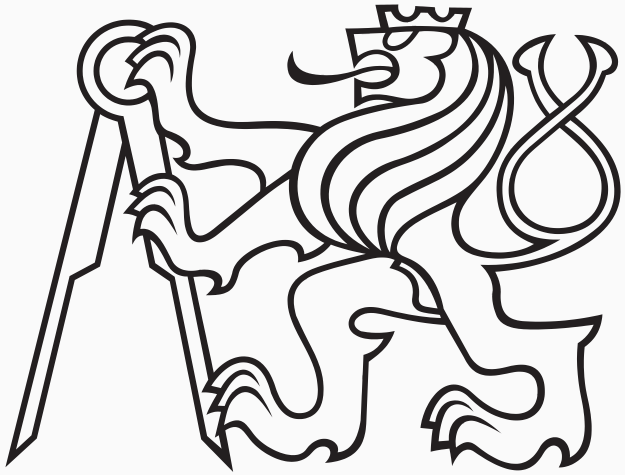
READS BAG SIMILARITY

Petr Ryšavý

August 14, 2016

IDA, Dept. of Computer Science, FEE, CTU



About myself

Name:	Petr Ryšavý
Date of birth:	July, 1991
Email:	rysavpe1@fel.cvut.cz
Affiliation:	IDA research group
	Department of Computer Science
	Faculty of Electrical Engineering
	Czech Technical University in Prague
	Prague, Czech Republic

Motivation

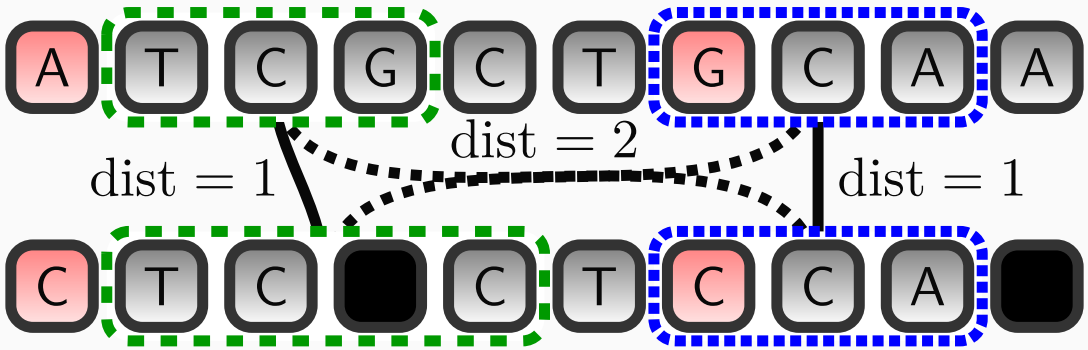
- Goal: cluster sequences given their read-set representation
- Classical approach:
  1. sequence assembly
  2. clustering
- Assembly is NP-hard, clustering is usually polynomial
- Idea: avoid the assembly step and cluster read sets directly

Basics

- Clustering algorithms like neighbor joining or UPGMA require as input only distance matrix
- Distance matrix on original sequences would be based on edit distance
- Main idea: approximate the edit distance based on read sets

Monge-Elkan distance

- Our approach is based on Monge-Elkan distance known from databases
- For each read from a read set we find the least distant read in the second read set



- Then we average over the reads pairs
- Further modifications so that the approach fits the motivation

Experimental results

- Our approach beats baseline method and first-assemble-then-cluster approach in:
  - Pearson's correlation coefficient between true distance and approximation
  - Fowlkes-Mallows index between expected tree and tree produced by a method
- Our approach is quadratic (vs. NP-hard assembly)
- Only α^2 slower than calculating the edit distance on the original sequences
- A lot of further research is needed

Thank you for your attention.

Time for questions!