAATTGAGTCAGGGTA	[186, 13933, 56528]	3

- All the predicted promoters had score of higher 80%
- The sequence length is 6o nucleobases
- The positions reflect the position in the start location in the genome for the predicted sequence



- Specificity : Ability to predict with expected parameters
- Accuracy: Matched identified published promoters
- Transferability: Predicting closely related organism using one organism to train and another to test

*All percentages were rounded down

N = 1655	Predicted: No	Predicted: Yes	Neg. results: 1%
Actual: No	TN: 48%	FP: 1%	Total: 49%
Actual: Yes	FN: 0%	TP: 49%	Total: 49%
	Total: 48%	Total: 50%	Pos. results: 97%

N = 1721	Predicted: No	Predicted: Yes	Neg results: 12%
Actual: No	TN: 40%	FP: 7%	Total: 47%
Actual: Yes	FN: 5%	TP: 46%	Total: 51%
	Total: 45%	Total: 53%	Pos. results: 86%

N = 1690	Predicted: No	Predicted: Yes	Neg. results: 13%
Actual: No	TN: 41%	FP: 6%	Total: 47%
Actual: Yes	FN: 7%	TP: 44%	Total: 51%
	Total: 48%	Total: 50%	Pos. results: 85%

N = 1750	Predicted: No	Predicted: Yes	Neg. results: 28%
Actual: No	TN: 33%	FP: 16%	Total: 49%
Actual: Yes	FN: 12%	TP: 38%	Total: 50%
	Total: 45%	Total: 54%	Pos results: 71%

N = 1761	Predicted: No	Predicted: Yes	Neg. results: 46%
Actual: No	TN: 48%	FP: 1%	Total: 49%
Actual: Yes	FN: 45%	TP: 4%	Total: 49%
	Total: 93%	Total: 5%	Pos results: 52%

- TP: True positives
- TN: True Negatives
- FN: False Negatives
- FP: False positives

- When trained using at 50% of the same data set as testing, the model prediction was 100% accurate
- When trained on a different organism genome data set than testing data set, the accuracy dropped up to 47%

Acknowledgements

- The opinions and assertions contained herein are those of the authors and are not to be construed as those of the U.S. Navy, military service at large or U.S. Government.
- Parts of this work were supported by the Office of the Assistant Secretary of Defense for Research and Engineering through the Applied Research for the Advancement of S&T Priorities (ARAP) Synthetic Biology for Military Environments (SBME) program, and parts by the Office of Naval Research via U.S. Naval Research Laboratory core (6.1) funds. Parts of the work were performed while the author GAE held an NRC Research Associateship award at the Naval Research Laboratory. Author WJ currently holds an NRC Research Associateship award at the Naval Research Laboratory.

U.S. NAVAL RESEARCH LABORATORY

- Literature

  - CNN: Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification, 1-9 . (https://arxiv.org/abs/1509.01626)
  - RNN: Schmidhuber, Jürgen (1993). Habilitation thesis: System modeling and optimization.
  - LSTM: Hochreiter & Schmidhuber (1997) http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf
  - BD-RNN: Schuster, Mike, and Kuldip K. Paliwal. (1997) Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on 45.11
  - Baldi, Pierre, et al. (1999) Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15.11: 937-946
  - https://github.com/dennybritz/cnn-text-classification-tf
  - https://www.tensorflow.org/tutorials/recurrent/