Scalable metagenomics of food borne pathogens and antimicrobial resistance using supercomputing with Bayesian and error correction methods

Steven Lakin and Zaid Abdo
Colorado State University
GenomeTrakr-2018

The Abdo Lab

- Our focus is on Microbiome and Microbial Genomics and Metagenomics Research

- Including Detecting Antimicrobial Resistance and Study of its Spread

The Abdo Lab

- This is the work of Steven Lakin, PhD student in my lab

The Problem

- Classification of metagenomics data is not a simple problem
- It is limited by available databases
- Also by the size of the dataset at hand
- And by the available computational power

Goals

- To speed up metagenomic classification
- To alleviate limitations of the available databases

Background

- Alignment Methods
  - Slow
  - Allow for missmatches and errors

Background

- K-mer (compositional) based methods
  - Fast (like kraken and centrifuge)
  - Looks for exact k-mer matches (no room for error)

Approach

- K-mer with error correction, the best of two worlds

Approach

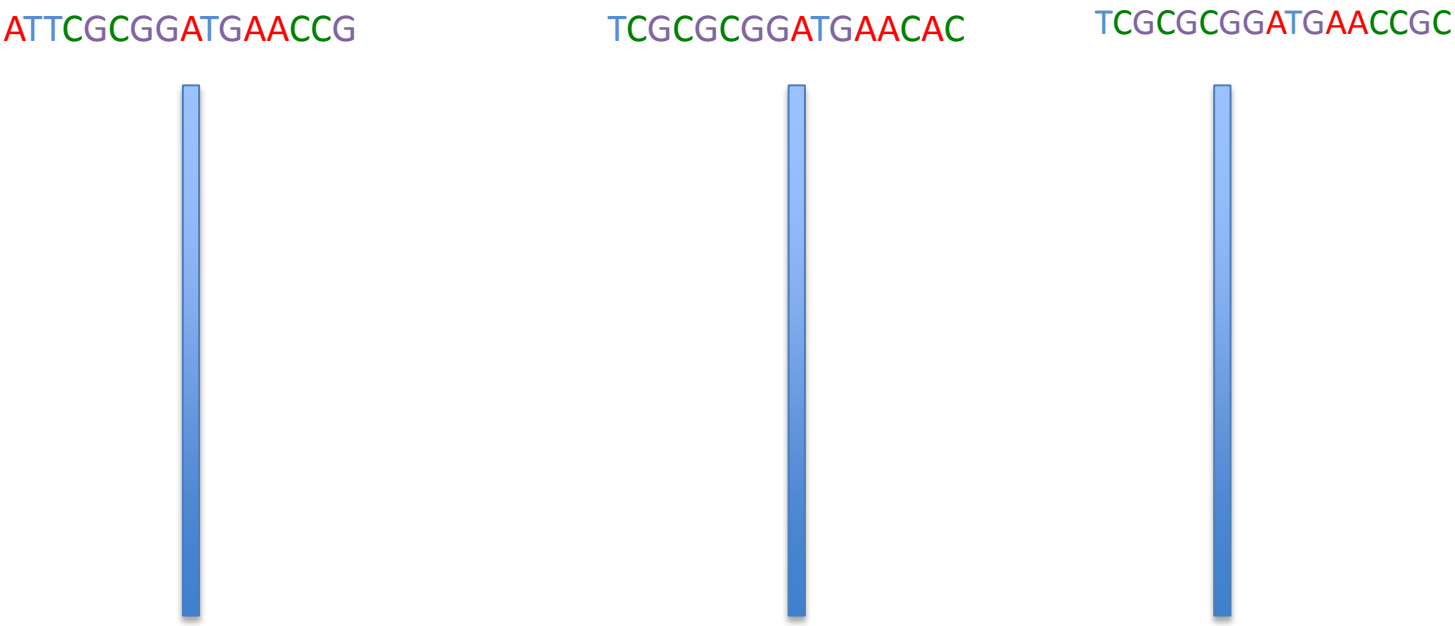
- K-mers

ATTCGCGGATGAACCG

A 16-mer

Approach

- K-mers classification (finding the spikes in the space)



Approach

- t-mers

ATTCGCGGATGAACCG

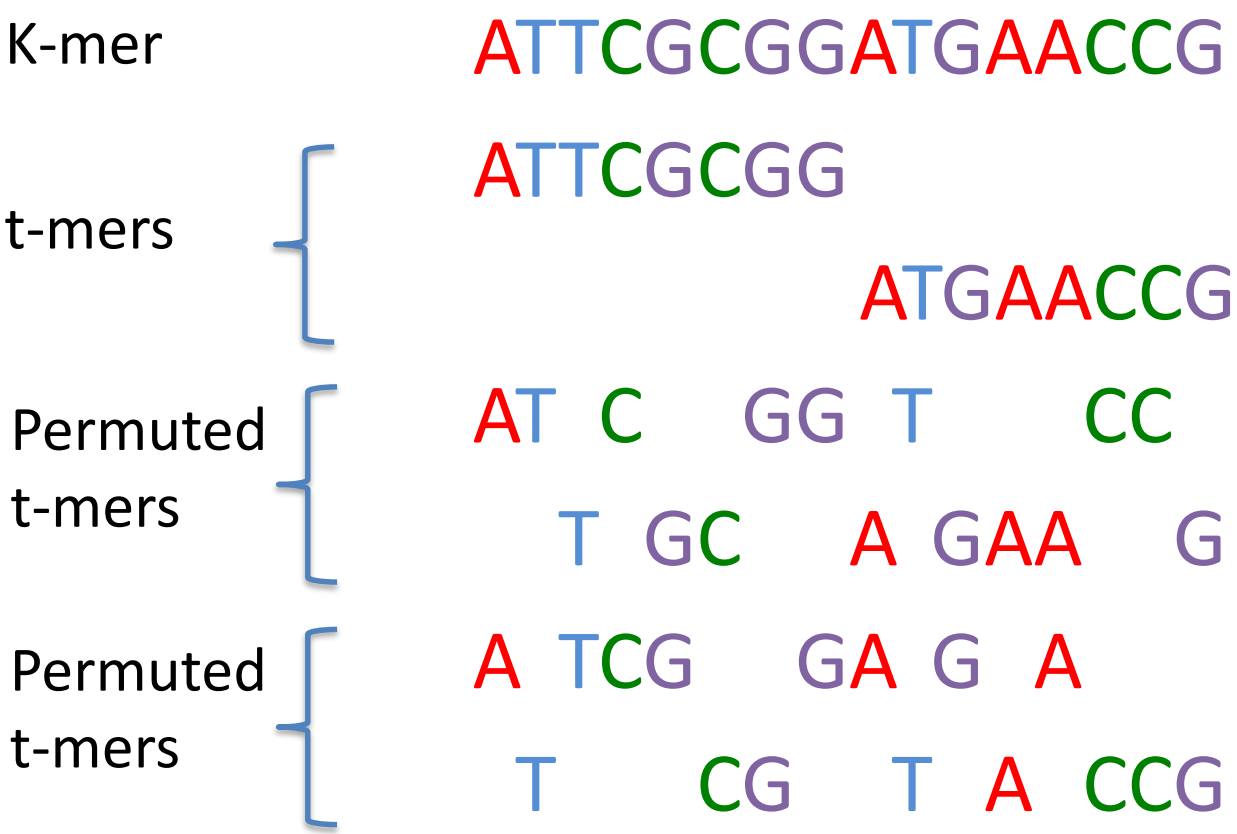
ATTCGCGG

ATGAACCG

A (16,8)-mer: 16 is the length of the k-mer and 8 is the length of the t-mer

Approach

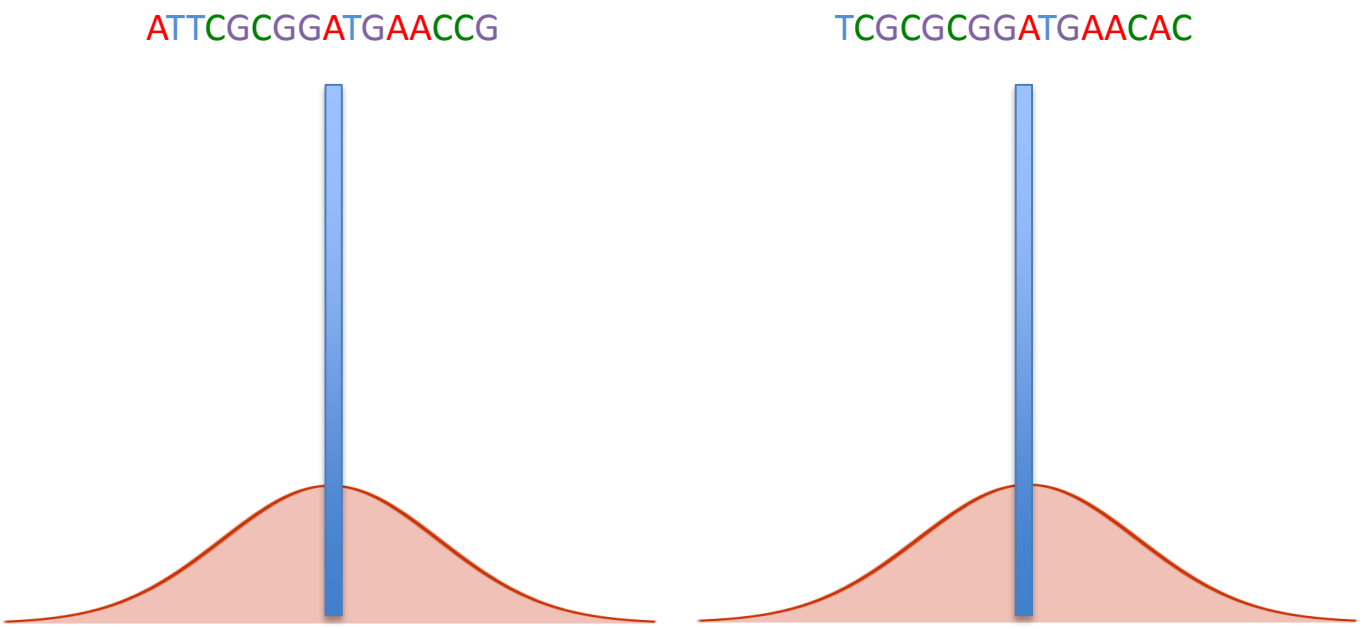
- The Gallagar-low density parity check (LDPC)



Per-base pair coverage = 3

Approach

- LDPC classification (allowing for error, flattening the space)



Approach

- The unbalanced Gallagar-low density parity check (ULDPC)

ATTCGCGGATGAACCG

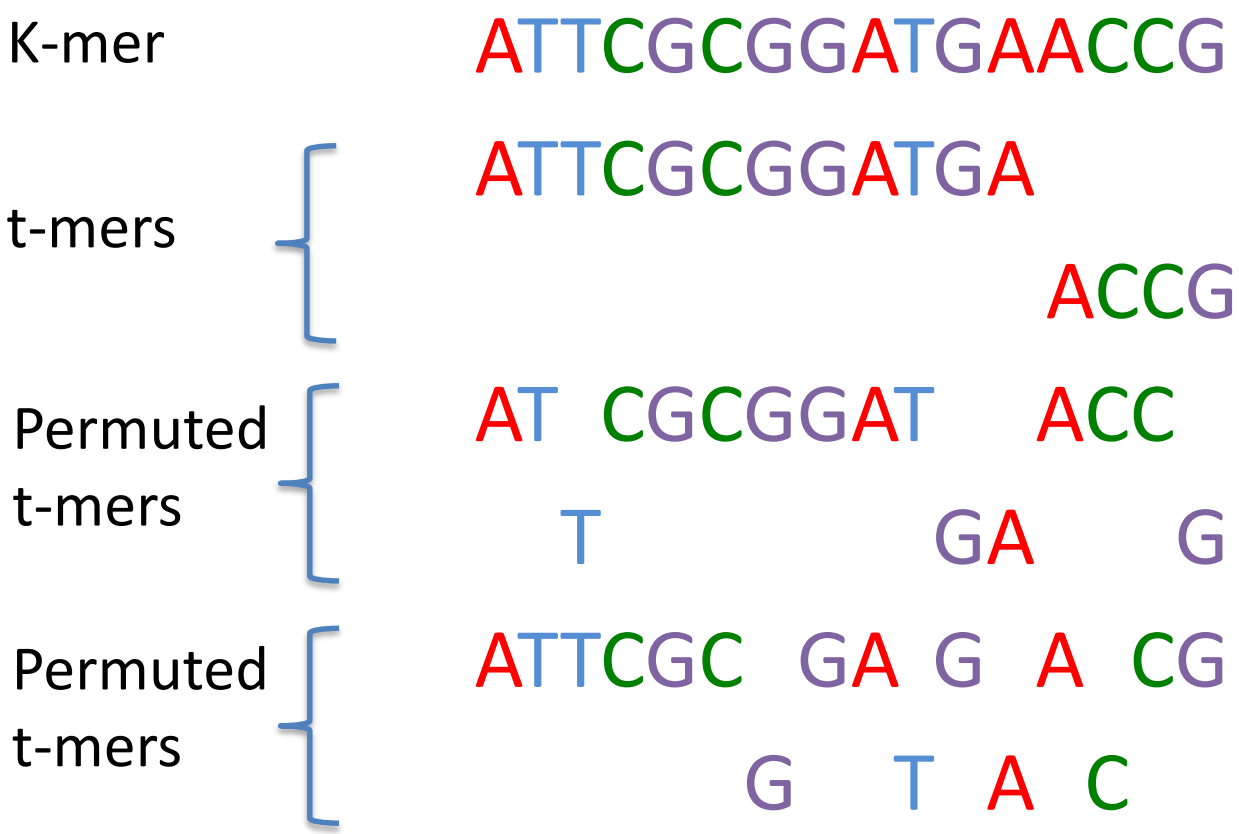
ATTCGCGGATGA

ACCG

A (16,12,4)-mer: 16 is the length of the k-mer and 8 is the length of the t-mer

Approach

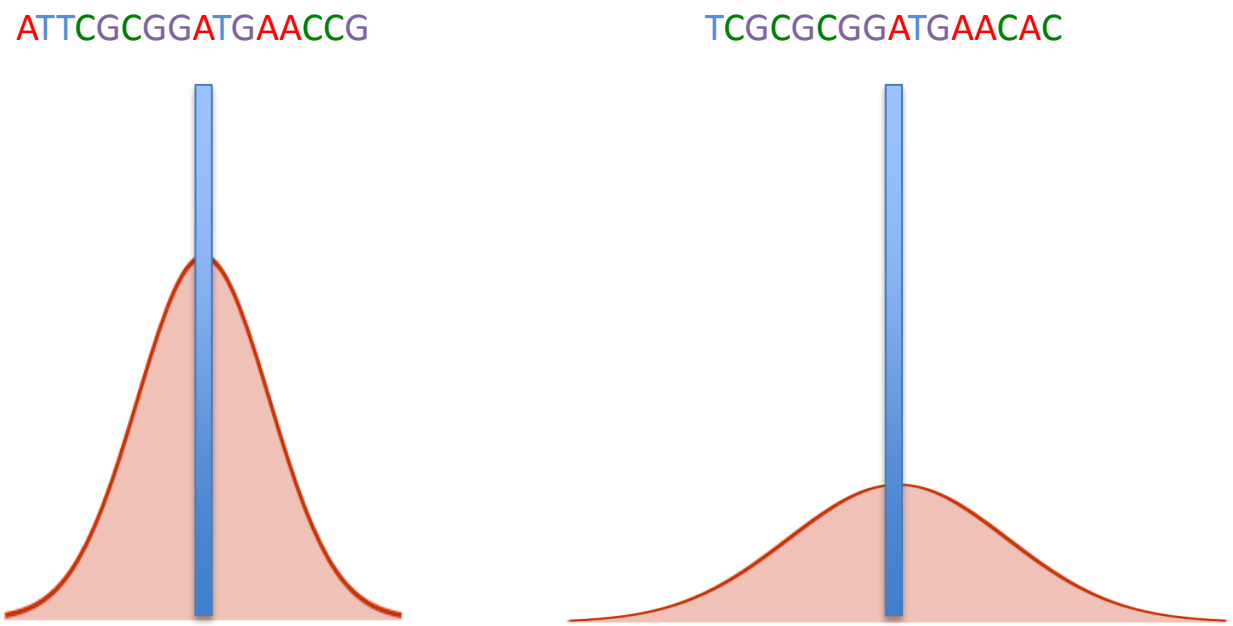
- The unbalanced Gallagar-low density parity check (ULDPC)



Per-base pair coverage = 3

Approach

- ULDPC classification (differential error allowance)



Approach

Trained t-mer table

t-mer\group	*group 1*	*group 2*	*group 3*
1	0.1	0.6	0.03
2	0.05	0.01	0.17
3	0.2	0.06	0
4	0	0.13	0
5	0.5	0	0
6	0.17	0	0
7	0.08	0.15	0.8

Sample

t-mer	*sample*
1	3
2	17
3	1
4	10
5	0
6	0
7	100

Approach

- Naïve Bayesian Classification
- Standard Dirichlet-Multinomial:

$$log\left(P(x|D_{g_j})\right) = \sum_{m=1}^{M}\left(\sum_{\forall i \in m} x_i\left(log(n_{g_j m} + \alpha_j) - log(N_{g_j} + M\alpha_j)\right)\right)$$

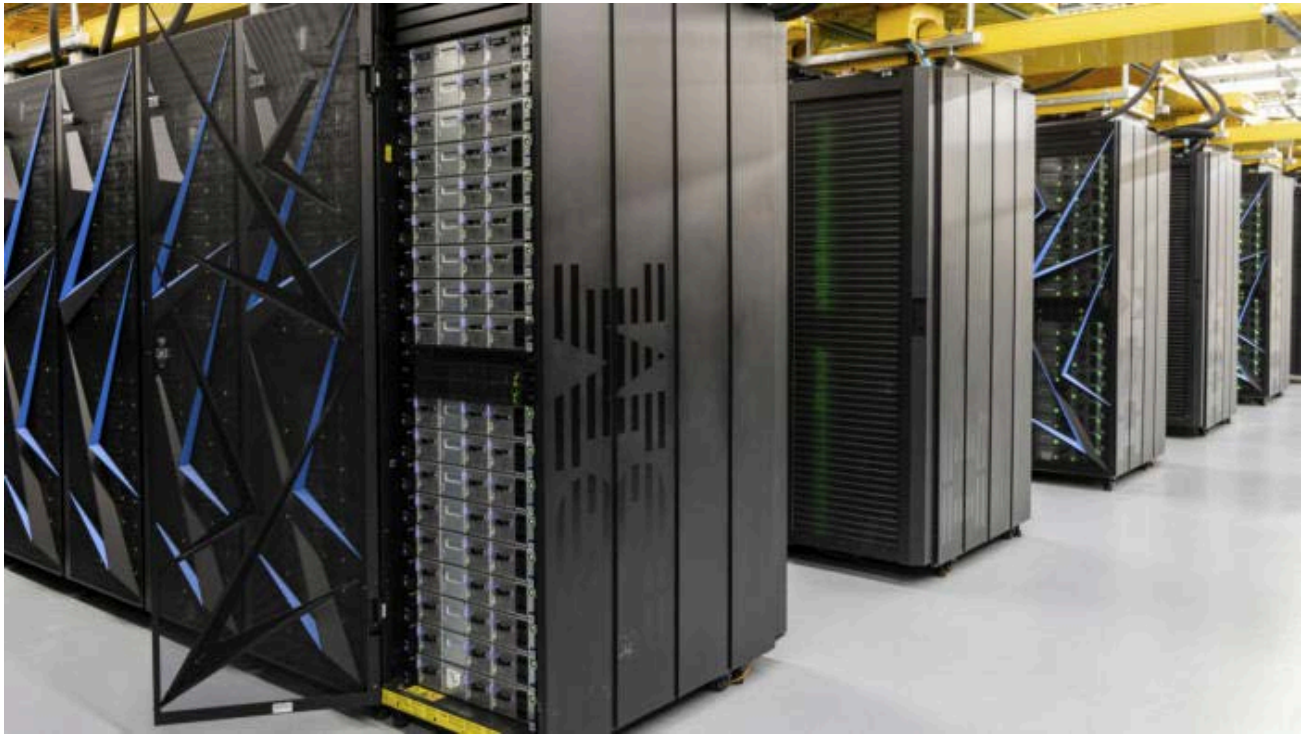
- All reduces to matrix multiplication and addition.

Implementation

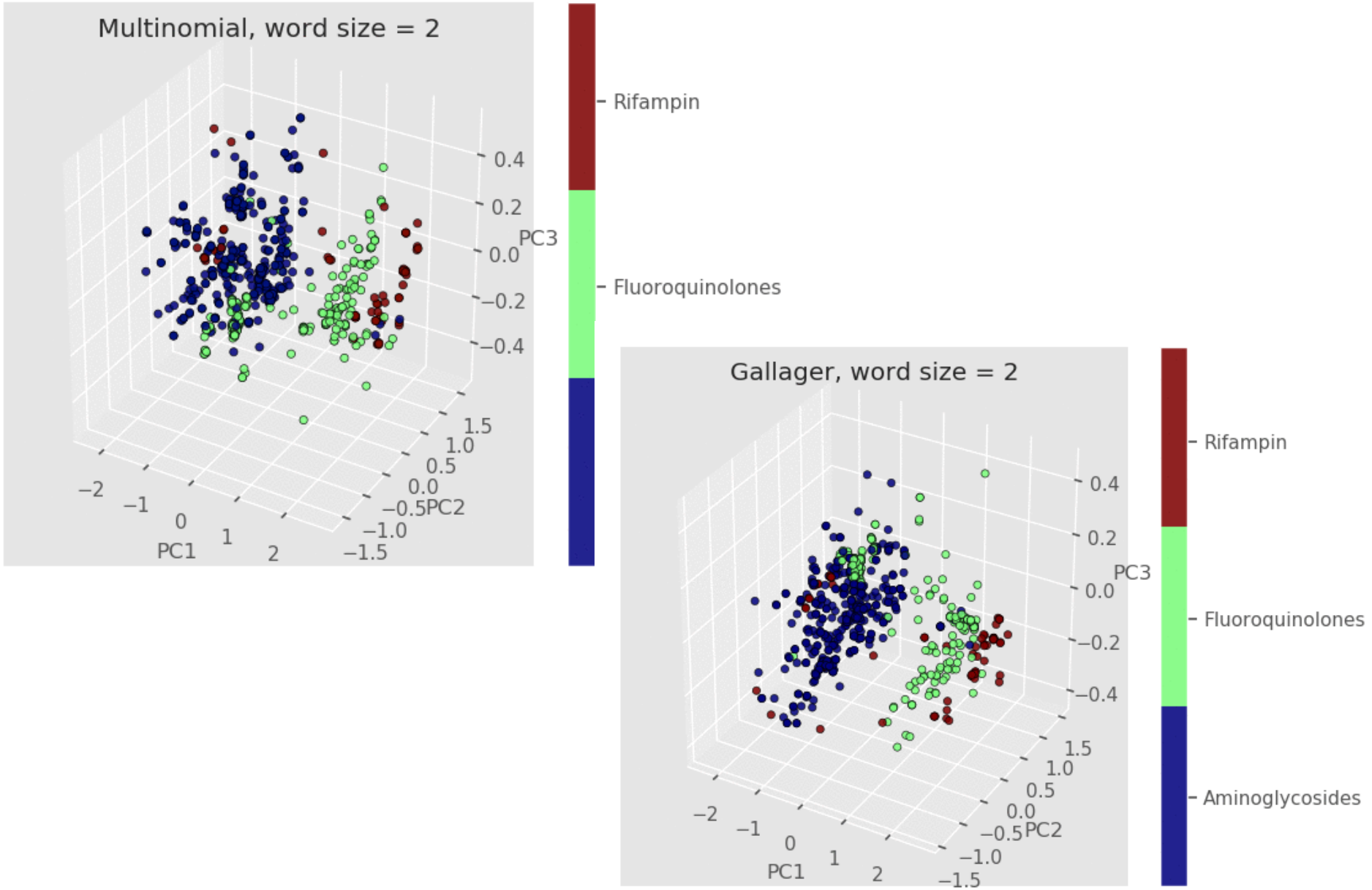
- GPU compute CUDA coding

Implementation

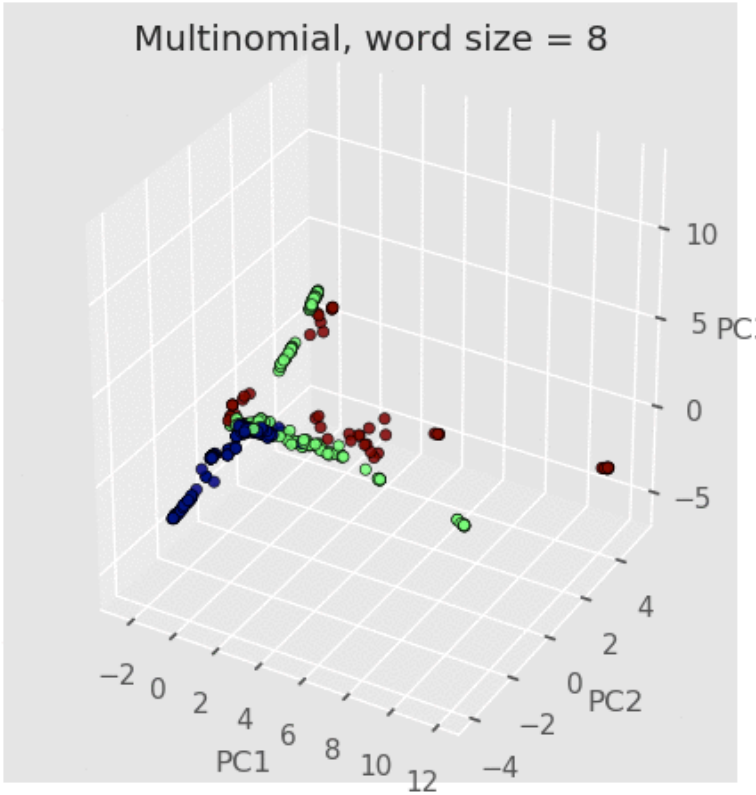
- Oakridge Summit Super Computer

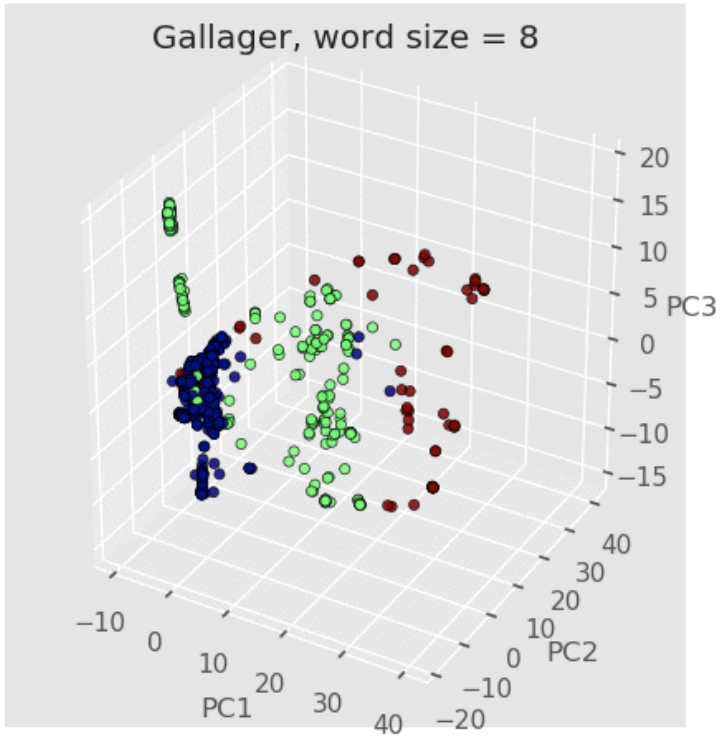


Some Results

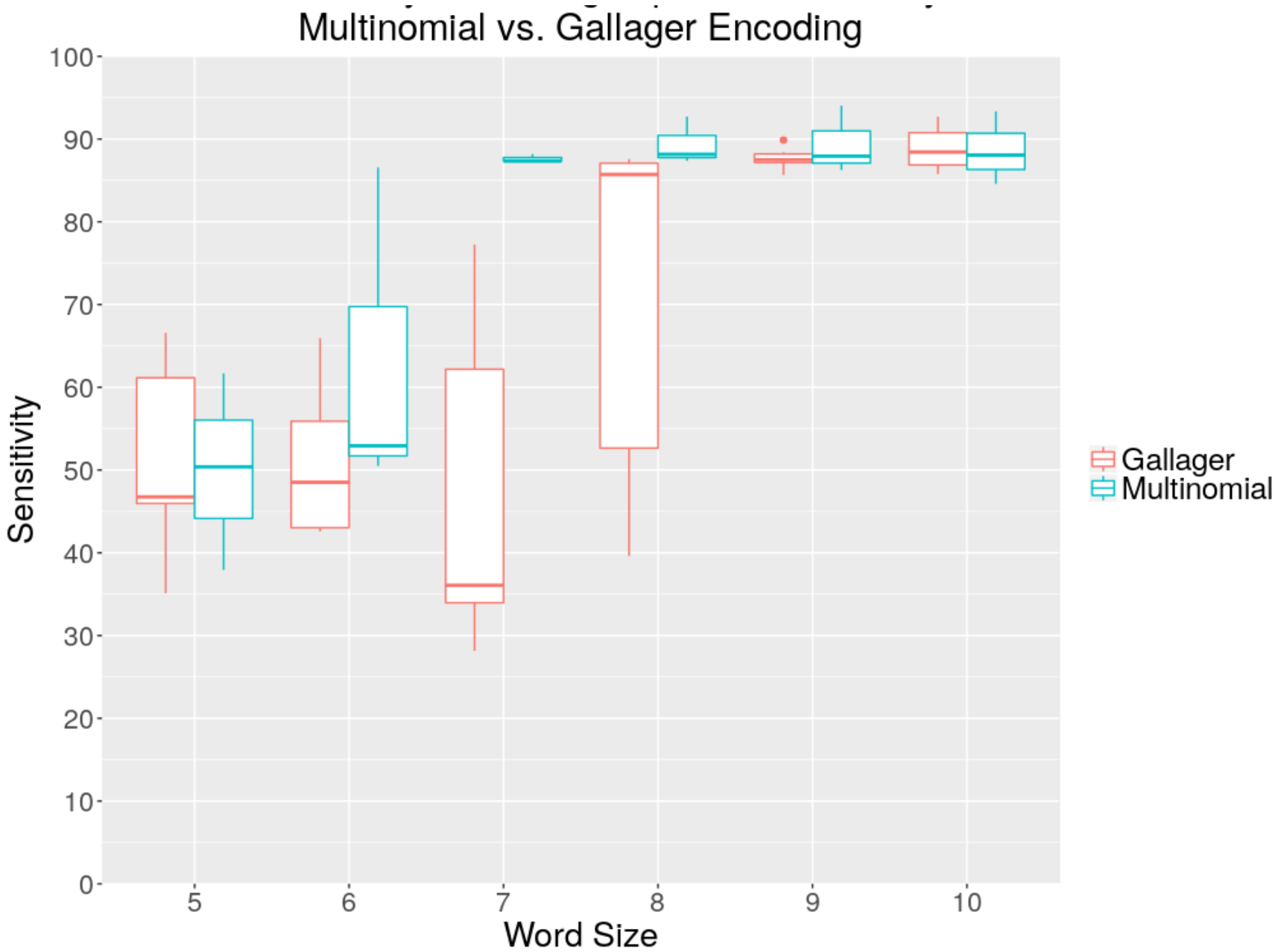


Some Results





Some Results



In Conclusion

- Our Naïve Bayes-ULDPC approach is promising
- The theoretical development and implementation are mostly complete
- Focus right now is on benchmarking and scaling

Aknowledgement

- Colorado State University Start up Funding to Zaid Abdo

- NSF-GAUSSI Fellowship to Steven Lakin