DeepRibo: precise gene annotation of prokaryotes using deep learning and ribosome profiling data

Jim Clauwaert
Ghent University
jim.clauwaert@ugent.be

Gerben Menschaert
Ghent University
gerben.menschaert@ugent.be

Willem Waegeman
Ghent University
willem.waegeman@ugent.be

1 INTRODUCTION

The vast number of sequenced prokaryotes, across all different phyla, has proven that the curation of the genome based on sequence alignments is not able to capture the full genomic complexity [13]. The delineation of the open reading frame (ORF) is an essential element in gene annotation and is mostly performed *in silico* [4, 5]. This makes novel *in silico* methods for genome annotations necessary. Recently, ribosome profiling (ribo-seq) was introduced, measuring mRNA that is associated with ribosomes by sequencing ribosome-protected fragments (RPFs) [6, 7]. We present DeepRibo, a novel neural network implementation applying ribosome profiling data for the precise annotation of transcription initiation sites in prokaryotes. It is trained on a combination of available experiments from different bacteria and has been tested to work equally well on *de novo* ribo-seq data of bacterial genomes.

2 MATERIALS AND METHODS

DeepRibo is a neural network specifically designed to process two types of data: strings (i.e. DNA-sequences) and floats (i.e. ribo-seq signal). The model first processes each type of data in parallel before combining the features created from both inputs into a set of fully-connected layers. Only a short DNA fragment of 30 nucleotides covering the Shine-Dalgarno region is processed, preventing the model from using gene sequences for training/testing. The DNA sequence is transformed into a binary image with 4 channels [1]. This image is consecutively processed by two convolutional layers. The first layer transforms the sparse matrix into a dense matrix using four 1x1 convolutional kernels. Afterwards, 32 kernels of 1x12 convolutions process the data in the second and last convolutional layer. The ribosome profiling data is fed into a double-layered, bidirectional Gated Recurrent Unit (GRU). Only the final hidden states of the memory cell are used for further processing, making the use of varied length inputs (i.e. candidate ORFs) possible. The output nodes of both networks are concatenated and fed into three fully-connected layers. Several databases have been included for training, covering both gram-negative (*S. typhimurium* [11], *E. coli* [9], *C. crescentus* [14]) and gram-positive bacteria (*B. Subtilis* [10], *S. coelicolor* [8], *S. aureus* [3]). Six models have been trained on each combination of five datasets, using the sixth as a test set. The dataset contains all possible ORFs from within each genome. Specifically, sequences larger than 30 nucleotides from their start codon (ATG, GTG, TTG) up to the stop codon (TAA, TGA or TAA) are considered candidate ORFs. The genome assembly for each species is used to label the data. Using the S-curve methodology proposed and described by Ndah et. al. [11], a cut-off is obtained determining the minimum signal count each candidate ORFs must cover to be considered for training/testing.

3 RESULTS

DeepRibo proved to provide noteworthy results, with PR-AUC as high as 0.955 for *E. coli* (Table 1). Two custom models have been trained on either the Shine-Dalgarno sequence (based on CNN) or the ribosome profiling data (based on RNN) for *E. coli*, achieving a PR-AUC score of 0.884 and 0.717, respectively. This proves that the model is able to combine both types of information in a meaningful way. Furthermore, predictions were evaluated against a set of proteins which have been verified and collected on EcoGene [15], and a mass spectrometry dataset created by Ndah et. al. [11]. In short, both datasets were in agreement with the model for more than 92% of their entries. The false positives of *E. coli* and *S. aureus* have been evaluated using pBLAST [2], aligning up to 63.5% and 78.0% of the false positives with a known protein at both their start and stop sites. The search was performed against the 'non-redundant protein database'.

Table 1: The ROC-AUC and PR-AUC performance values for the different experimental set-ups.

	S. typhimurium	*E. coli*	*C. crescentus*	*B. subtilis*	*S. coelicolor*	*S. aureus*
ROC-AUC	0.991	0.995	0.973	0.993	0.973	0.995
PR-AUC	0.910	0.943	0.842	0.922	0.863	0.965

4 DISCUSSION

DeepRibo learns from information contained in both DNA sequence of the Shine-Dalgarno region and ribo-seq, using a novel architecture which combines both convolutional layers and recurrent memory cells. The performance of DeepRibo is consistent on all six test sets, with a difference of 0.11 in PR-AUC score between the best and worst performing model. It furthermore outperforms REPARATION [11], reporting PR-AUC values of 0.74, 0.80 and 0.89 after 10-fold cross validation on the *S. typhimurium*, *E. coli* and *B. subtilis* dataset, respectively. DeepRibo is the first tool for the precise delineation of ORFs in prokaryotes trained and validated on various datasets. Identification of small open reading frames (sORFs) through *in silico* techniques is hindered, as the size of the ORFs influences the power of the statistical methods [12]. The novel ORF predictions given by the models have a median length of 165 and 63 for *E. coli* and *S. aureus*, well below the median length of the annotated genes of each species (807 and 723). Many of the novel predictions were furthermore situated within a pseudogene. These cover a nucleotide sequence in which multiple stop codons can be present, and are thus not in the samples, creating a hot-spot for 'novel' predictions. Corroborated by the results obtained from the pBLAST, it is plausible that false positives observed are due to a genome annotation which does not fully map the translational complexity of the organisms. The full article is hosted on the bioRxiv repository. The predictions are hosted at http://www.kermit.ugent.be/files/gwips_hub/index.html

REFERENCES

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33, 8 (2015), 831–838. https://doi.org/10.1038/nbt.3300 arXiv:cs/9605103

[2] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215, 3 (1990), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 arXiv:arXiv:1611.08307v1

[3] A. R. Davis, D. W. Gohara, and M.-N. F. Yap. 2014. Sequence selectivity of macrolide-induced translational attenuation. *Proceedings of the National Academy of Sciences* 111, 43 (2014), 15379–15384. https://doi.org/10.1073/pnas.1410356111 arXiv:arXiv:1408.1149

[4] A. Delcher. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* 27, 23 (1999), 4636–4641. https://doi.org/10.1093/nar/27.23.4636

[5] Doug Hyatt, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11 (2010). https://doi.org/10.1186/1471-2105-11-119 arXiv:1401.7457

[6] Nicholas T. Ingolia, Sina Ghaemmaghami, John R.S. Newman, and Jonathan S. Weissman. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 5924 (2009), 218–223. https://doi.org/10.1126/science.1168978 arXiv:arXiv:1408.1149

[7] Nicholas T. Ingolia, Liana F. Lareau, and Jonathan S. Weissman. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 4 (2011), 789–802. https://doi.org/10.1016/j.cell.2011.10.002

[8] Yujin Jeong, Ji Nu Kim, Min Woo Kim, Giselda Bucca, Suhyung Cho, Yeo Joon Yoon, Byung Gee Kim, Jung Hye Roe, Sun Chang Kim, Colin P. Smith, and Byung Kwan Cho. 2016. The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2). *Nature Communications* 7 (2016). https://doi.org/10.1038/ncomms11605

[9] Gene Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* 157, 3 (2014), 624–635. https://doi.org/10.1016/j.cell.2014.02.033 arXiv:cs/9605103

[10] Gene Wei Li, Eugene Oh, and Jonathan S. Weissman. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484, 7395 (2012), 538–541. https://doi.org/10.1038/nature10965 arXiv:NIHMS150003

[11] Elvis Ndah, Veronique Jonckheere, Adam Giess, Eivind Valen, Gerben Menschaert, and Petra Van Damme. 2017. REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic acids research* 45, 20 (2017), e168. https://doi.org/10.1093/nar/gkx758 arXiv:1705.11170

[12] Andrea Pauli, Eivind Valen, and Alexander F. Schier. 2015. Identifying (non-)coding RNAs and small peptides: Challenges and opportunities. *BioEssays* 37, 1 (2015), 103–112. https://doi.org/10.1002/bies.201400103

[13] Emily J. Richardson and Mick Watson. 2013. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics* 14, 1 (2013), 1–12. https://doi.org/10.1093/bib/bbs007

[14] Jared M. Schrader, Bo Zhou, Gene Wei Li, Keren Lasker, W. Seth Childers, Brandon Williams, Tao Long, Sean Crosson, Harley H. McAdams, Jonathan S. Weissman, and Lucy Shapiro. 2014. The Coding and Noncoding Architecture of the Caulobacter crescentus Genome. *PLoS Genetics* 10, 7 (2014). https://doi.org/10.1371/journal.pgen.1004463

[15] Jindan Zhou and Kenneth E. Rudd. 2013. EcoGene 3.0. *Nucleic Acids Research* 41, D1 (2013). https://doi.org/10.1093/nar/gks1235