CAFE:

aCcelerated Alignment-FrEe sequence analysis

Name: Elahe Badali
Student No: 98209072

INTRODUCTION

A general look

- Problem Definition
- Alignment-Free Methods
- CAFE, A Good Solution

Problem Definition

- The dominant tools for sequence comparison are alignment-based methods, including global and local sequence alignments.

- But these tools have challenges in some cases:
    - Gene regulatory regions → not highly conserved
    - large amounts of short reads from NGS → assembling challenges
    - virus-host infectious associations

- Solution: Alignment-free sequence comparison

Alignment-free methods

- Several types of alignment-free approaches based on:
  - counts of k mers
  - longest common subsequences
  - shortest absent patterns
- This paper is based on k-mer counts.
- Example: d2* and d2S
- Advantages: perform well theoretically, solve 3 challenges of alignment-based methods.
- Disadvantages: relatively slow due to the requirement of calculating the expected k-mer counts.

CAFE, A Good Solution

- speeds up the calculation of recently developed measures.
- reduced memory requirement.

Features:

- Can work with:
  - assembled genomic sequences .
  - unassembled shotgun sequence reads from NGS technologies.
- Counts k-mers by JELLYFISH.
- The resulting pairwise dissimilarities: a symmetric matrix in PHYLIP format.
- four types of built-in downstream visualized analyses:
  - dendrograms using the UPGMA algorithm
  - heatmap visualization of the matrix
  - projecting the matrix to a 2D space using principal coordinate analysis (PCoA)
  - network display

MATERIALS

CAFE in detail

- 3 Type Of Methods
- Measures
- Workflows

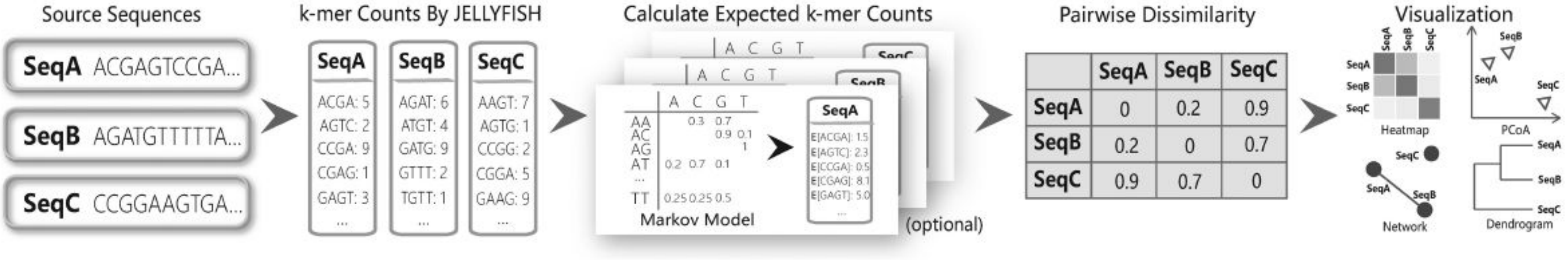
3 Type Of Methods

- Conventional k-mer counts: normalize k-mer counts into the k-mer frequencies.
- Adjusted k-mer counts: the Markov models for the sequences are assumed as the underlying generative models.
- presence/absence of k-mers: binarize k-mer counts into presence/absence indicators.

Measures

Conventional k-mer counts	**Adjusted k-mer counts**	**presence/absence of k-mers,**
1. Canberra	1. CVTree	1. Anderberg
2. Ch	2. d∗2	2. Antidice
3. Cosine	3. d2S	3. Dice
4. Co-phylog		4. Gower
5. D2		5. Hamman
6. Eu,		6. Hamming
7. FFP,		7. Jaccard
8. JS,		8. Kulczynski
9. Ma		9. Matching
10. Pearson		10. Ochiai
		11. Phi
		12. Russel
		13. Sneath
		14. Tanimoto
		15. Yule.

Workflows



	A	C	G	T
AA		0.3	0.7	
AC			0.9	0.1
AG				1
AT	0.2	0.7	0.1	
...				
TT	0.25	0.25	0.5	

	SeqA	SeqB	SeqC
SeqA	0	0.2	0.9
SeqB	0.2	0	0.7
SeqC	0.9	0.7	0

Figure 1. The workflow of CAFE. The JELLYFISH software parses the input sequence files (in Fasta format), counts *k*-mers and saves compressed information into separate databases. CAFE subsequently loads the databases and constructs a symmetric dissimilarity matrix among the inputs. CAFE also integrates four types of visualized downstream analysis, including dendrograms, heatmap, principal coordinate analysis (PCoA) and network display.

RESULTS

Evaluation in 3 Cases

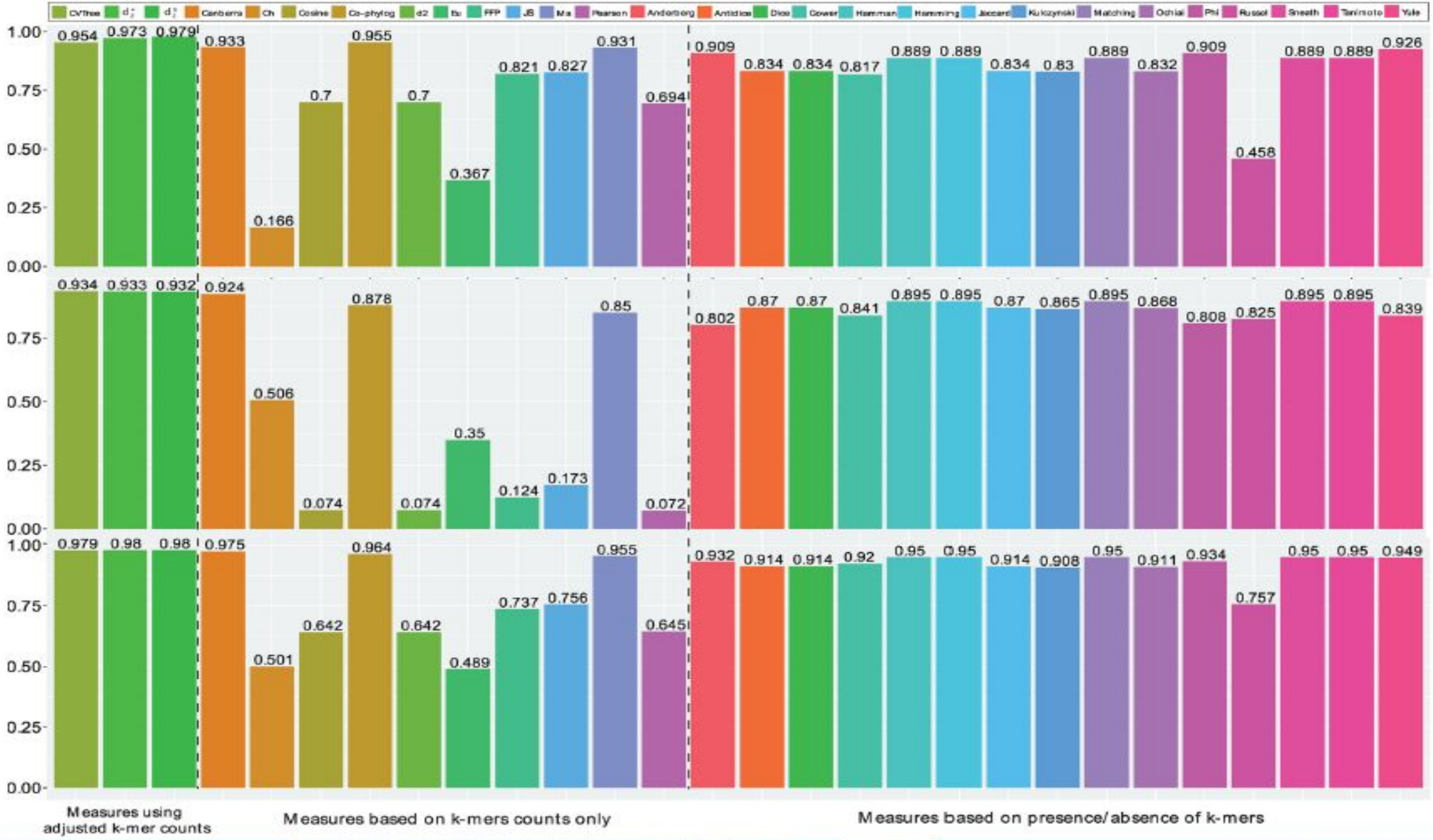
- Application to primate and vertebrate genomic sequences
- Application to microbial genomic sequences
- Application to metagenomic samples

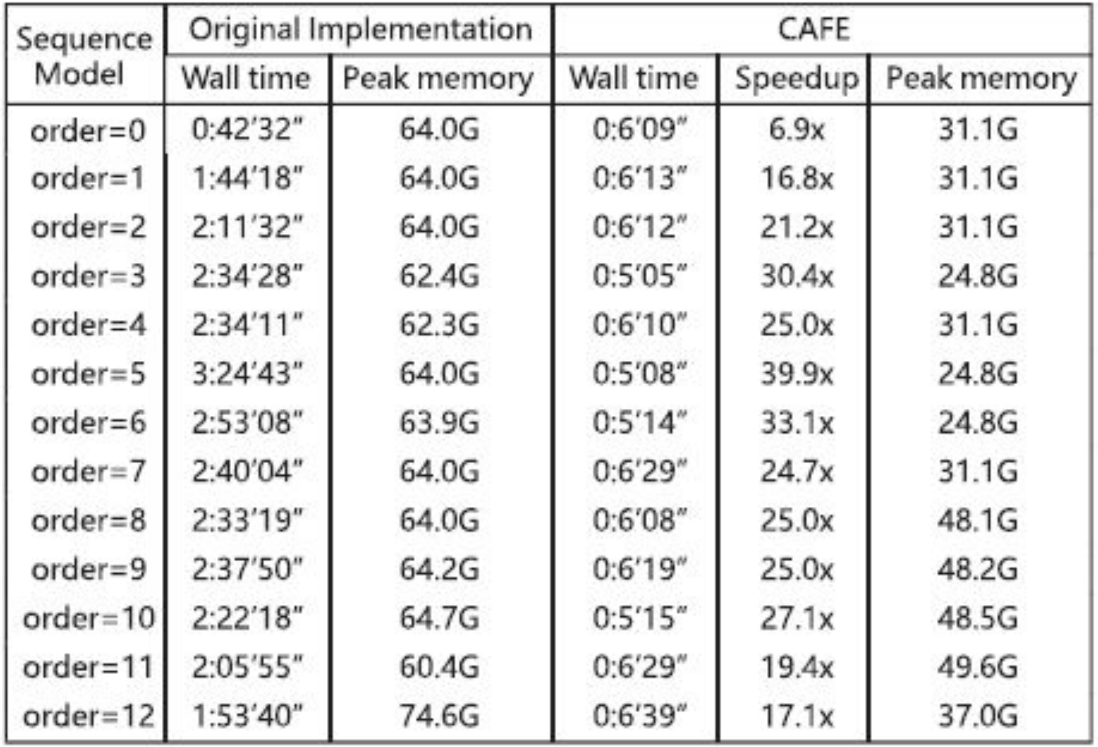
Application to microbial genomic sequences

1. Dataset 1: 21 primates species
2. Dataset 2: 28 vertebrate species
3. Dataset 3: Combined the two datasets

How to Evaluate:

The Spearman correlation of various dissimilarity measures with the evolutionary distances using maximum likelihood approach across many genomic regions based on above datasets.


CVTree
d_2^*
d_2^s
Canberra
Ch
Cosine
Co-phylog
d2
Eu
FFP
JS
Ma
Pearson
Anderberg
Antidice
Dice
Gower
Hamman
Hamming
Jaccard
Kulczynski
Matching
Ochiai
Phi
Russel
Sneath
Tanimoto
Yule
1.00
0.75
0.50
0.25
0.00
0.954 0.973 0.979 0.933 0.166 0.7 0.955 0.7 0.367 0.821 0.827 0.931 0.694
0.909 0.834 0.834 0.817 0.889 0.889 0.834 0.83 0.889 0.832 0.909 0.458 0.889 0.889 0.926
0.934 0.933 0.932 0.924 0.506 0.074 0.878 0.074 0.35 0.124 0.173 0.85 0.072
0.802 0.87 0.87 0.841 0.895 0.895 0.87 0.865 0.895 0.868 0.808 0.825 0.895 0.895 0.839
0.979 0.98 0.98 0.975 0.501 0.642 0.964 0.642 0.489 0.737 0.756 0.955 0.645
0.932 0.914 0.914 0.92 0.95 0.95 0.914 0.908 0.95 0.911 0.934 0.757 0.95 0.95 0.949
Measures using adjusted k-mer counts
Measures based on k-mers counts only
Measures based on presence/absence of k-mers



Sequence Model	Original Implementation		CAFE		
	Wall time	Peak memory	Wall time	Speedup	Peak memory
order=0	0:42′32″	64.0G	0:6′09″	6.9x	31.1G
order=1	1:44′18″	64.0G	0:6′13″	16.8x	31.1G
order=2	2:11′32″	64.0G	0:6′12″	21.2x	31.1G
order=3	2:34′28″	62.4G	0:5′05″	30.4x	24.8G
order=4	2:34′11″	62.3G	0:6′10″	25.0x	31.1G
order=5	3:24′43″	64.0G	0:5′08″	39.9x	24.8G
order=6	2:53′08″	63.9G	0:5′14″	33.1x	24.8G
order=7	2:40′04″	64.0G	0:6′29″	24.7x	31.1G
order=8	2:33′19″	64.0G	0:6′08″	25.0x	48.1G
order=9	2:37′50″	64.2G	0:6′19″	25.0x	48.2G
order=10	2:22′18″	64.7G	0:5′15″	27.1x	48.5G
order=11	2:05′55″	60.4G	0:6′29″	19.4x	49.6G
order=12	1:53′40″	74.6G	0:6′39″	17.1x	37.0G

Figure 4. Wall time, peak memory usage and speedup ratio comparison between CAFE and the original implementation to calculate d_2^* dissimilarity between a pair of genomes for $k = 14$.

Application to microbial genomic sequences

- Dataset: 27 E. coli and Shigella genomes dataset
- 6 E. coli reference (ECOR) groups: A, B1, B2,D, E and S.
- Used UPGMA to cluster the samples based on the calculated pairwise dissimilarity matrix
- The Markov order 1
- k = 14
- Result here based on CVTree, d2* and d2S

Result

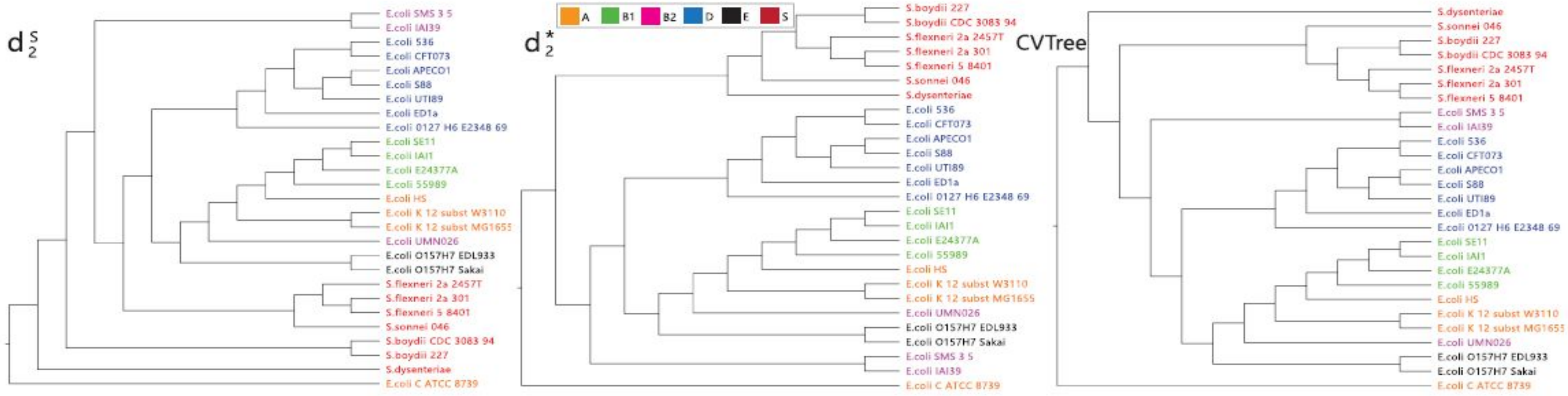


Figure 5. The clustering results of 27 *Escherichia coli* and *Shigella* genomes using measures based on background adjusted 14-mer counts: d_2^S, d_2^* and *CVTree*. The Markov order of the sequences were set at 1. The colors indicate the six different *E. Coli* reference groups.

Application to metagenomic samples

- Dataset: mammalian gut metagenomic dataset comprised of NGS short reads from 28 samples.
- 3 groups: 8 hindgutfermenting, herbivores, 13 foregut-fermenting herbivores 7 simple-gut carnivores.
- UPGMA to cluster
- TheMarkov order 0
- k = 5
- Result here based on CVTree, d2* and d2S

Result

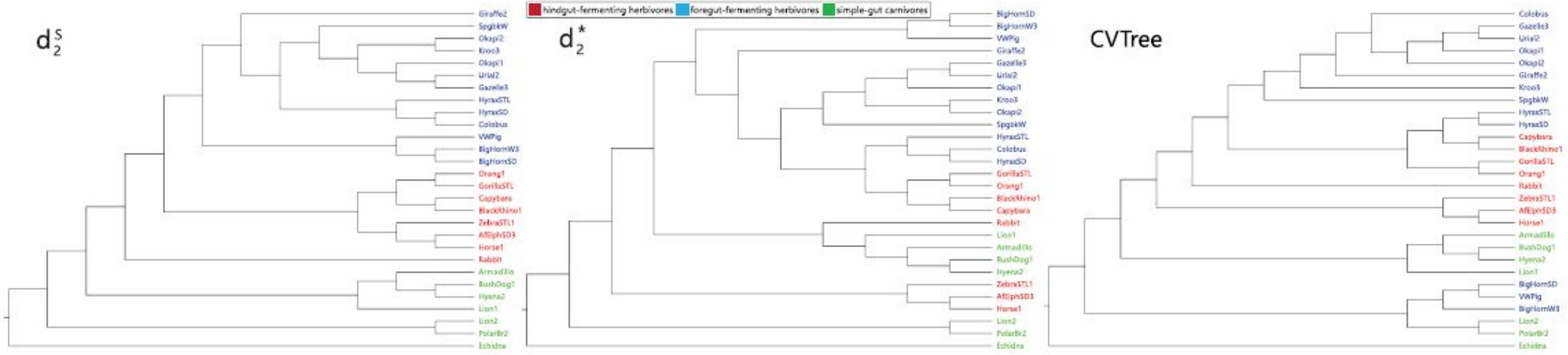


Figure 6. The clustering results of the mammalian gut samples using measures based on background adjusted k-mer counts: d_2^S, d_2^* and *CVTree*.

Other Articles

- This article Cited by 14 articles
- For Example:
  - **Alignment-Free Sequence Analysis and Applications**
    - an updated review
  - **Reads Binning Improves Alignment-Free Metagenome Comparison**
    - imperfection in d2S and d2* ⟹ neglect the heterogeneity
    - different reads bins
  - **Afann: Bias Adjustment for Alignment-Free Sequence Comparison Based on Sequencing Data Using Neural Network Regression**
    - Alignment-free methods ⟹ more time and memory efficient
    - However, dissimilarity can be overestimated
    - Afann ⟹ adjusts this bias

Other Works of Author

- **Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences**
  - Viruses and their host genomes often share similar oligonucleotide frequency (ONF) pattern
  - k=6 and d_2^*
- **Improving contig binning of metagenomic data using *dS*2 oligonucleotide frequency dissimilarity**
  - model contigs using relative *k*-tuple composition, followed by measuring dissimilarity between contigs using d2s
- .**VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data**
  - the first *k*-mer frequency based, machine learning method for virus contig identification that entirely avoids gene-based similarity searches.

Thank You For Your Attention