

4CAC: 4-class classification of metagenomic assemblies using machine learning and assembly graphs

Lianrong Pu and Ron Shamir

The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.

lianrongpu@mail.tau.ac.il, rshamir@tau.ac.il

Supplementary Material

1 Supplementary tables

Table S1. Performance of four-class classifiers on simulated metagenomes without using graph information. Dvf+PC+Tiara, VV+Tiara, and PM+Tiara represent combined classifiers Dvf+PlasC+Tiara, viralVerify+Tiara, and PPR-Meta+Tiara, respectively. Sim_SG and Sim_SF are datasets of short reads while Sim_LG and Sim_LF are datasets of long reads.

Datasets	Tools	Phage			Plasmid			Prokaryote			Eukaryote		
		precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score
Sim_SG	XGBoost	0.81	0.60	0.69	0.81	0.36	0.50	0.99	0.73	0.84	1.00	0.89	0.94
	VV+Tiara	0.46	0.49	0.48	0.54	0.47	0.50	0.92	0.76	0.83	0.97	0.70	0.81
	PM+Tiara	0.34	0.81	0.48	0.31	0.78	0.45	0.98	0.71	0.82	0.96	0.74	0.84
	Dvf+PC+Tiara	0.24	0.80	0.36	0.31	0.50	0.38	0.92	0.78	0.85	0.96	0.74	0.84
Sim_SF	XGBoost	0.56	0.61	0.58	0.79	0.37	0.50	0.97	0.54	0.69	0.99	0.81	0.89
	VV+Tiara	0.37	0.49	0.42	0.60	0.48	0.53	0.92	0.41	0.56	0.98	0.37	0.54
	PM+Tiara	0.10	0.79	0.17	0.05	0.78	0.09	0.89	0.56	0.69	0.97	0.38	0.55
	Dvf+PC+Tiara	0.11	0.53	0.18	0.04	0.68	0.08	0.80	0.76	0.78	0.97	0.38	0.55
Sim_LG	XGBoost	0.89	0.86	0.87	0.83	0.77	0.80	0.99	0.73	0.84	1.00	0.82	0.90
	VV+Tiara	0.62	0.85	0.72	0.58	0.87	0.69	0.96	0.80	0.87	0.91	0.80	0.85
	PM+Tiara	0.52	0.94	0.67	0.39	0.87	0.54	0.98	0.75	0.85	0.91	0.54	0.68
	Dvf+PC+Tiara	0.24	0.03	0.06	0.35	0.78	0.48	0.89	0.85	0.87	0.89	0.91	0.90
Sim_LF	XGBoost	0.96	0.85	0.91	0.88	0.77	0.82	0.99	0.77	0.87	1.00	0.97	0.98
	VV+Tiara	0.73	0.89	0.80	0.73	0.83	0.78	0.95	0.88	0.92	1.00	0.91	0.95
	PM+Tiara	0.73	0.90	0.81	0.56	0.87	0.68	0.99	0.66	0.79	0.97	1.00	0.99
	Dvf+PC+Tiara	0.43	0.71	0.54	0.60	0.68	0.63	0.94	0.78	0.85	0.97	1.00	0.99

Table S2. Performance of four-way classifiers and three-way classifiers on the Gut_HiFi dataset. The Gut_HiFi dataset is assembled from long reads and only consists of contigs from phages, plasmids, and prokaryotes. VV+Tiara and PM+Tiara represent combined classifiers viralVerify+Tiara and PPRMeta+Tiara, respectively. 3CAC(VV) and 3CAC(PM) represent running 3CAC based on the solution of viralVerify and PPR-Meta, respectively.

Tools	Phage			Plasmid			Prokaryote			All		
	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score
4CAC	0.03	0.40	0.05	0.00	0.00	0.00	0.99	0.83	0.91	0.90	0.83	0.86
3CAC+Tiara	0.07	0.17	0.10	0.04	0.30	0.07	0.99	0.93	0.96	0.93	0.92	0.93
VV+Tiara	0.09	0.30	0.14	0.02	0.44	0.04	0.99	0.84	0.91	0.86	0.83	0.85
PM+Tiara	0.07	0.30	0.11	0.02	0.63	0.04	0.99	0.80	0.89	0.80	0.79	0.80
3CAC(VV)	0.13	0.27	0.17	0.02	0.22	0.04	0.99	0.71	0.83	0.91	0.71	0.79
3CAC(PM)	0.07	0.17	0.10	0.04	0.30	0.07	0.99	0.58	0.73	0.91	0.58	0.71
viralVerify	0.09	0.30	0.14	0.02	0.44	0.04	0.99	0.56	0.72	0.82	0.56	0.66
PPR-Meta	0.07	0.30	0.11	0.02	0.63	0.04	0.99	0.67	0.80	0.77	0.67	0.72

Table S3. Performance of 4CAC on simulated metagenomes with initial classifications generated by different score thresholds. To generate an initial four-way classification by XGBoost classifier, we tested score thresholds 0.8, 0.85, 0.9, and 0.95 to classify contigs as phages and plasmids. Note that this table presents the results of the full 4CAC algorithm.

Datasets	Score threshold	Phage			Plasmid			Prokaryote			Eukaryote			All		
		precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score	precision	recall	F1 score
Sim_SG	0.95	0.79	0.79	0.79	0.74	0.66	0.70	0.96	0.94	0.95	1.00	0.95	0.97	0.95	0.92	0.93
	0.9	0.70	0.84	0.77	0.65	0.75	0.69	0.96	0.92	0.94	1.00	0.94	0.97	0.93	0.91	0.92
	0.85	0.64	0.87	0.73	0.60	0.79	0.68	0.97	0.91	0.94	1.00	0.93	0.96	0.92	0.91	0.91
	0.8	0.60	0.88	0.71	0.56	0.83	0.67	0.97	0.89	0.93	1.00	0.93	0.96	0.91	0.90	0.91
Sim_SF	0.95	0.60	0.80	0.69	0.82	0.76	0.79	0.94	0.90	0.92	0.98	0.85	0.91	0.95	0.88	0.91
	0.9	0.42	0.85	0.56	0.63	0.81	0.71	0.94	0.90	0.92	0.98	0.85	0.91	0.94	0.88	0.90
	0.85	0.32	0.87	0.47	0.50	0.84	0.63	0.94	0.89	0.91	0.98	0.85	0.91	0.92	0.87	0.90
	0.8	0.25	0.89	0.39	0.40	0.87	0.55	0.94	0.87	0.91	0.98	0.85	0.91	0.90	0.86	0.88
Sim_LG	0.95	0.91	0.91	0.91	0.84	0.79	0.82	0.98	0.97	0.97	1.00	0.90	0.95	0.96	0.94	0.95
	0.9	0.90	0.92	0.91	0.79	0.83	0.81	0.98	0.96	0.97	1.00	0.90	0.95	0.95	0.94	0.95
	0.85	0.86	0.93	0.89	0.76	0.89	0.82	0.98	0.95	0.96	1.00	0.90	0.95	0.95	0.94	0.94
	0.8	0.84	0.93	0.88	0.70	0.92	0.79	0.99	0.93	0.96	1.00	0.90	0.95	0.94	0.93	0.93
Sim_LF	0.95	0.96	0.90	0.93	0.88	0.81	0.84	0.97	0.95	0.96	1.00	0.97	0.98	0.96	0.93	0.95
	0.9	0.93	0.91	0.92	0.85	0.85	0.85	0.97	0.95	0.96	1.00	0.97	0.98	0.95	0.93	0.94
	0.85	0.91	0.91	0.91	0.78	0.90	0.84	0.97	0.93	0.95	1.00	0.97	0.98	0.94	0.93	0.94
	0.8	0.89	0.92	0.90	0.76	0.92	0.83	0.98	0.92	0.95	1.00	0.97	0.98	0.94	0.93	0.93

2 Supplementary figures

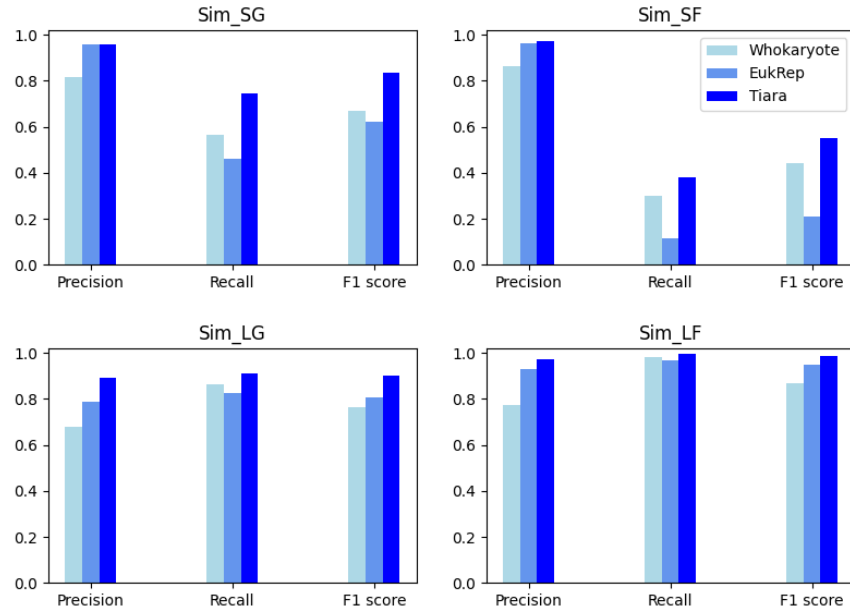


Fig.S1. Performance of eukaryote classifiers on simulated metagenomes. Sim_SG and Sim_SF are datasets of short reads while Sim_LG and Sim_LF are datasets of long reads.

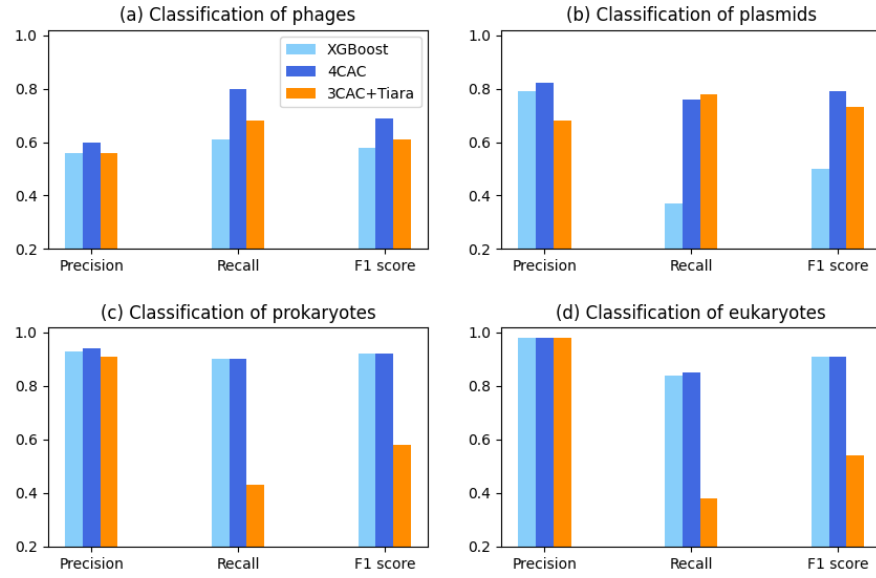


Fig. S2. Performance on each class for simulated short read dataset Sim_SF.

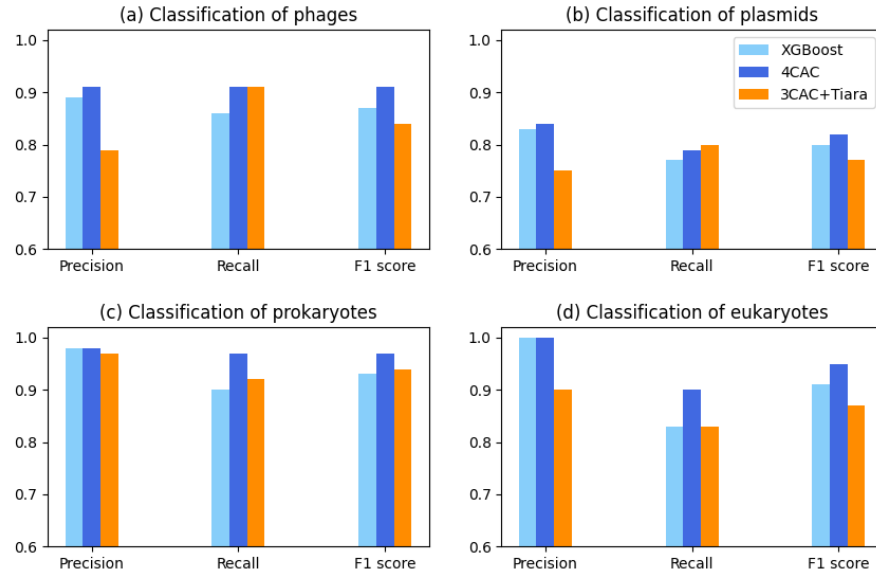


Fig. S3. Performance on each class for simulated long read dataset Sim_LG.

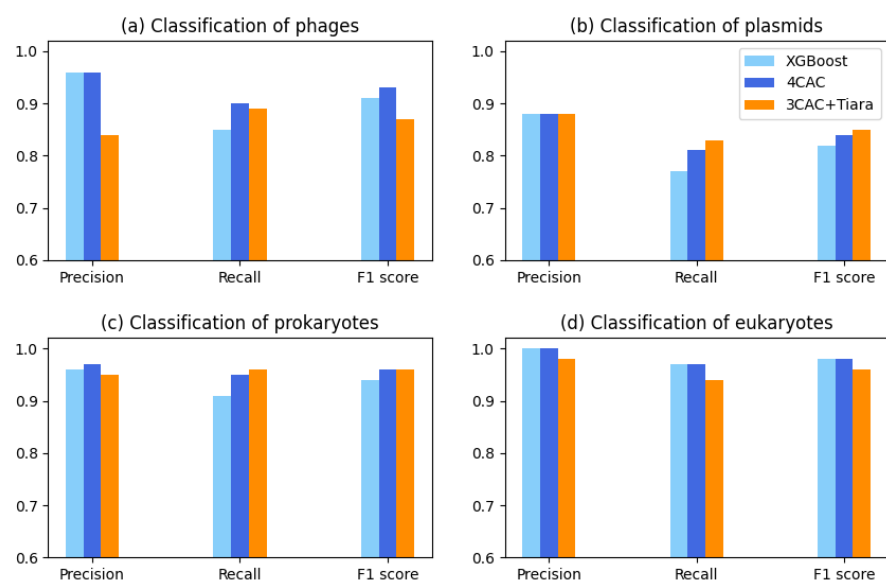


Fig. S4. Performance on each class for simulated long read dataset Sim.LF.

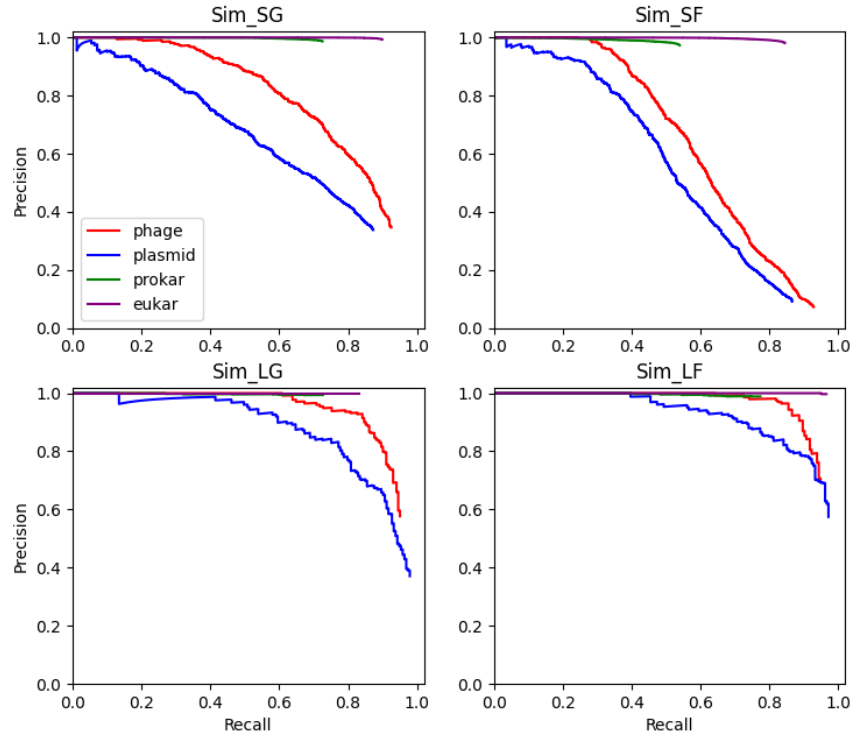


Fig.S5. The precision-recall curve of our XGBoost classifier on simulated metagenomes. Lines colored red, blue, green, and purple represent precision-recall curves for phage, plasmid, prokaryote, and eukaryote classification, respectively. Sim_SG and Sim_SF are datasets of short reads while Sim_LG and Sim_LF are datasets of long reads.