BENCHMARKING BLAST ACCURACY OF GENUS/PHYLA CLASSIFICATION OF METAGENOMIC READS UNDER A QUERY SEQUENCE ERROR MODEL AND AN INCOMPLETE TRAINING DATABASE

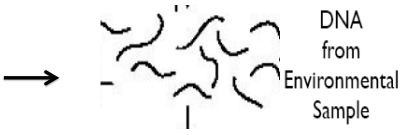
Steven D. Essinger, Gail L. Rosen

Electrical & Computer Engineering, Drexel University, 3141 Chestnut Street
Philadelphia, PA 19104, US

Summary

Goal: Benchmark the taxonomic performance of BLAST on databases of varying coverage and with the introduction of query sequence error

We aim to answer: "Who is in this metagenomic sample?"



Examples of metagenomic applications:

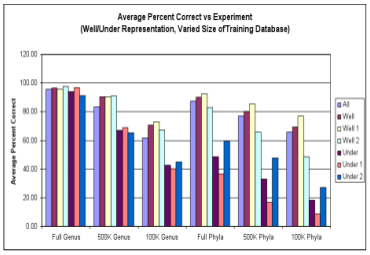
human health, soil fertility and forensics

Sequence alignment tools (e.g. BLAST) are employed to annotate fragments of DNA obtained from these samples

*Accuracy of this method for taxonomy is currently unknown

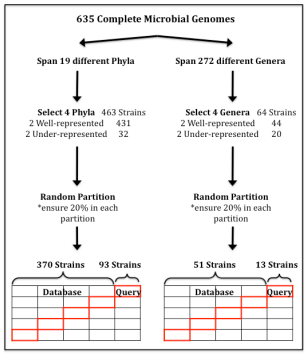
Results

Varying Database Coverage

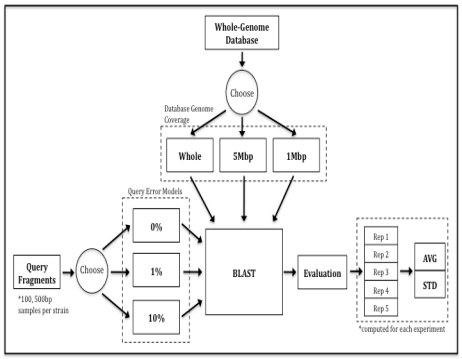


Percentages		All	Well	Well 1	Well 2	Under	Under 1	Under 2
Whole Genome Genus	AVG	95.87	96.60	95.65	97.87	94.15	96.90	91.40
	STD	2.10	3.10	4.91	3.03	3.57	4.56	8.51
5Mbp Genus	AVG	83.18	90.48	90.43	90.78	67.10	68.80	65.40
	STD	3.49	4.16	5.09	4.08	3.06	1.20	6.22
1Mbp Genus	AVG	61.83	70.52	72.95	67.57	42.70	40.40	45.00
	STD	1.76	1.96	3.97	4.21	2.06	2.51	3.66
Whole Genome Phyla	AVG	87.21	90.06	92.67	83.01	48.74	36.80	59.38
	STD	2.29	2.30	0.79	7.80	9.64	16.43	14.52
5 Mbp Phyla	AVG	76.76	80.00	85.31	65.64	33.11	16.67	47.80
	STD	1.44	1.56	1.15	7.16	6.88	5.17	10.85
1 Mbp Phyla	AVG	65.70	69.21	76.86	48.48	18.38	8.74	26.98
	STD	0.71	0.87	1.24	4.66	3.10	2.59	5.24

Experimental Dataset

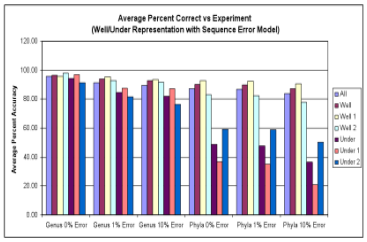


Experimental Setup



Results

Query Sequence Error



Percentages		All	Well	Well1	Well2	Under	Under1	Under2
Genus 0% Error	AVG	95.87	96.60	95.65	97.87	94.15	96.90	91.40
	STD	2.10	3.10	4.91	3.03	3.57	4.56	8.51
Genus 1% Error	AVG	91.17	94.07	95.29	92.65	84.55	87.60	81.50
	STD	8.48	4.38	5.41	10.60	19.97	20.73	21.05
Genus 10% Error	AVG	89.37	92.62	93.63	91.65	81.90	87.30	76.50
	STD	7.99	4.69	6.41	10.89	20.14	21.23	21.63
Phyla 0% Error	AVG	87.21	90.06	92.67	83.01	48.74	36.80	59.38
	STD	2.29	2.30	0.79	7.80	9.64	16.43	14.52
Phyla 1% Error	AVG	86.93	89.84	92.57	82.46	47.64	35.20	58.83
	STD	2.28	2.31	0.71	8.03	9.76	15.94	15.08
Phyla 10% Error	AVG	83.74	87.24	90.70	77.89	36.54	20.93	50.42
	STD	2.54	2.58	0.76	8.72	9.33	10.83	18.73

Conclusions

• Using whole genome training, under-representation of examples of taxa degrades BLAST's accuracy at the phyla-level more than the genus-level. **Classification of "broadly-defined" taxa are more sensitive to underrepresentation than "finer-resolution" taxa.**

• BLAST accuracy at the genus-level degrades faster than phyla by reduced coverage (partial-genomes in the database) and sequence query error. **Accuracy for "finer-resolution" taxa degrades faster with database and query error than "broadly-defined" taxa. And Underrepresented phyla accuracies are extremely sensitive to database and query errors.**

• Users of BLAST should be aware that **insufficient taxa representation in the database may skew BLAST's** ability to correctly label a next-generation read **taxonomically**. This effect may compound when examining higher-level taxa and with sequence/database error and incomplete training.

This work was supported in part by National Science Foundation award #0845827