After acquiring a sample and using next-generation technology to perform shotgun sequencing, the next step in metagenomic analysis is to assess the taxonomic content of the sample. This methodology, also known as phylogenetic analysis, gives a simple look at ‘Who is in this sample?’ The first tool ever used (which is still widely used) for taxonomic assessment is Basic Local Alignment Search Tool (BLAST; Altschul et al., 1990). In recent years, several specialized webservers have been made available to the public to ease the process of taxonomically classifying reads, namely Phylopythia (McHardy et al., 2007), CAMERA (Seshadri et al., 2009), WebCARMA (Gerlach et al., 2009), MG-RAST (Meyer et al., 2008) and Galaxy (Pond et al., 2009). Unlike BLAST, Phylopythia and WebCARMA return more specific taxonomic information and assign reads to higher level taxonomic levels using a consensus of models (HMMs) to match the EGTs against protein families during homology-based approach that matches environmental gene tags to protein families and reports good results for long and ultrashort 35-bp reads using (i) BLASTX to find candidate environmental gene tags (EGTs) and (ii) using Pfam (protein family) hidden Markov models to match the EGTs against protein families during an EGT candidate selection process. MG-RAST offers a homology-based approach, SEED (Overbeck et al., 2005), CAMERA and Galaxy provide high-throughput implementations and custom databases for BLASTN. BLASTN yields best hit sequence matches and is known to have reasonable accuracy (Rosen et al., 2009).

Previously, Rosen et al. have explored a machine learning method, naive Bayes classifier (NBC), as a possible way to classify fragments that can annotate more sequences than BLAST (Rosen et al., 2008). We now implement the algorithm on a webservice for public use and benchmark it against other web sites.

1 METHODS AND MATERIALS

We selected a previously benchmarked dataset (Gerlach et al., 2009): the Biogas reactor dataset (Schiöller et al., 2008), composed of 355,213 reads of average 230-bp length. We selected a real dataset as opposed to a synthetic one because we did not want to tailor the dataset to any specific database, since the database will vary on each web site. This comparison fairly assesses each webserver’s performance on a ‘real’ dataset containing known and novel organisms.

We conducted our tests against NBC and five other webservers in July and August of 2010. WebCARMA and MG-RAST require no parameters. Phylopythia requires the type of model to match against. MG-RAST requires an E-value cutoff under the SEED viewer (which we selected the highest). We selected default BLAST parameters for the NT database for Galaxy. For NBC, we used an E-value of 15 and the default 1032 organism genome list. For CAMERA, we only retained the best top-hit organism for each read and used the ‘All Prokaryotes’ BLASTN database (and used the default parameters for the rest).

We implement the NBC approach in Rosen et al. (2008) that assigns each read a log-likelihood score. We introduce two functions of NBC: (i) the
The naïve Bayes classification tool is implemented on a web site database and therefore, it did not find any matches. Thermosinus is not in NBC’s completed microbial training set. Therefore, NBC potentially missed Thermosinus. Since this is an agricultural bioreactor, it tend to agree for over 70% of the genera shown while MG-RAST agrees for 50% of the genera shown while MG-RAST agrees with CAMERA and NBC near 50%. WebCARMA bins fewers reads, and Galaxy has high variability. For the first 5602 reads (1.5 Mb web site limit), Phylopythia only classifies eight reads to the phylum level and is not included in the graph due to its inability to make assignments at the genus level.

3 DISCUSSION

In Figure 1, we show the percentage of reads (out of the whole dataset) that ranked in the top eight genera for each algorithm. We see that all methods are in unanimous agreement for Clostridium and Bacillus, while most methods (except Galaxy) agree for prominence of Methanoculleus. CAMERA supports NBC’s findings of Pseudomonas and Burkholderia, known to be found in sewage treatment plants (Vinners et al., 2006). The biogas reactor contained ~2% chicken manure so it can have the traits of sludge waste (Schlüter et al., 2008). In Hery et al. (2010), Pseudomonas and Sorangium have been found in sludge wastes. Streptosporangium and Streptomyces are commonly found in vegetable gardens (Nolan et al., 2010), which is quite reasonable since this is an agricultural bioreactor. Therefore, NBC potentially has found significant populations of genera that other classifiers have missed. Thermosinus is not in NBC’s completed microbial training database and therefore, it did not find any matches.

NBC took 21 h to run and classified all 100% of the reads compared with 12h23% for WebCARMA, 5h99% for CAMERA, 2–3h140% for Galaxy 1, and a few weeks 756.2% for MG-RAST. NBC runs on a 4-core Intel machine and speed would linearly increase with distributed computing in the future.

4 CONCLUSION

The naïve Bayes classification tool is implemented on a web site for public use. We demonstrate that the tool can handle a complete pyrosequencing dataset, and it gives the full taxonomy for each read, so that users can easily analyze the taxonomic composition of their datasets. NBC classifies every read unlike other tools and is easy to use, runs an entire dataset in a reasonable amount of time and yields competitive results.

ACKNOWLEDGEMENT

We thank Christopher Cramer for the scoring code and binary packages.

Funding: Supported in part by the National Science Foundation CAREER award #0845827 and Department of Energy award DE-SC0004335.

Conflict of Interest: none declared.

REFERENCES

Ahmed,N.F. et al. (2007) Monitoring of bacterial communities during low temperature thermal treatment of activated sludge combining dna phylochip and respirometry techniques. Water Res., [Epub ahead of print, doi: 10.1016/j.watres.2010.07.003.].
Huson.D.E. et al. (2007) Megan analysis of metagenomic data. Genome Res., 17, 377–386.

McHardy,A.C. et al. (2007) Accurate phylogenetic classification of variable-length dna fragments. Nat. Methods, 4, 63–72.

Meyer,F. et al. (2008) The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9, 386.

Nolan,M. et al. (2010) Complete genome sequence of streptosorangium roseum type strain (ni 9100t). Stand. Genomic Sci., 2, 1.

Overbeek,R. et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res., 33, 5691–5702.

Pond,S.K. et al. (2009) windshield splatter analysis with the galaxy metagenomic pipeline. Genome Res., 19, 2144–2153.

Rosen,G.L. et al. (2009) Signal processing for metagenomics: extracting information from the soup. Curr. Genomics, 10, 493–510.

Rosen,G.L. et al. (2008) Metagenome fragment classification using n-mer frequency profiles. Adv. Bioinformatics, 2008, Article ID 205969.

Schlüter,A. et al. (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J. Biotechnol., 136, 77–90.

Seshadri,R. et al. (2007) Camera: A community resource for metagenomics. PLoS Biol., 5

Vinneras,B. et al. (2006) Identification of the microbiological community in biogas systems and evaluation of microbial risks from gas usage. Sci. Total Environ., 367, 606-615.