CLUSTERING OF NEXT-GENERATION SEQUENCING DATA

Petr Ryšavý, supervised by Filip Železný

Thursday 25th April, 2019

IDA, Dept. of Computer Science, FEE, CTU

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

RESEARCH CENTER FOR INFORMATICS
rci.cvut.cz

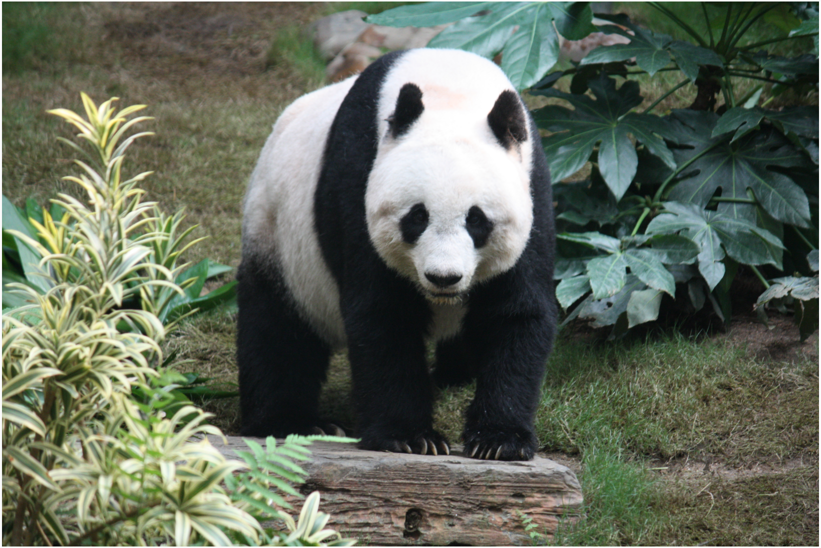
EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
OP Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

IDA
Intelligent Data Analysis
RESEARCH GROUP

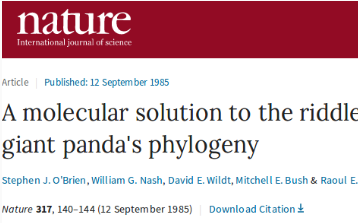
INTRODUCTION

Bear or raccoon?



[J. Patrick Fischer, CC BY-SA 3.0,
`https://commons.wikimedia.org/wiki/File:Grosser_Panda.JPG`]

Why evolutionary trees?



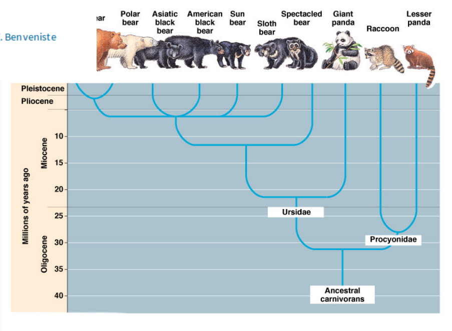
nature
International journal of science

Article | Published: 12 September 1985

A molecular solution to the riddle of the giant panda's phylogeny

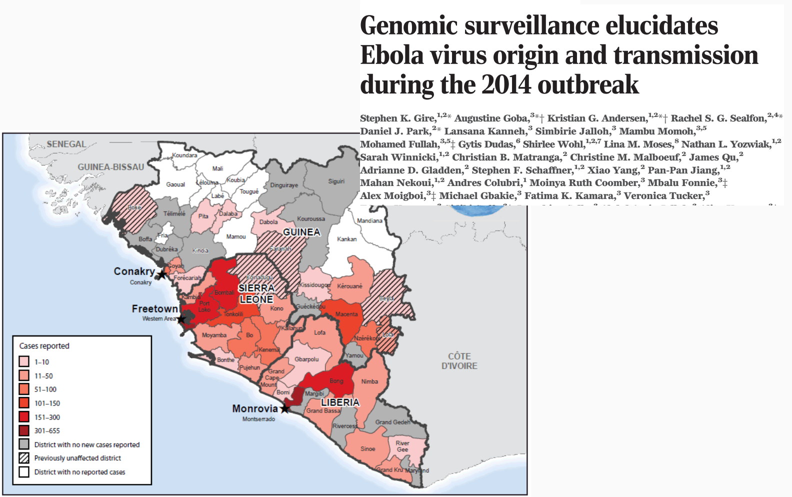
Stephen J. O'Brien, William G. Nash, David E. Wildt, Mitchell E. Bush & Raoul E. Benveniste

Nature **317**, 140–144 (12 September 1985) | Download Citation ⤓



[Reece, Jane B., et al. Campbell biology. No. s 1309. Boston: Pearson, 2014.]

Another example



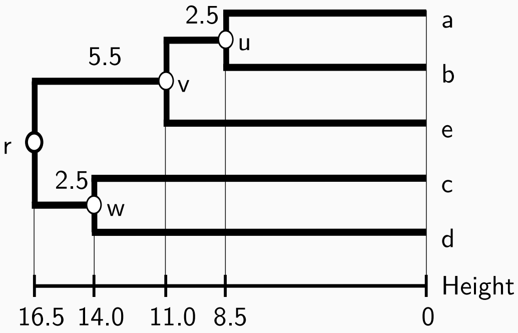
Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak

Stephen K. Gire,[1,2*] Augustine Goba,[3*†] Kristian G. Andersen,[1,2*†] Rachel S. G. Sealfon,[2,4*] Daniel J. Park,[2*] Lansana Kanneh,[3] Simbirie Jalloh,[3] Mambu Momoh,[3,5] Mohamed Fullah,[3,5‡] Gytis Dudas,[6] Shirlee Wohl,[1,2,7] Lina M. Moses,[8] Nathan L. Yozwiak,[1,2] Sarah Winnicki,[1,2] Christian B. Matranga,[2] Christine M. Malboeuf,[2] James Qu,[2] Adrianne D. Gladden,[2] Stephen F. Schaffner,[1,2] Xiao Yang,[2] Pan-Pan Jiang,[1,2] Mahan Nekoui,[1,2] Andres Colubri,[1] Moinya Ruth Coomber,[3] Mbalu Fonnie,[3‡] Alex Moigboi,[3‡] Michael Gbakie,[3] Fatima K. Kamara,[3] Veronica Tucker,[3]

[Nolen, Leisha et al. "Incidence of Hansen's Disease — United States, 1994–2011." MMWR. Morbidity and mortality weekly report (2014).]

Output - hierarchical clustering

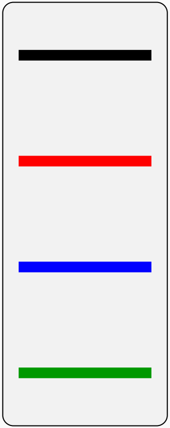
- Output is a dendogram of the species



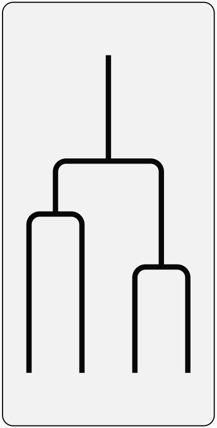
[By Manudouz (Own work) [CC BY-SA 4.0], via Wikimedia Commons]

Clustering algorithms

- The only input of hierarchical clustering algorithms is a distance matrix
- This includes UPGMA and neighbor-joining

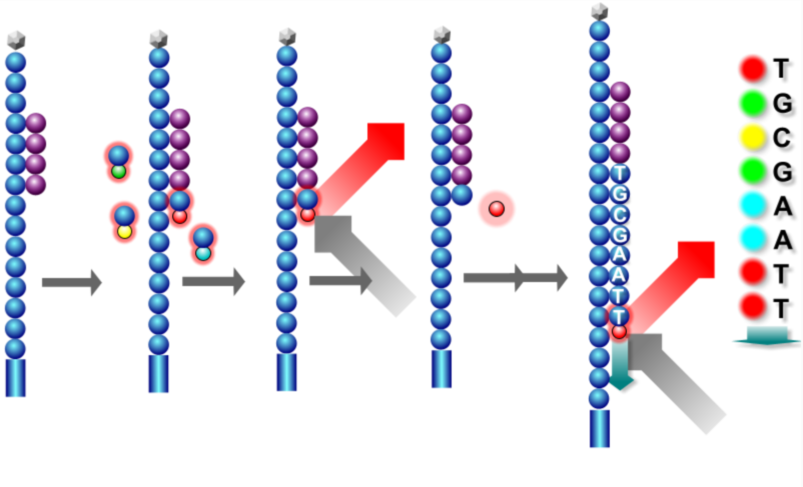


$$\Rightarrow \begin{pmatrix} 0 & 5 & 9 & 9 \\ 5 & 0 & 10 & 10 \\ 9 & 10 & 0 & 9 \\ 9 & 10 & 9 & 0 \end{pmatrix} \Rightarrow$$



That simple?

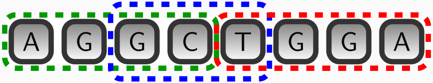
Sequencing by synthesis

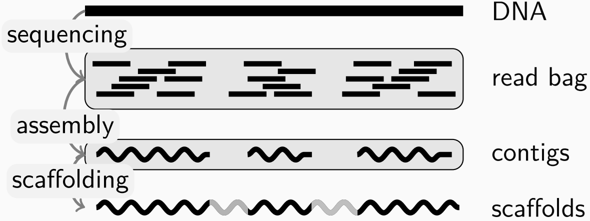


[By Abizar Lakdawalla, CC BY-SA 3.0, `https://en.wikipedia.org/wiki/File:Sequencing_by_synthesis_Reversible_terminators.png`]

Sequencing, read-sets

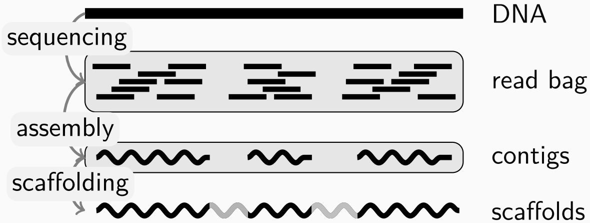
- Product of sequencing is not a long sequence, but short substrings called reads
- Reads have length of 10s to 100s of symbols
- Sequence AGGCTGGA is represented by set $\{\mathrm{AGGC}, \mathrm{TGGA}, \mathrm{GCT}\}$.





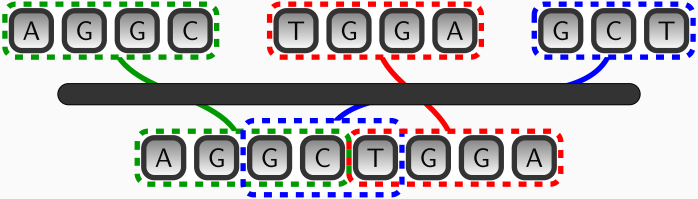
Contigs

- Assembly does not produce a single putative sequence, but several contigs
- Process of scaffolding and gap filling requires some additional wet-lab work
- Contigs are approximate substrings with unknown location and orientation



Classical approach - first assemble

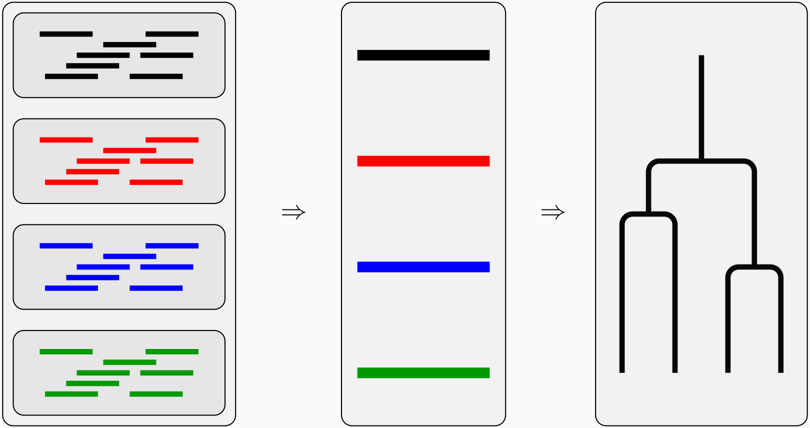
- Classical approach is to reconstruct the original sequence first



- Genome assembly
- NP-hard problem

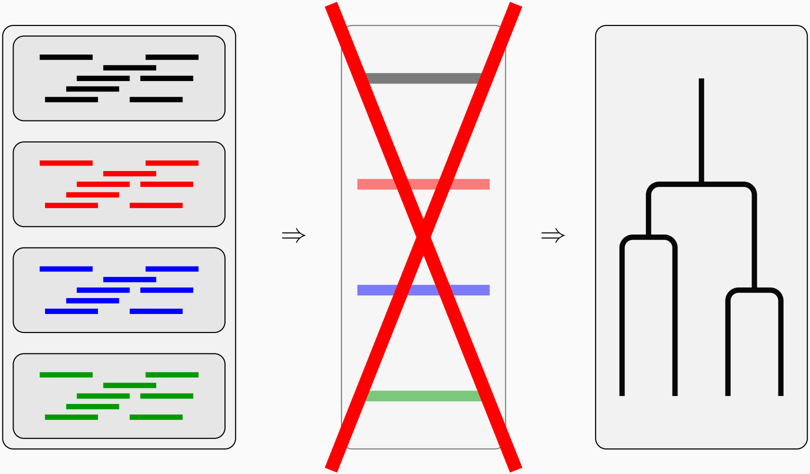
Classical approach - then cluster

- Hierarchical clustering algorithm is used to build a dendogram
- Dendogram is based on edit distance



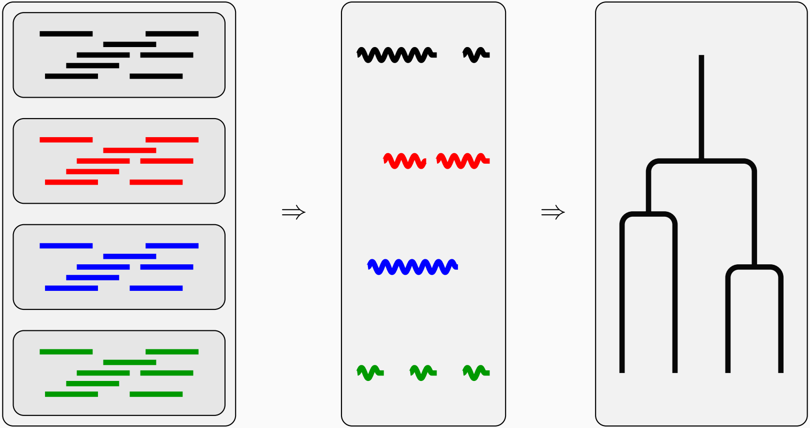
Our approach - skip assembly.

- Goal is to build dendrogram directly from the read sets



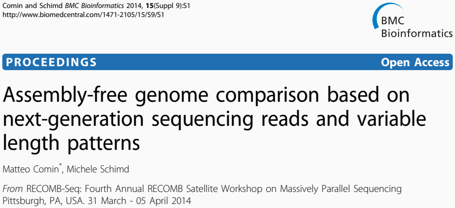
Our approach - skip assembly.

- Do not skip the assembly, do only the easy parts.



Alignment-free approaches

- Originally designed do avoid alignment step for genome comparison
- Genome broken into k-mers
- Some approaches work with read data



Comin and Schimd *BMC Bioinformatics* 2014, **15**(Suppl 9):S1
http://www.biomedcentral.com/1471-2105/15/S9/S1

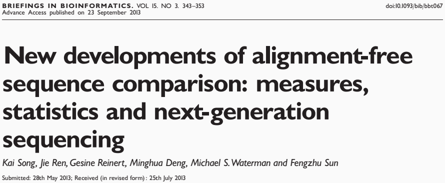
BMC Bioinformatics

PROCEEDINGS **Open Access**

Assembly-free genome comparison based on next-generation sequencing reads and variable length patterns

Matteo Comin*, Michele Schimd

From RECOMB-Seq: Fourth Annual RECOMB Satellite Workshop on Massively Parallel Sequencing
Pittsburgh, PA, USA. 31 March - 05 April 2014



BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 3. 343–353 doi:10.1093/bib/bbt067
Advance Access published on 23 September 2013

New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing

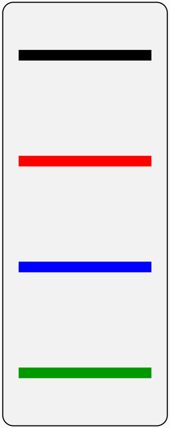
Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S. Waterman and Fengzhu Sun

Submitted: 28th May 2013; Received (in revised form): 25th July 2013

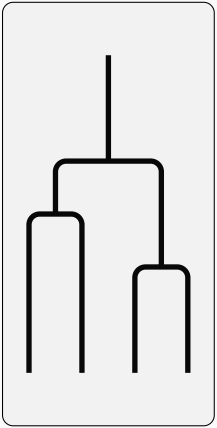
DISTANCE FUNCTION DESIGN

Clustering algorithms

- The only input of hierarchical clustering algorithms is a distance matrix
- This includes UPGMA and neighbor-joining

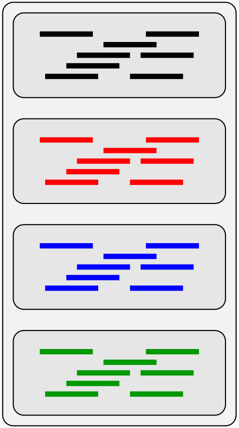


$$\Rightarrow \begin{pmatrix} 0 & 5 & 9 & 9 \\ 5 & 0 & 10 & 10 \\ 9 & 10 & 0 & 9 \\ 9 & 10 & 9 & 0 \end{pmatrix} \Rightarrow$$

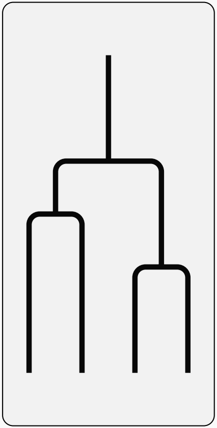


Key observation

- To build dendogram we need to approximate the distance matrix
- Measure that approximates edit distance needed



$$\Rightarrow \begin{pmatrix} 0 & 5 & 9 & 9 \\ 5 & 0 & 10 & 10 \\ 9 & 10 & 0 & 9 \\ 9 & 10 & 9 & 0 \end{pmatrix} \Rightarrow$$



Problem reformulation

- Approximate edit distance between two sequences from their read-set/contig-set representations

Assumptions:

- All reads have the same length l.
- Reads are sampled i.i.d. with replacement from the uniform distribution on all substrings of length l of the sequences.

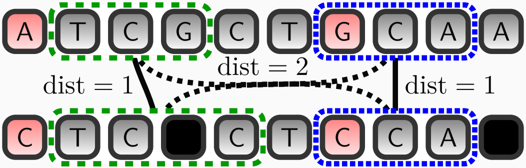
Key terms:

- Read length l.
- Coverage α.

USING READ-SETS

Monge-Elkan distance

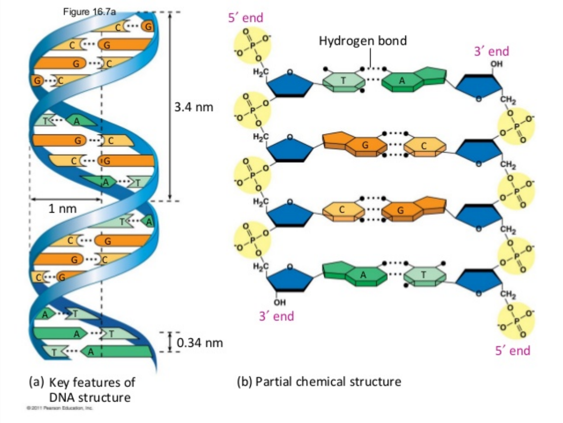
- Our approach is based on Monge-Elkan distance known from databases
- For each read from a read set we find the least distant read in the second read set



- Then we average over the read pairs

Strand and orientation

- In practical setting we do not know which strand do the reads come from.
- Sometimes we do not know whether a read starts on 5’-end.



(a) Key features of DNA structure

(b) Partial chemical structure

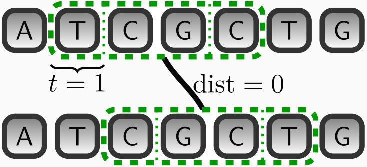
[https://www.slideshare.net/jenuerz/replication-transcription-translation2012]

Distance scale and symmetry

- Our measure should be symmetric
- Monge-Elkan distance has upper bound l
- Bring distance to proper scale

Margin gaps

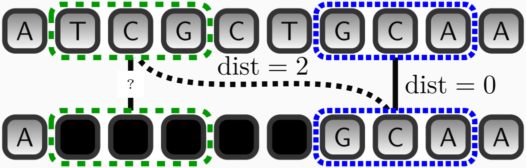
- Special treatment of leading and trailing gaps
- They may be caused by random positions of the reads



- Modification to edit distance

Threshold

- Read can match gaps in the sequence alignment
- If distance is an outlier, it is forced to be l



Too slow ? - sample

- Coverage α around 2 provides results that are good enough.
- For high coverage data downsample to $\alpha = 2$.

Too slow ? - use embedding

- We do not need exact minimum in Monge-Elkan distance.
- We use embedding to identify good candidates.
- q-gram profile is vector of counts of all possible q-grams, i.e. strings from Σ^q.
- q-gram distance of two strings is Manhattan distance of their q-gram profiles.
- Inspiration by BLAST and dictionary search, $q = 3$.
- We evaluate edit distance only on reads minimizing the q-gram distance.
- q-gram distance is LB on edit distance.

USING CONTIG-SETS

Three step procedure

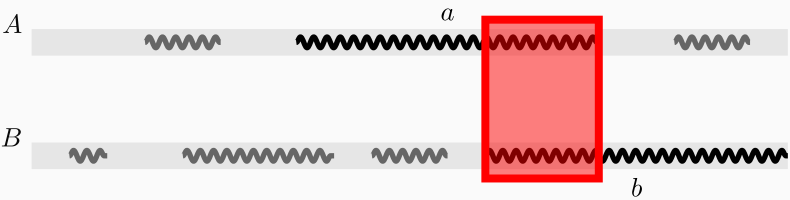
1. Calculate expected overlaps of contig pairs.
2. Select appropriate overlaps for each contig.
3. Average the distances over overlaps.

1) Estimating overlaps for contig pairs

- Consider two contigs a and b and assume they overlap in the optimal alignment
- Select overlap that minimizes the post-normalized edit distance

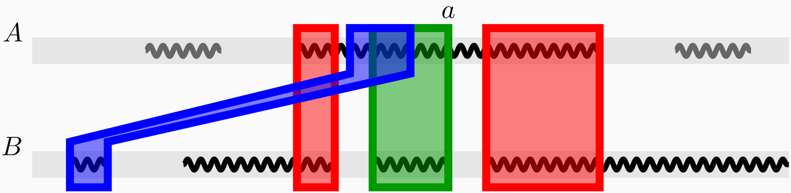
$$\overline{\mathsf{dist}}(a,b) = \frac{\mathsf{dist}(a,b)}{\max\{|a|,|b|\}}. \tag{1}$$

- Heuristic approach based on modification of Smith-Waterman algorithm



2) Estimating overlaps for contig sets

- For one contig we have overlaps with the other contig set
- Select non-overlapping regions that maximize the total value (post-normalized edit distance)
- Reduction to *weighted interval schedulling problem*

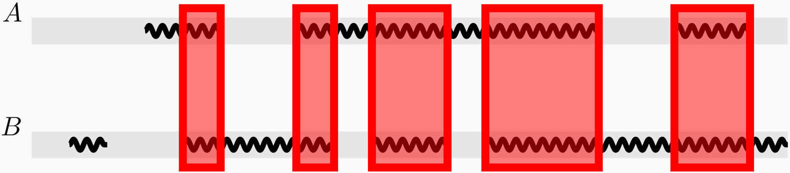


3) Combining the Results

- Sum distances of overlap pairs

$$d(C_A, C_B) = \sum_{(c,d) \in \mathsf{overlap}(C_A, C_B)} \mathsf{dist}(c, d).$$

- The sum does not capture contig size w.r.t. genome size

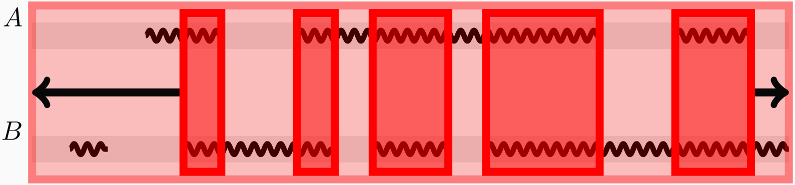


3) Combining the Results

- Normalize
- Divide by maximum possible distance of all overlaps ...
- ... and multiply by genome maximum distance

$$d(C_A, C_B) = \frac{\sum_{(c,d)\in \mathsf{overlap}(C_A,C_B)} \mathsf{dist}(c,d)}{\sum_{(c,d)\in \mathsf{overlap}(C_A,C_B)} \max\{|c|,|d|\}} \cdot \frac{l \max\{|R_A|, |R_B|\}}{\alpha}.$$

- The resulting measure is not symmetric ...



3) Combining the Results

- … average both directions

$$\mathsf{Dist}(C_A, C_B) = \frac{d(C_A, C_B) + d(C_B, C_A)}{2}$$

EXPERIMENTAL RESULTS

Experimental setup

- Two real-world and three artificial datasets
- Original DNA sequences used as a reference (if available)
- Two clustering algorithms (Neighbor-joining and UPGMA)
- Comparison with 5 common de novo assemblers (ABySS, edena, SSAKE, SPADes, velvet)

Measured characteristics:

- time (assembly time, distance matrix time, clustering time)
- Pearson's correlation coefficient measuring similarity of the distance matrix to the reference one
- Fowlkes-Mallows index measuring similarity of the clusterings
- Averaging over α and l values.

Results

- Pearson's correlation between distance matrices is close to one

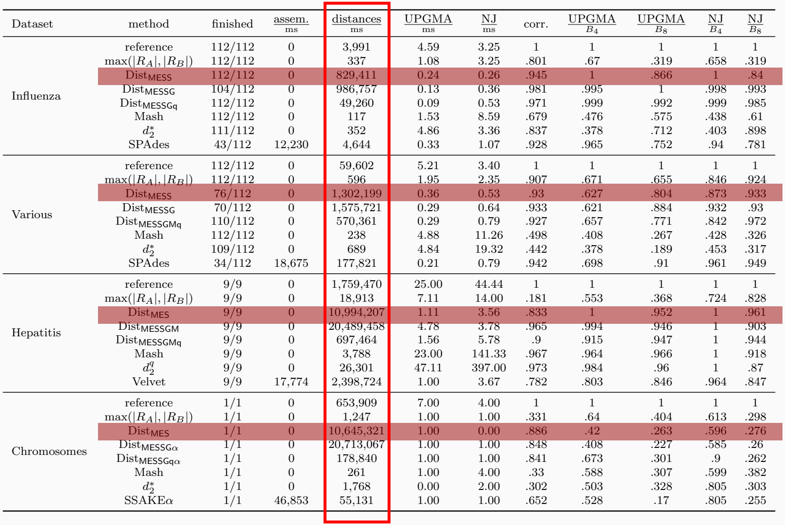
Table 4 Runtime, Pearson's correlation coefficient between distance matrices and Fowlkes-Mallows index for $k = 4$ and $k = 8$. The 'reference' method calculates distances from the original sequences. We show only assembly algorithm that gave the highest correlation, the best d-type method, and the better algorithm of pairs MES/MESS, MESSG/MESSGM, and MESSGq/MESSGMq.

Dataset	method	finished	assem.	distances	UPGMA	NJ	corr.	UPGMA	UPGMA	NJ	NJ
			ms	ms	ms	ms		B_4	B_8	B_4	B_8
Influenza	reference	112/112	0	3,991	4.59	3.25	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	112/112	0	337	1.08	3.25	.801	.67	.319	.658	.319
	$\text{Dist}_{\mathsf{MESS}}$	112/112	0	829,411	0.24	0.26	.945	1	.866	1	.84
	$\text{Dist}_{\mathsf{MESSG}}$	104/112	0	986,757	0.13	0.36	.981	.995	1	.998	.993
	$\text{Dist}_{\mathsf{MESSGq}}$	112/112	0	49,260	0.09	0.53	.971	.999	.992	.999	.985
	Mash	112/112	0	117	1.53	8.59	.679	.476	.575	.438	.61
	d_2^*	111/112	0	352	4.86	3.36	.837	.378	.712	.403	.898
	SPAdes	43/112	12,230	4,644	0.33	1.07	.928	.965	.752	.94	.781
Various	reference	112/112	0	59,602	5.21	3.40	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	112/112	0	596	1.95	2.35	.907	.671	.655	.846	.924
	$\text{Dist}_{\mathsf{MESS}}$	76/112	0	1,302,199	0.36	0.53	.93	.627	.804	.873	.933
	$\text{Dist}_{\mathsf{MESSG}}$	70/112	0	1,575,721	0.29	0.64	.933	.621	.884	.932	.93
	$\text{Dist}_{\mathsf{MESSGMq}}$	110/112	0	570,361	0.29	0.79	.927	.657	.771	.842	.972
	Mash	112/112	0	238	4.88	11.26	.498	.408	.267	.428	.326
	d_2^*	109/112	0	689	4.84	19.32	.442	.378	.189	.453	.317
	SPAdes	34/112	18,675	177,821	0.21	0.79	.942	.698	.91	.961	.949
Hepatitis	reference	9/9	0	1,759,470	25.00	44.44	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	9/9	0	18,913	7.11	14.00	.181	.553	.368	.724	.828
	$\text{Dist}_{\mathsf{MES}}$	9/9	0	10,994,207	1.11	3.56	.833	1	.952	1	.961
	$\text{Dist}_{\mathsf{MESSGM}}$	9/9	0	20,489,458	4.78	3.78	.965	.994	.946	1	.903
	$\text{Dist}_{\mathsf{MESSGMq}}$	9/9	0	697,464	1.56	5.78	.9	.915	.947	1	.944
	Mash	9/9	0	3,788	23.00	141.33	.967	.964	.966	1	.918
	d_2^q	9/9	0	26,301	47.11	397.00	.973	.984	.96	1	.87
	Velvet	9/9	17,774	2,398,724	1.00	3.67	.782	.803	.846	.964	.847
Chromosomes	reference	1/1	0	653,909	7.00	4.00	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	1/1	0	1,247	1.00	1.00	.331	.64	.404	.613	.298
	$\text{Dist}_{\mathsf{MES}}$	1/1	0	10,645,321	1.00	0.00	.886	.42	.263	.596	.276
	$\text{Dist}_{\mathsf{MESSG}\alpha}$	1/1	0	20,713,067	1.00	1.00	.848	.408	.227	.585	.26
	$\text{Dist}_{\mathsf{MESSGq}\alpha}$	1/1	0	178,840	1.00	1.00	.841	.673	.301	.9	.262
	Mash	1/1	0	261	1.00	4.00	.33	.588	.307	.599	.382
	d_2^*	1/1	0	1,768	0.00	2.00	.302	.503	.328	.805	.303
	SSAKEα	1/1	46,853	55,131	1.00	1.00	.652	.528	.17	.805	.255

Results

- Exact evaluation of Monge-Elkan distance is too slow for real-world

Table 4 Runtime, Pearson's correlation coefficient between distance matrices and Fowlkes-Mallows index for $k = 4$ and $k = 8$. The 'reference' method calculates distances from the original sequences. We show only assembly algorithm that gave the highest correlation, the best d-type method, and the better algorithm of pairs MES/MESS, MESSG/MESSGM, and MESSGq/MESSGMq.



Dataset	method	finished	assem. ms	distances ms	UPGMA ms	NJ ms	corr.	UPGMA B_4	UPGMA B_8	NJ B_4	NJ B_8
Influenza	reference	112/112	0	3,991	4.59	3.25	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	112/112	0	337	1.08	3.25	.801	.67	.319	.658	.319
	$\text{Dist}_{\text{MESS}}$	112/112	0	829,411	0.24	0.26	.945	1	.866	1	.84
	$\text{Dist}_{\text{MESSG}}$	104/112	0	986,757	0.13	0.36	.981	.995	1	.998	.993
	$\text{Dist}_{\text{MESSGq}}$	112/112	0	49,260	0.09	0.53	.971	.999	.992	.999	.985
	Mash	112/112	0	117	1.53	8.59	.679	.476	.575	.438	.61
	d_2^*	111/112	0	352	4.86	3.36	.837	.378	.712	.403	.898
	SPAdes	43/112	12,230	4,644	0.33	1.07	.928	.965	.752	.94	.781
Various	reference	112/112	0	59,602	5.21	3.40	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	112/112	0	596	1.95	2.35	.907	.671	.655	.846	.924
	$\text{Dist}_{\text{MESS}}$	76/112	0	1,302,199	0.36	0.53	.93	.627	.804	.873	.933
	$\text{Dist}_{\text{MESSG}}$	70/112	0	1,575,721	0.29	0.64	.933	.621	.884	.932	.93
	$\text{Dist}_{\text{MESSGMq}}$	110/112	0	570,361	0.29	0.79	.927	.657	.771	.842	.972
	Mash	112/112	0	238	4.88	11.26	.498	.408	.267	.428	.326
	d_2^*	109/112	0	689	4.84	19.32	.442	.378	.189	.453	.317
	SPAdes	34/112	18,675	177,821	0.21	0.79	.942	.698	.91	.961	.949
Hepatitis	reference	9/9	0	1,759,470	25.00	44.44	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	9/9	0	18,913	7.11	14.00	.181	.553	.368	.724	.828
	Dist_{MES}	9/9	0	10,994,207	1.11	3.56	.833	1	.952	1	.961
	$\text{Dist}_{\text{MESSGM}}$	9/9	0	20,489,458	4.78	3.78	.965	.994	.946	1	.903
	$\text{Dist}_{\text{MESSGMq}}$	9/9	0	697,464	1.56	5.78	.9	.915	.947	1	.944
	Mash	9/9	0	3,788	23.00	141.33	.967	.964	.966	1	.918
	d_2^q	9/9	0	26,301	47.11	397.00	.973	.984	.96	1	.87
	Velvet	9/9	17,774	2,398,724	1.00	3.67	.782	.803	.846	.964	.847
Chromosomes	reference	1/1	0	653,909	7.00	4.00	1	1	1	1	1
	$\max(\lvert R_A\rvert, \lvert R_B\rvert)$	1/1	0	1,247	1.00	1.00	.331	.64	.404	.613	.298
	Dist_{MES}	1/1	0	10,645,321	1.00	0.00	.886	.42	.263	.596	.276
	$\text{Dist}_{\text{MESSG}\alpha}$	1/1	0	20,713,067	1.00	1.00	.848	.408	.227	.585	.26
	$\text{Dist}_{\text{MESSGq}\alpha}$	1/1	0	178,840	1.00	1.00	.841	.673	.301	.9	.262
	Mash	1/1	0	261	1.00	4.00	.33	.588	.307	.599	.382
	d_2^*	1/1	0	1,768	0.00	2.00	.302	.503	.328	.805	.303
	SSAKEα	1/1	46,853	55,131	1.00	1.00	.652	.528	.17	.805	.255

Results

- Embedding and scaling puts runtime between assembly and alignment-free approaches

Table 1 Runtime on "E. coli" dataset. Assembly time (without distance matrix calculation) on the same dataset is 18,844 s (ABySS), 18,606 s (Edena), 33,545 s (SPAdes), and 17,701 s (Velvet).

Method	Time (in seconds)
$\text{Dist}_{\mathsf{MESSG(M)q}\alpha}$	11,073
co-phylog	583
Mash	480
d_2	3,221
d_2^*	3,235
d_2^q	3,228
d_2^{q*}	3,225
D_2	3,235
D_2^*	3,301
D_2^q	3,224
D_2^{q*}	3,227

Results

- Our approach requires lower coverage than assembly

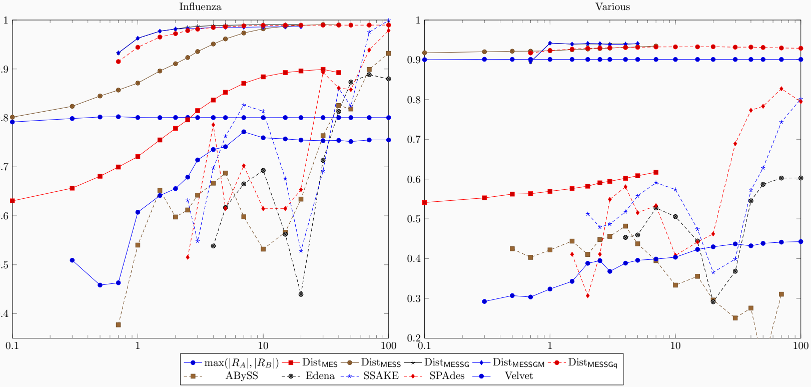


Figure 2: Plot of average Pearson's correlation coefficient for several choices of coverage values.

Results

- Our approach works better for short reads than assembly

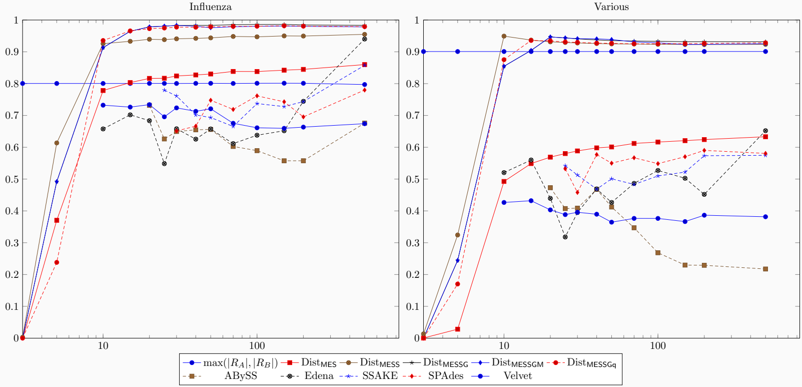


Figure 3: Plot of average Pearson's correlation coefficient for several choices of read length.

Conclusion

- We have seen two methods for estimating sequence similarity form read/contig sets
- Only single approximation step
- Adapts advantages of both alignment-free approaches and alignment similarity
- Further work needed

Thank you for your attention.

Time for questions!

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

RESEARCH CENTER FOR INFORMATICS
rci.cvut.cz

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
OP Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

IDA
Intelligent Data Analysis
RESEARCH GROUP