PlasClass improves plasmid sequence classification

David Pellow[1], Itzik Mizrahi[2], and Ron Shamir[1]

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, 6997801, Israel
[2] Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva, 8410501,Israel

Abstract. Many bacteria contain plasmids, but separating between contigs that originate on the plasmid and those that are part of the bacterial genome can be difficult. This is especially true in metagenomic assembly, which yields many contigs of unknown origin. Existing tools for classifying sequences of plasmid origin give less reliable results for shorter sequences, are trained using only some of the known plasmids, and can be difficult to use in practice.
Results: We present PlasClass, a new plasmid classifier. It uses a set of standard classifiers trained on the most current set of known plasmid sequences for different sequence lengths. PlasClass outperforms the state-of-the-art plasmid classification tool on shorter sequences, which constitute the majority of assembly contigs, while using significantly less time and memory
Availability: `https://github.com/Shamir-Lab/PlasClass`
Contact: {dpellow,rshamir}@tau.ac.il

1 Introduction

When using high-throughput sequencing to study the presence and dynamics of plasmids in their bacterial hosts, it is often necessary to classify sequences as being of plasmid or bacterial origin. This is especially true in the case of metagenomic sequencing, which can include many sequences of unknown origin and varying lengths. We focus on the challenge of classifying contigs in a metagenomic assembly in order to identify which are of plasmid origin. The current state-of-the-art classifier of plasmid sequences is PlasFlow [3], a neural network based algorithm. While PlasFlow is successful in classifying small sets of long sequences, it produces less reliable results for short sequences and has difficulty with very large metagenomic sequence datasets due to memory constraints. Here we present PlasClass, a new plasmid sequence classifier that is implemented as an easy to use Python package.
PlasClass consists of a set of logistic regression classifiers each trained on sequences of a different length sampled from reference sequences of plasmid and bacterial origin. When applying PlasClass on a set of sequences, the appropriate length-specific classifier is used for each sequence.
We tested PlasClass on simulated data, on bacterial isolates, on waste water plasmidome, and on human gut microbiome samples. For shorter sequences, which are the majority of contigs in an assembly, PlasClass achieves better F1 scores than PlasFlow. It is also faster and uses significantly less RAM and disk memory.
PlasClass is provided at `https://github.com/Shamir-Lab/PlasClass`.

2 Methods

2.1 Training databases

We used reference sequence databases to obtain the training sequences for our classifiers. For the plasmid references we used plasmid sequences listed in PLSDB [2], an up-to-date

curated plasmid database. After filtering out duplicate sequences this database contained 13469 reference plasmids (median length: 53.8kbp).
For the bacterial references we downloaded all complete bacterial genome assemblies from NCBI (download date January 9, 2019). We removed plasmid sequences and filtered out duplicates, leaving 13491 reference chromosomes (median length: 3.7Mbp).
One quarter of the sequences were randomly removed from the databases before training in order to provide a held-out test set (Section 3.1). PlasClass was retrained on the full databases and this version was used for testing on assembled data (Sections 3.2 – 3.5).

2.2 Training the classifiers

We sampled sequence fragments of different lengths from the reference sequences with replacement and constructed a k-mer frequency vector for each fragment. Canonical k-mers of lengths 3 – 7 were used resulting in a feature vector of length 10952 for each fragment. Fragment lengths were 500k, 100k, 10k, and 1k. For the two shorter lengths 90,000 training fragments were used from each class. For 500k and 100k, since there were not enough long plasmids to do the same, we sampled enough fragments to cover all plasmids of sufficient length to a depth of 5.
For each length, a logistic regression classifier was trained on the plasmid and bacterial fragments' k-mer frequency vectors using the scikit-learn machine learning library in Python.

2.3 Length-specific classification

PlasClass uses four logistic regression models to classify sequences of different length scales. Given a dataset, each sequence is classified using the length-specific model with the closest length. Further implementation details can be found in the Supplement (SectionS1).

2.4 Classification with PlasClass

PlasClass is available at `https://github.com/Shamir-Lab/PlasClass`. It has been retrained using the full set of database references. PlasClass can be used as a command-line tool to classify sequences in an input fasta file or it can be imported as a module into the user's code to classify sequences in the user's program. It can be run in parallel mode to achieve faster runtimes. PlasClass is fully documented at the url provided above, see the Supplement for details (Section S7).

3 Results

We compared performance of PlasFlow and PlasClass on both simulated and real data. See the Supplement for technical details (S2).

3.1 Classifying sequences from held-out references

We sampled overlapping L-long fragments covering the held out plasmids with an overlap of $L/2$ for $L = 100$k, 10k and 1k. A matching number of L-long fragments were sampled from the held out bacterial genomes for each length L. Table 1 summarizes the classification results. PlasClass improved precision at the cost of slightly lower recall and had better overall F1 on the shorter sequence lengths. These short sequences can make up the majority of contigs in metagenomic assemblies (see Tables S1, S3 and Figure S1), allowing PlasClass to outperform PlasFlow in many settings.

	PlasFlow			PlasClass		
Length (bp)	Precision	Recall	F1	Precision	Recall	F1
100k	95.6	88.4	91.9	96.9	85.4	90.8
10k	83.1	87.7	85.3	88.7	86.4	87.6
1k	59.7	79.1	68.1	75.1	74.6	74.8

Table 1: Performance of PlasFlow and PlasClass on fixed length sequence fragments sampled from the held out references.

3.2 Performance on bacterial isolates

We compared the performance of PlasFlow and PlasClass on the isolate assemblies from the benchmark in [1]. For both classifiers precision was very low due to the very small number of contigs of plasmid origin. PlasClass outperformed PlasFlow with an F1 of 1.49 compared to 0.80. Further details can be found in Supplement S3.

3.3 Performance on simulated metagenome assemblies

We simulated two metagenomes, assembled them, and classified the contigs. Performance improved for both methods as short sequences were excluded, and PlasClass outperformed PlasFlow for all size thresholds, achieving F1 scores of 12.35 and 10.86 on the two simulations. See the Supplement for further details (Section S4).

3.4 Performance on plasmidome sample

We assembled a waste water plasmidome [5] and classified the contigs (see Supplement S5 for details). Although the plasmid-enriched setting favors PlasFlow, which sacrifices precision for higher recall, PlasClass still had a higher combined F1. Table 2 compares the performance and resource usage of the methods. PlasClass was more than twice as fast and can be even faster when using multiprocessing. PlasClass used much less RAM than PlasFlow. These differences become more manifest as the dataset size grows (results not shown). PlasFlow writes the feature matrices to disk while PlasClass does not.

	Precision	Recall	F1 score	Runtime	RAM	Disk
PlasFlow	16.4	**89.1**	27.7	3.4	15.4	3.1
PlasClass	**22.8**	65.2	**33.8**	**1.5**	**6.1**	**0**

Table 2: Results on plasmidome sample. Runtime is in minutes, memory usage is in GB, RAM reported is the maximum resident set size.

3.5 Classifying metagenomic plasmid assemblies

We assembled six publicly available human gut microbiome samples (accessions listed in the Supplement, Section S6) and found plasmid sequences in the assemblies using Recycler [4]. Recycler assembles plasmid sequences based on coverage and circularity - features that are not used by the classifiers. 16–27 plasmids were assembled per sample (median length: 3.4kb).

We classified each of the plasmids generated by Recycler to determine the extent of agreement between the sequence classifiers and this orthogonal approach. As seen in Figure 1, PlasClass agreed with Recycler on the same number or more plasmids than PlasFlow in all samples. This suggests that PlasClass can correctly identify more plasmids in real datasets that contain many previously unknown plasmid sequences.

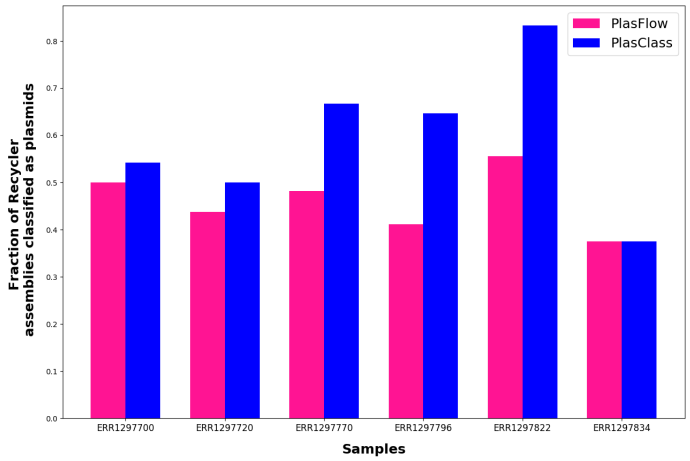

Fig. 1: Agreement of PlasFlow and PlassClass classifications with the plasmids generated by Recycler.

Acknowledgements

We thank the members of the Shamir and Mizrahi Labs for their help and advice.

Funding

DP is supported in part by an Edmond J Safra PhD fellowship and in part by a Ministry of Immigrant Absorption PhD fellowship. Study supported in part by the Israel Science Foundation (grant 1339/18), and by grant 2016694 from the United State - Israel Binational Science Foundation (BSF), Jerusalem, Israel and the United States National Science Foundation (NSF). The research described here was supported by grants from the Israel Science Foundation (grant no. 1947/19) and the European Research Council under the European Unions Horizon 2020 research and innovation program (grant agreement no. 640384)

References

1. Arredondo-Alonso, S., Willems, R.J., van Schaik, W., Schürch, A.C.: On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microbial genomics **3**(10), e000128 (2017)
2. Galata, V., Fehlmann, T., Backes, C., Keller, A.: PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Research **47**(D1), D195–D202 (2018)

3. Krawczyk, P.S., Lipinski, L., Dziembowski, A.: PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Research **46**(6), e35 (2018)
4. Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., Shamir, R.: Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics **33**(4), 475–482 (2017)
5. Shi, Y., Zhang, H., Tian, Z., Yang, M., Zhang, Y.: Characteristics of arg-carrying plasmidome in the cultivable microbial community from wastewater treatment system under high oxytetracycline concentration. Applied microbiology and biotechnology **102**(4), 1847–1858 (2018)