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: $i = 1, ..., G$
    Do: $j = 1, ...4^N$ (# of diff. motif perm.)
        $P(M_j|genome_i) = \frac{Freq.\ of\ M_j\ in\ genome_i}{Total\ M\ in\ genome_i}$   (1)
    End
End
```

B. Score fragments, evaluate fragment, f using NBC

```
Do: $f = 1, ..., F$ (# of fragments)
    1. Identify $J$ $(N-1)$ overlapping motifs each of length $N$ in fragment, $f$:
       $[M_1, M_2, M_3, ..., M_J]^T$
    2. Calculate probability of fragment belonging to $genome_i$ in TDB:
       $Score, S_{f,i} = P(f|genome_i) = \prod_{j=1}^{J} P(M_j|genome_i)$   (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

$$A_{unity} = \frac{1}{F}\sum_{p=1}^{P}\left[\underset{f_{c_p}}{argmax}(f_{c_p}|p)\right] \quad (3)$$

- Accuracy of algorithm to isolate dissimilar classes

$$A_{isolate} = \sum_{c=1}^{C}\frac{f_c}{F}\left[\frac{\underset{f_t'}{argmax}(f_t'|c)}{f_c}\right] = \frac{1}{F}\sum_{c=1}^{C}\left[\underset{f_t'}{argmax}(f_t'|c)\right] \quad (4)$$

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