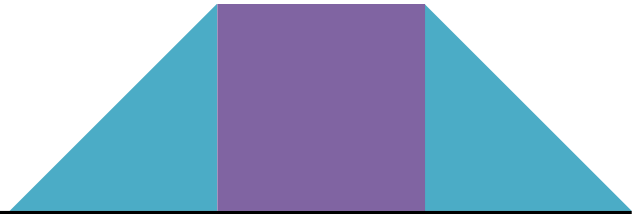
CAMIL: Clustering and Assembly with Multiple Instance Learning for Phenotype Prediction

Nathan LaPierre (me), Mohammad Arifur Rahman, and Huzefa Rangwala
George Mason University, Computer Science Department

Goal: Phenotype Prediction with Metagenomic Data

- Predict disease state (“phenotype”) of patient based on metagenome data

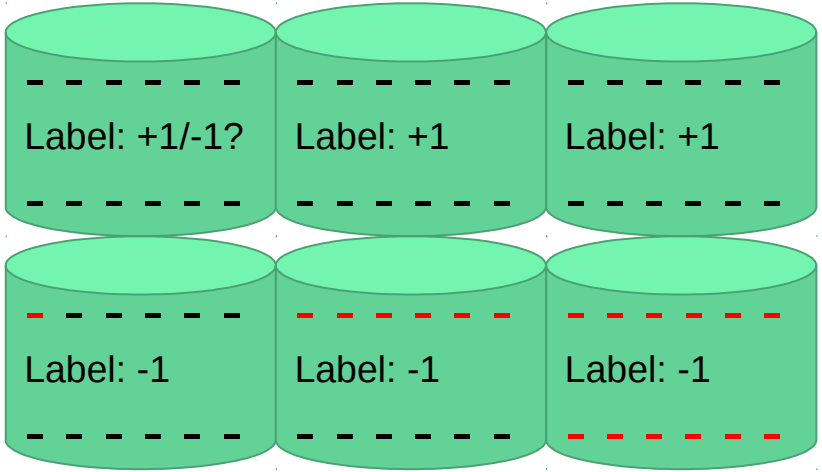
- Qin et al. dataset* (2012)
  - Type 2 Diabetes
  - 367 Chinese patients
  - Stool samples



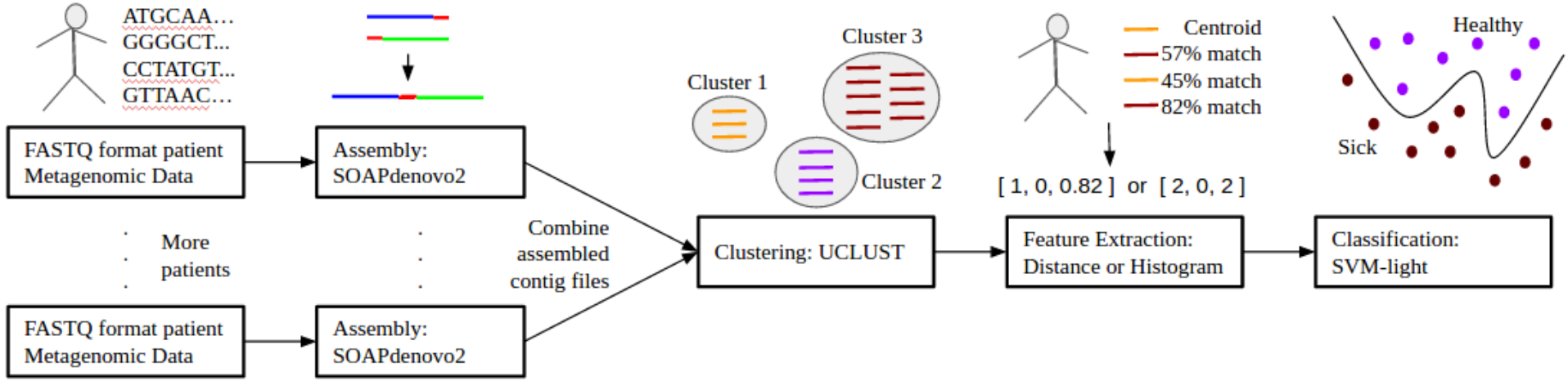
* J. Qin et al., “A metagenome-wide association study of gut microbiota in type 2 diabetes,” Nature, vol. 490, no. 7418, pp. 55–60, 2012.

Background: Multiple Instance Learning

- Labeled bags containing unlabeled instances
- In this context: labeled patients and unlabeled reads



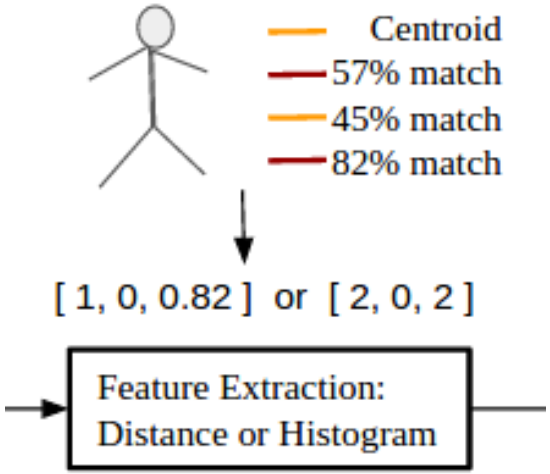
Pipeline



FASTQ Data → Assembly → Clustering → Feature Extraction → SVM Classification

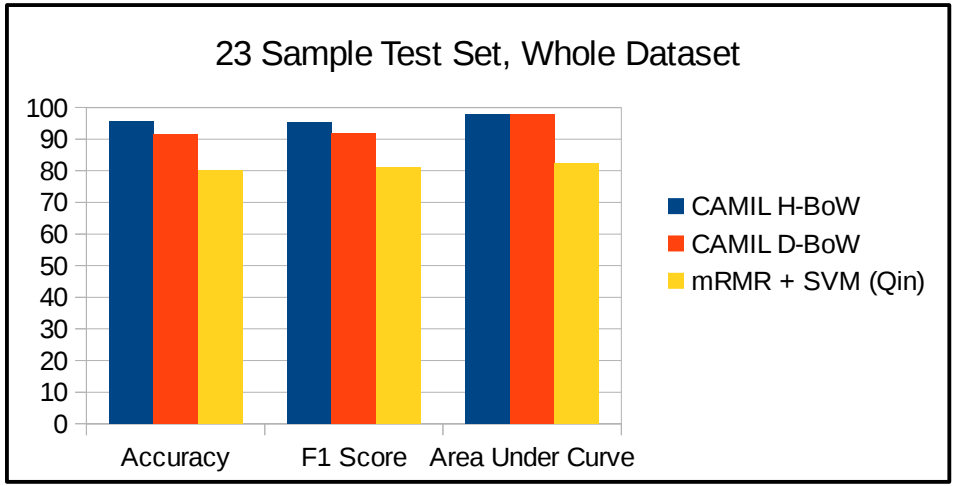
CAMIL Feature Extraction

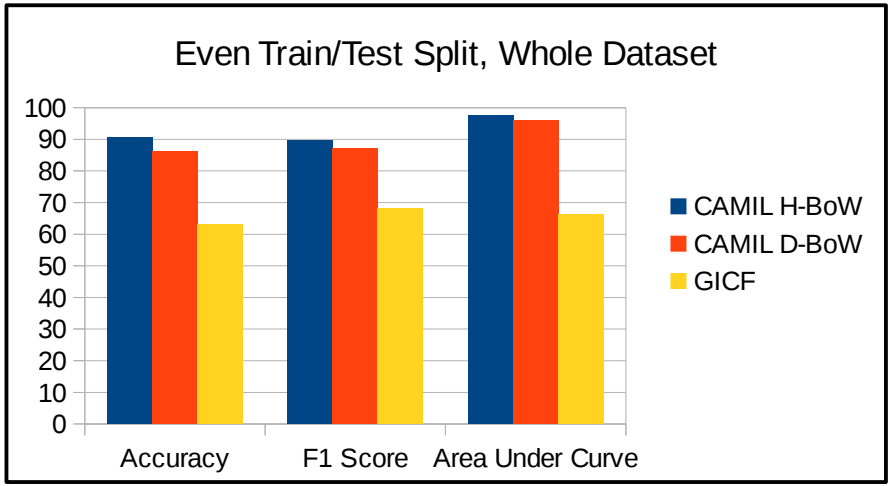
- MIL: “Bag of Words” (Amores)*
  - 1. Cluster instances
  - 2. Map to feature vectors
  - 3. Standard classifier
- Histogram Bag of Words (H-BoW)
- Distance Bag of Words (D-BoW)

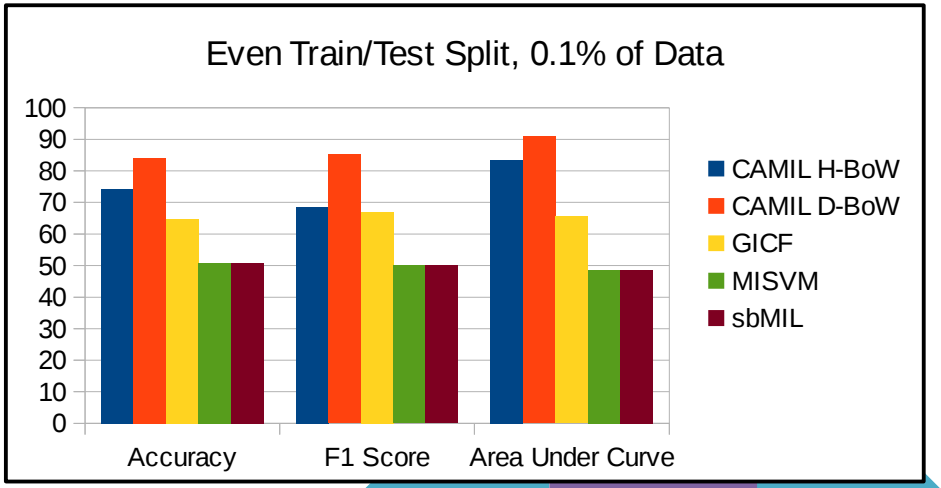


* J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artificial Intelligence, vol. 201, no. 1, pp. 81–105, 2013.

Results


23 Sample Test Set, Whole Dataset
100
90
80
70
60
50
40
30
20
10
0
Accuracy
F1 Score
Area Under Curve
CAMIL H-BoW
CAMIL D-BoW
mRMR + SVM (Qin)


Even Train/Test Split, Whole Dataset
100
90
80
70
60
50
40
30
20
10
0
Accuracy
F1 Score
Area Under Curve
CAMIL H-BoW
CAMIL D-BoW
GICF


Even Train/Test Split, 0.1% of Data
100
90
80
70
60
50
40
30
20
10
0
Accuracy
F1 Score
Area Under Curve
CAMIL H-BoW
CAMIL D-BoW
GICF
MISVM
sbMIL

Conclusion and Future Work

- MIL can be an effective approach towards phenotype prediction
- CAMIL is a general example of this kind of method
- Future: different clustering & assembly algorithms, different MIL-based feature extraction methods, instance labels