metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

map sequence reads to taxa

tree of life

gut

Tuesday, May 15, 12
metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

compare sequence reads to genomes in databases

NCBI RefSeq
NCBI Taxonomy

acid mine
mammoth
metagenomic sample

...ACGCTAGCATGTACGCGCGCCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq

NCBI Taxonomy

compare sequence reads to genomes in databases
metagenomic sample

compare sequence reads to genomes in databases

various algorithmic methods
metagenomic sample

...ACGCTAGCATGTACGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAA...
...ATACCAACAGCTACTACTCATAACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq
NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity
metagenomic sample

...ACGCTAGCATGTACGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAA...
...ATACCAACAGCTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

compare sequence reads to genomes in databases

NCBI RefSeq
NCBI Taxonomy

various algorithmic methods

metagenomic sample

ACGCTAGCATGTACGCGCGCTTCCTATG...
ACGTCATGACTACGACTCAGAT...
CCTCCTCTCTAAGAGAGAGAAAA...
ATACCAACAGCTACTCATACGACGACTAC...
CTAAGCTACGACTCATACTGACGCA...

Tuesday, May 15, 12
metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq
NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity
BLAST
metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq
NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity

BLAST

composition
metagenomic sample

compare sequence reads to genomes in databases

various algorithmic methods

similarity

BLAST

composition

NCBI RefSeq

NCBI Taxonomy
metagenomic sample

...ACGCTAGCATGTACGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATAACAAACAGCTACTCTACTACGTAGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq

NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity
BLAST

composition
HMMs
metagenomic sample

...ACGCTAGCATGTACCGCGCCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq

NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity
compositional
phylogeny

BLAST
HMMs
metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACTACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

compare sequence reads to genomes in databases

NCBI RefSeq
NCBI Taxonomy

various algorithmic methods

similarity | composition | phylogeny

BLAST | HMMs
metagenomic sample

...ACGCTAGCATGTACGCGCGCGCTTCCTATG...
...ACGTCATGACTACGACTCAGAT...
...CCTCCTCTCTAAGAGAGAGAAAAA...
...ATACCAACAGCTACTACTCATACGACGACTAC...
...CTAAGCTACGACTCATACTGACGCA...

NCBI RefSeq
NCBI Taxonomy

compare sequence reads to genomes in databases

various algorithmic methods

similarity
BLAST
composition
HMMs
phylogeny
evolutionary models
### Program Attributes and Characteristics

#### Similarity-based Methods

<table>
<thead>
<tr>
<th>Program</th>
<th>Similarity Method</th>
<th>LCA</th>
<th>Specific Genes Req'd</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARMA</td>
<td>BLAST, HMM</td>
<td>LCA</td>
<td>multiple alignment</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>FACS</td>
<td>other</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line</td>
</tr>
<tr>
<td>jMOTU/Taxonerator</td>
<td>BLAST, other</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line, GUI</td>
</tr>
<tr>
<td>MARTA</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>MEGAN</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>MetaPhyler</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line</td>
</tr>
<tr>
<td>MG-RAST</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line</td>
</tr>
<tr>
<td>MTR</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>SOrt-ITEMS</td>
<td>BLAST</td>
<td>LCA-like</td>
<td>marker genes</td>
<td>command line</td>
</tr>
</tbody>
</table>

#### Composition-based Methods

<table>
<thead>
<tr>
<th>Program</th>
<th>Composition Method</th>
<th>Machine Learning</th>
<th>Confidence Method</th>
<th>Specific Genes Req'd</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes Classifier</td>
<td>NBC</td>
<td>supervised</td>
<td>other</td>
<td>command line, web-based</td>
<td></td>
</tr>
<tr>
<td>PhyloPhylicaS</td>
<td>other</td>
<td>supervised</td>
<td>other</td>
<td>command line, web-based</td>
<td></td>
</tr>
<tr>
<td>Phyml</td>
<td>IMM</td>
<td>supervised</td>
<td>other</td>
<td>web-based</td>
<td></td>
</tr>
<tr>
<td>RAiPhy</td>
<td>other</td>
<td>semi-supervised</td>
<td>other</td>
<td>command line, GUI</td>
<td></td>
</tr>
<tr>
<td>RDP</td>
<td>k-means/kNN, NBC</td>
<td>supervised</td>
<td>bootstrap</td>
<td>16S rRNA</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>Scimm</td>
<td>IMM</td>
<td>semi-supervised</td>
<td>bootstrap</td>
<td>16S rRNA</td>
<td>command line</td>
</tr>
<tr>
<td>TACOA</td>
<td>k-means/kNN</td>
<td>supervised</td>
<td></td>
<td></td>
<td>command line</td>
</tr>
</tbody>
</table>

#### Phylogeny-based Methods

<table>
<thead>
<tr>
<th>Program</th>
<th>Phylogeny Method</th>
<th>Confidence Method</th>
<th>Specific Genes Req'd</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPA</td>
<td>ML</td>
<td>bootstrap, other</td>
<td>multiple alignment</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>FastTree</td>
<td>other</td>
<td>bootstrap</td>
<td>16S rRNA</td>
<td>web-based</td>
</tr>
<tr>
<td>green genes (NAST, Simrank)</td>
<td>other</td>
<td></td>
<td>16S rRNA</td>
<td>web-based</td>
</tr>
<tr>
<td>pplacer</td>
<td>ML, Bayesian</td>
<td>posterior probability, other</td>
<td>multiple alignment</td>
<td>command line</td>
</tr>
</tbody>
</table>

#### Combined Similarity and Composition-based Methods

<table>
<thead>
<tr>
<th>Program</th>
<th>Similarity Method</th>
<th>Composition Method</th>
<th>Machine Learning</th>
<th>Specific Genes Req'd</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPHINX</td>
<td>BLAST</td>
<td>k-means/kNN</td>
<td>supervised</td>
<td></td>
<td>web-based</td>
</tr>
</tbody>
</table>

#### Combined Similarity and Phylogeny-based Methods

<table>
<thead>
<tr>
<th>Program</th>
<th>Similarity Method</th>
<th>Phylogeny Method</th>
<th>Confidence Method</th>
<th>Specific Genes Req'd</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMPHORA</td>
<td>HMM</td>
<td>other</td>
<td>bootstrap</td>
<td>marker genes</td>
<td>command line</td>
</tr>
<tr>
<td>MLTreeMap</td>
<td>BLAST, HMM</td>
<td>ML</td>
<td>bootstrap, other</td>
<td>marker genes</td>
<td>command line, web-based</td>
</tr>
<tr>
<td>SAP</td>
<td>BLAST</td>
<td>Bayesian, other</td>
<td>posterior probability, other</td>
<td>marker genes</td>
<td>command line</td>
</tr>
<tr>
<td>TreePhyler</td>
<td>HMM</td>
<td>other</td>
<td>bootstrap</td>
<td>marker genes</td>
<td>command line</td>
</tr>
</tbody>
</table>
two basic metrics

“sensitivity” = \frac{\text{number of sequences correctly classified}}{\text{total number of sequences}}

“precision” = \frac{\text{number of sequences correctly classified}}{\text{number of sequences classified}}
the idea

create an ensemble classifier that outperforms any single classifier by means of complementation and voting
implementation

- use random forest to determine the optimal combination of classifiers

- the classifiers are the predictor variables; the correct call is the response variable

- use random forest’s measure of variable importance to weight the call each program makes

- combine these weighted calls into a single call
PhyloPythia 961 bp “simulated medium complexity” data set, 114,457 reads

random forest

variable importance
<table>
<thead>
<tr>
<th></th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>carma</td>
<td>0.1523643</td>
<td>1244.2580</td>
</tr>
<tr>
<td>megan</td>
<td>0.1535782</td>
<td>2420.7659</td>
</tr>
<tr>
<td>metaphyler</td>
<td>0.1421925</td>
<td>83.0732</td>
</tr>
<tr>
<td>mgrast</td>
<td>0.1520907</td>
<td>1034.0064</td>
</tr>
<tr>
<td>nbc</td>
<td>0.1535500</td>
<td>1207.7562</td>
</tr>
<tr>
<td>phylopythias</td>
<td>0.1486372</td>
<td>275.1587</td>
</tr>
<tr>
<td>phymmbbl</td>
<td>0.1535986</td>
<td>2098.2432</td>
</tr>
<tr>
<td>raiphy</td>
<td>0.1521719</td>
<td>907.4872</td>
</tr>
<tr>
<td>mltreemap</td>
<td>0.1456021</td>
<td>103.3918</td>
</tr>
<tr>
<td>treephyler</td>
<td>0.1514593</td>
<td>530.2817</td>
</tr>
</tbody>
</table>

PhyloPythia 961 bp “simulated medium complexity” data set, 114,457 reads
example

-bash-3.2$ ./assess_rf_performance.pl (using MDA)
correct: 41599 classified: 44434 total: 114457
percentage classified: 38.8%
sensitivity: 36.3%
precision: 93.6%

-bash-3.2$ ./assess_rf_performance.pl (using MDG)
correct: 83161 classified: 93225 total: 114457
percentage classified: 81.4%
sensitivity: 72.7%
precision: 89.2%

PhyloPythia 961 bp “simulated medium complexity” data set, 114,457 reads
results

- so far, have worked with two different data sets, evaluating assignments at the genus level with various combinations of classifiers and schemes for integrating confidence of assignment

- the ensemble classifier using random forest has not been able to outperform the single best classifier standalone, but this is probably due to some technical reasons (hard limit on number of predictor classes = 32, few observations per response class at the genus level)

- this approach may still prove fruitful in the more general case (higher taxonomic ranks = fewer predictor classes, and more observations per response class)