The Effect of Sequence Error and Partial Training Data on BLAST Accuracy

Steven D. Essinger & Gail L. Rosen
Electrical & Computer Engineering
Drexel University
Philadelphia, PA USA
sessinger@drexel.edu

Abstract - **Metagenomics is the study of environmental samples. Because few tools exist for metagenomic analysis, a natural step has been to utilize the popular homology tool, BLAST, to search for sequence similarity between DNA reads and an administered database. Most biologists use this method today without knowing BLAST's accuracy, especially when a particular taxonomic class is under-represented in the database. The aim of this paper is to benchmark the performance of BLAST for taxonomic classification of metagenomic datasets in a supervised setting; meaning that the database contains microbes of the same class as the 'unknown' query DNA reads. We examine well- and under-represented genera and phyla in order to study their effect on the accuracy of BLAST. We investigate the degradation in BLAST accuracy when genome coverage is reduced in the training database as well as the performance when errors are introduced into the query DNA reads. We conclude that on fine-resolution classes, such as genera, the accuracy of BLAST does not degrade very much with under-representation, but in a highly variant class, such as phyla, performance degrades significantly when whole genomes are used in the training database. BLAST accuracy at the genus level is affected greater than phyla when coverage in the training database is reduced or when 1% sequence error is introduced into the query DNA reads. Our analysis includes five-fold cross validation to substantiate our findings.**

Keywords – Metagenomics, BLAST, Error analysis

I. INTRODUCTION

The relatively new field of metagenomics has been rapidly expanding over the past several years [1, 2]. This field focuses on DNA obtained from an environmental sample rather than from pure cultures in a laboratory. This markedly substantial difference from conventional microbial genomics poses a unique set of problems that are now gaining attention. Instead of asking the question "How does one organism work?" we are now interested in "Who is here in this sample and what are they doing?". Since greater than 99% of microbes cannot be cultivated in isolation [3], metagenomics is a necessity if we wish to understand the microbial diversity of our planet.

Examples of metagenomic applications include human health, soil fertility and forensics. The National Institute of Health has created an initiative called The Human Microbiome Project to examine microbes associated with health in several areas of the human body [2]. For example it is hypothesized that the human gastrointestinal tract contains microbes that outnumber human cells 10 to 1 [2]. Many of these microbes are believed to be involved with the digestive process. Most of these microbes cannot be isolated in the laboratory. Therefore they cannot be cultured for abundance so that their DNA can be extracted and amplified for genomic analysis. Instead we turn to metagenomics where we obtain the DNA of the environmental sample, extract and amplify the DNA, sequence the samples, assemble the samples and finally attempt to annotate the sequences. Annotation is certainly an elusive task since we do not know which microbes are in the sample to begin with. So we turn to sequence alignment tools such as BLAST [4, 5] which aid us in answering a fundamental question in metagenomics, namely "Who is here?". Before we can fully trust the results of BLAST for taxonomic classification, we seek to benchmark how database representation affects its performance.

II. BACKGROUND ON TAXONOMY

Answering the question "Who is here?" is an issue of taxonomy. Taxonomy refers to the science of naming and classifying organisms. The National Center for Biotechnology Information (NCBI) maintains a taxonomy database, which is considered a well-respected source by the scientific community for taxonomic information [4]. The standard hierarchy of the taxonomy used in this paper is Phyla, Order, Family, Genus, Species, and Strains as recommended by the NCBI. As of September 2009 there are over 339,500 taxa represented in the database. Of these taxa 968 are completely sequenced genomes of microbial organisms. Clearly, this is only a small fraction of the microbes inhabiting our planet today, however, the databases are expanding rapidly and as the field of metagenomics becomes more pervasive we shall see substantial increases in the number of taxa maintained in these databases.

When an organism's DNA or metagenomic sample has been sequenced it is a natural step to compare this new sequence to existing, annotated sequences in the databases for similarity [6, 7]. BLAST (Basic Local Alignment Search Tool) is both a web based and standalone tool developed by the NCBI for comparing sequence similarity between two nucleotide or protein sequences [5]. The most popular way researchers use the tool is to input a sequence as a query against the public sequence databases, which include NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/). BLAST returns sequences that are similar to the input query. BLAST will attempt to align the query with the sequences in the

978-0-7695-4083-2/10 $26.00 © 2010 IEEE
DOI 10.1109/BIBE.2010.49

databases and then issue a statistical report to provide a level of confidence in the alignment. BLAST is actively maintained by the NCBI and can be found here (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

The first alignment in the report returned by BLAST is supposedly the sequence in the database with the greatest similarity to the query sequence. When the query sequence is small (e.g. < 500bp), BLAST tends to produce multiple ambiguous top-hits. It has been found that the closest BLAST hit is often not the nearest neighbor [8]. Generally speaking, microbiologists rely on the BLAST results without question [9, 10, 11]. Researchers have now begun to analyze and compare the performance of BLAST for metagenomic datasets. The findings are indicating that classifying genome sequence fragments based on the best BLAST hit only yield reliable results if there are close relatives represented in database for comparison [12, 13].

III. METHODS

A total of 635 distinct microbial strains downloaded in 2008 from the NCBI GenBank database were considered for our experiments. We have found that each of the 635 strains in our database can be classified to one of 19 different phyla and 272 different genera. In order to partition the database for our experiments we decided to focus on two well represented and two under-represented classes each for the levels of phyla and genus. Thus two separate experiments were performed: one for the level of phyla and the other for genus. Table 1 shows the composition of each class for each experiment.

The two well-represented classes were chosen to be the two classes at each level that contained the greatest amount of microbial strains. For example, the phyla class Proteobacteria contained 315 strains out of the 635 strains in the overall database. The two under-represented classes were chosen arbitrarily so that they each contain no more than 20 strains. Many classes in the database contained only 1 strain; however the five-fold cross validation statistical measure necessarily requires that we have a minimum of 5 strains. We chose under-represented classes containing 10 to 17 strains as shown in Table 1.

The five-fold cross validation experiments proceeded as the following for phyla using 500bp DNA reads which we herein refer to as query fragments. The identical procedure was followed for genus thus yielding a total of twelve separate experiments. Six experiments were dedicated to varying the coverage in the BLAST training database. The second set of six experiments focused on introducing error into the query fragments. The distribution of the classes for the experiments can be found in Table 1.

We randomly partitioned the strains from each class into five groups as necessitated by five-fold cross validation. The first group from each class was combined to create a set of query strains. To simulate a metagenomics dataset obtained using the next generation of 454 pyrosequencing technology [14], each query strain's genome was randomly sampled extracting 100 fragments each 500bp in length. Each fragment was annotated with its membership class so that we could determine if BLAST correctly matched the fragment. These sampled fragments were used as queries for BLAST sequence alignment. The whole-genomes of the remaining strains were used to construct the BLAST training database in which BLAST would attempt to align against the query sequences. For example, in the phyla experiment, 93x100 (20%) query fragments were BLAST against a database of 370 (80%) whole-genomes comprised of the remaining strains belonging to the 4 phyla. The percent accuracy is calculated as the number of query fragments correctly identified by BLAST over the total number of query fragments. This procedure was repeated a total of five repetitions so that each strain was in the query test set once. The results from the five partitions were averaged and the standard deviation was calculated. A survey of cross validation methods can be found from these sources [15, 16, 17].

A. Varying the Training Database Coverage

The varying database coverage experiments tested BLAST's resiliency in accuracy when the coverage per genome in the database was reduced. We randomly sampled each whole-genome 100x taking contigs 50Kbp (10Kbp) in length to construct the training database. The experiment proceeded identically as described above upon which we compared the results of whole-genome, 5Mbp and 1Mbp coverage in the training database.

TABLE I.

Phyla					
Total Strains – 463		***Database (80%) – 370***		***Query (20%) – 93***	
Well-Represented			***Under-Represented***		
Class	***# Of Strains***	***# Queries Sampled***	***Class***	***# Of Strains***	***# Queries Sampled***
Proteobacteria (well1)	315 (68%)	63	Crenarchaeota (under1)	15 (3%)	3
Fermicutes (well2)	116 (25%)	23	Tenericutes (under2)	17 (4%)	4
Genus					
Total Strains – 64		***Database (80%) – 51***		***Query (20%) – 13***	
Well-Represented			***Under-Represented***		
Class	***# Of Strains***	***# Queries Sampled***	***Class***	***# Of Strains***	***# Queries Sampled***
Steptococcus (well1)	26 (40%)	5	Yersinia (under1)	10 (16%)	2
Staphylococcus (well2)	18 (28%)	4	Synechococcus (under2)	10 (16%)	2

The class composition for the phyla and genus five-fold cross validation experiments are provided below. A total of 463 strains were included in the phyla experiment. We chose to use two phyla having well-representation and two having under-representation in the database. For example, Proteobacteria (well) accounted for 315 (68%) of the 463 strains included in the experiment. These strains were partitioned into five groups each containing 63 strains. The remaining three classes were partitioned in the same manner ensuring that approximately 20% of the strains belonging to the class were in each group. The first group from all four classes was combined and BLAST against the remaining four groups. This procedure was repeated five times so that each group was used for query once. An identical procedure was used at for the genus experiment.

B. *Varying Error in the Query Sequence*

The query sequence error model experiments were developed to quantify BLAST's accuracy in classifying fragments at both the genus and phyla levels when errors have been introduced in the query fragments. It has been shown that Roche pyrosequencing has a per-base accuracy of 96%, however, this figure depends on a number of factors [18]. To simulate sequencing error we randomly changed 1% (10%) of the bases in each 500bp fragment to a randomly chosen nucleotide. The rest of the experiment proceeded identically as described above. We compare the results of 0%, 1% and 10% sequencing error.

BLAST may potentially return multiple ambiguous hits meaning that all of the top scores returned have the same statistical expect value (e-value). In these instances all of the aligned sequences must be from the true taxonomic class otherwise the BLAST result was marked incorrect for the corresponding query sequence. Additionally, BLAST may not return a report for a query sequence that it has determined to be a low-complexity region. In these few instances we marked the query as incorrect. While this filter may be turned off we've found that BLAST consumes significantly more resources; therefore we've chosen to leave it in the default setting. Corresponding, we chose to use all BLAST default settings including an e-value cutoff of 10.

IV. RESULTS

A. *Varying Database Coverage Experiments*

First, we aim to show how full and partial training data effects BLAST's ability to classify fragments into their taxonomy. This is important because for every complete microbial genome project in GenBank, there are two projects: "in-progress", or those having incomplete coverage of the actual genome. Therefore, we show how using 100 random 50kbp (5 Mbp total genomic data) and 10kbp contigs (1 Mbp genomic data) compares to having the full-genome for training.

The results of the five-fold cross-validation experiments with well/under representation for assessing training database coverage are summarized in Table 2. BLAST accuracy was evaluated for classification both at the genus and phyla levels. The coverage of each genome included in the training databases was varied from whole-genomes to genomes consisting of 100 random samples each 50Kbp (10Kbp) in length. Therefore we performed three separate experiments at each level.

Each experiment had four classes; two classes that were well represented by strains in the dataset and two classes that were under-represented. The percent accuracy is the number of strain fragments that BLAST matched with the correct class over the total number of fragments. The average score reported is the average of all five repetitions of the five-fold cross validation experiment. The standard deviation is calculated in a similar manner. Individual scores for each repetition, for all experiments are provided in the appendix.

In addition to the percent accuracy of BLAST across all strains for each experiment, Table 2 lists the accuracy of BLAST on the four individual classes as well as the accuracy on the combined well and under-represented classes. Each of these combined groups contains two classes.

All seven different scores for percent accuracy are plotted against the six experiments shown in Figure 1. Phyla generally had lower accuracy than genus for all experiments with the exception when there was 1Mbp coverage in the training database. Figure 1 clearly shows that the percent accuracy of all strains for each experiment is highly dependent on the BLAST's ability to correctly identify the fragments belonging to strains having membership in the under-represented classes. For example, the genus level accuracy using whole genome training scored similarly across all classes while under-represented phyla classes using whole-genome training scored nearly 40% less than the phyla well-represented classes. This disparity results in an overall score for phyla 10% less than that of genus for the whole genome experiments.

As the genome coverage in the training database is reduced we see from Figure 1 that the average percent accuracy also

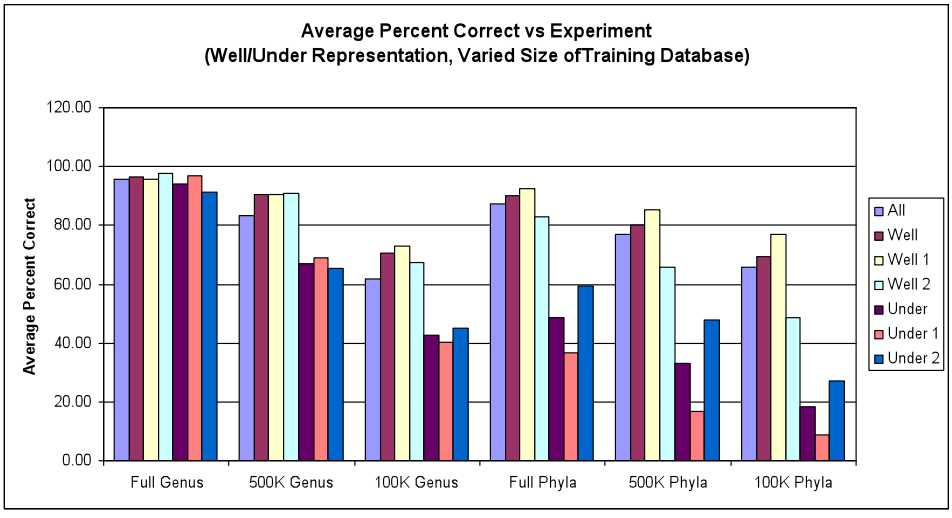

Figure 1. This bar graph illustrates the data provided in Table 2. All four classes in the full genus experiment exhibited similar percent accuracy scores. However, this percent accuracy degrades when BLAST is trained on partial genomes. This trend is also evident in the phyla experiments. There is also a greater difference in percent accuracy between the well- and under-represented classes in the phyla experiments than the genus experiments. We've found that this is due in part to the genus level having less diversity than the phyla level. The percent accuracy decreased 34% (genus) and 22% (phyla) when partial genomes of 1 Mbp were included in the training database as opposed to whole genomes.

TABLE II.

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

The percent accuracy scores of BLAST for the genus and phyla experiments are provided below. BLAST was marked correct if it matched the query fragment to the correct class and incorrect otherwise. It was also marked incorrect if it provided multiple ambiguous hits whereupon these hits belonged to two or more different classes. The percent accuracy for each five-fold cross validation repetition is the number of correct matches over the total number of query fragments. The percent accuracy scores over all five repetitions were average and are provided below along with the standard deviation of scores. The whole genome caption indicates that the database was comprised of whole genomes representing each class. 5Mbp (1Mbp) indicates the database was comprised of partial genomes 5Mbp (1Mbp) in length representing each class.

decreases. The genus level was most significantly affected decreasing coverage from whole genome to 1Mbp of each genome. The percent accuracy decreased 34% while the decrease was 22% for phyla. The genus under-represented classes experienced nearly a 50% decrease in accuracy while the phyla under-represented classes had an approximately 30% reduction in accuracy.

B. Query Sequence Error Model Experiments

In this experiment, we aim to show how error in a DNA sequence affects BLAST's ability to taxonomically classify the fragment. Error can be viewed as coming from the DNA sequencing method [19, 20], or divergence in sequence due to mutation. In either scenario, BLAST should be robust to error in the sequence, and we will investigate this effect.

The results of the five-fold cross validation experiments with well/under representation assessing BLAST's performance on query sequences with error is summarized in Table 3. BLAST accuracy was evaluated for classification both at the genus and phyla levels. In contrast to the prior experiments the coverage of each genome included in the training databases remained the same using whole-genomes. They query-error experiments differ from the training database experiments because error is introduced into the query fragments while the training database size is constant: while the converse is true in the former. Three separate experiments were performed at both the genus and phyla levels. The first used unaltered query fragments. The second and third experiments used fragments each having 1% and 10% randomly changed base pairs, respectively.

The introduction of 1% error into the query fragments resulted in a nearly 1% decrease in average percent accuracy across all classes for phyla. This difference was much higher for the genus experiments with the under-represented classes suffering a loss of nearly 10% accuracy. However, when the error introduced in the query fragments was raised to 10% the difference in accuracy for genus was negligible compared to the 1% error experiment. Phyla accuracy continued to decrease in a near linear fashion moving from 0% to 10% error in the query fragments. On average this decrease was about 5% across of the classes for phyla.

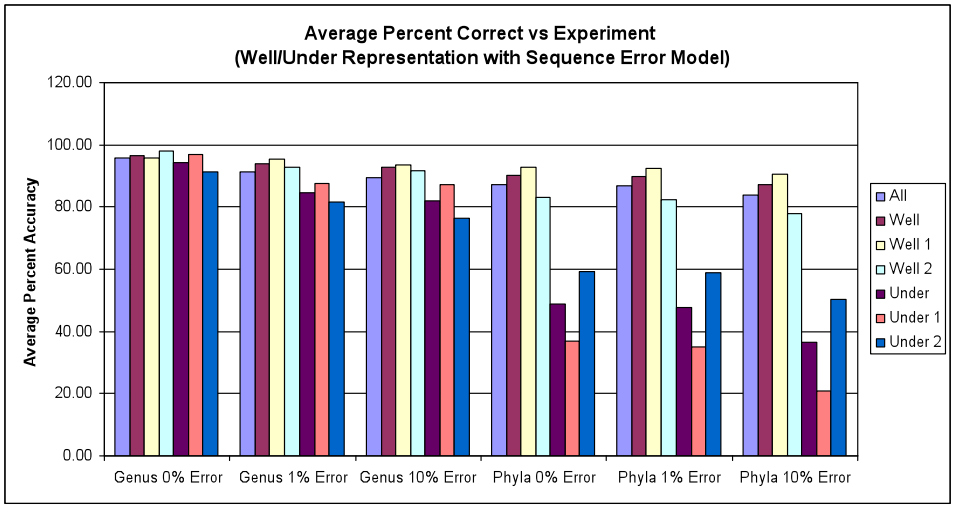

Figure 2. This bar graph illustrates the data provided in Table 3. The average percent accuracy for all strains in the genus experiments decreases 6% when 10% sequence error is introduced in the query fragments. The average percent accuracy for all strains in the phyla experiments decreases 4% when 10% sequence error is introduced in the query fragments. Under-represented classes experience larger decreases in percent accuracy than well-represented classes. Both under-represented classes in the genus and phyla experiments experienced a decrease of approximately 12% with the introduction of 10% sequence error.

TABLE III.

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

The results of the sequence error model five-fold cross validation experiments are provided below. To simulate sequencing error we randomly changed 1% (10%) of the 500bp in each of the query fragments. The first score column is the percent accuracy of BLAST for all four classes while the four columns to the right, labeled with the phyla's abbreviated name, refer to the individual scores for each class considered in this experiment. The results show that accuracy at the genus level was more susceptible to sequencing error than phyla. However, while there was approximately 10% decrease in accuracy for the under-represented genus class with 1% sequencing error, the decreases were marginal between 1% and 10% error. Phyla's decrease in accuracy appears to have a near linear response to the introduction of sequencing error.

V. DISCUSSION

Previously, we have conducted an experiment to measure BLAST's accuracy of taxonomic classification at the levels of genus and phyla consisting of four classes; two that are well represented and two that are under-represented using whole genomes in the training database [18]. Our findings indicate that BLAST is able to classify much better than chance, however, BLAST assigns misclassified queries by chance. Overall, the greater the representation in the database, the greater the accuracy of BLAST is on taxonomic classification. This is especially true for phyla, which we believe to have more diversity than genus based on the findings in this reference [21]. Therefore, it is no surprise that the decrease in BLAST accuracy is greater for under-represented classes than those that are well represented in the training database.

It is clear from these experiments that the accuracy of BLAST is highly dependent on the composition of the training database. The whole genome phyla experiment confirmed that the well-represented classes have nearly 40% higher accuracy than the under-represented classes. Still BLAST is performing much better than chance on all classes. For phyla (whole-genome) we see that Proteobacteria (well) scored 92.67%. With a database composition of 252/370 we confirm that this score is much higher than chance, which would be about 68%. This can also be verified for under-represented classes. For instance BLAST scored 59.38% for Tenericutes (under). Given its database composition we expect a percent accuracy of 14/370 or 3.7% by chance.

As the coverage in the database was reduced the average percent accuracy scores decreased for all classes for both the genus and phyla experiments. Furthermore, the disparity in accuracy between the well and under-represented classes increased with the database reduction from whole-genome to 5Mbp down to 1Mbp. This was apparent in both the phyla and genus experiments. Still BLAST was capable of scoring better than chance with even with 1Mbp coverage in the database. Phyla Crenarchaeota (under 1) suffered the greatest decrease in accuracy scoring at 8%, yet this was still higher than its chance of 12/370 or 3.2%.

Upon further examination of the database's composition we observe the ratio of well to under-represented strains in the phyla database is nearly 14:1. Incidentally, by chance, if we rolled a die we would expect BLAST to classify a strain to a well-represented class 14/15 or 93.3% percent of the time. While we found through our experiments that BLAST is able to classify much better than chance, the allocation of BLAST misclassifications follows a different trend. For example, in the phyla experiments when BLAST misclassified a fragment approximately 95% (nearly chance) of the fragments were assigned to a well-represented class.

These trends are also reflected in the genus experiments. For example for genus (whole-genome) we find that Streptococcus (well) scored 96.6%. By chance we would observe 21/51 or 41.1% accuracy. For Yersinia (under) BLAST scored 91.4% while a score by chance would be 8/51 or 15.6%. The genus database composition is about 2.2:1 predicting that BLAST would classify a strain to a well-represented class about 69% of the time by chance. This is reflected in the allocation of BLAST misclassifications where about 76% of the BLAST misclassified fragments went to a well-represented class.

The sequence error model experiments have shown that errors in the query fragments affect classification at the genus level greater than at phyla. Additionally, the affect was greater on the under-represented classes. The decreases in average percent accuracy at the genus level were incremental between 1% and 10% error while a substantial decrease of 10% for under-represented classes was observed between the introduction of 0% and 1% error.

The much greater decrease in accuracy at the genus level with partial-training data or query sequence error can be explained by sequence divergence. It is well known that 16S rRNA has 3% divergence at the species level and goes up to 6% at the genus level [21]. Extending this finding we expect a greater divergence at the phyla level. Therefore we find that the finer the taxonomic resolution, the more susceptible to error the taxonomy will be.

VI. CONCLUSION

Twelve five-fold cross validation experiments were examined in this study spanning the taxonomic levels of genus and phyla. We showed how whole and partial training data effects BLAST's ability to classify fragments into their taxonomy. Additionally, we showed how error in a DNA sequence affects BLAST's ability to taxonomically classify the

fragment. Maximizing the coverage in the training database and reducing the amount of error in the query fragments increases the reliability of BLAST, however, there is always the potential for missing data in the training database and error in the query fragments so it's important to understand how BLAST is affected by these deficiencies.

Figure 1 intuitively indicates that genome coverage in the database affects BLAST's ability to correctly classify fragments at both the genus and phyla levels. When coverage is 1Mbp per genome BLAST is still able to correctly classify fragments better than chance, if only marginal, but generally greater by tens of percentage points. Table 4 highlights the decrease in accuracy when the training database coverage is reduced from whole genome sequences.

The results of the sequencing error study as described by Figure 2 show that BLAST accuracy decreases at the genus level by several percent with the introduction of 1% error. Classification at the phyla level is also affected by sequencing error, but not as great as the decrease exhibited by genus with 1% error. Further introduction of error appears to have minimal affect on genus, however, classification accuracy at the phyla level continues to decreases with a near linear trend. Table 5 highlights the decrease in accuracy observed when error is increased in the query sequence fragment from 0%.

Our study has shown that if a class is under-represented in the training database and only contains uncompleted genome projects, meaning lots of partial-training data, then BLAST performance may be severely limited. Users of BLAST should be aware of faulty classifications if they suspect they are querying taxa that have only a few examples in the training database. Overarching, our findings show that the higher up the phylogenetic tree we classify fragments into; the more robust they are to both partial training-data and test-data error, with genus-level accuracy more susceptible to both than phyla-level accuracy.

TABLE IV.

Table 4	**Genus**		**Phyla**	
Partial Genome Training Size	***5Mbp***	***1Mbp***	***5Mbp***	***1Mbp***
All	-12.69	-34.04	-10.45	-21.51
Well	-6.12	-26.08	-10.06	-15.81
Under	-28.1	-51.45	-15.63	-30.36

TABLE V.

Table 5	**Genus**		**Phyla**	
Query Error %	***1%***	***10%***	***1%***	***10%***
All	-4.7	-6.5	-0.28	-3.47
Well	-2.53	-3.98	-0.22	-2.82
Under	-9.6	-12.25	-1.1	-12.2

Percent Change in Accuracy for Partial Training (Relative to whole-genome case) [Table 4]. Percent Change in Accuracy for Query-Sequence Error Model (Relative to 0% error case) [Table 5]. The column *All* includes the performance across all four classes in each experiment. The *Well* and *Under* columns each include their two respective classes in each experiment. BLAST decreases in accuracy both for the levels of genus and phyla when the coverage in the training database is reduced or with the introduction of error in the query sequence fragment. Overall, BLAST accuracy at the genus level is more susceptible to reductions in coverage in the training database or introduction of error in the query sequence than at the phyla level.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. #0845827.

REFERENCES

1. V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis and P. Hugenholtz, *Micro Mol Biol Rev.* **72**, 4 (2008).
2. G. Rosen, B. Sokhansanj, R. Polikar, M. Bruns, J. Russell, E. Garbarine, S. Essinger, and N. Yok, *Current Genomics*. **10**, 7 (2009).
3. J. Handelsman, Committee on Metagenomics: Challenges and Functional Applications, N. R. Council, Ed. *The National Academies Press*, (2007).
4. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, *J. Mol Biol*. **215**, 3 (1990).
5. T. Madden, *The NCBI Handbook.* Ch. 16, 1-17 (2003).
6. J. Venter, K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelson, W. Nelson, D. Fouts, S. Levy, A. Knap, M. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Tillson, C. Pfannkoch, Y. Rogers, and H. Smith, *Science*. **304**, 5667 (2004).
7. M. Tress, D. Cozzetto, A. Tramontano, and A. Valencia, *BMC Bioinformatics*. **7**, 213 (2006).
8. L. Koski and G.Golding, *J. Mol Evol.* **52**, 6 (2001).
9. A. Anderson, M. Lindberg, H. Jakobsson, F. Backhed, P. Nyren, and L. Engstrand, *PLoS One.* **3**, 7 (2008).
10. D. Huson, A. Auch, J. Qi, and S. C. Schuster, *Genome Res*. **17**, 3 (2007).
11. S. Havre, B. Webb-Roberston, A. Shah, C. Posse, B. Gopalan, and F. Brockman, *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, 341–350 (2005).
12. K. E. Wommack, J. Bhavsar, and J. Ravel, *Appl Environ Microbiol.* **74**, 5 (2008).
13. C. Manichanh, C. Chapple, L. Franguel, K. Gloux, R. Guigo and J. Dore, *Nucleic Acids Res*. **36**, 16 (2008).
14. L. Krause, N. Diaz, A. Goesmann, S. Kelley, T. Nattkemper, F. Rohwer, R. Edwards, and J. Stoye, *Nucleic Acids Res*. **36**, 7 (2008).
15. G. L. Rosen, E. M. Garbarine, D. A. Caseiro, R. Polikar, and B. A. Sokhansanj, *Hindawi Adv Bioinfo*. **2008**, (2008).
16. R. Kohavi, *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 1137–1143 (1995).
17. P. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach, *Prentice-Hall*, London. (1982).
18. S. Essinger and G. Rosen, *Pacific Symposium on Biocomputing*, (2010).
19. S. Huse, J. Huber, H. Morrison, M. Sogin and D. Welch, *Genome Biology*. **8**, 143 (2007).
20. J. Petrosino, S. Highlander, R. Luna, R. Gibbs and J. Versalovic, *Clinical Chemistry*, **55**, 856-866 (2009).
21. J. Clarridge, *Clin Microbiol Rev*. **17**, 4 (2004).