Bioinformatics, 36(10), 2020, 3239–3241
doi: 10.1093/bioinformatics/btaa103
Advance Access Publication Date: 24 February 2020
Applications Note

Genome analysis

PlasGUN: gene prediction in plasmid metagenomic short reads using deep learning

Zhencheng Fang , Jie Tan, Shufang Wu, Mo Li, Chunhui Wang, Yongchu Liu and Huaiqiu Zhu *

State Key Laboratory for Turbulence and Complex Systems, Department of Biomedical Engineering, College of Engineering and Center for Quantitative Biology, Peking University, Beijing 100871, China

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on July 22, 2019; revised on January 16, 2020; editorial decision on February 8, 2020; accepted on February 10, 2020

Abstract

Summary: We present the first tool of gene prediction, PlasGUN, for plasmid metagenomic short-read data. The tool, developed based on deep learning algorithm of multiple input Convolutional Neural Network, demonstrates much better performance when tested on a benchmark dataset of artificial short reads and presents more reliable results for real plasmid metagenomic data than traditional gene prediction tools designed primarily for chromosome-derived short reads.

Availability and implementation: The PlasGUN software is available at http://cqb.pku.edu.cn/ZhuLab/PlasGUN/ or https://github.com/zhenchengfang/PlasGUN/.

Contact: hqzhu@pku.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Plasmids are among the most important components of mobile genetic elements. Recently, experimental and computational approaches (Fang *et al.*, 2019; Jones and Marchesi, 2007; Krawczyk *et al.*, 2018; Zhou and Xu, 2010) have been developed to enrich plasmid DNA from the metagenome, leading to the discovery of a large number of novel plasmids. Identifying genes, particularly novel genes, in plasmid metagenomic data is a fundamental method for elucidating the mechanisms by which plasmids regulate the microbial community, especially the spread of resistance genes.

However, gene prediction in plasmid metagenomic data may be much more difficult than that in untargeted metagenomic data, in which chromosome-derived DNA is dominant. Generally, sequence assembly facilitates gene prediction because fewer genes are cut off in longer fragments. Unfortunately, sequence assembly for plasmid metagenomic short reads is still a difficult task owing to the mobile nature of plasmids, such as the presence of repetitive elements (Rozov *et al.*, 2017). Although some tools have been designed to address the assembly challenge caused by repetitive regions (Ji *et al.*, 2017; Shi *et al.*, 2017), other unique genomic patterns of plasmids, like the existence of shared genes and the mosaic structure, may prevent the usage of state-of-the-art assembly tools. Research has shown that the sensitivity of finding resistance genes on assembled contigs is lower than that on short reads directly, which may be caused by misassembly (Clausen *et al.*, 2016). Therefore, identifying genes in metagenomic short reads may be a better choice for plasmid studies than assembled contigs. Since a short read often contains at least one functional domain of a gene, even if the gene is cut-off in the fragment, this approach is widely used in metagenomic studies (Sharpton, 2014). However, traditional gene prediction tools for metagenomic short reads are primarily constructed using data from bacterial chromosomes. It has been shown that plasmids often contain different sequence signatures with chromosomes (Krawczyk *et al.*, 2018). Therefore, although traditional tools present good results for bacterial chromosomes, they may not adapt to plasmid metagenomic short reads.

In this work, we present PlasGUN, the first tool for gene prediction in plasmid metagenomic short reads, which is constructed over a large-scale plasmid dataset. Using deep learning, PlasGUN first extracts all candidate ORFs (Open Reading Frame) from the input short reads and then judges each ORF as a coding or non-coding ORF. The tool demonstrates high performance when tested on both benchmark datasets of artificial short reads and real plasmid metagenomic data.

2 Materials and methods

Because we lack experimental metagenomic data with precise gene annotation, constructing a benchmark dataset of artificial short reads is the primary approach for developing a gene prediction tool (Liu *et al.*, 2013). Herein we downloaded 4395 complete genomes of prokaryotic plasmids from the RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/). The accession list of the genomes is provided in Supplementary Material S1. To test whether

Table 1. Performance comparison of PlasGUN and related tools

Group	Tool	Sequence without genes of uncertain function			Sequence with genes of uncertain function		
		Sp (%)	*SnC* (%)	*SnP* (%)	*Sp* (%)	*SnC* (%)	*SnP* (%)
GroupS	PG versus PD	86.65	91.65	97.05	88.01	79.95	86.60
	PG versus MG	90.06	89.51	95.80	90.42	72.52	84.38
	PG versus MM	90.21	83.97	95.73	89.20	68.11	85.61
	PG versus FG	82.45	91.10	97.85	84.13	80.16	88.89
	PG versus MA	84.54	90.19	97.53	85.69	79.21	88.05
	PG versus OP	82.79	81.47	97.80	85.36	68.75	88.21
GroupL	PG versus PD	91.67	94.16	96.27	88.54	80.59	84.95
	PG versus MG	93.48	91.74	95.17	91.45	72.99	81.21
	PG versus MM	92.48	91.42	95.85	89.04	75.87	84.45
	PG versus FG	84.28	92.84	97.93	79.64	82.21	89.90
	PG versus MA	89.68	93.08	96.92	86.66	80.33	86.45
	PG versus OP	92.74	90.50	95.71	90.13	75.08	83.02

Note: PG, PD, MG, MM, FG, MA and OP represent PlasGUN, Prodigal, MetaGUN, MetaGeneMark, FragGeneScan, MetaGeneAnnotator and Orphelia, respectively.

PlasGUN could identify genes in novel genomes, which is an important task for a metagenomic study, we used genomes released before 2013 to build the training set and the rest to build the test set. We then extracted DNA fragments from all genomes to construct an artificial short-read dataset. The lengths of the fragments were between 100 and 900 bp, which encompasses the lengths of different shotgun sequencing technologies. All candidate ORFs (both complete and partial) longer than 60 bp were extracted from the six phases of each DNA fragment, and all ORFs were labelled as coding or non-coding according to the RefSeq annotation.

Considering that the plasmid sequence often contains mixed patterns from its wide range of hosts, we used the 'one-hot' encoding form as a mathematical model to characterize an ORF. Unlike the global statistics used by previous tools, such encoding form provided a more detail characterization for the sequence with mixed patterns. We further designed miCNN (multiple input Convolutional Neural Network) to detect the functional domain and sequence pattern from the ORF represented by the 'one-hot' encoding form through the convolution operation. For each input ORF, miCNN outputs a likelihood score (0–1) reflecting whether the ORF is a coding ORF. We trained two miCNNs for sequences between 100–400 bp (GroupS) and 401–900 bp (GoupL), respectively. More details about the dataset construction and the structure of the miCNN are provided in Supplementary Material S2.

3 Results and discussion

Against the test set with gene annotation, PlasGUN was evaluated and achieved AUCs (Area Under Curve) of 98.72% and 98.81% for GroupS and GroupL, respectively, indicating that PlasGUN can present high performance. We further compared the prediction of PlasGUN with that of Prodigal (Hyatt *et al.*, 2012), MetaGUN (Liu *et al.*, 2013), MetaGeneMark (Zhu *et al.*, 2010), FragGeneScan (Rho *et al.*, 2010), MetaGeneAnnotator (Noguchi *et al.*, 2008) and Orphelia (Hoff *et al.*, 2009). The evaluation criteria were defined as sensitivity $Sn=TP/(TP+FN)$ and specificity $Sp=TP/(TP+FP)$. For PlasGUN, an ORF with a score higher than a given threshold would be regarded as coding ORF. To make the comparison convincing, we adjusted the threshold to allow PlasGUN to achieve the same *Sp* as that of the comparative tools. Under the same *Sp*, we compared the *Sn* of the comparative tools (shown as *SnC* in Table 1) with the *Sn* of PlasGUN (shown as *SnP* in Table 1). Notably, a large percentage of genes (41.09%) from the complete genomes of the dataset were labelled 'hypothetical' or 'putative' or lacked exact product annotation, which were primarily annotated by the computational pipeline, and had unknown functions. Although these genes are less convincing to serve as benchmarks, gene annotation in complete genomes is less challenging than that in metagenomic genomes and is generally reliable because the metagenomic genomes contain too many incomplete genes, leading to less information used for gene prediction (Liu *et al.*, 2013). Since identifying novel genes is an important goal in metagenomic studies, it was necessary to determine the performance of the gene prediction tools on sequences with genes of uncertain function because these sequences might contain a large number of novel genes. Thus, we compared sequences without genes of uncertain function and sequences with genes of uncertain function separately. The results are shown in Table 1. Under the same *Sp*, PlasGUN consistently achieved a higher *Sn* than other gene prediction tools, and this advantage was more remarkable in GroupS than in GroupL. Considering that the short reads obtained from the most widely used sequencing technologies, such as Illumina, often contain only hundreds of bases, the improved performance of PlasGUN for short fragments makes this tool highly powerful to analyse plasmid metagenomic short reads. Additionally, we found that compared with other prediction tools, PlasGUN had an obvious advantage when tested on sequences with genes of uncertain function. Among the comparative tools, Prodigal was the best performing software, and the *Sn* of PlasGUN was 5.40% and 6.65% higher than that of Prodigal on sequences without genes of uncertain function and sequences with genes of uncertain function in GroupS, respectively. This showed that PlasGUN was more competent in finding novel genes. In the released version of the PlasGUN tool, users can adjust the threshold according to their own requirements. In Supplementary Material S2, we evaluated the performance of PlasGUN under different thresholds. By default, the threshold with the highest harmonic mean $Hm=(2\times Sn\times Sp)/(Sn+Sp)$ will be selected.

In addition, we used real plasmid metagenomic short reads from a wastewater treatment plant sample (Szczepanowski *et al.*, 2008) to evaluate the reliability of each tool. PlasGUN was run under default settings. All predictions of each tool were searched both against the RefSeq plasmid protein database using PSI-BLAST, a homology search strategy that is more sensitive for novel genes with low similarity to the database, and against the Conserved Domain Database using DELTA-BLAST, a sensitive protein search strategy for novel genes. PlasGUN achieved the highest proportion of predictions that contained BLAST hits, indicating that the predictions of PlasGUN were more reliable and might contain more novel genes than other prediction tools. We also evaluated related tools using plasmid artificial short reads with 5% contamination of chromosome-derived short reads, and PlasGUN was still the best performing tool. See Supplementary Material S2 for more details about the related analysis.

Funding

This work was supported by the National Key Research and Development Programme of China [2017YFC1200205] and the National Natural Science Foundation of China [31671366].

Conflict of Interest: none declared.

References

Clausen,P.T. *et al.* (2016) Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.*, **71**, 2484–2488.

Fang,Z. *et al.* (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, **8**, giz066.

Hoff,K.J. *et al.* (2009) Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.*, **37**, W101–W105.

Hyatt,D. *et al.* (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28**, 2223–2230.

Ji,P. *et al.* (2017) MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat. Commun.*, **8**, 14306.

Jones,B.V. and Marchesi,J.R. (2007) Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods*, **4**, 55–61.

Krawczyk,P.S. *et al.* (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.

Liu,Y. *et al.* (2013) Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinformatics*, **14**, S12.

Noguchi,H. *et al.* (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.*, **15**, 387–396.

Rho,M. *et al.* (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.

Rozov,R. *et al.* (2017) Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*, **33**, 475–482.

Sharpton,T.J. (2014) An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.*, **5**, 209.

Shi,W. *et al.* (2017) The combination of direct and paired link graphs can boost repetitive genome assembly. *Nucleic Acids Res.*, **45**, e43.

Szczepanowski,R. *et al.* (2008) Insight into the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to antimicrobial drugs analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, **136**, 54–64.

Zhou,F. and Xu,Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, **26**, 2051–2052.

Zhu,W. *et al.* (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.