A Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy

Qiong Wang[1], George M. Garrity[1], James M. Tiedje[1], James R. Cole[1]

Keywords: classification, naïve Bayesian classifier, ribosomal RNA, bacterial taxonomy

1 Introduction.

Starting in the mid '80s, Carl Woese revolutionized the field of microbiology with his ribosomal RNA-based phylogenetic comparisons delineating the three main branches of life [5]. Today, rRNA based analysis remains a central method in microbiology, used not only to explore microbial diversity, but as a day-to-day method for bacterial identification. rRNA identification (classification) methods, as opposed to phylogenetic (clustering) methods have been hindered due to the lack of a consistent higher-level bacterial classification structure (taxonomy). This situation changed recently, when in 2002, Bergey's Trust published a revised higher-order taxonomy attempting to reconcile bacterial taxonomy with rRNA based phylogeny [2].

We have developed a naïve Bayesian classifier for classifying bacterial rRNA sequences into the new Bergey's bacterial taxonomy. This classifier is fast, does not require sequence alignment and works well with partial sequences. (The vast majority of rRNA sequences in the public databases are partial.) This classifier is currently being used internally by the Ribosomal Database Project [1] (RDP; http://rdp.cme.msu.edu) to organize its publicly available sequence library.

2 Data and Methods.

Small subunit ribosomal RNA sequences from approximately 4400 bacterial species type strains in 900 genera were obtained from Bergey's Trust, along with associated taxonomic assignment information [2]. The sequences averaged 1459 bases in length with a range of 1200 - 1775 bases. These training sequences were each labeled with a set of taxa, from domain to genus. The classifier uses a feature space consisting of all possible eight-base subsequences (words) in the query molecule. Word-specific priors were calculated from their frequency in the entire training set. As with text-based Bayesian classifiers, only those words occurring in the query contribute to the score [3]. A similar word-based classification scheme has been used to search for horizontal gene transfer events in whole-genome sequences [4].

To classify a query, the joint probability of observing the words in the query was calculated separately for each genus from the training set probability values. For bootstrap analysis, the collection of all overlapping unique words in the query was first calculated. Then a subset of these words was randomly chosen (with replacement) and the words in this subset were then used to calculate the joint probability. (Since overlapping words are highly dependent, we conservatively chose only one-eighth of the words for each trial.) The number of times a genus was selected out of 100 bootstrap trials was used as an estimate of confidence in the assignment to that genus. For higher-rank assignments, we sum the results for all genera under each taxon.

---

[1] Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, E-mail: {wangqion, garrity, tiedjej, colej}@msu.edu

This research was supported by DOE-OBER grant DE-FG02-99ER62848 & NSF grant DBI-0328255.

3 Results.

We tested the classifier by exhaustive leave-one-out testing. For each test, we reserved a single training set sequence as query and re-trained the classifier on the remaining sequences. The process was repeated for all sequences in the training set. In addition to the near-full-length sequences, we also tested the classifier on small contiguous regions of 400 and 200 bases chosen at random from the test sequences (Table 1). For the near-full-length and 400 base partial rRNA sequences, the classifier was highly accurate down to the genus level, while with 200 base partial sequences the classifier was accurate at the phylum and class levels. The bootstrap provided a good estimate of classification reliability (Table 2). Overall, 90.4% of taxon assignments matched in 95 or more of the 100 bootstrap trials and these assignments were correct 98.7% of the time.

This classifier is fast enough to handle large sample volumes. On a 1Ghz Apple G4 processor, it can classify approximately 5 sequences per second (with 100 bootstrap samples of each). The new taxonomy is still evolving as species are reevaluated and discrepancies are resolved. As these changes occur, it has proved relatively simple to re-train the classifier and update the assignments of the greater than 86,000 sequences in the RDP library.

length	phylum (%)	class (%)	order (%)	family (%)	genus (%)
1459	99.4	98.7	97.2	94.2	91.0
400	99.2	98.4	96.6	93.0	87.7
200	91.6	87.4	77.3	60.4	46.5

Table 1: Classifier accuracy at different taxonomic ranks for varying query lengths.

	Number of bootstrap assignments out of 100 trials					
rank	100-95	94-90	89-80	79-70	69-60	59-50
phylum	4332/4340[†]	51/51	19/19	10/10	6/6	6/13
class	4186/4214	40/50	41/42	18/20	15/18	4/6
order	4005/4058	79/81	55/61	31/37	27/30	18/26
family	3705/3786	110/117	98/126	54/72	47/68	37/59
genus	3025/3113	192/207	157/191	97/126	60/95	69/108
overall	98.7%	93.2%	84.3%	79.2%	71.4%	63.2%

[†] Number of correct assignments over total number of assignments.

Table 2: Classifier accuracy versus bootstrap confidence estimate.

References

[1] Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarrell, D.M., Schmidt, T.M., Garrity, G.M. and Tiedje, J.M. 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*, 31(1):442-3.

[2] Garrity, G.M., Bell, J.A. and Lilburn, T.G. 2003. Taxonomic outline of the prokaryotes. *Bergey's Maunal of Systematic Bacteriology*, Second Edition. Release 4.0. New York: Springer-Verlag. DOI: 10.1007/bergeysmanual.

[3] Li, Y.H. and Jain, A.K. 1998. Classification of text documents. *The Computer Journal*, 41(8):537-546.

[4] Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. and Coster, J. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Research*, 11(8):1404-9.

[5] Woese, C.R., Kandler, O. and Wheelis, M.L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci*, 87(12):4576-9.