A best-score strategy: The following sequence was submitted to the NCBI BLAST web server (and separately, MARTA):

CAGAACCAAGAGATCCGTTGTTGAAAGTTTTGTTTAATTTGCTTAAACTCCGACGCAGAGATGCAGGGTTGGAGG
GCCTCCGGGGGCGCTCGCCGTCGAGACGGCAGGGTCGCCCCCGAAGCAACAAGTGTGTTCACAGAGGGTGGGA
GGTCGGGCCCGGGGGCCCTCACTCGGTAATGATCCCTCCGCAGGTTCACCTACGGAGACCTTGTTATGACTT

This resulted in the following:

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
EU167607.1	Mycosphaerella brassicicola strain CBS 174.88 small subunit ribosomal RNA gene, inte	390	390	100%	1e-105	98%
AF297236.1	Mycosphaerella brassicicola IPO99157 18S ribosomal RNA gene, partial sequence; 5.8	390	390	100%	1e-105	98%
AF297227.1	Mycosphaerella brassicicola IPO99156 18S ribosomal RNA gene, partial sequence; 5.8	390	390	100%	1e-105	98%
AF297223.1	Mycosphaerella brassicicola IPO99510 18S ribosomal RNA gene, partial sequence; 5.8	385	385	100%	6e-104	98%
GQ511723.1	Uncultured fungus clone OTU#3475-62-3988_3800 internal transcribed spacer 2, part	340	340	100%	1e-90	94%
EU167603.1	Mycosphaerella berberidis strain CBS 324.52 small subunit ribosomal RNA gene, inter	340	340	100%	1e-90	94%
GQ510303.1	Uncultured fungus clone OTU#2053-68-3943_0310 internal transcribed spacer 2, part	335	335	100%	6e-89	94%
GQ509652.1	Uncultured fungus clone OTU#1400-38-3601_1728 internal transcribed spacer 2, part	335	335	100%	6e-89	94%
GQ509227.1	Uncultured fungus clone OTU#974-44-3853_2085 internal transcribed spacer 2, partia	335	335	100%	6e-89	94%
GQ508979.1	Uncultured fungus clone OTU#726-62-3634_1028 internal transcribed spacer 2, partia	335	335	100%	6e-89	94%
GQ515838.1	Uncultured fungus clone Unisequence#62-3688_0476 internal transcribed spacer 2, p	335	335	100%	6e-89	94%
GQ519006.1	Uncultured fungus clone Unisequence#68-3374_3755 internal transcribed spacer 2, p	335	335	100%	6e-89	94%
GQ526313.1	Uncultured fungus clone Unisequence#39-3567_2389 internal transcribed spacer 2, p	335	335	100%	6e-89	94%

Note that the top score is 390, and that three hits share this same bitscore.

The BLAST web server and MARTA use a phylogenetically informative word-size of 28 (which is easily modified; arguments for MARTA follow the form: *–ws=28*). MARTA imposes an additional set of default requirements (arguments shown in parentheses):

1. A candidate accession's coverage must at minimum be 80% of the length of the query sequence (*-cov*=80).

2. The percent-identity >= 97% (*-p=97*)

3. The default consensus (%) requirements for the eight major taxonomic ranks are:

   - Domain, Kingdom, Phylum, Class, Order and Family: 100% consensus

   - Genus, Species: 66% consensus.

   To enforce 80% consensus at the genus and species levels, one would call MARTA with the argument: *-cutoffs=1,1,1,1,1,1,.8,.8*

Because there is a tie for the best score (390) the default consensus requires at least two of the three accessions to share the same 'taxonid' at the species level node and, if consensus isn't found there, 2 of 3 sequences must share the same taxonid at the genus level node. If consensus isn't found at either the species or the genus level, the software will step back through the hierarchy, from the family level node to the domain level, while requiring 100% consensus at each taxonomic rank until consensus is found. Because 3 of 3 accessions share the value: *Mycosphaerella brassicicola* (actually the taxonid: 161656), the value is assigned to the sequence.

A slip-score strategy:

The following sequence was submitted to the BLAST web server:

CAGAGCCAAGAGATCCGTTGTTGAAAGTTTTAAATATTAAATATTTTACTCAGACGACACTAATAATTCAGGGTTTT
GGGTTCTCTGGCGGGCACTTACCAGCCGAAGCCAGTAGCTAGGCGGCCCGCCAAAGCAACAAAGGTATAGTATAC
AAAGGGTGGGAGATCTACCCCGAAGGGCATGAACTCGGTAATGATCCTTCCGCAGGTTCACCTACGGAAACTCTT
GTACTTACTTTCTCTACTA

This resulted in the following output:

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
GU055705.1	Uncultured Tetracladium clone NG_R_H03 18S ribosomal RNA gene, partial sequence;	416	416	93%	2e-113	99%
GU055702.1	Uncultured Tetracladium clone NG_R_G08 18S ribosomal RNA gene, partial sequence;	416	416	93%	2e-113	99%
GU055653.1	Uncultured Tetracladium clone NG_R_A10 18S ribosomal RNA gene, partial sequence;	416	416	93%	2e-113	99%
FJ776910.1	Uncultured fungus clone Contig893-158-1085_2477 18S ribosomal RNA gene, partial	416	416	93%	2e-113	99%
DQ350129.1	Tetracladium sp. AR-5 small subunit ribosomal RNA gene, partial sequence; internal t	412	412	90%	3e-112	100%
GU055701.1	Uncultured Tetracladium clone NG_R_G07 18S ribosomal RNA gene, partial sequence;	411	411	93%	1e-111	99%
EF434086.1	Uncultured fungus clone TF22_OTU156 small subunit ribosomal RNA gene, partial seq	411	411	93%	1e-111	99%
DQ182426.1	Uncultured ascomycete isolate 1 18S ribosomal RNA gene, partial sequence; internal	411	411	90%	1e-111	100%
AJ890435.1	Tetracladium sp. CBS 118523 18S rRNA gene (partial), ITS1, 5.8S rRNA gene, ITS2,	411	411	93%	1e-111	99%
DQ068996.1	Tetracladium maxilliforme clone NS170D 18S ribosomal RNA gene, partial sequence;	407	407	89%	1e-110	100%

At first glance, there is some support for the classification: *Tetracladium*

By default, MARTA does not determine the taxonomic status of the query sequence when the best-scoring hits are all "Uncultured". This is because Uncultured or "Unidentified" assignments are rarely phylogenetically informative (e.g. Uncultured fungus clone ###). However, if MARTA is run in "slip-score" mode, one can still recover some phylogenetic information from this sequence, by stepping down (iteratively) to the next set of scores until a classification is achieved. For example, the top score in this result is 416. If the user implements a slip-score strategy and a tolerance criterion, for example, of 98% of the top-score (*-tile=98*) the software:

(1) Will try to vote using a best-score strategy while considering the hits whose scores are 416.
(2) Because the database does not have a classification for the hits at 'shelf' 416, the software steps-down to the next shelf at which a single accession has the score 412.
(3) Here, the score still is within the window the user specified. A slip-score of 98 with a top-score of 416 results in a lower limit score of 407.68 (because 416*0.98 == 407.68). The single accession does have a taxonomic assignment, and *Tetracladium* is assigned to the sequence.
(4) If the accession on shelf 412 did not have a classification in the database, the program would consider the next shelf at 411, which similarly has one accession with taxonomic information.

In practice it is preferable to run MARTA using a permissive slip-score strategy (e.g. *–tile=75*), rather than a best-score strategy, and to then use post-hoc approaches to distinguish between scores that were in reality the 'best-scores' and those that slipped to some extent. Determining an appropriate slip-score for a wide range of taxa is an area of ongoing research.

Sample Output from MARTA:

When MARTA finishes, the output is a tab-delimited file with the following fields:

ID	S000006665
TAXONID	1883
LEVEL	GENUS
TAXON	Streptomyces
EVALUE	0.00E+00
SLIPSCORE	2920
TOPSCORE	2920
PERCENT-	99.93

IDENTITY	
COVERAGE	1
FULL-TAXONOMY	Bacteria::Actinobacteria:Actinobacteria (class):Actinomycetales:Streptomycetaceae:Streptomyces:
VOTES-FOR	20
VOTES-ALL	20

The 'Id' field is the user-assigned sequence Id. The TaxonId is the numeric value from the NCBI Taxonomy database that points to the Genus level node: *Streptomyces*. This is the winning taxon, and the full-taxonomy is colon, ':', delimited and reports each of the eight major taxonomic ranks. Twenty total votes were cast, and all twenty votes agreed, at the genus level, on the taxonomic assignment. Regardless of whether MARTA was executed in slip-score or best-score mode, we know that the winning taxon had the highest (best) score, since the slip-score and the top-score are the same (2920).

SOFTWARE EVALUATION – TEST CASE

The program CARMA is here:

http://webcarma.cebitec.uni-bielefeld.de/cgi-bin/webcarma.cgi

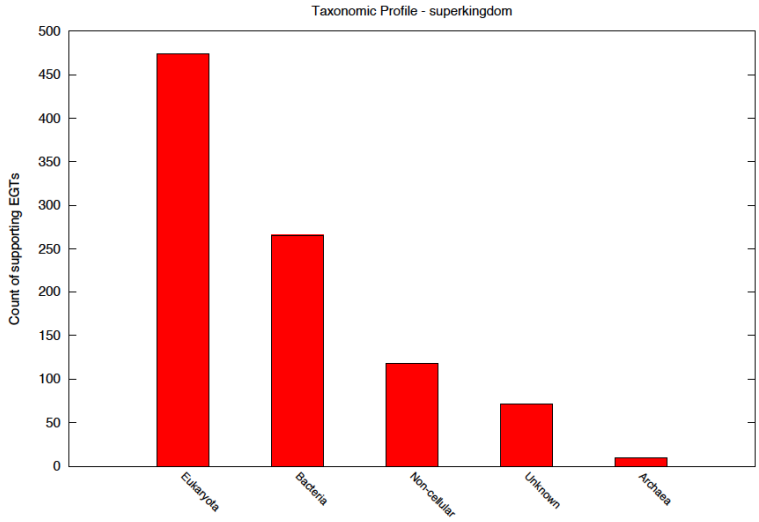
The RDP-II Classifier is here:

http://rdp.cme.msu.edu/classifier/classifier.jsp

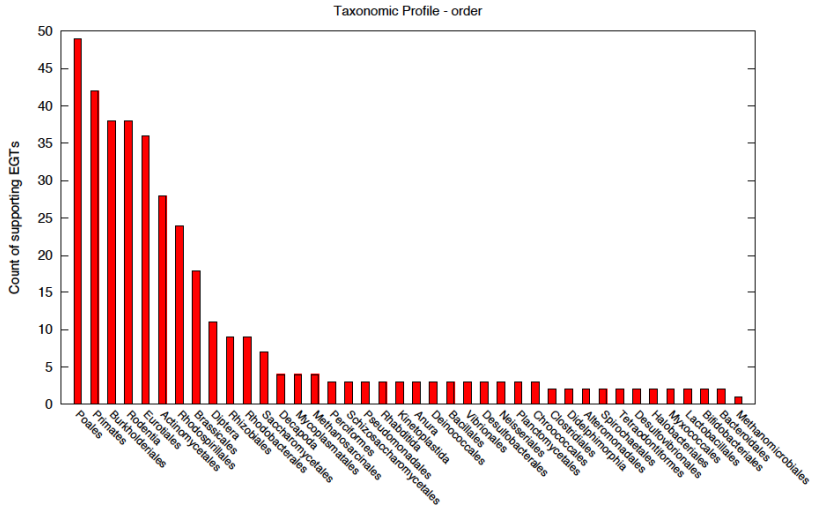
The test case corpus was constructed using the RDP-II browser found here:

http://rdp.cme.msu.edu/hierarchy/hb_intro.jsp

The phylogenetic assignments provided by CARMA show high error rates. The test case corpus is composed entirely of Bacterial Type strain sequences, yet the 'environmental gene tag' (EGT) category of highest abundance is labeled *Eukaryota* in the superkingdom profile provided by CARMA:



CARMA's phylogenetic classification at the order level is shown here, and it similarly shows erroneous classifications (e.g. *Poales, Primates, Rodentia, etc.*):



MARTA was run with the following parameters:

Strategy	Argument

Name of voting bout.	*-group=minscore*
Number of candidates*	*-top=100***
Percentage-Identity*	*-p=97***
Word-size*	*-ws=28***
NCBI Database*	*-db=nt***
Use Sun's Grid Engine (qsub) to run Marta in distributed mode	*-parallel*
Number of CPUs used during the alignment*	*-co=3***
Query-coverage	*-cov=80***
Voting strategy	*-minscore* (best-score strategy; omit this argument to run MARTA in slip-score mode)
Slip-Score	*-tile=98*** (omitted in this example to permit best-score strategy)
Taxonomic-thresholds (levels shown below):	*-cutoffs=1,1,1,1,1,1,2/3,2/3***, ***
DOMAIN/SUPERKINGDOM	100%***
KINGDOM	100%***
PHYLUM	100%***
CLASS	100%***
ORDER	100%***
FAMILY	100%***
GENUS	66.6%***
SPECIES	66.6%***

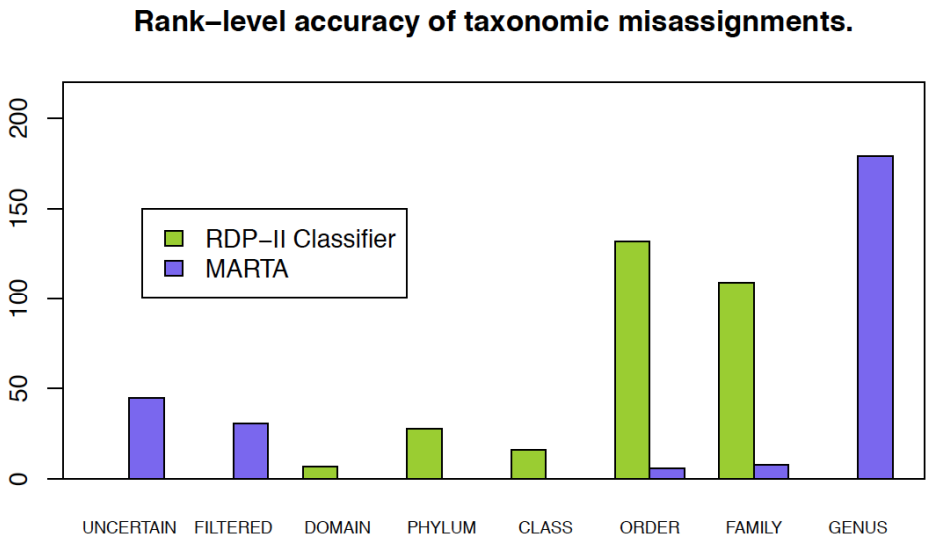
the arguments used to assign taxonomic status to the RDP-II corpus of 5,148 'Type' sequences.
* An argument that Marta forwards to megablast.
** sample argument is the value that Marta uses in its default configuration. It isn't necessary to call MARTA with these arguments if the user agrees with the setting.
*** requisite agreement among bitscore ties to cause Marta to use the given taxonomic level.

THE CONGRUENCE OF RDP-II CLASSIFIER AND MARTA

The assignments from MARTA and the RDP-II Classifier were both highly congruent with the corpus' assignments at the species and genus levels, respectively (by design, the RDP-II Classifier restricts its most specific assignment to the genus level, while MARTA attempts to resolve sequences to the species level). The assignments provided by the RDP-II Classifier were congruent, at the genus level, for 4,852 of 5,148 sequences (94.25%).

MARTA attempts to phylogenetically classify sequences to the species-level, and 4,939 of 5,148 of its assignments were congruent at the species-level (95.9%).

When either the RDP-II Classifier or MARTA provided an assignment that did not match at either the genus-level (RDP-II Classifier) or species-level (MARTA), the assignments still tended to find congruence within the taxonomic classification provided by the corpus:



Of the 209 mismatches found in MARTA's assignments, 125 of these were nevertheless congruent with the corpus at the genus level. For 89 of these mismatches, MARTA purposefully restricted its assignments to the genus-level because of the settings used to run the program (above). MARTA's high congruence at the species-level (95.9 %) and the results shown above illustrate that MARTA is highly congruent at the genus-level (98.3%).

Using the default settings, MARTA finds higher congruence at the species and genus-level than the RDP-II Classifier does at the genus-level, though that is likely to be due to the default parameters that the Classifier uses to analyze sequence data. What is interesting to note is that MARTA performs well during phylogenetic classification. This gives us confidence in using the software to classify sequences from taxa that are not considered by the RDP-II Classifier.

It is possible for MARTA to assign a sequence to one of three other possible categories. **(1.)** "No significant results" occurs when the BLAST utility returns no hits for consideration. A sequence is labeled **(2.)** "Filtered" when the user-defined requirement "Query coverage" (defaulting to 80%) is not satisfied by the candidate hits.

A sequence's taxonomic assignment is labeled **(3.)** Uncertain when (a.) none of the considered accessions have a classification in the Taxonomy database (e.g. all "Uncertain" or "Uncultured") or (b.) when a discrepancy is detected in the Taxonomy database. For example, the majority of the "Uncertain" sequences shown in this figure have an annotation error in the Taxonomy database (no genus-level node). We include code on the website to review and repair the Taxonomy database and will continue to do so as we recognize discrepancies in the database.