SIMCOMP: A hybrid Soft Clustering of Metagenome Reads

By

Shruthi Prabhakara

Raj Acharya

What is metagenomics?

- A study of genetic material recovered directly from the environment, bypassing cultivation
  - Who is there?
  - How many are there?

  → Sequence-driven

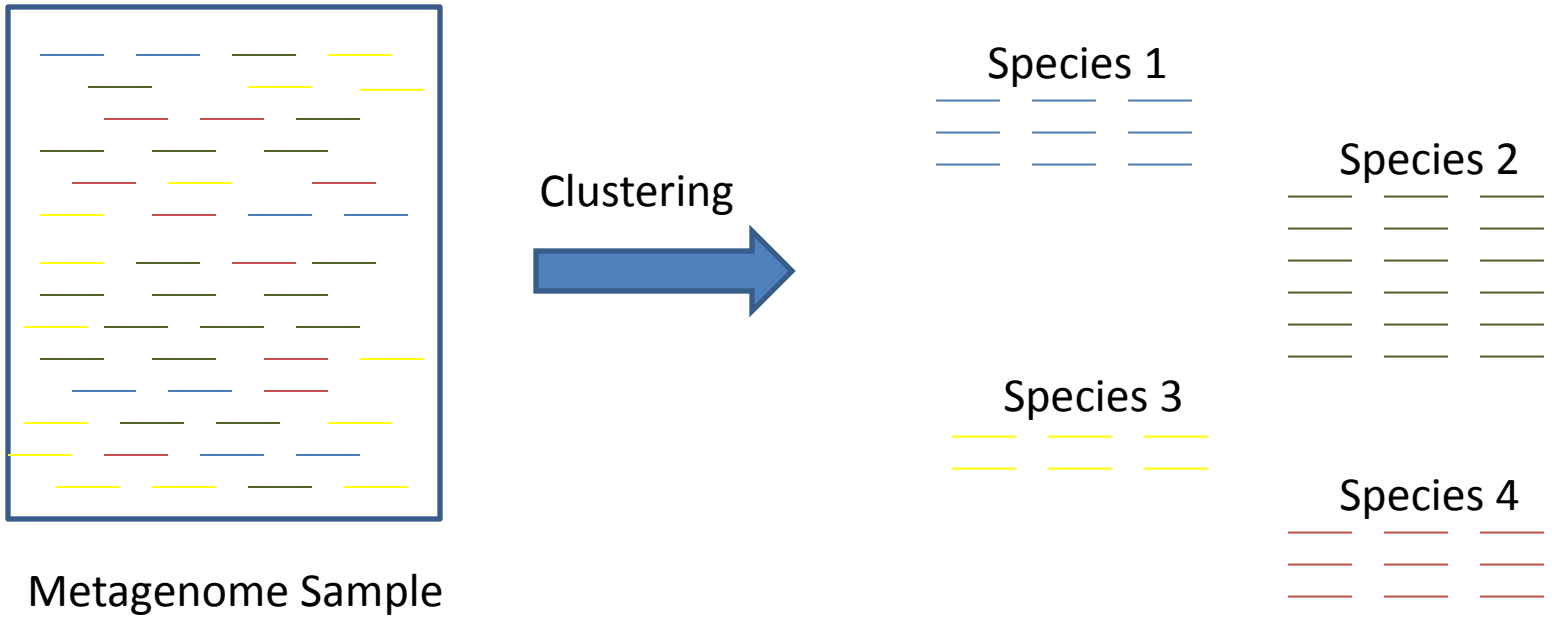
  - What are they doing?
  - Who does what? (difficult!)

  → Function-driven

- Metagenomes:
  - Acid-mine drainage, Hot springs, Termite Gut

Metagenomics

- Taxonomic characterization of metagenome
  - Clustering/Binning



Related Work

- Similarity/Comparative-based: Align to reference set
  - MEGAN: A representative example
  - Limitations:
    - Incomplete database
    - Fails to find homologs for new species

- Composition-based: GC content, kmer frequency, codon usage
  - CompostBin, TETRA, Phylopythia, Phymm, TACOA
  - Limitation:
    - Requires longer reads

BIOINFORMATICS

Vol. 24 ECCB 2008, pages i7–i13
doi:10.1093/bioinformatics/btn276

Annotation of metagenome short reads using proxygenes

Daniel Dalevi[1], Natalia N. Ivanova[2], Konstantinos Mavromatis[2], Sean D. Hooper[2], Ernest Szeto[1], Philip Hugenholtz[3], Nikos C. Kyrpides[2] and Victor M. Markowitz[1,*]

- Propose a clustering method based on protein-hits obtained by BLASTx of reads

> "High sequence similarity between the read and the proxygene implies phylogenetic proximity of the organisms from which the read and the proxygene have originated "

PRIB 2009, LNBI 5780, pp. 102–112, 2009.

Evidence-based Clustering of Reads and Taxonomic Analysis of Metagenomic Data

Gianluigi Folino[1], Fabio Gori[2], Mike S.M. Jetten[2], Elena Marchiori[2*]

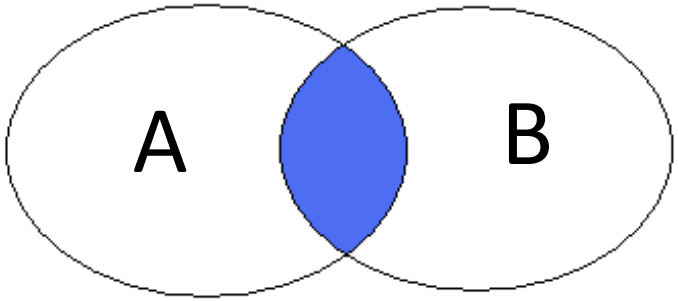
- Comparative clustering approach:
  1. Use BLASTx to associate a list of proteins with each read
  2. Cluster the reads using a method based on weighted proteins

- What happens when we have reads from novel species?

Why Soft clustering?

- Nature of metagenomic dataset:
  - Homologous sequences shared between species
    - Increased polymorphism, horizontal gene transfer
  - Incomplete and fragmentary nature of dataset

- Solution: Overlapping clusters
  - Capture homologous reads in soft boundaries
  - Allows for tolerance of errors in fragment matching



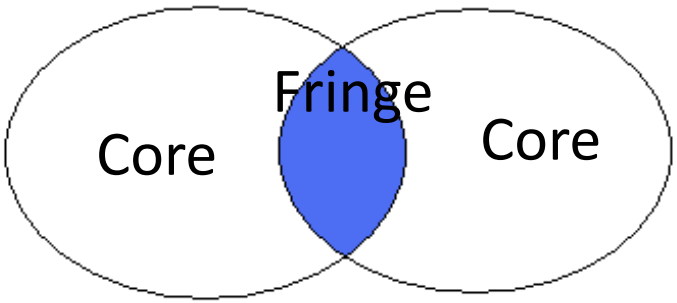
Chan et al.

An adaptive rough fuzzy single pass algorithm for clustering large data sets

S. Asharaf[a] , M. Narasimha Murty[b,*]

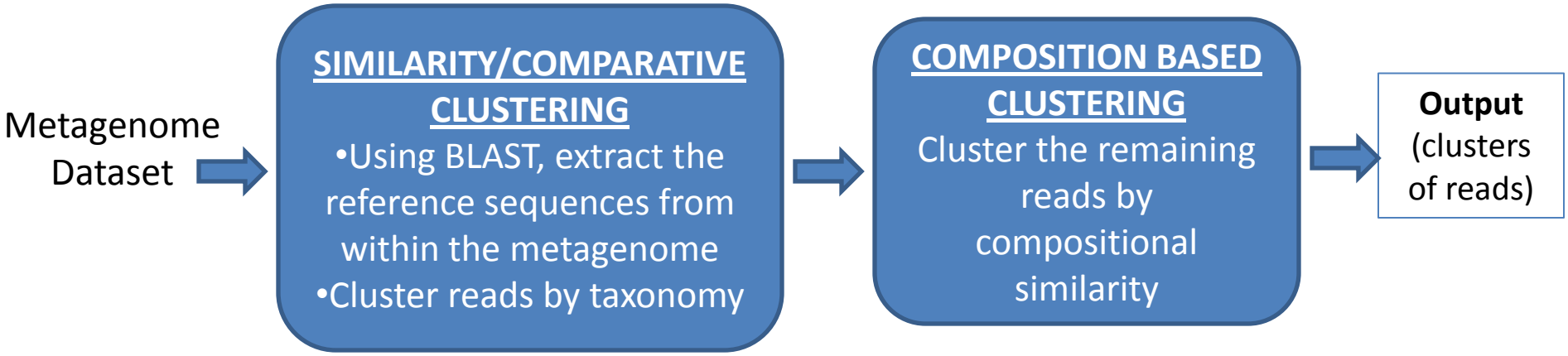
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

- Fuzzy adaptation of incremental leader clustering algorithm
- Each cluster has a core, a fringe and a leader



- Depending on the distance between the read and existing leaders:
  - Read is added to core of a cluster
  - Read is added to fringes of several clusters
  - Read gets elected as the leader

SIMCOMP


Metagenome Dataset
SIMILARITY/COMPARATIVE CLUSTERING
•Using BLAST, extract the reference sequences from within the metagenome
•Cluster reads by taxonomy
COMPOSITION BASED CLUSTERING
Cluster the remaining reads by compositional similarity
Output (clusters of reads)

Similarity/Comparative Clustering

1. BLASTx the reads against the NR protein database
2. Assign weights to all protein hits
3. Cluster fraction of reads with significant hits by taxonomy of proteins
4. Find read leader for each cluster

- **Output**: Similarity-based Clusters with read leaders

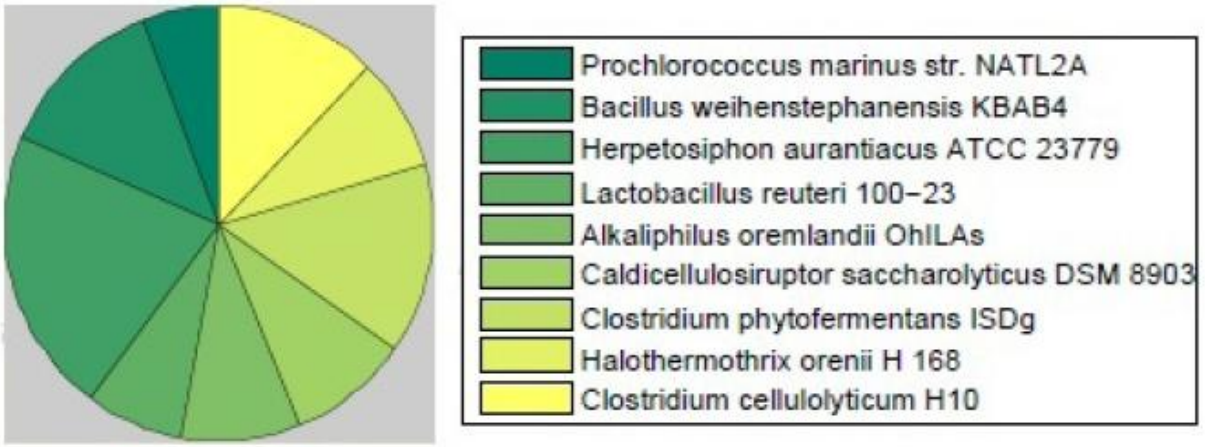
Composition based Clustering

For each remaining read, determine its Read Similarity (RS) with each existing read leader:

a. If RS > Core_Threshold, the read gets added to the core of cluster
b. Else if RS > Fringe_Threshold, the read gets added to the fringes of one or more clusters
c. Else, the read forms a new cluster with itself as the read leader

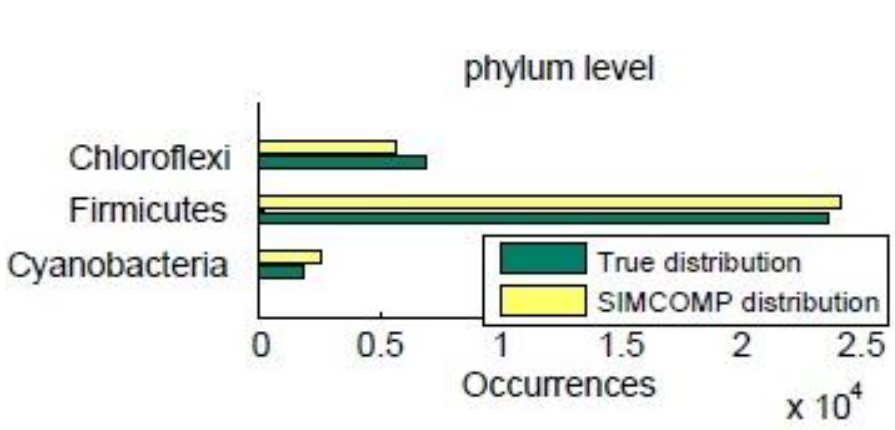
Dataset

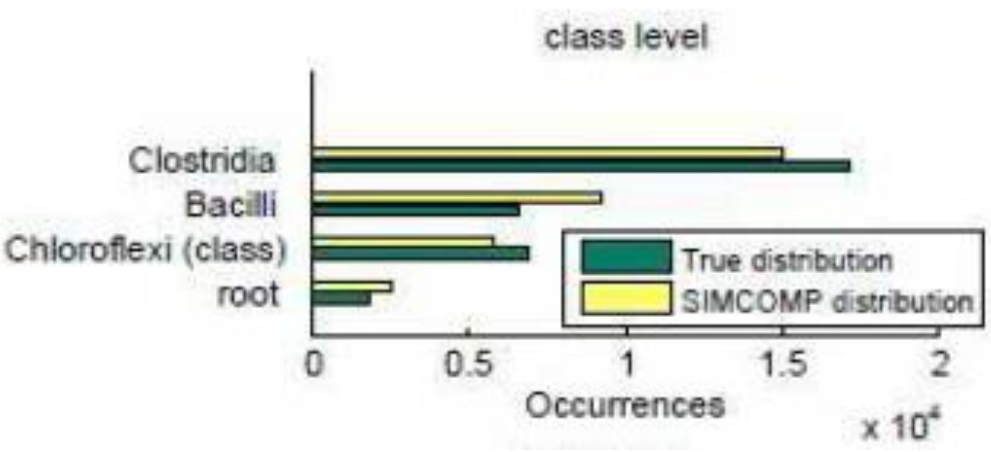
- Mavromatis et. al: (12 data sets)
  - A simulated data set : 454 reads (~100 bp) from 9 organisms at a coverage of 0.1X, 1X, 2X and 4X



Taxonomic distribution

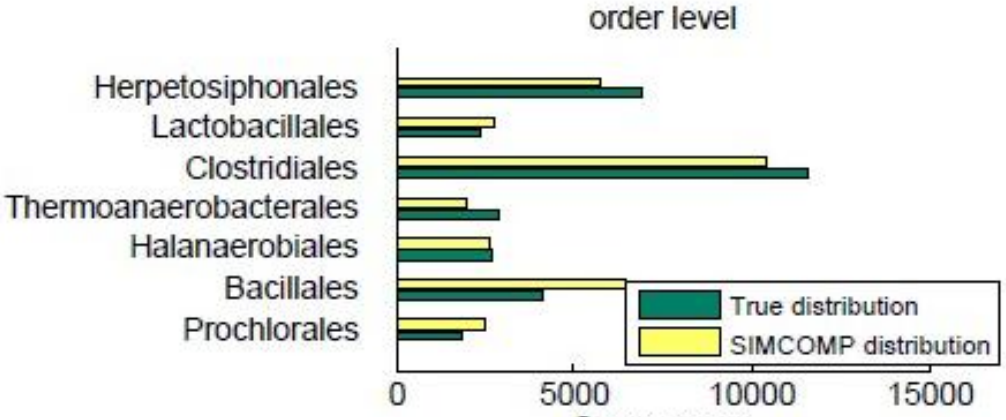
- Enriches the dataset into a small number of clusters
- Reads in a clusters are assigned the same taxa as that of the leader

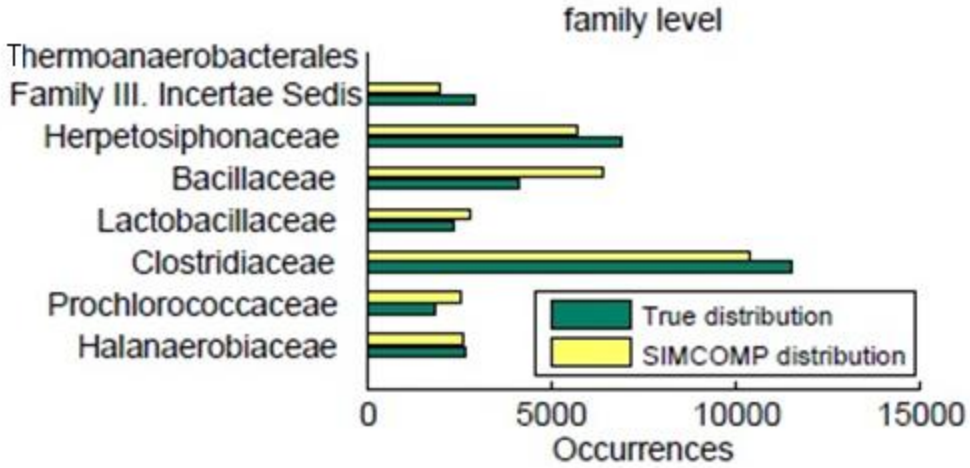




Core Threshold = 15, Fringe Threshold = 12, nmer = 6

Taxonomic distribution



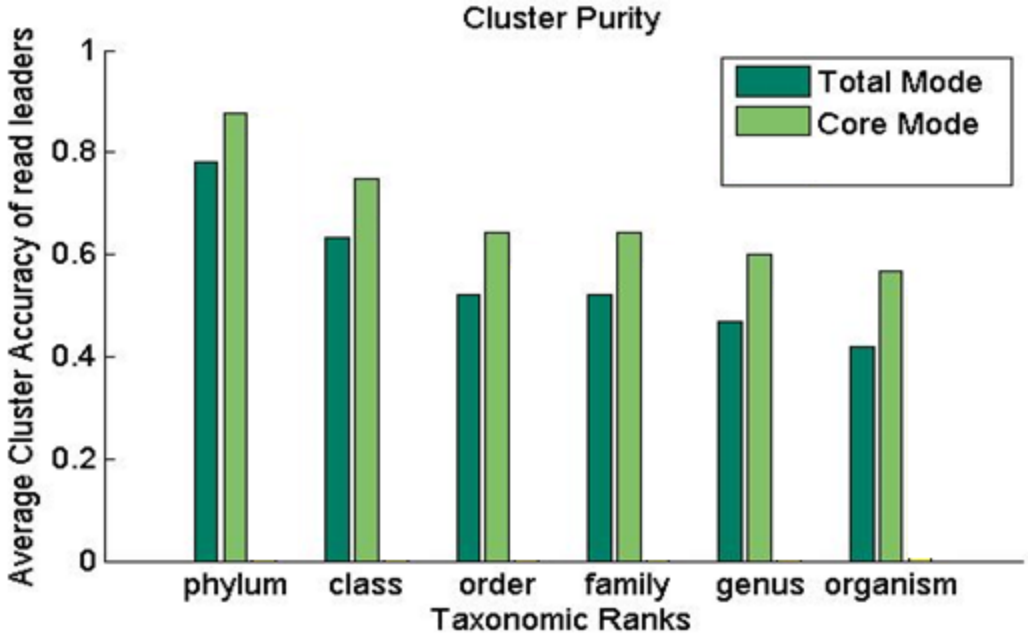


- Leaders can be used as an accurate estimate of the taxonomic distribution

Core Threshold = 15, Fringe Threshold = 12, nmer = 6

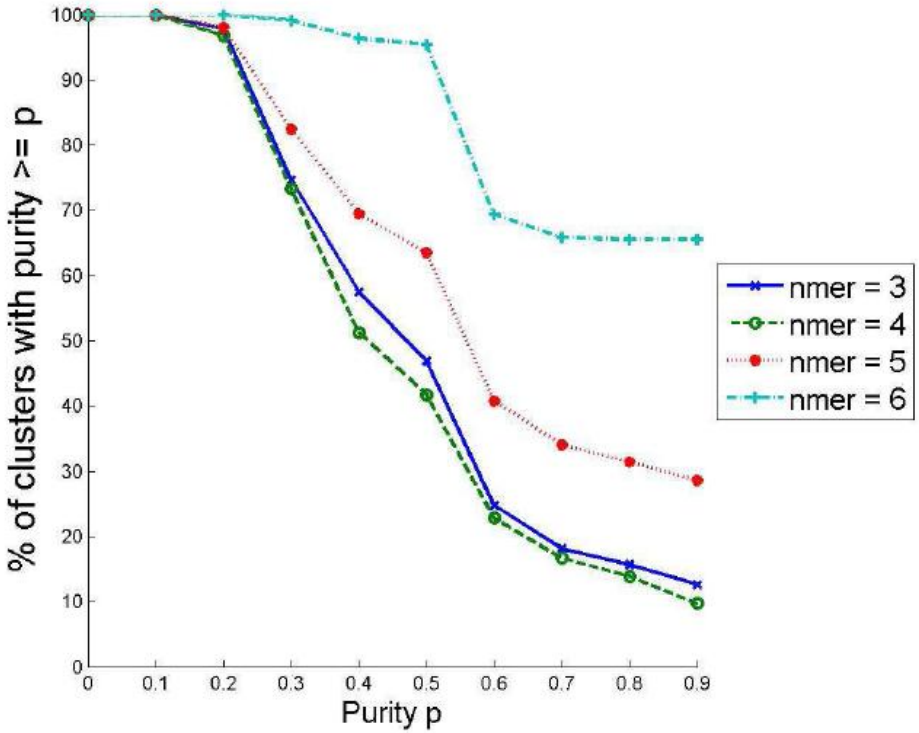
Cluster purity

- Homogeneity of the cluster: Fraction of reads within a set (cluster or core) belonging to the same taxon

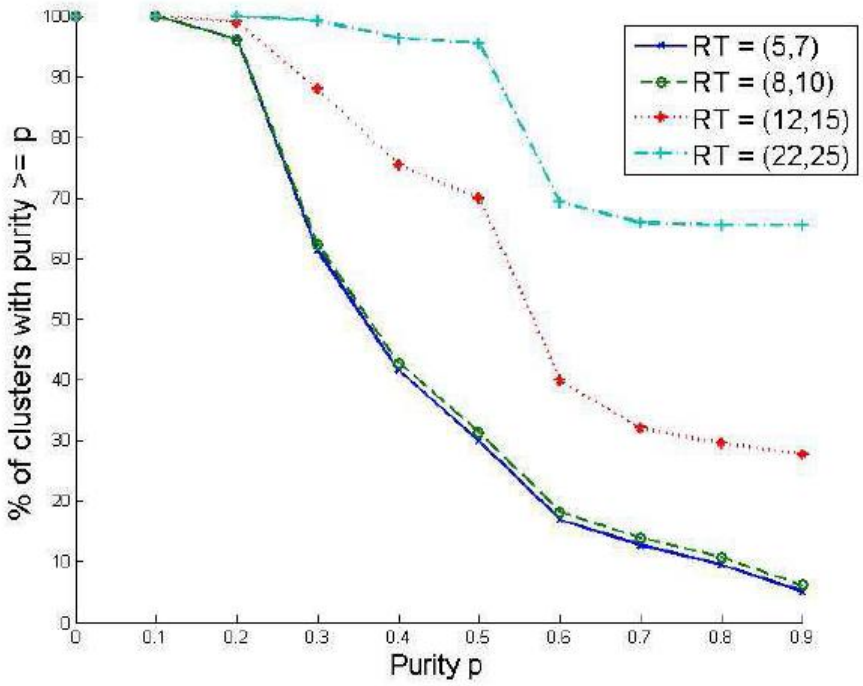


Core Threshold = 15, Fringe Threshold = 12, nmer = 6

Non-singleton clusters vs. purity



$RT_c = 25$, $RT_f = 22$



nmer = 6

- Hexamers have the best discriminatory power

Conclusion

- Semi-supervised, hybrid soft clustering algorithm
- Time consuming part: BLASTx of read
- Linear composition pass: avoids all-to-all comparisons
- Future Work:
  - Investigate the scope of soft boundaries of clusters
  - Use database other than NR
  - Mask the best BLAST hit to test for robustness

Thank you!
Questions?

Extra

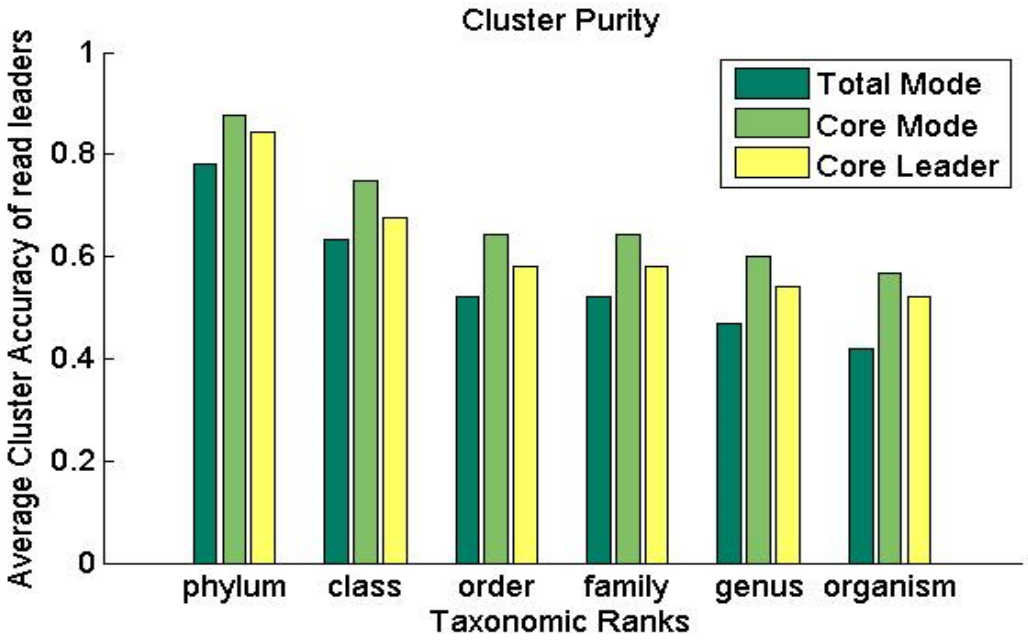
Assign weights to each protein

- For each read r, extract $h = (p; S_B; Id; E)$
- For each protein p, H_p be hits containing p
- Weight of the protein,

$$w_p = 1 + \lceil \frac{1}{|H_p|} \sum_{h \epsilon H_p} (100 \frac{max\_score - S_B(h)}{max\_score - min\_score} + 100 - Id(h)) \rceil$$

Cluster purity

- Homogeneity of the cluster: Fraction of reads within a set (cluster or core) belonging to the same taxa



Core Threshold = 15, Fringe Threshold = 12, nmer = 6