Neural Network-based Taxonomic Clustering for Metagenomics

Steven Essinger, Robi Polikar, Gail Rosen
Electrical & Computer Engineering, Drexel University, 3141 Chestnut Street
Philadelphia, PA 19104, US

Summary

Goal: Predict the taxonomic classification of organisms based on the fragments obtained from an environmental sample that may include many previously unidentified organisms.

Conclusion:
- Compared to other unsupervised and semi-supervised approaches, we cluster shorter reads (500bp) and more strains (200 to 400) than any other method, to show the clustering method’s feasibility on real metagenomics datasets.
- We demonstrate that adaptive resonance theory is able to cluster novel phyla better than K-means when there are a large number of fragments to cluster. This is due to the incremental learning capability of ART and its ability to learn non-spherical clusters.
- On an extremely challenging dataset of grouping 500bp reads from 204 strains spanning 17 phyla, ART is able to accomplish this with 43% accuracy (5.9% by chance)

Challenge

- The challenge we face is that we cannot simply cluster fragments together that are similar in composition as many clustering methods tend to do.
- While two strains may be similar inter-genomically, each generally will vary greatly intra-genomically. Since the fragments we are clustering represent short samples of each strain’s genome, we expect that the fragments in each cluster will vary greatly.
- Current methods do not address next-generation sequencing technology
  - LikelyBin: successful only for low complexity samples (2-10 species)
  - GSSOM: successful when read lengths are greater than 8kb
  - CompositBin: successfully tested only for low-complexity samples

Test Data

- 635 microbe genomes obtained from National Center for Information Biotechnology
- Dataset spans 19 different phyla: We selected this level since it is comprised of microbes that are much more diverse than those belonging to the levels of genus or species
- Whole-genomes used in training database
- Test fragments obtained from test strains by random sample 500 bp in length, 300x

Table 1

<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Phyla</td>
<td>17</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>Test Phyla</td>
<td>431</td>
<td>204</td>
<td>320</td>
</tr>
<tr>
<td>Test Strains</td>
<td>204</td>
<td>431</td>
<td>315</td>
</tr>
</tbody>
</table>

Results

Table I. The results of clustering 20400 fragments spanning 17 different phyla to cluster 17 smaller phyla.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Performance</th>
<th>Phyla</th>
<th>K-M</th>
<th>ART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Unity</td>
<td>Avg</td>
<td>Std</td>
<td>Avg</td>
<td>Std</td>
</tr>
<tr>
<td>Class Isolation</td>
<td>0.25</td>
<td>0.05</td>
<td>0.43</td>
<td>0.05</td>
</tr>
<tr>
<td># of Clusters</td>
<td>17</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II. Results of clustering 43100 fragments spanning 2 different phyla to cluster 2 large phyla.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Performance</th>
<th>Phyla</th>
<th>K-M</th>
<th>ART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Unity</td>
<td>Avg</td>
<td>Std</td>
<td>Avg</td>
<td>Std</td>
</tr>
<tr>
<td>Class Isolation</td>
<td>0.73</td>
<td>0.03</td>
<td>0.73</td>
<td>0.03</td>
</tr>
<tr>
<td># of Clusters</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Algorithm

Proposed Algorithm

A. Metagenomic reads (fragments) from next generation sequencing technology
B. Training database (TDB) - consists of G labeled genomes, previously acquired
C. Unsupervised clustering algorithms (e.g. ART, K-means)
D. Set free parameters (e.g. D is K-means and c in ART)

Algorithm:

A. Train Naïve Bayes Classifier (NBC) model H of G-genome probability profiles
   - Train NBC model using Chimera chieks and high coverage shotgun data
   - Classify each fragment using corresponding feature vector of dimension G

B. Score Fragments, evaluate fragment, Assess NBC
   - Do $j = 1, ..., C$ (of fragments)
   - Identify (N-j) overlapping motifs
   - Calculate probability of fragment belonging to genome, in TDB
     - $p_j = \frac{\sum_{c=1}^{C} p(c) p_j(c)}{\sum_{c=1}^{C} p(c)}$
   - End

C. Build feature vectors for unsupervised classifier

Test Figures of Merit

- Accuracy to group similar classes together
  - $\frac{\sum_{c=1}^{C} (TP_c + TN_c)}{\sum_{c=1}^{C} (TP_c + TN_c)}$
- Accuracy of algorithm to isolate dissimilar classes
  - $\frac{\sum_{c=1}^{C} (FP_c + FN_c)}{\sum_{c=1}^{C} (FP_c + FN_c)}$

This work was supported in part by National Science Foundation award #0845827