Are You My Neighbor? Bringing Order to Neighbor Computing Problems.

David C. Anastasiu[1,2] , Huzefa Rangwala[3], and Andrea Tagarelli[4]

[1]Computer Engineering, San Jose State University, CA

[1]Computer Science & Engineering, Santa Clara University, CA

[2]Computer Science & Engineering, George Mason University, VA

[3]DIMES, University of Calabria, Italy

25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 2019

Part II: Neighbors in Genomics, Proteomics, and Bioinformatics

David C. Anastasiu, San José State University [david.anastasiu@sjsu.edu]

Huzefa Rangwala, George Mason University [rangwala@cs.gmu.edu]

Tutorial Outline

- **Part I: Problems and Data Types**
  - Dense, sparse, and asymmetric data
  - Bounded nearest neighbor search
  - Nearest neighbor graph construction
  - Classical approaches and limitations
- **Part II: Neighbors in Genomics, Proteomics, and Bioinformatics**
  - Mass spectrometry search
  - Microbiome analysis
- **Part III: Approximate Search**
  - Locality sensitive hashing variants
  - Permutation and graph-based search
  - Maximum inner product search
- **Part IV: Neighbors in Advertising and Recommender Systems**
  - Collaborative filtering at scale
  - Learning models based on the neighborhood structure
- **Part V: Filtering-Based Search**
  - Massive search space pruning by partial indexing
  - Effective proximity bounds and when they are most useful
- **Part VI: Neighbors in Learning and Mining Problems in Graph Data**
  - Neighborhood as cluster in a complex network system
  - Neighborhood as influence trigger set

Exact Filtering Open Modification Spectral Library Search

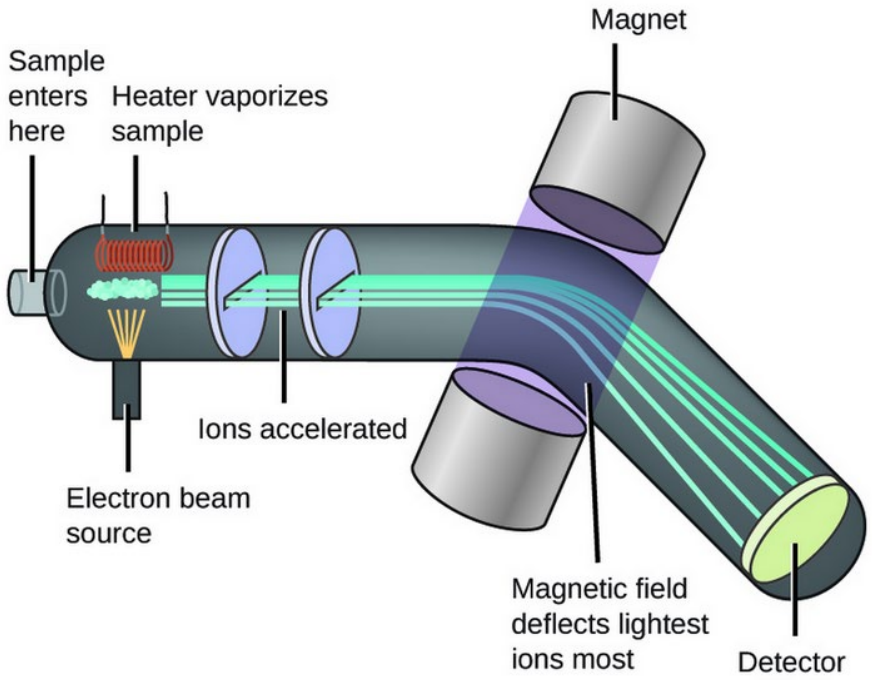
David C. Anastasiu, San José State University [david.anastasiu@sjsu.edu]

Starting September:
Department of Computer Science and Engineering
Santa Clara University

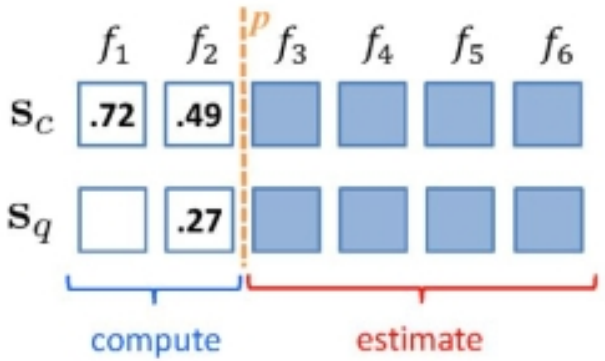
Open Modification Spectral Library Search

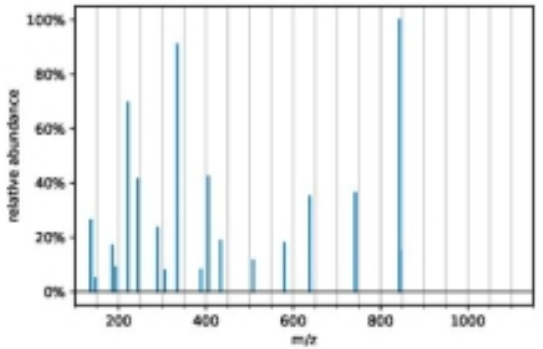
w/ William Stafford Noble
Genome Sciences, UW

- Methods for characterizing the protein composition of biological samples
  - Mass spectrometers output relative abundance histograms (spectra)
  - Massive databases exist for protein-associated spectra (spectral libraries)
  - Task is to match unknown spectra against nearest neighbor in library



- Challenges
  - Imperfect ionization/spectrometry
  - Size of databases (10's to 100's or million)





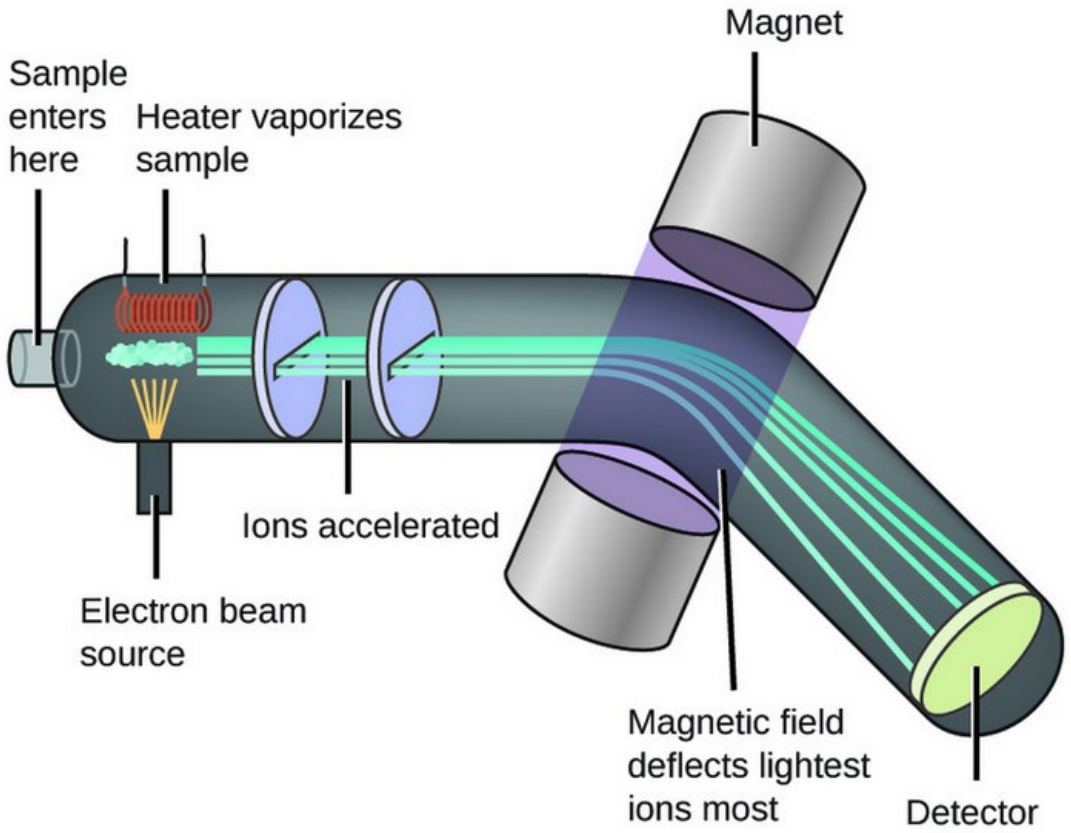
- Challenged By:

Eran Halperin,
CS @ UCLA

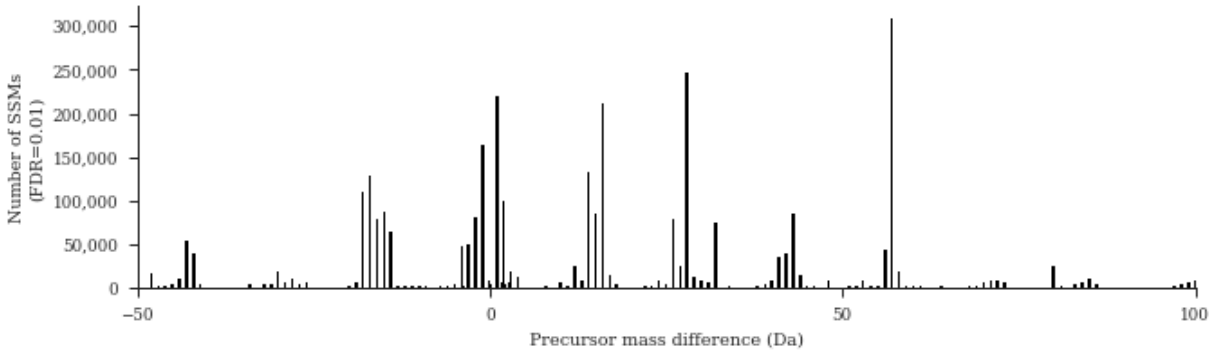
Image: https://i.stack.imgur.com/iVYVY.png

What Are Spectra?

- Mass spectrometry (MS)
  - **mass**-to-**charge** ratio of ions



Spectrum

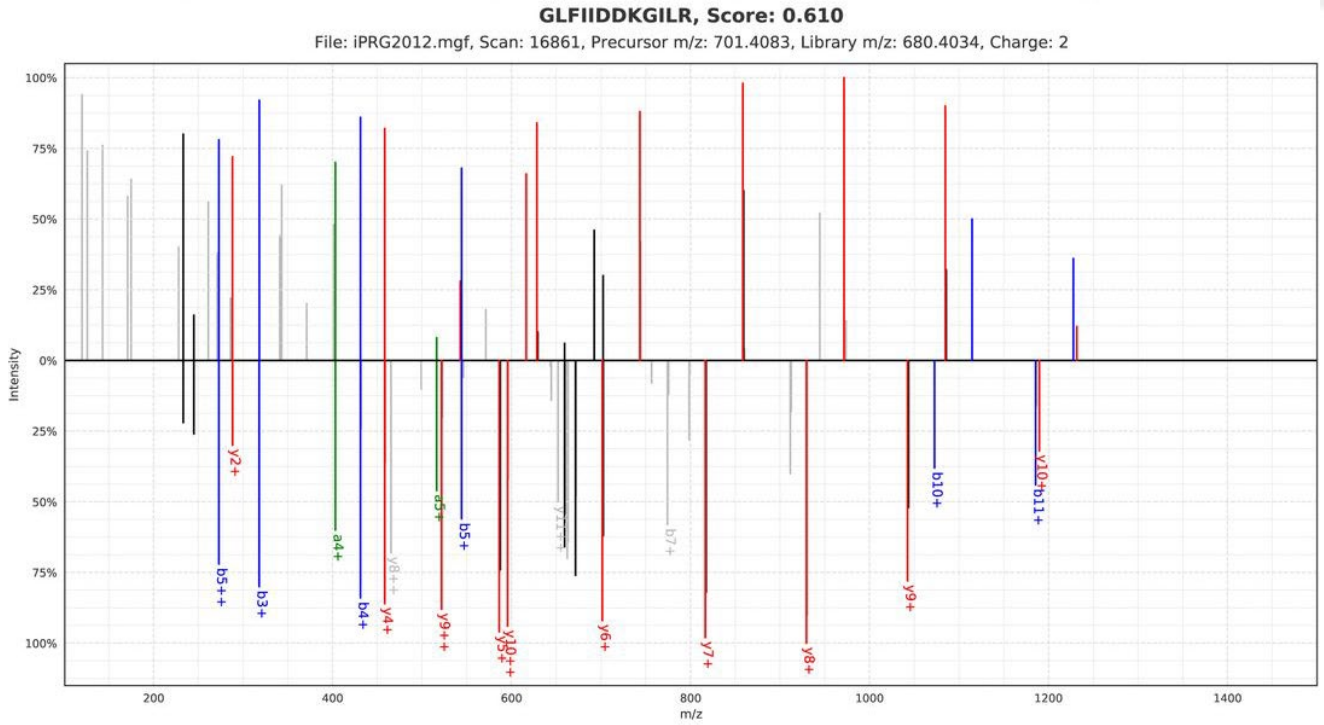


Shows relative abundance of chunks

Image: https://i.stack.imgur.com/iVYVY.png

Matching Spectra to Peptides

- Database search
  - How to represent spectra?
  - Is simple matching appropriate?



(b) The shifted dot product correctly matches both unmodified and modified fragments.

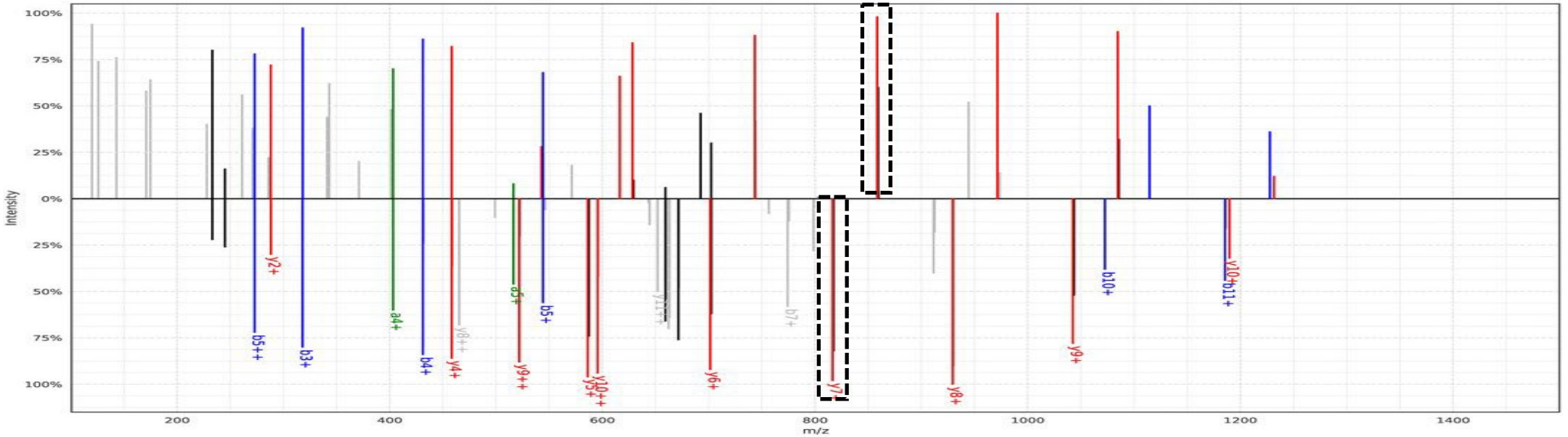
https://www.biorxiv.org/content/biorxiv/early/2019/05/05/627497.full.pdf

Account for:

- Post Translational Modification (PTM)
- Amino-acid mutations
- Precursor mass

Shifted Dot-Product Proximity

- Use peak charge information to locate potential shifted peaks in the query
  - Greedily match against peak with highest product
  - Discount weight by α (2/3) if not direct match
- For peaks with unknown charge (no annotation), test most common charges (0, 2, 3)



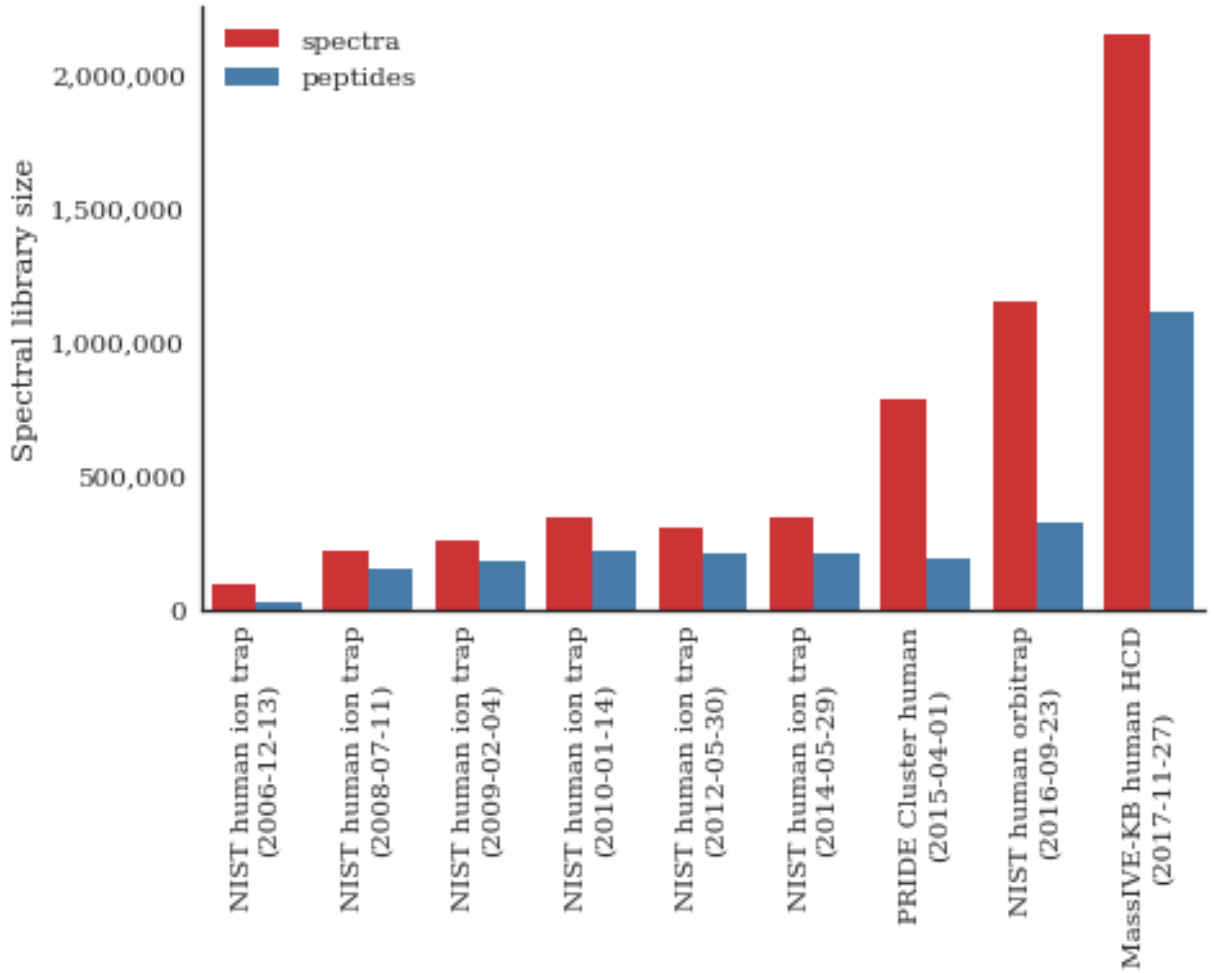
https://www.biorxiv.org/content/biorxiv/early/2019/05/05/627497.full.pdf

Massive Datasets

- Spectral libraries are growing exponentially
- Query sets also

<u>Dataset Specs:</u>

- Library
  - MassIVE-KB peptide SL
    - 4,226,826 spectra (incl. decoys)
  - Derived from 30TB human MS/MS proteomics data
- Queries
  - Human draft proteome
    - 30 samples, 2212 raw files
    - 24,033,575 spectra
    - LTQ-Orbitrap Velos & LTQ-Orbitrap Elite

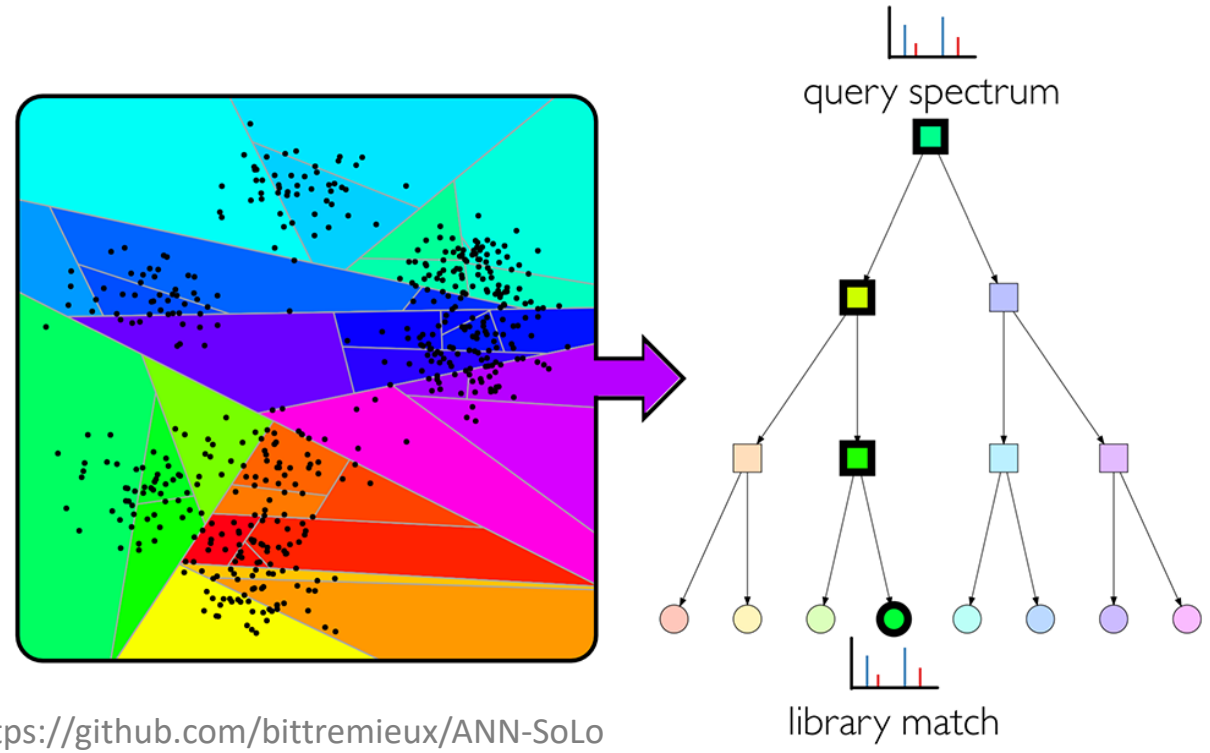


https://github.com/bittremieux/ANN-SoLo/blob/master/notebooks/spec_lib_size.ipynb

Current State-Of-The-Art

- ANN-SoLo (Wout Bittremieux et al., Bill Noble)
  - Embed spectra in Euclidean space
  - Existing approximate nearest neighbor search methods
  - Verify candidates using shifted dot-product (SDP) proximity

Met Bill @ CGSI
COMPUTATIONAL GENOMICS SUMMER INSTITUTE



Latest version uses GPUs to further accelerate search

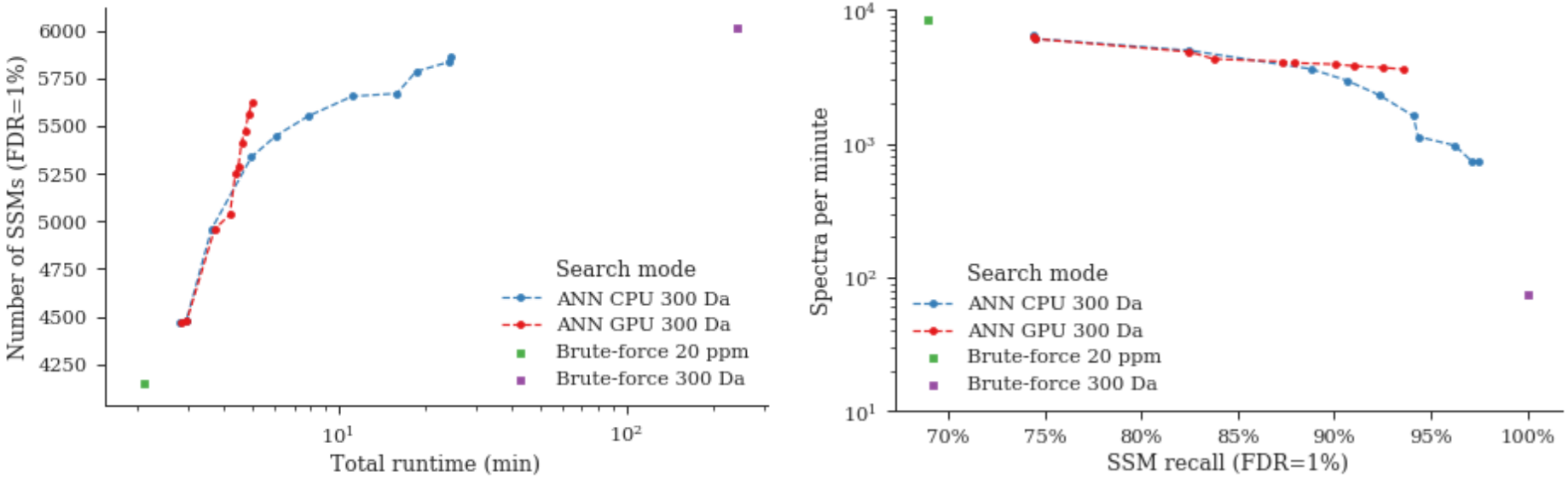
https://github.com/bittremieux/ANN-SoLo

Results

- Total time: 1,177,305 s, i.e., **1 week, 6 days, 15 hours, 1 minute, 45 seconds**
- Num. queries: 24,033,575
- Total matches: 14,032,494
- Cosine matches: 9,760,497
- SDP matches: 4,271,997

Servers:
Intel Xeon E5-2660 v4, 28 cores
256 GB RAM
NVIDIA Tesla P100 GPU

Experiments on the iPRG2012 data set (human HCD SL)



https://github.com/bittremieux/ANN-SoLo/blob/master/notebooks/iprg2012_ann_hyperparameters.ipynb

Next Steps

- Phase I: Improve quality of results through exact cosine similarity candidate generation
  - Still requires retrieving a large number of candidates, since the gap between the Cosine and SDP scores can be quite large
  - Use of efficient filtering-based searcher can mitigate efficiency concerns

- Phase II: Filtering-based SDP searcher
  - Focus directly on the SDP 1-NN (or low k-NN) problem
  - Eliminate potential matches whose SDP score cannot be higher than the lowest SDP score of current neighbors

- Phase III: Effective data decompositions for distributed parallel SDP filtering

Acknowledgements

CRII: III: RUI: Effective Protein Characterization via Fast Exact Open Modification Searching
Grant No: 185055

Students:


Ging Gonzalez, BS


Ryan Shu, MS


Xueer (Cindy) Xue,BS

Parners:

- Wout Bittremieux, UW
- Prof. William Stafford Noble, UW

References

[1] Wout Bittremieux, Kris Laukens, & William Stafford Noble. Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units. bioRxiv (2019).

[2] Wout Bittremieux, Pieter Meysman, William Stafford Noble & Kris Laukens. Fast open modification spectral library searching through approximate nearest neighbor indexing. Journal of Proteome Research, 17, 3463–3474 (2018).

Microbiome Analysis

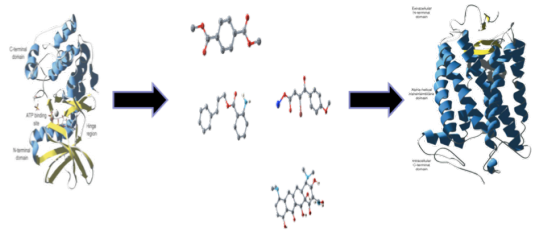
Huzefa Rangwala, Ph.D.

Computer Science

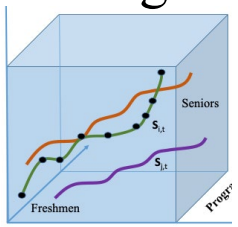
Research Area: Data Mining

Develop novel and practical computational solutions towards inter-disciplinary applications.

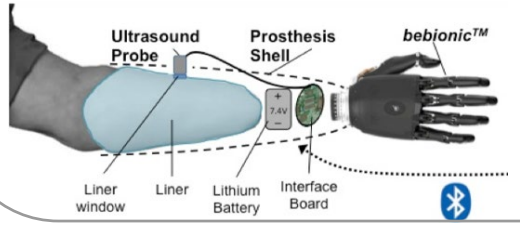
Biology

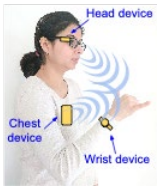


Educational Data Mining

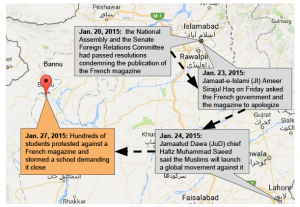


Cyber-Physical Sciences

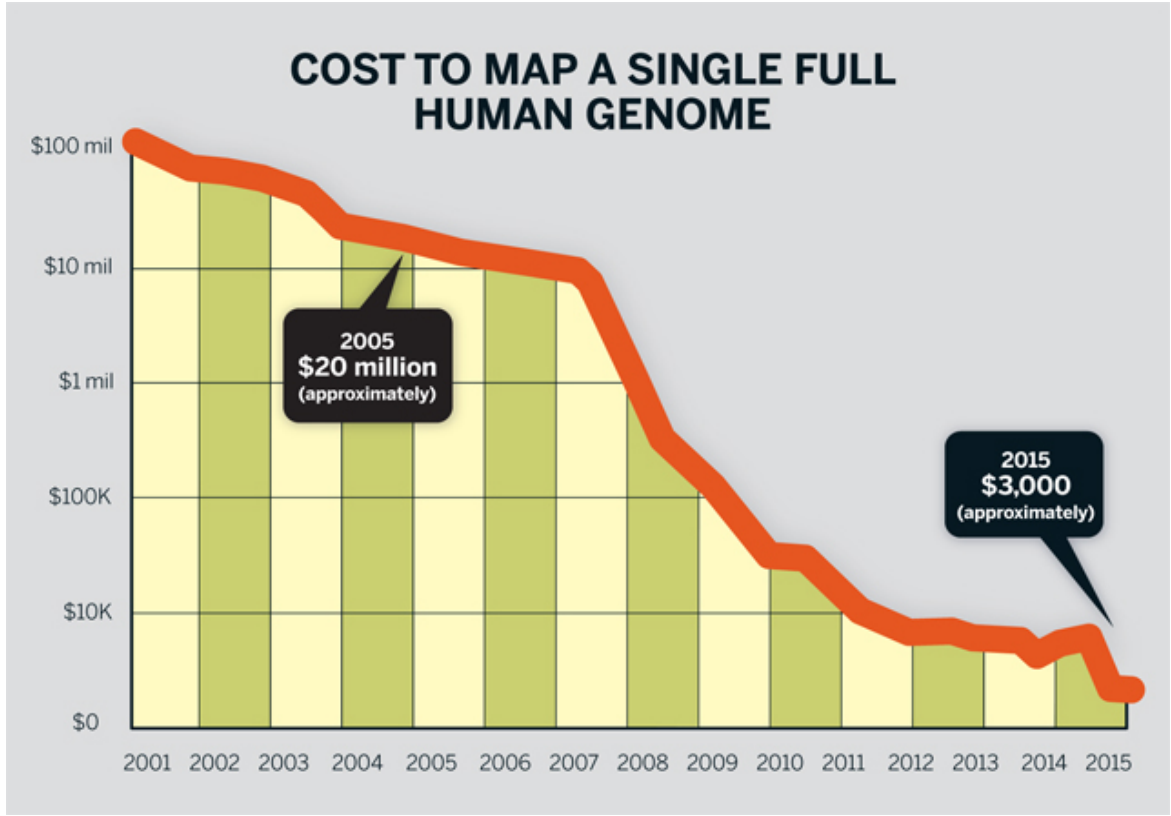




Social Forecasting



Sequencing Technology Advances



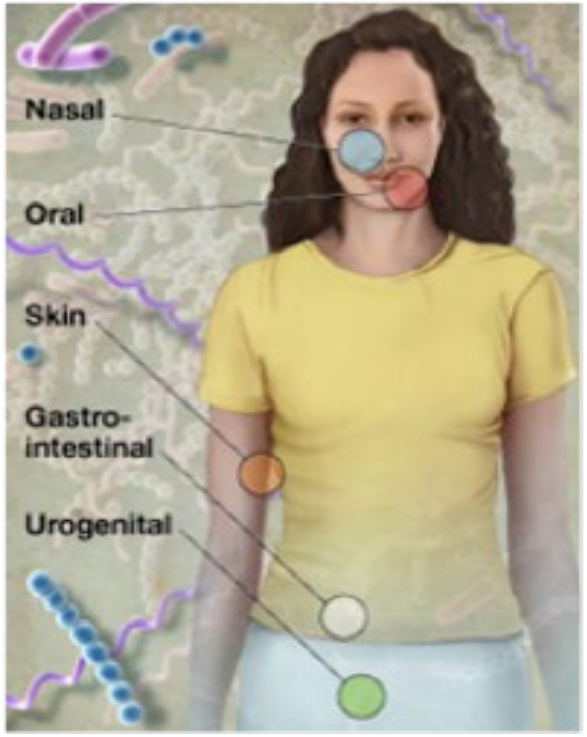
Ion Torrent

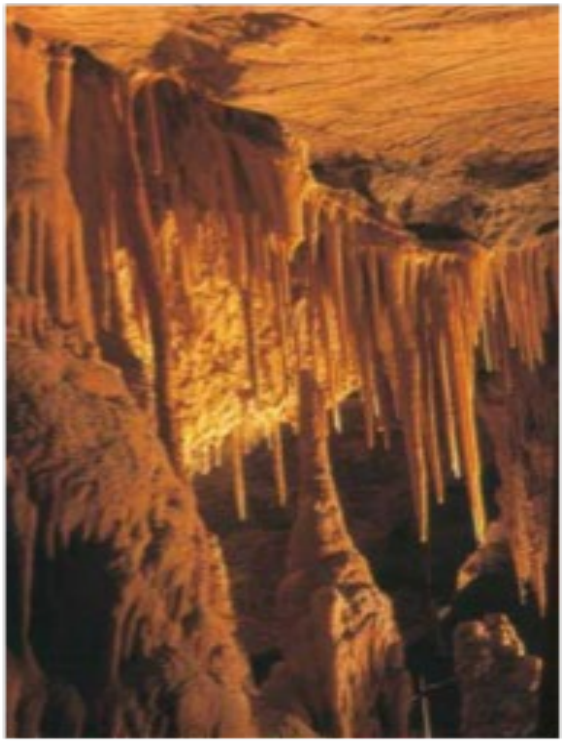
600GB Per Run

Human + Microbial Cells = Microbiome

"Our Self-Portrait: the Human Microbiome" by Joana Ricou. Illustration by Steven H. Lee

Microbial Communities Everywhere





E. Grice, H. Kong, S. Conlan, C. Deming, J. Davis, A. Young, G. Bouard, R. Blakesley, P. Murray, E. Green et al., “Topographical and temporal diversity of the human skin microbiome,” science, vol. 324, no. 5931, pp. 1190-1192, 2009

Metagenome Assembly and Annotation

Input Reads From Sequencing Technologies

N reads: Read Length L ≈ 100-1000 bp

Challenges

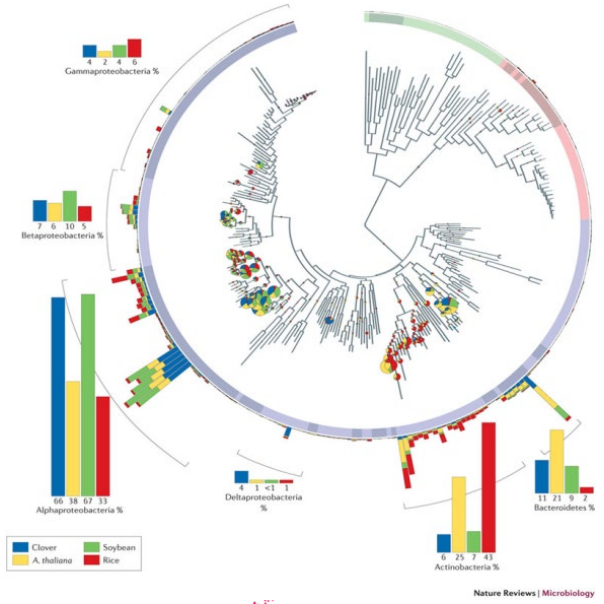
- Microbes vary in abundance across samples.
  - Distribution is unknown
  - Samples may have one species dominating whereas others may have several with uniform distribution. (also referred to as complexity)
  - Low abundance species overlooked (Need High coverage)
- Microbial genomes vary in length
- Unrelated microbes may have similar sequence reads
- Lack of reference genomes
  - Not lab-culturable or individually sequenced.
- Sequencing Technologies Issues
  - High throughput, Short reads, TB of data.
  - Error profiles.

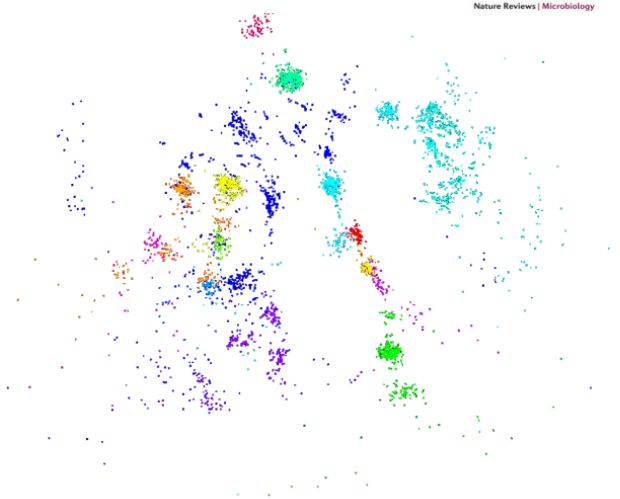
Research Objectives

- To build efficient computational algorithms for metagenome analysis using both supervised and unsupervised learning.
  - Classification methods assist in identifying the taxonomic classification of reads with the metagenome samples (supervised)
  - Clustering methods lead to species-specific groupings and assists in the identifying the content and abundance of microbial species within the metagenome samples (unsupervised)
- To analyze and annotate large volumes of available sequence data (require efficient tools and algorithms)

Problem Description

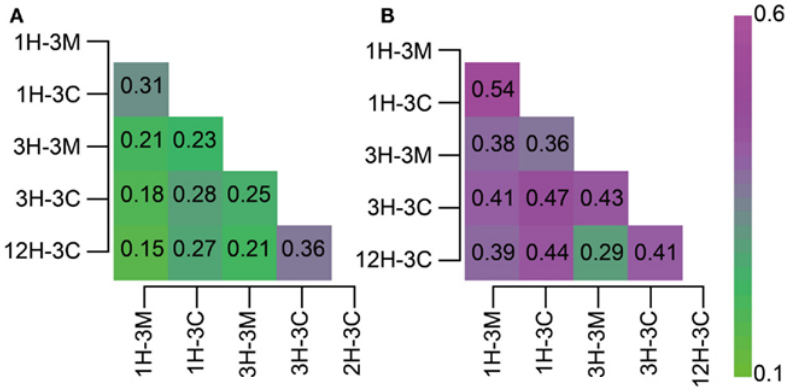
- Most metagenome projects follow sequencing of **16S genes** (rather than entire genomes) to identify different communities in an environment
- Different groups/species in a sample are called Operational Taxonomic Units (OTUs)
- Gives an approximation of species diversity in a sample.
- Clustering methods lead to species-specific groupings and give abundance of microbial species (unsupervised)

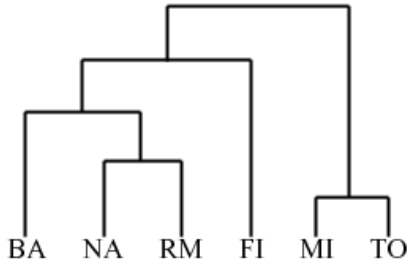


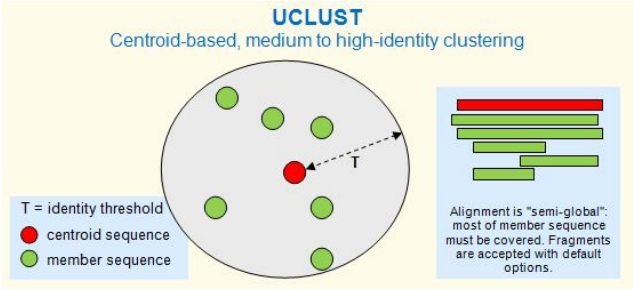


Related Work

- Mothur and DOTUR uses a pairwise distance matrix and perform hierarchical clustering to determine OTUs.
- ESPRIT computes w-mer distance and perform hierarchical clustering to define OTUs.
- UCLUST and CD-HIT use cluster representative approach.
- CROP uses bayesian clustering approach to define OTUs.
- Memory and time intensive algorithms.







Our Proposed Approach

- The key characteristic of new algorithm is the use of an efficient randomized search technique called “locality sensitive hashing”.

- Incorporate the use of fixed-length gapless subsequences, commonly referred to as w-mer to improve the sensitivity of matching pairs of sequences.

Locality Sensitive Hashing (LSH)

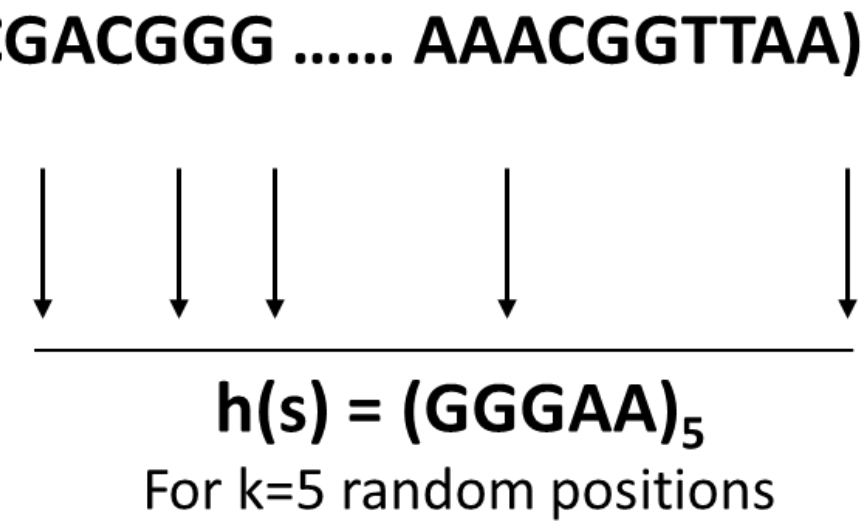
- Finding very similar items can be computationally demanding.

- **Idea**: Construct hash function h: Rd → U such that for any pair of points p,q :
- If D(p,q) ≤ r, then Pr[h(p) = h(q)] is high
- If D(p,q) > r, then Pr[h(p) = h(q)] is small

- Example: Hamming Distance
    - LSH function: h(p) = pi , i.e. the i-th bit of p
    - Probabilities: Pr[h(p) = h(q)] = 1 - D(p,q) / d

- Thus somewhat similar can be efficient.

LSH-Div Framework



- Given a nucleotide string **s** of length **n**, we construct a randomized hash function. We choose **k** uniform, random indices $i_1 \ldots i_k$ in the range {1…n} to define a hash function h(s) given by: h(s) = < s $[i_1]$, s$[i_2]$ … s$[i_k]$ >

Use of w-mers per position

$s = (ACGACGGG \ldots\ldots AAACGGTTAA)_n$

- Given a string **s** of length **n**

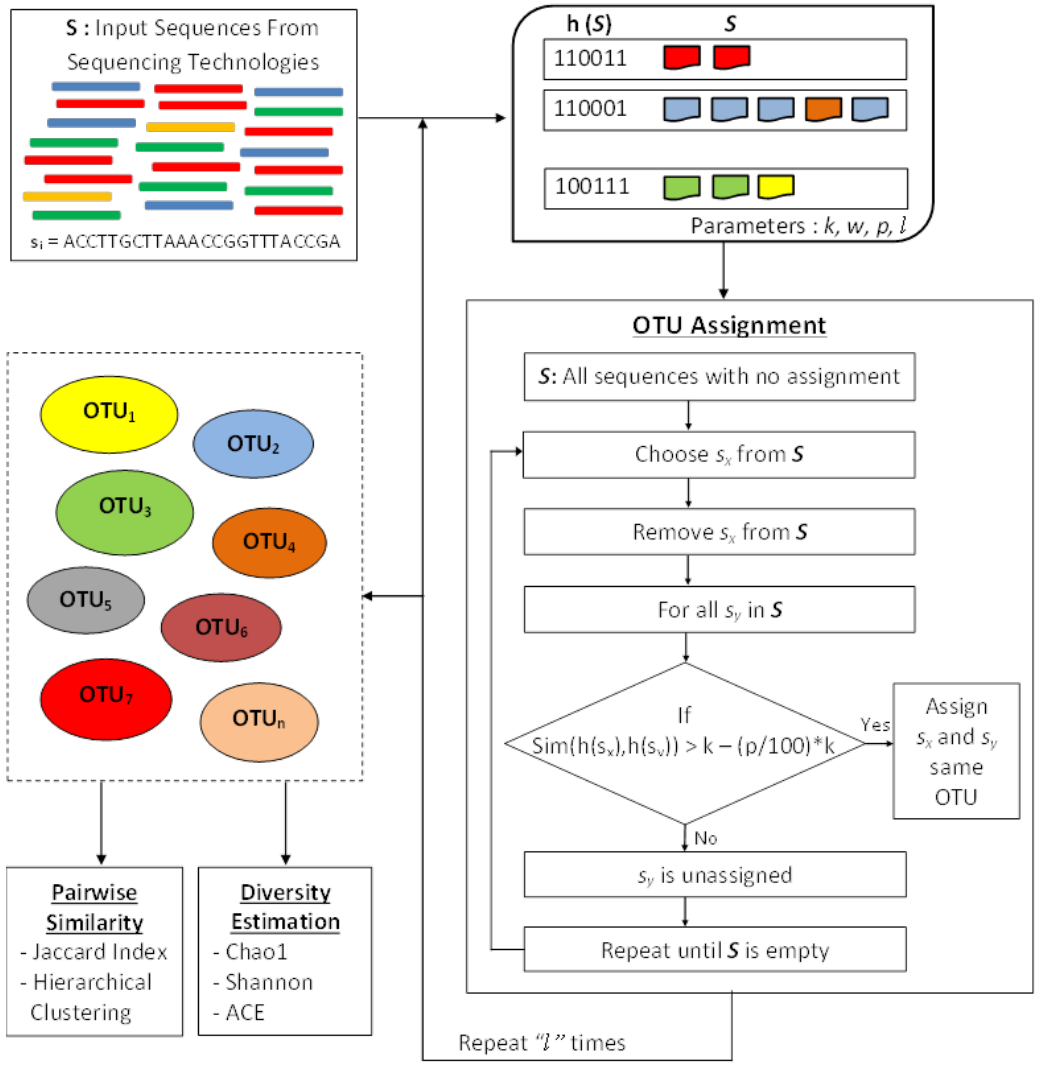
$s = (ACGACGGG \ldots\ldots AAACGGTTAA)_n$

- Define h(s) with k=4 random positions

$s = (ACGACGGG \ldots\ldots AAACGGTTAA)_n$

- Define h(s) with k=4 random positions and choose w characters to the left and right

LSH-Div Process Flow Diagram



Experimental Protocol

- Environmental samples
  - Eight seawater samples (give a global in-depth description of the diversity of microbes and their relative abundance in the ocean)
  - Human skin data (covers 21 different locations)

- Synthetic Dataset
  - 43 reference gene sequence data
  - Fourteen simulated whole metagenome datasets with varying proportions of microbes

- Evaluation Metrics
  - Number of OTUs (groups)
  - Chao Estimate, Shannon diversity, Abundance-Based Coverage (ACE) indices.
  - Sequence Similarity (Global Sequence Alignment Score)
  - Cluster Accuracy
  - Computation time and Memory

Species Richness Metrics

- **Chao Index**: Chao Index is based on the number of OTUs with only one sequence called “singletons” and the number of OTUs with only two sequences called “doubletons”.

$$S_{chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)},$$

- **Shannon Diversity**: Shannon Diversity index uses the number of sequences in each OTU and the total number of sequences in the community.

$$H' = -\sum_{i=1}^{S_{obs}} \frac{n_i}{N} \ln \frac{n_i}{N},$$

- **ACE Index:** Abundance-based Coverage Estimator Index is based on an “abundant” threshold which sets a limit on the number of assigned sequences in an OTU. The number of OTUs with “abundant” or fewer sequences are referred to as rare OTUs

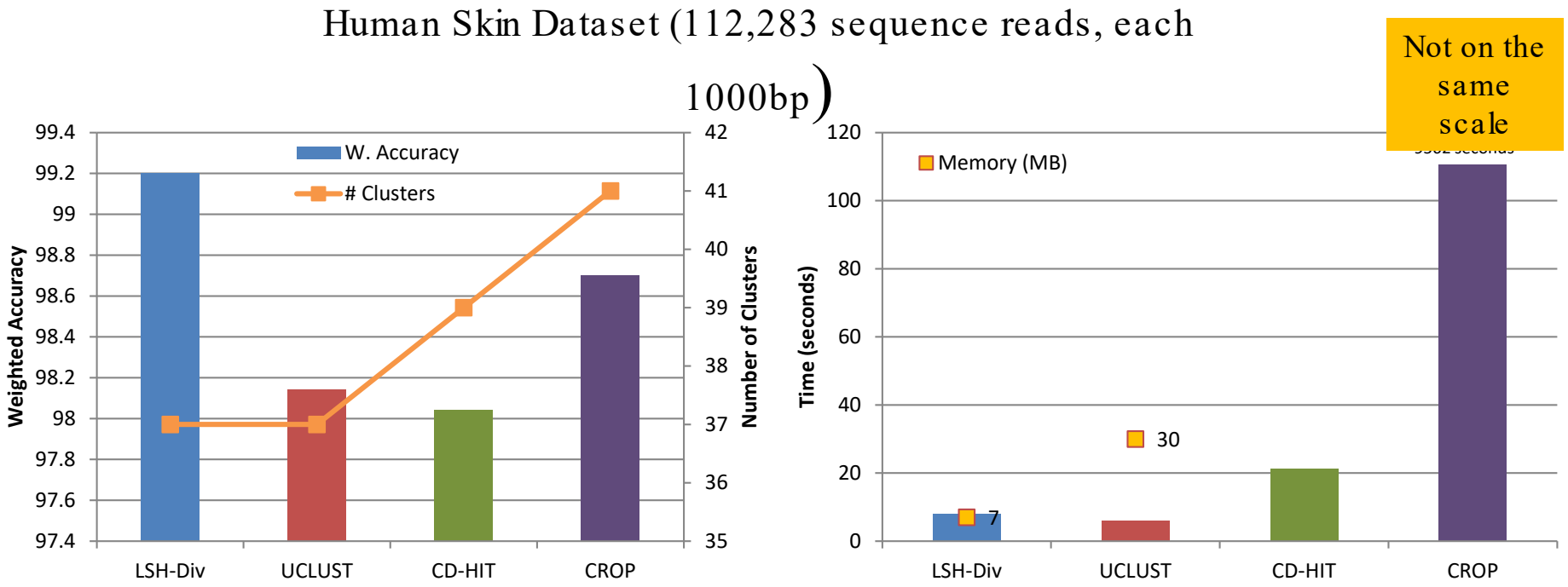
$$N_{rare} = \sum_{i=1}^{abund} in_i$$

$$C_{ACE} = 1 - \frac{n_1}{N_{rare}}$$

$$\gamma^2_{ACE} = max\left[\frac{S_{rare}}{C_{ACE}} \frac{\sum_{i=1}^{abund} i(i-1)n_i}{N_{rare}(N_{rare}-1)} - 1, 0\right]$$

$$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{n_1}{C_{ACE}}\gamma^2_{ACE},$$

Performance Evaluation

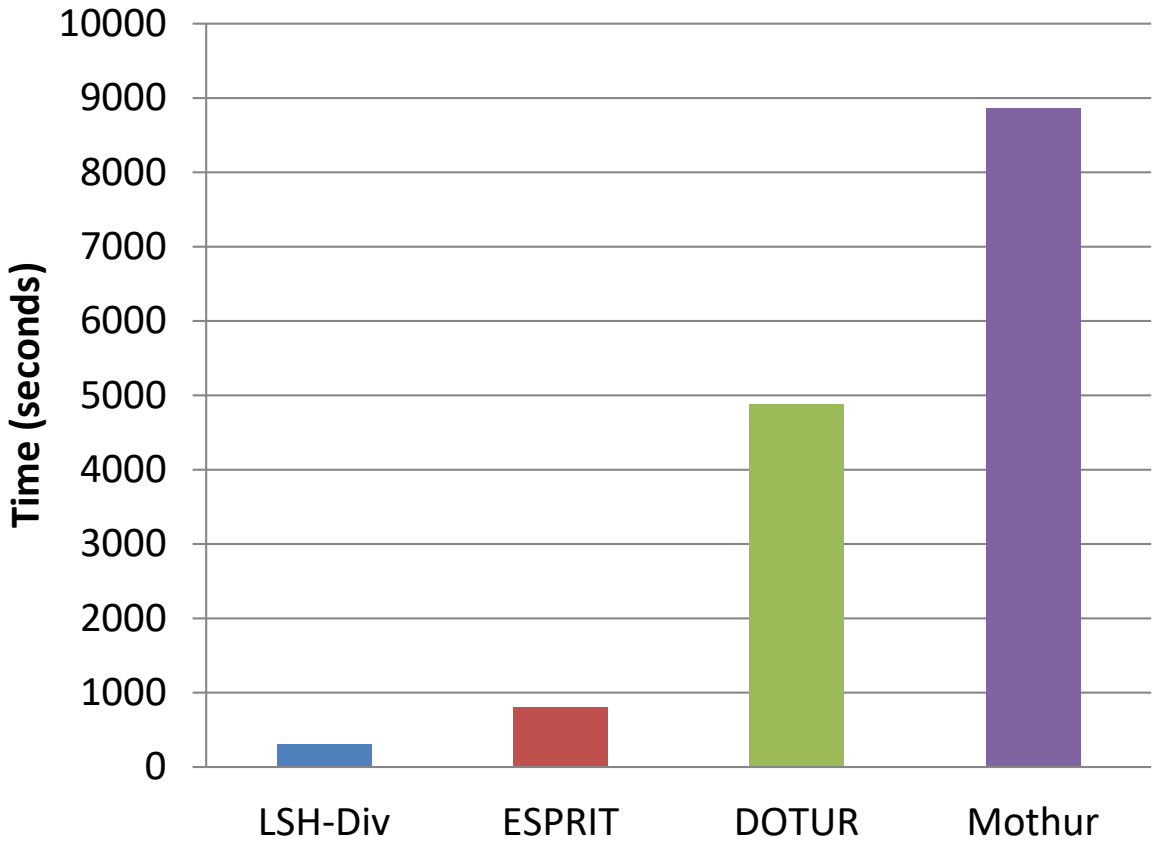


- LSH-Div produces smaller number of clusters with a higher weighted accuracy
- LSH-Div is time and memory efficient

E. Grice, H. Kong, S. Conlan, C. Deming, J. Davis, A. Young, G. Bouard, R. Blakesley, P. Murray, E. Green et al., "Topographical and temporal diversity of the human skin microbiome," science, vol. 324, no. 5931, pp. 1190-1192, 2009

Run Time Comparison

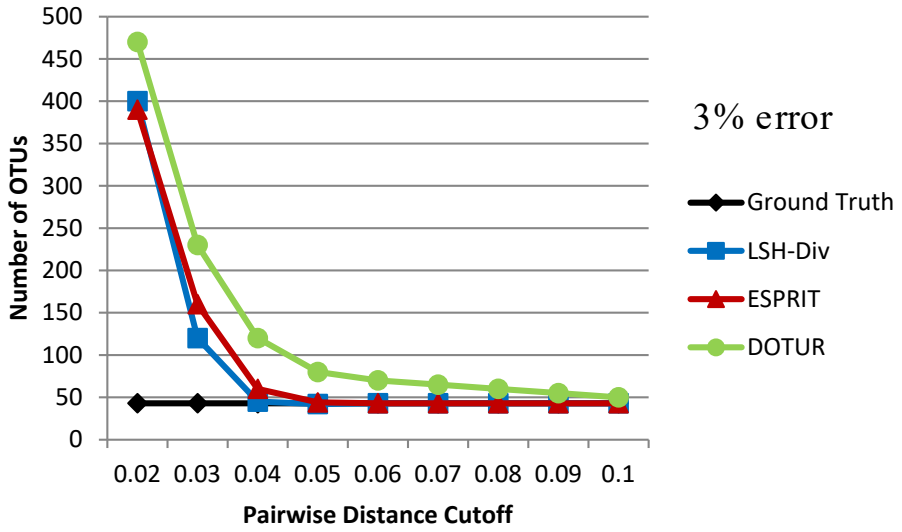
- Environmental samples containing 100,000 sequence reads per sample (each 60 bp)
- Average computational time across eight environmental samples.
- LSH-Div is computationally efficient compared to other methods

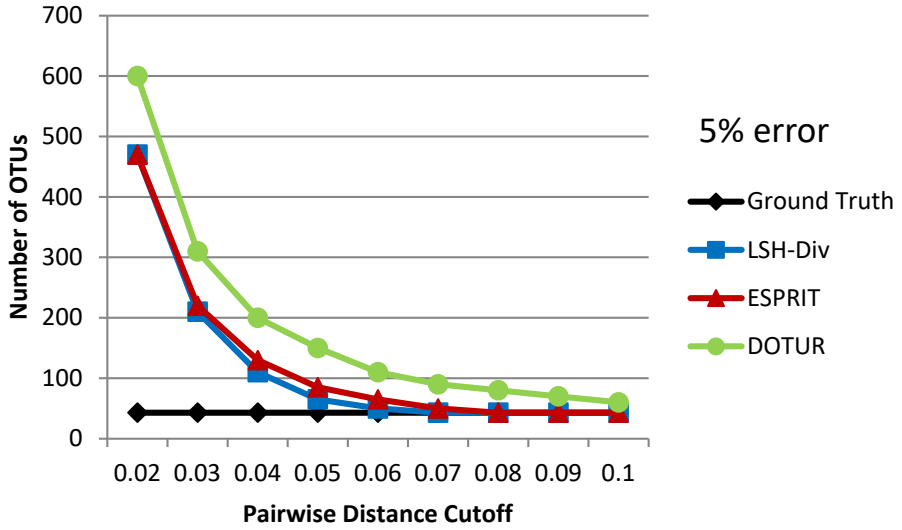


- S. M. Huse, L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin, “Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing,” PLoS genetics, vol. 4, no. 11, p. e1000255, 2008.

Synthetic Data

- 345,000 sequence reads representing 43 reference gene sequences
- Proven performance of LSH-Div for OTU estimation



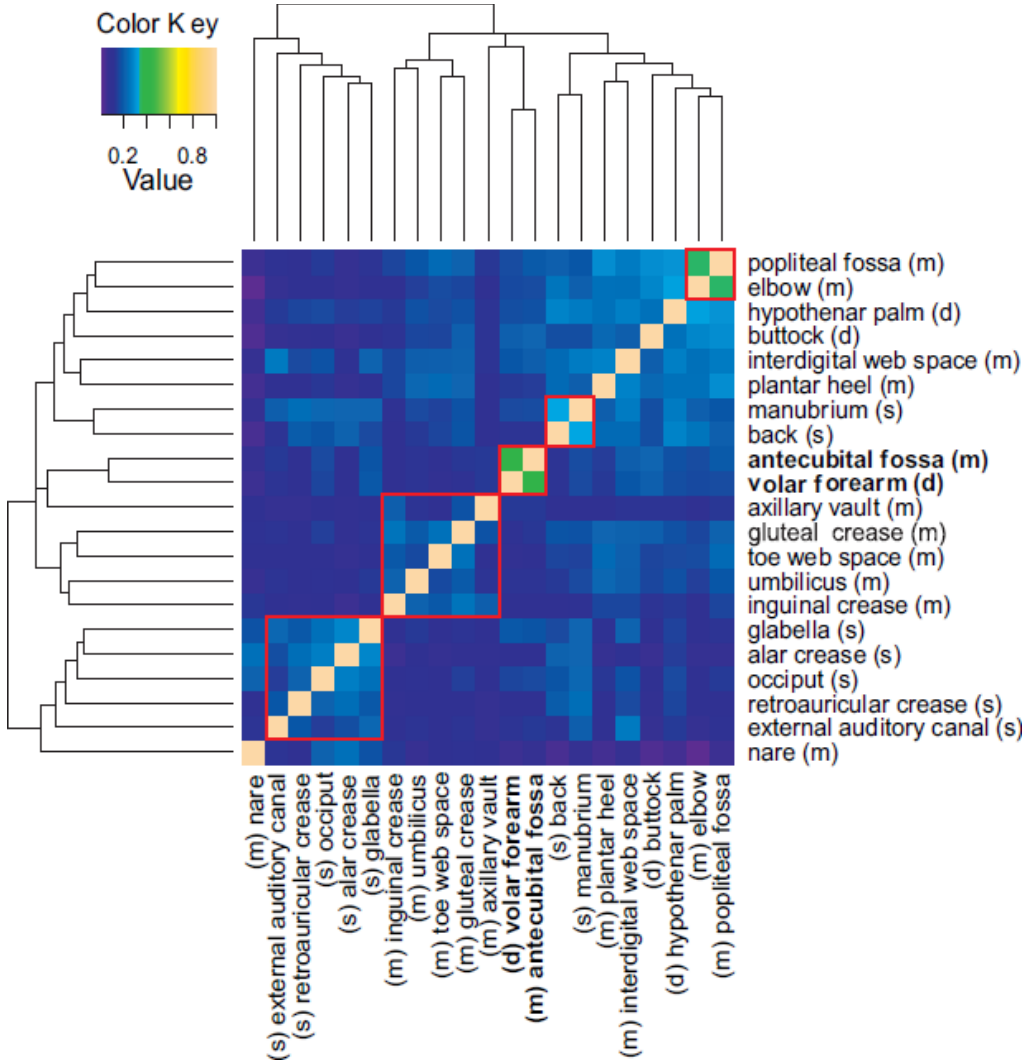


- S. M. Huse, L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin, “Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing,” PLoS genetics, vol. 4, no. 11, p. e1000255, 2008.

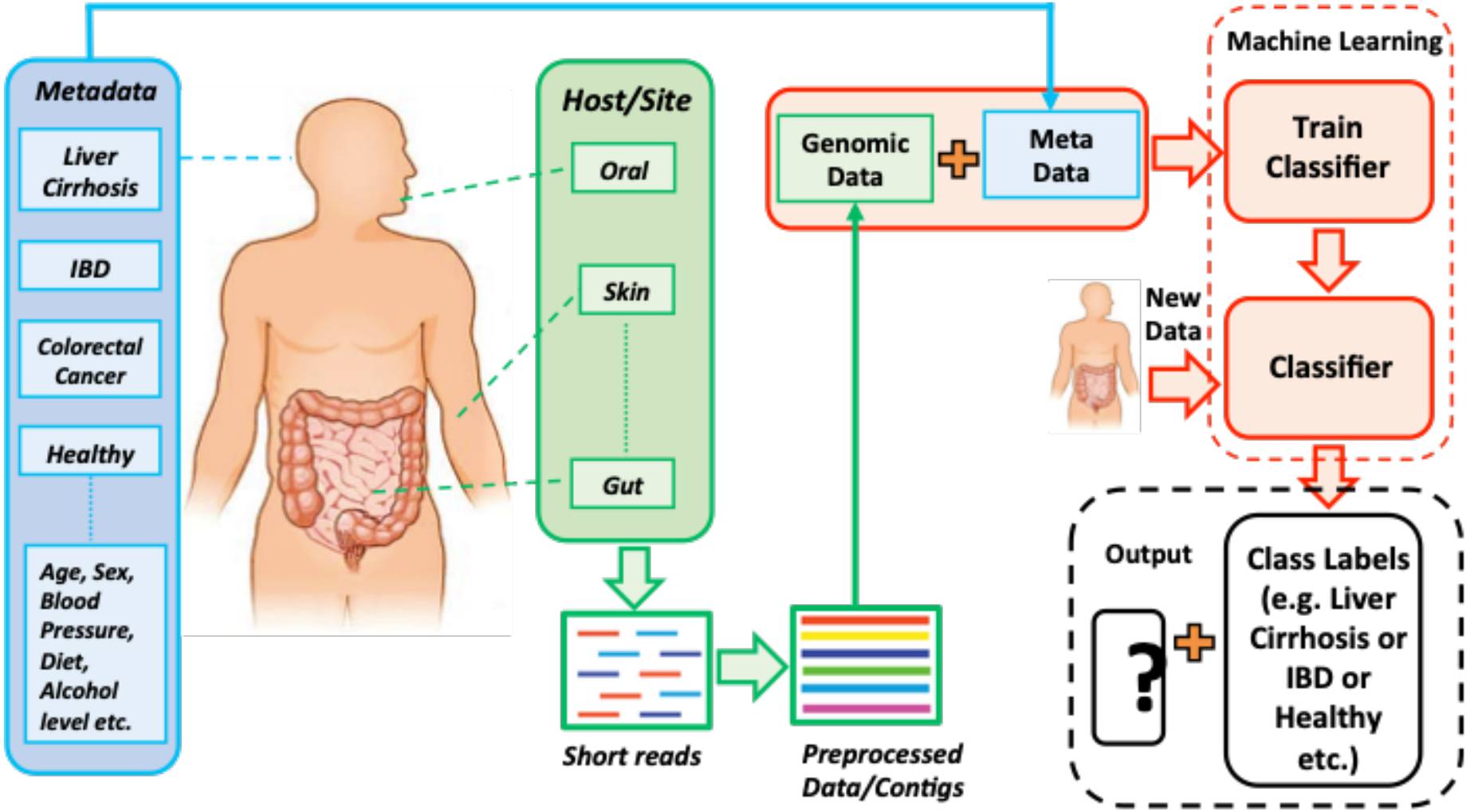
Use Case Scenario (Application)

- 21 different skin locations.
- Computes Jaccard coefficient using OTUs as features per skin sample / location.
- The Jaccard coefficient measures the membership using the proportion of shared OTUs between the two samples / locations
- Validated by previous study by Costello et al.

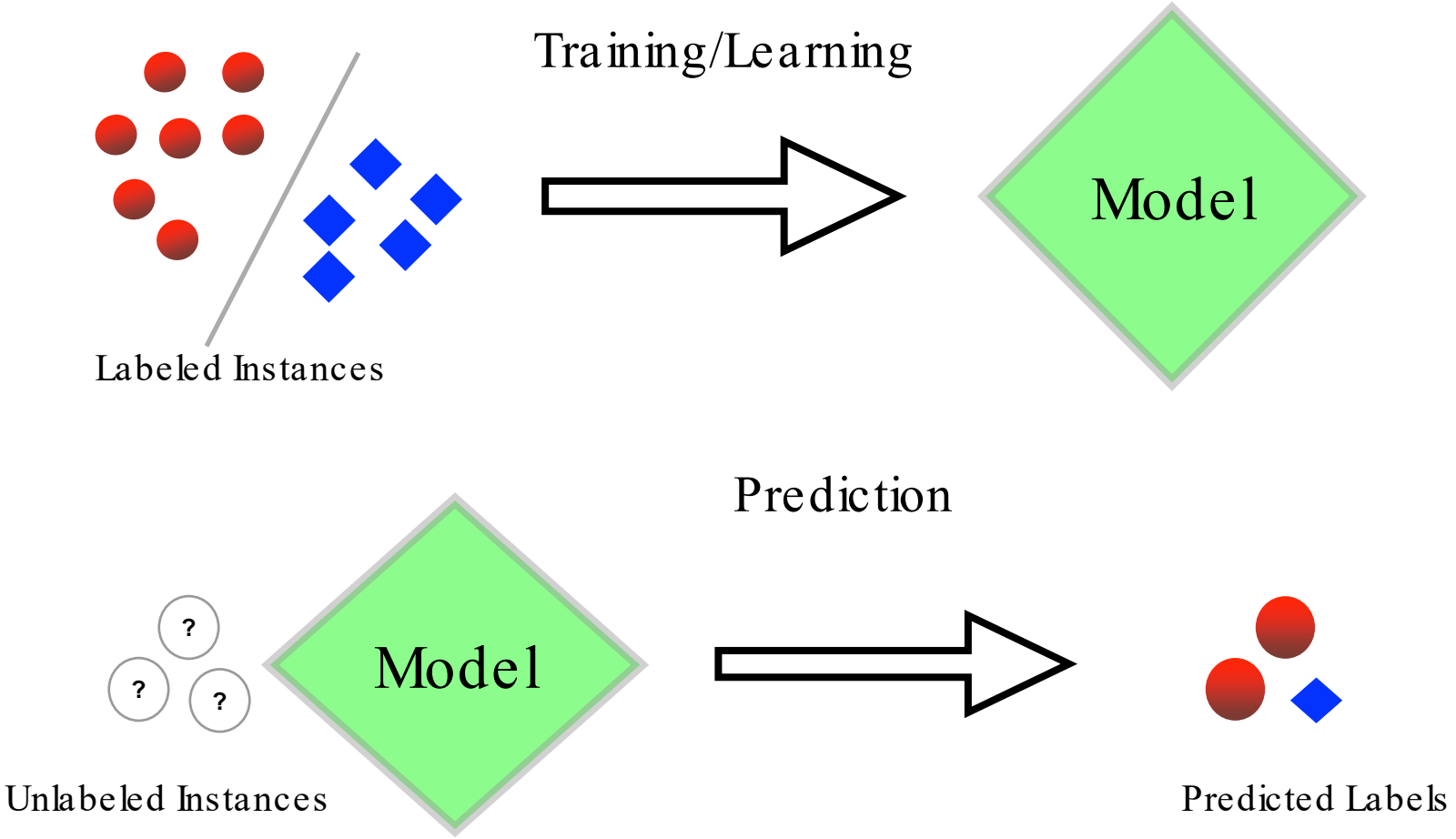
- E. K. Costello and et al., "Bacterial community variation in human body habitats across space and time," Science, vol. 326, no. 5960, pp. 1694-1697, 2009.



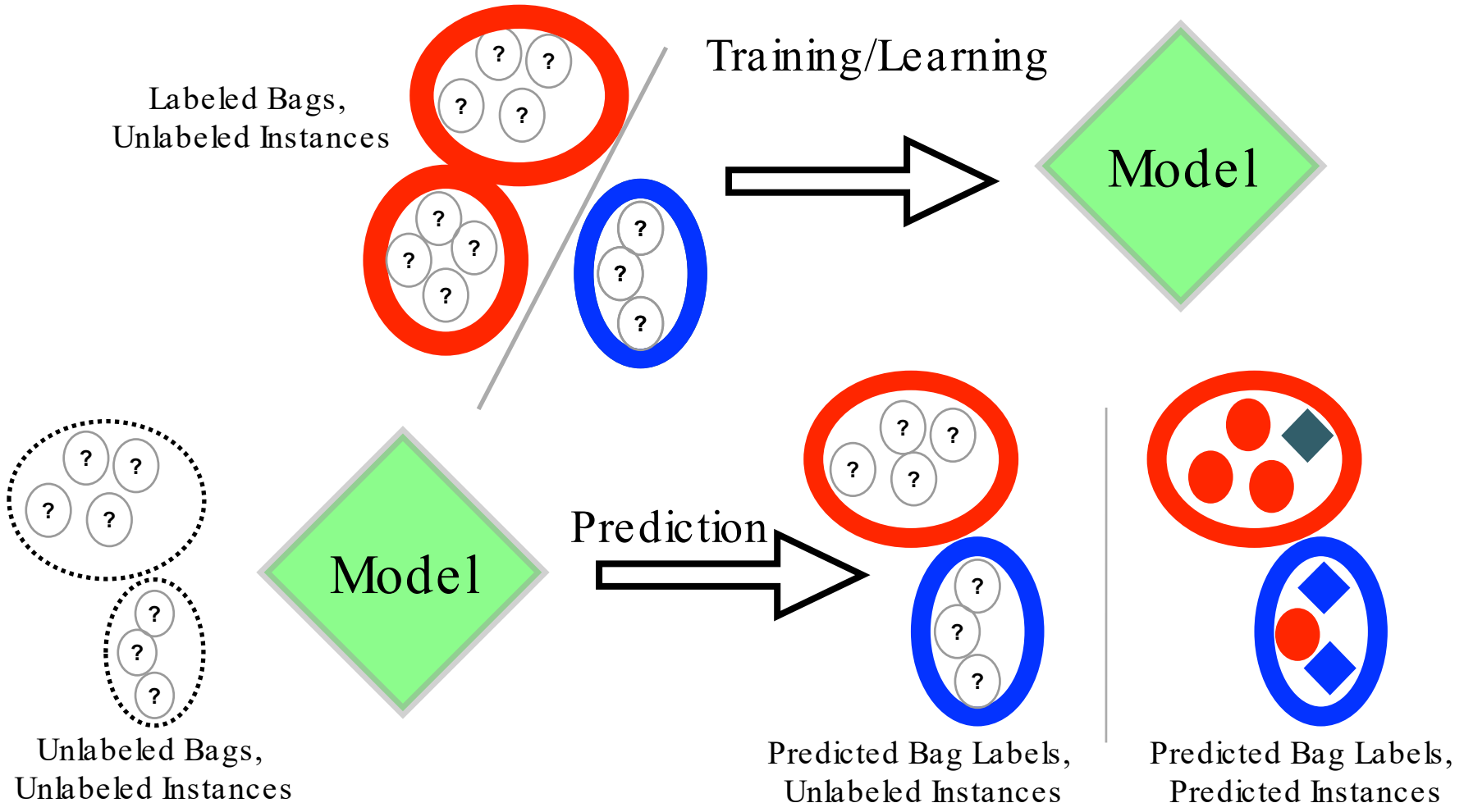
Phenotype Prediction Workflow



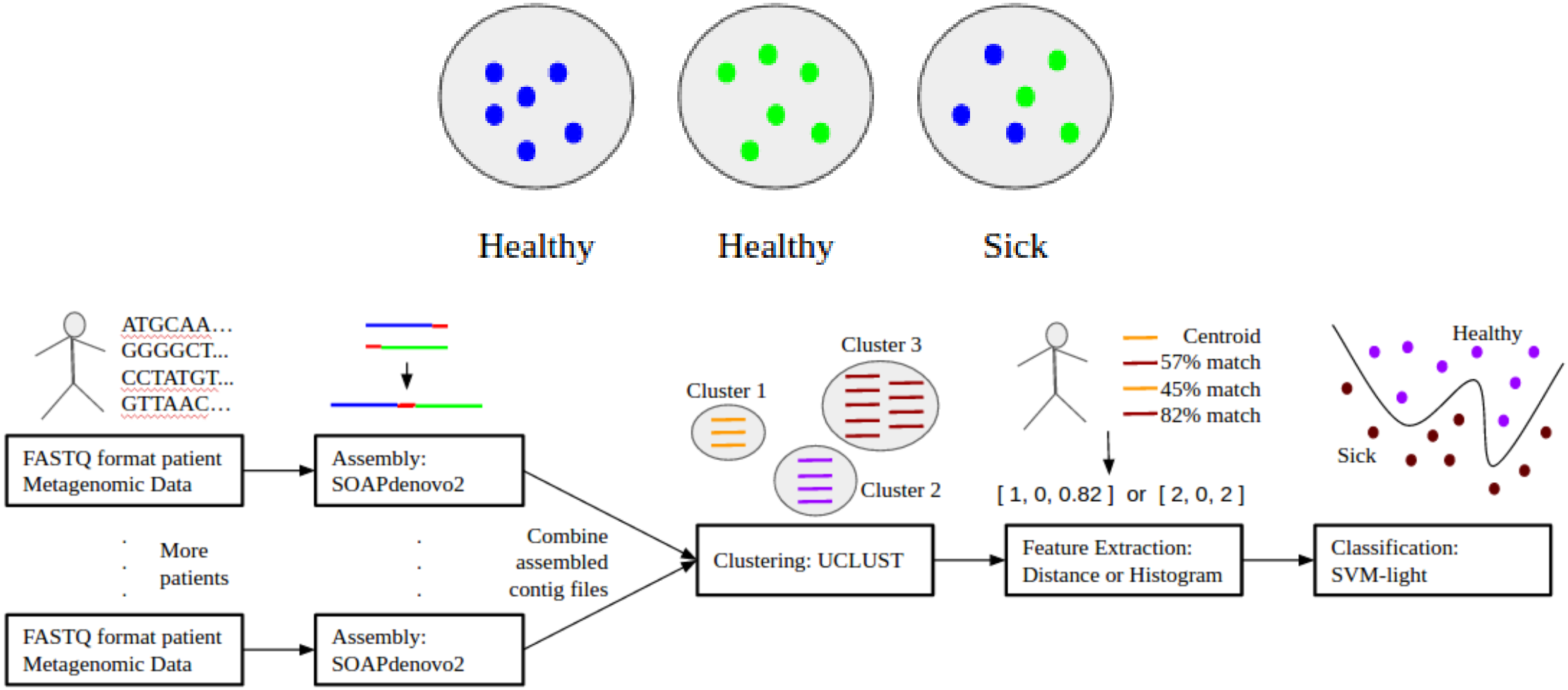
Supervised Learning



Multiple Instance Learning



CAMIL (Clustering & Assembly with Multiple Instance Learning)



Rahman et. al. ACM TCBB 2017

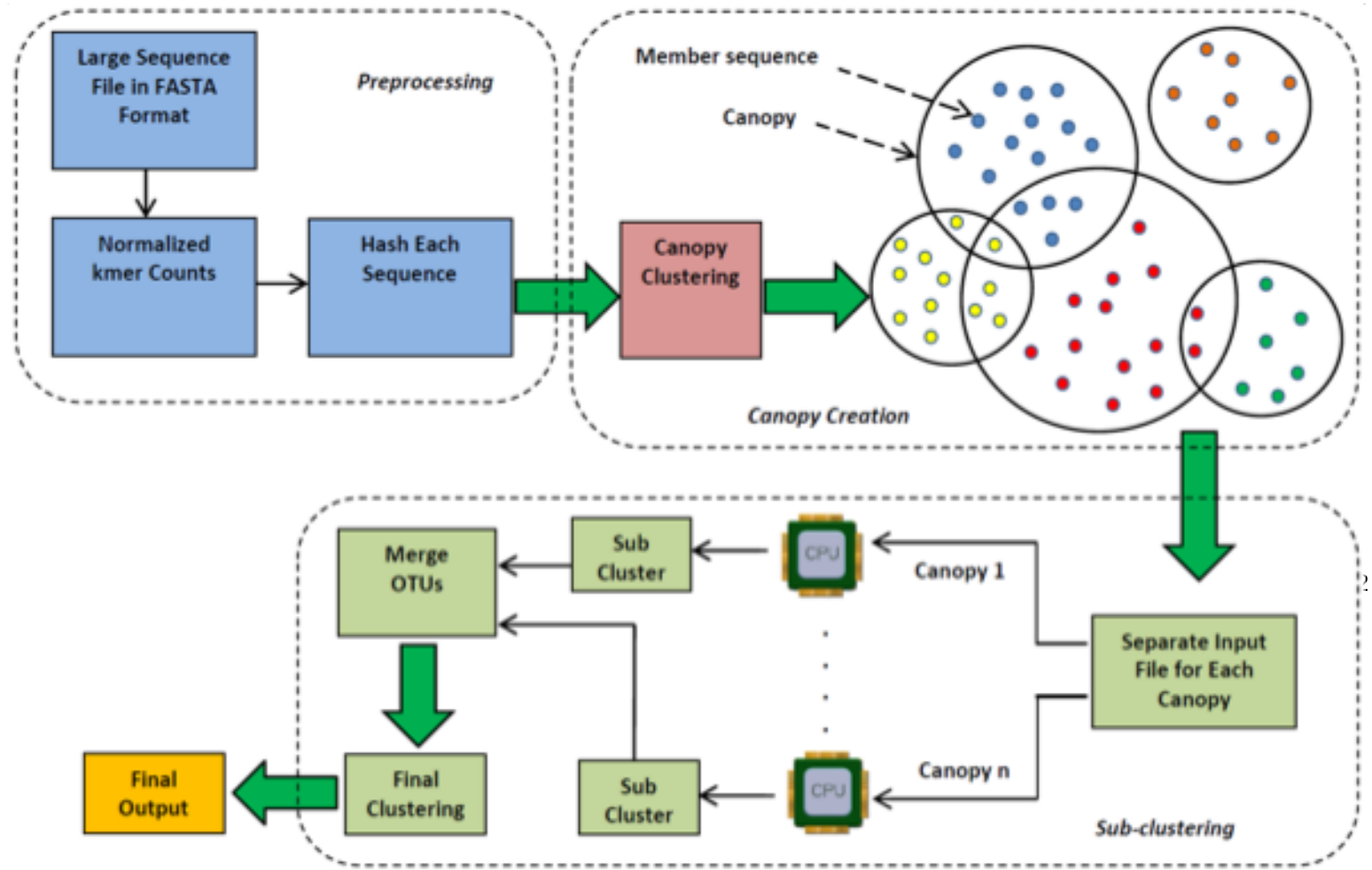
Multiple Instance Learning Challenges

- Standard Multiple Instance Learning Assumptions
  - Bag is considered negative if at least one instance within the bag is negative [Deittrich et. al. 1997]
- Key instance Discovery
  - We are interested in which instances are associated with the phenotype label
- Data Size
  - Large number of reads per clinical sample (bag)
  - Total Data Size Ranges from GB to TB
    - Prior MIL Algorithms worked on few hundreds of bags with 1000 instances per bag.

Key Contributions

- CAMIL: Incorporate Clustering Solutions within the MIL pipeline
- Clustering Algorithms Applied to Large Genome Data Sets
  - Large Computational Run Time due to pairwise sequence comparisons
- Proposed Two-Phased Approximate Clustering Solution
  - Greedy Approach
  - Use of Fast Neighborhood Search Techniques for Fast Sequence Comparison
  - Distributed Implementation
    - Breaks down to be highly concurrent
  - Speeds Up Other Metagenome Clustering Algorithms

Canopy Clustering



Rahman et. al. JBCB 2017

Canopy Clustering

Canopy Clustering (Mcallum et. al. 2000):

Input: N data points

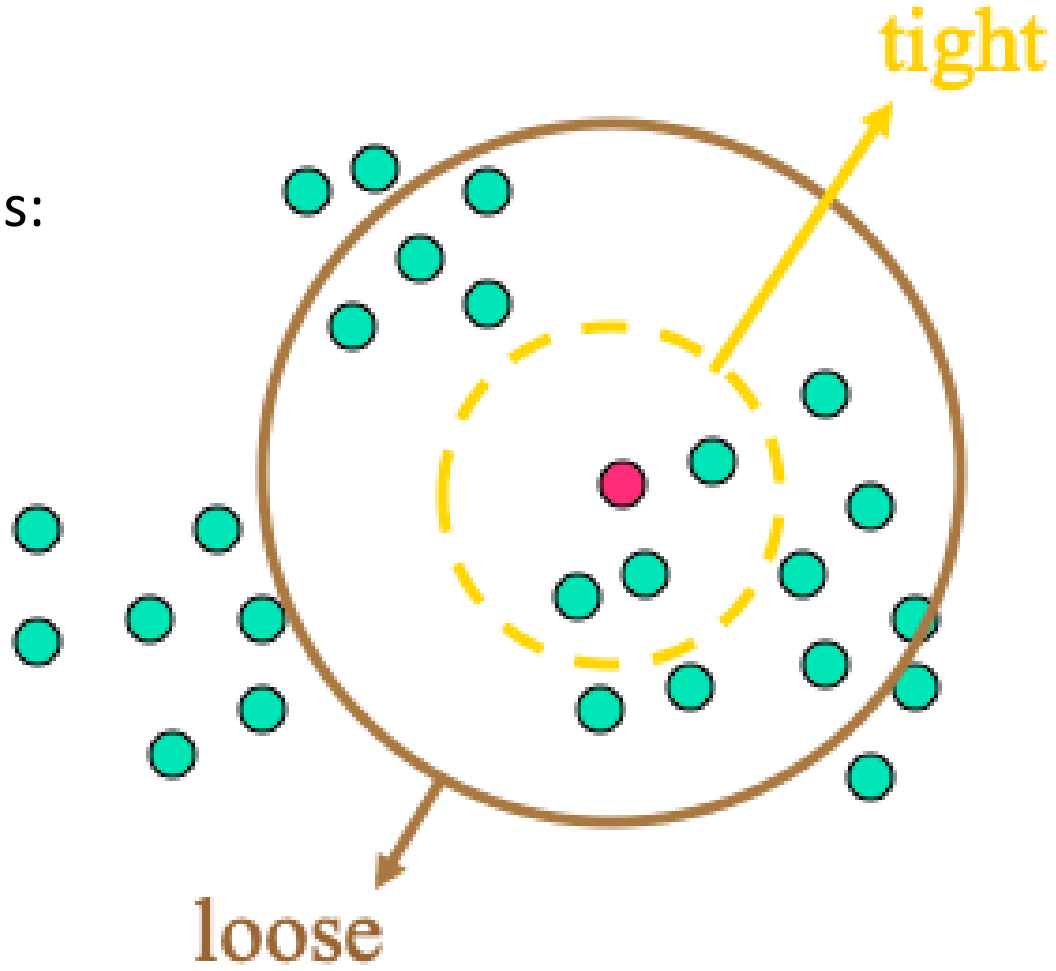
Output: Clusters (called Canopies)

Input Parameters: Two distance thresholds:

loose threshold, T1 and

tight threshold, T2

T1 > T2



Results (Improved Run Times)

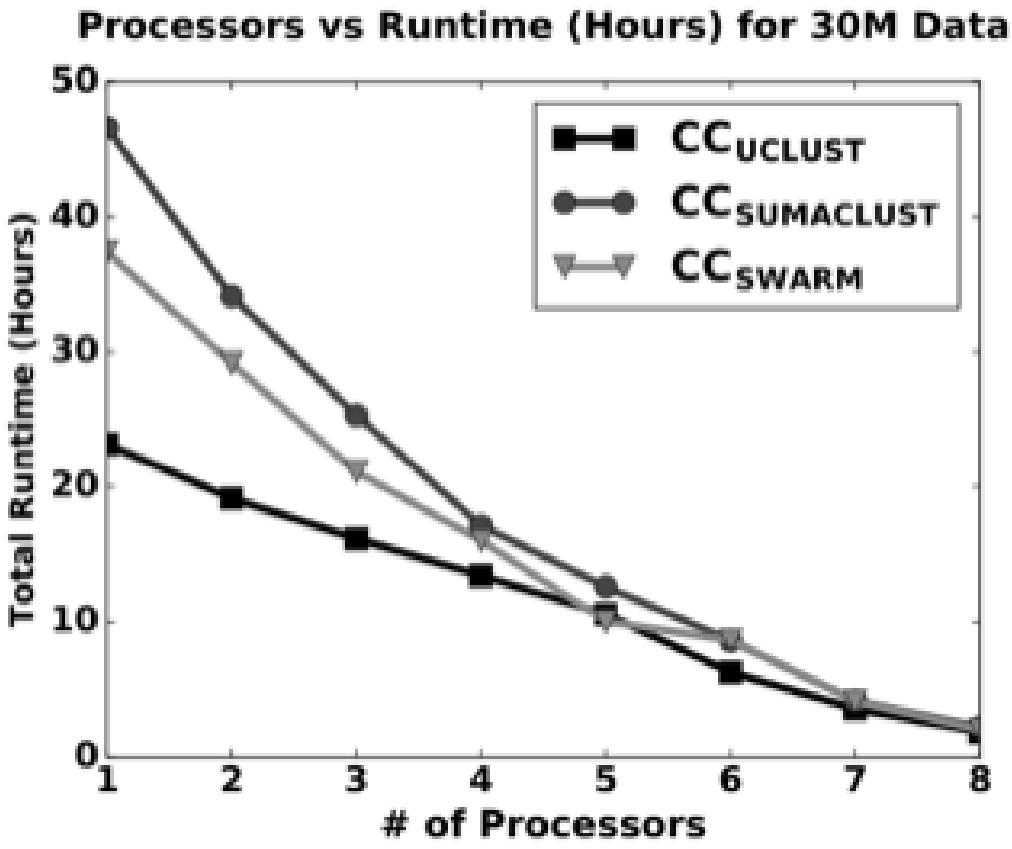
Integrate Canopy-Based Clustering with Prior Clustering Algorithms for Improved Runtime

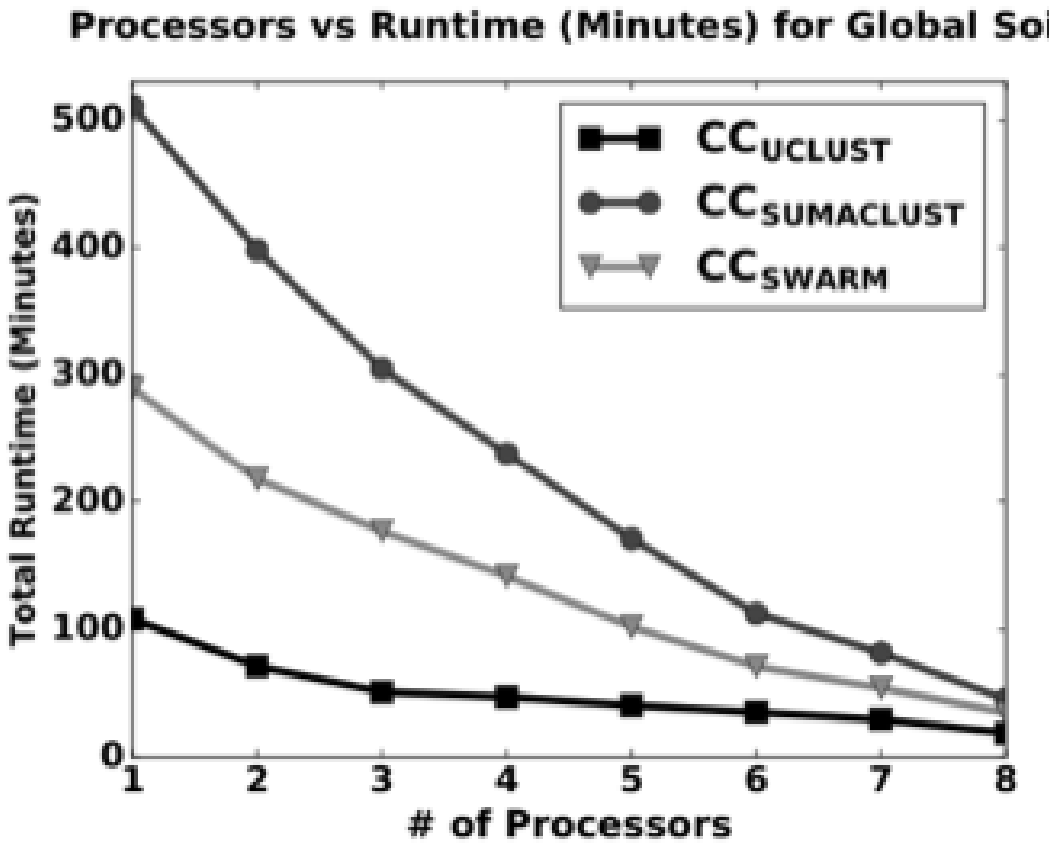
Speedup on Different Benchmarks Integrating CC

Dataset	#Reads	CC+UCLUST	CC+SUMACLUST	CC+SWARM
Bokulich	6.9M	4.1x	8.2x	7.4x
Canadian Site	2.9M	3.2x	6.4x	5.6x
Global Site	9.2M	5.7x	11.2x	8.3x
Liver Cirrhosis	30M	12.1x	21.1x	18.6x

Rahman et. al. JBCB 2017

Results (Scalability)





Experiments on Intel i7 64-bit processor with 8 core CPUs and 12GB RAM

Rahman et. al. JBCB 2017

Results (Type-2 Diabetes Dataset)

CAMIL Phenotype Classification Performance

Method	Classification Time	Memory Usage	F1-Score
MISVM	-	Error	-
sbMIL	-	Error	-
GICF	8h, 44min	2.6 GB	68.33%
CAMIL	**10 min**	**695 MB**	**74.31%**

Instance Level Results (Liver Cirrhosis)

Top-Instance Level Predictions
Streptococcus Salivarius
Clostridium Bolteae
Veillonella Parvula
Haemophilus Parain
Ruminococcus Gnavus
Lachnoclostridium
Prevotella Melaninogenica
Ruminococcus Torques
Klebsiella Pneumoniae

Verified Top-Instance Predictions based on BLAST (Google-like) Hits to Annotated Databases Associated with Bacterial Gene Sequences

Summary & Outcomes

- Developed clustering algorithms to scale to metagenome datasets.
  - LSH-Div (Locality-Sensitive Hashing) [Rasheed et. al. 2012 BMC Genomics]
  - Mc-MinH (Min-wise Hashing) [Rasheed and Rangwala 2013 SIAM SDM]
  - MrMc-MinH (Map-Reduce based) [Rasheed and Rangwala 2013 IPDSW]
  - Canopy Clustering [Rahman et. al. 2017 JBCB]
- Developed hierarchical classification methods
  - Given a metagenome sample, identify taxa, function and metabolic potential [Rasheed et. al. 2012 JBCB]
- Clinical Outcomes related to Alcoholism and Inflammatory Bowel Disease [Mutlu et. al. 2012 Gut, Bajaj et. al. 2013 Plos One]
- Multiple Instance Learning based Pipeline (Deep Learning Based)
  - Scaling [Rahman et. al 2018 In Review]
  - Instance-level Classification [Rahman et. al. 2017 TCBB]

Acknowledgement:

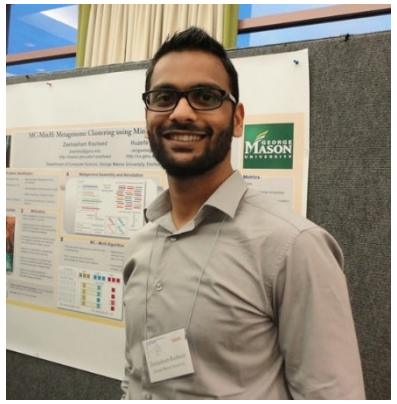
National Science Foundation (NSF), National Institutes of Health (NIH) and US Dept. of Agriculture (USDA)

CAREER: Annotating the Microbiome using Machine Learning Methods

- Contributors:

Mohammad Arifur Rahman
Grad. Student
Dept. of CS, GMU

Nathan LaPierre
Undergrad. Student
Dept. of CS, GMU



Zeehasham Rasheed, PhD
Data Scientist, AOL

Dr. Huzefa Rangwala
Professor
Dept. of CS, GMU