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.

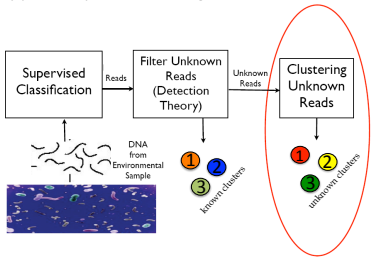

Fig. 1

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 feasibilities 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)
  - *GSOM:* successful when read lengths are greater than 8kbp
  - *CompostBin:* 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, 100x

Experiment	1	2	3
Training Phyla	2	17	19
Test Phyla	17	2	19
Training Strains	431	204	320
Test Strains	204	431	315

Table 1

Results

Experiment 1 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.25	0.05	0.43	0.05
Class Isolation	0.31	0.04	0.34	0.03
# of Clusters	17		17	

Experiment 1: Training on 2 large phyla to cluster 17 smaller phyla

Experiment 2 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.73	0.03	0.73	0.03
Class Isolation	0.74	0.04	0.86	0.04
# of Clusters	2		4	

Experiment 2: Training on 17 smaller phyla to cluster 2 large phyla

Experiment 3 Performance	Phyla			
	K-M		ART	
	Avg	Std	Avg	Std
Class Unity	0.52	0.04	0.51	0.05
Class Isolation	0.22	0.06	0.53	0.05
# of Clusters	19		18	

Experiment 3: Training on examples of each phyla to cluster the rest

Algorithm

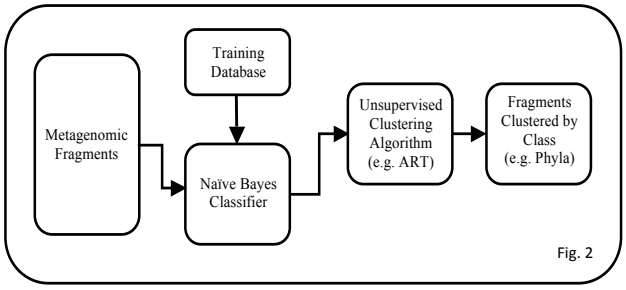

Fig. 2

Proposed Algorithm

Input:

- Metagenomic reads (fragments) from next-gen sequencing technology
- Training database (TDB) – consists of *G* labeled genomes, previously acquired
- Unsupervised clustering algorithm (e.g. ART, K-means)
- Set free parameters (e.g. *K* in K-means and *v* in ART)

Algorithm:

A. Train Naïve Bayes Classifier (NBC) motifs, *M* of *G* genome probability profiles
   Do: $i = 1, \ldots, G$
   Do: $j = 1, \ldots 4^N$ (# of diff. motif perm.)

$$P(M_j|genome_i) = \frac{Freq.\ of\ M_j\ in\ genome_i}{Total\ M\ in\ genome_i} \quad (1)$$

   End
   End

B. Score fragments, evaluate fragment, f using NBC
   Do: $f = 1, \ldots, F$ (# of fragments)
   1. Identify *J (N-1)* overlapping motifs each of length *N* in fragment, *f*: $[M_1, M_2, M_3, \ldots, M_J]^T$
   2. Calculate probability of fragment belonging to *genome*, in TDB:

$$Score.S_{fi} = P(f|genome_i) = \prod_{j=1}^{J} P(M_j|genome_i) \quad (2)$$

   End

C. Build feature matrix for unsupervised classifier

	NBC Scores	Features			
		$genome_1$	$genome_2$	...	$genome_G$
Objects	Frag1	S1,1	S1,2	...	S1,G
	Frag2	S2,1	S2,2	...	...
	...	...	...	...	...
	FragF	SF,1	...	...	SF,G

D. Call unsupervised clustering algorithm
   - Cluster each fragment using corresponding feature vector of dimension *G*

Output:

- Fragments clustered by taxonomic class (e.g. Phyla, Genus, Strain, etc.)

Test: Figures of Merit

- Accuracy to group similar classes together

$$A_{unity} = \frac{1}{P}\sum_{p=1}^{P}\left[\frac{argmax(f_{c,p})}{f_{c,p}}\right] \quad (3)$$

- Accuracy of algorithm to isolate dissimilar classes

$$A_{isolate} = \sum_{c=1}^{C}\frac{f_c}{F}\left[\frac{argmax(f_{c,p})}{f_{c,p}}\right] - \frac{1}{P}\sum_{p=1}^{P}\left[\frac{argmax(f_{c,p})}{f_{c,p}}\right] \quad (4)$$

C: # of clusters
P: # of taxonomic classes (e.g. phyla)
f_c: # of frag. in cluster, c
$f_{c,p}$: # of frag. in cluster, c belonging to taxonomic class, p
f'_c: # of fragments from taxonomic class, p
F: total number of fragments in all phyla

Fig. 3

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