Clustering metagenomic contigs based on composition and coverage

Brynjar Smári Bjarnason

KTH – School of Computer Science and Communication

June 24, 2013

Code

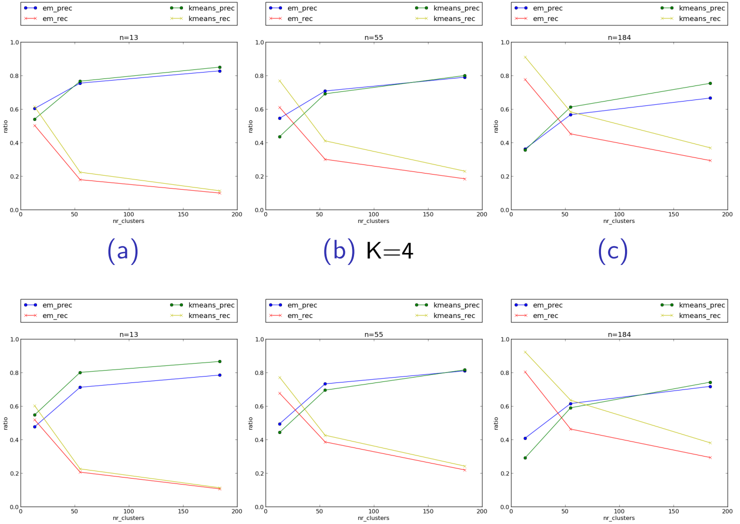
Go through code

Composition of in silico contigs from NCBI

Data:

1. 184 species from 55 genera and 13 families
2. ≈100 contigs from genera gave roughly 5.500 contigs
3. 100bp, 1.000bp, 10.000bp

Precision and recall, contig length 10.000, kmer 4 and 5



(a) (b) K=4 (c)

(d) (e) K=5 (f)

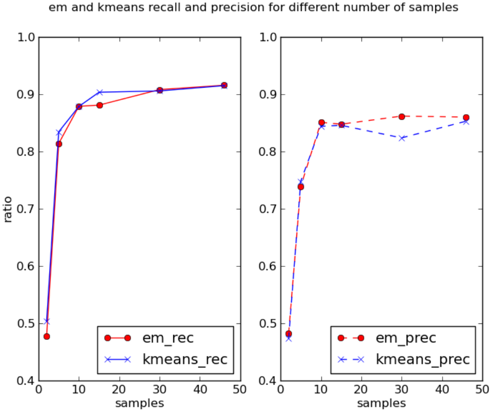
Composition of assembled Mock contigs

1. 50.000 contigs
2. 40 species
3. 4 kmer length

The results on species level: precision 0.616, recall 0.586

Coverage of in silico Mock timeseries

1. 50.000 contigs
2. 40 species
3. 2, 5, 10, 15, 30, 46 samples



What needs improvement

- One EM and Kmeans for all types of models
- Joint model for composition and coverage
- Memory and execution efficiency
- Estimate number of clusters
- Different models for composition