EFFICIENT GENOME SIMILARITY ESTIMATION FOR LEARNING FROM SEQUENCING DATA

Petr Ryšavý, supervised by Filip Železný

Tuesday 3rd October, 2023

IDA, Dept. of Computer Science, FEE, CTU

INTRODUCTION

Topics of this talk

Estimating Sequence Similarity from Read Sets for Clustering Sequencing Data

Petr Ryšavý & Filip Železný

First Online: 21 September 2016

Citations

lotes in Computer Science book series (LNISA,volume 9897)

Published: 04 August 2018

Estimating sequence similarity from read sets for clustering next-generation sequencing data

Petr Ryšavý & Filip Železný

owledge Discovery **33**, 1–23 (2019) | Cite this article

Estimating Sequence Similarity from Contig Sets

Petr Ryšavý & Filip Železný

Conference paper | First Online: 04 October 2017

938 Accesses | **1** Citations

Part of the Lecture Notes in Computer Science book series (LNISA,volume 1

An Algorithm to Calculate the p-value of the Monge-Elkan Distance *

Petr Ryšavý[0000−0002−6597−6616] and Filip Železný[0000−0001−9780−3376]

Department of Computer Science,
lectrical Engineering, Czech Technical University in Prague,
Prague, Czech Republic
{petr.rysavy,zelezny}@fel.cvut.cz

Research | Open Access | Published: 27 March 2023

Reference-free phylogeny from sequencing data

Petr Ryšavý & Filip Železný

BioData Mining **16**, Article number: 13 (2023) | Cite this article

1108 Accesses | **5** Altmetric | Metrics

Research | Open Access | Published: 27 September 2022

circGPA: circRNA functional annotation based on probability-generating functions

Petr Ryšavý, Jiří Kléma & Michaela Dostálová Merkerová

BMC Bioinformatics **23**, Article number: 392 (2022) | Cite this article

1101 Accesses | **1** Citations | **4** Altmetric | Metrics

Why should we care about genome similarity?

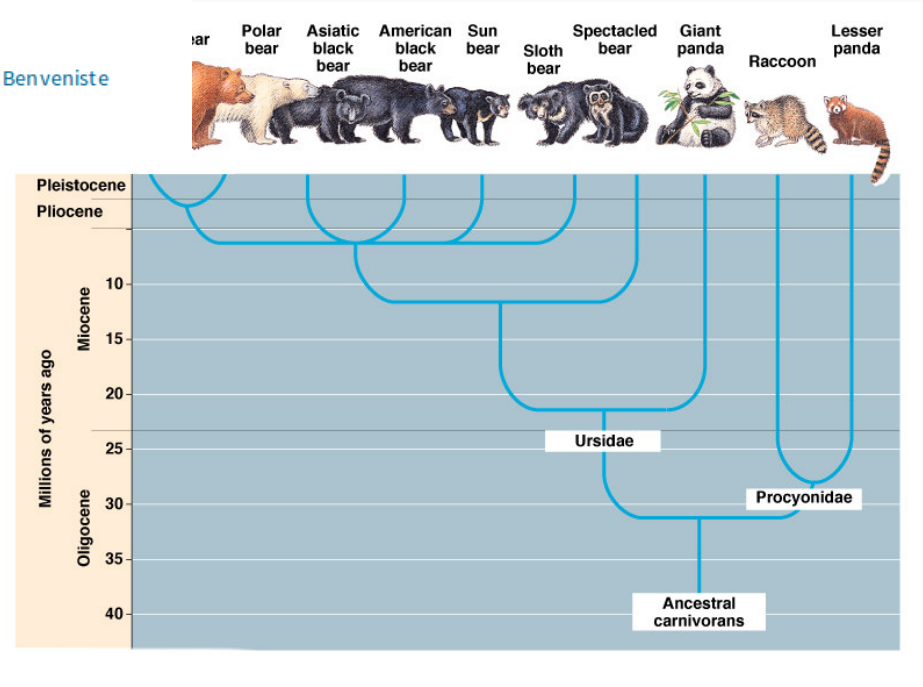
nature
International journal of science

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

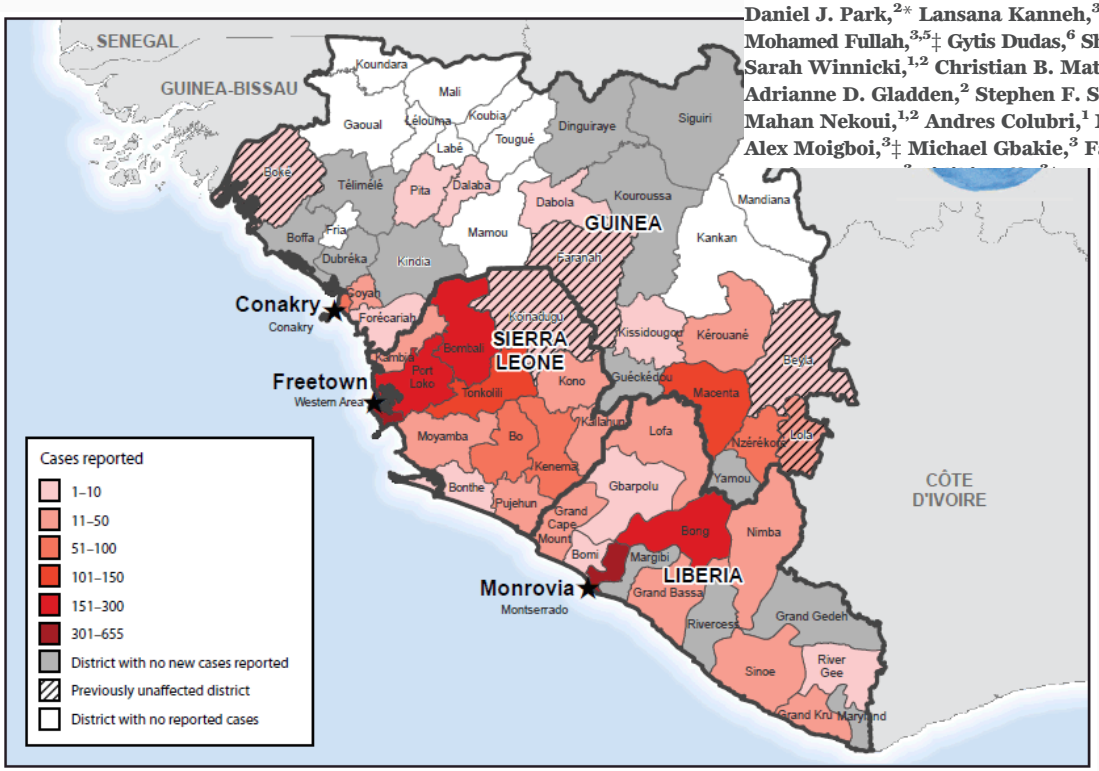


[Reece, Jane B., et al. Campbell biology. No. s 1309. Boston: Pearson, 2014.]

Another example

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

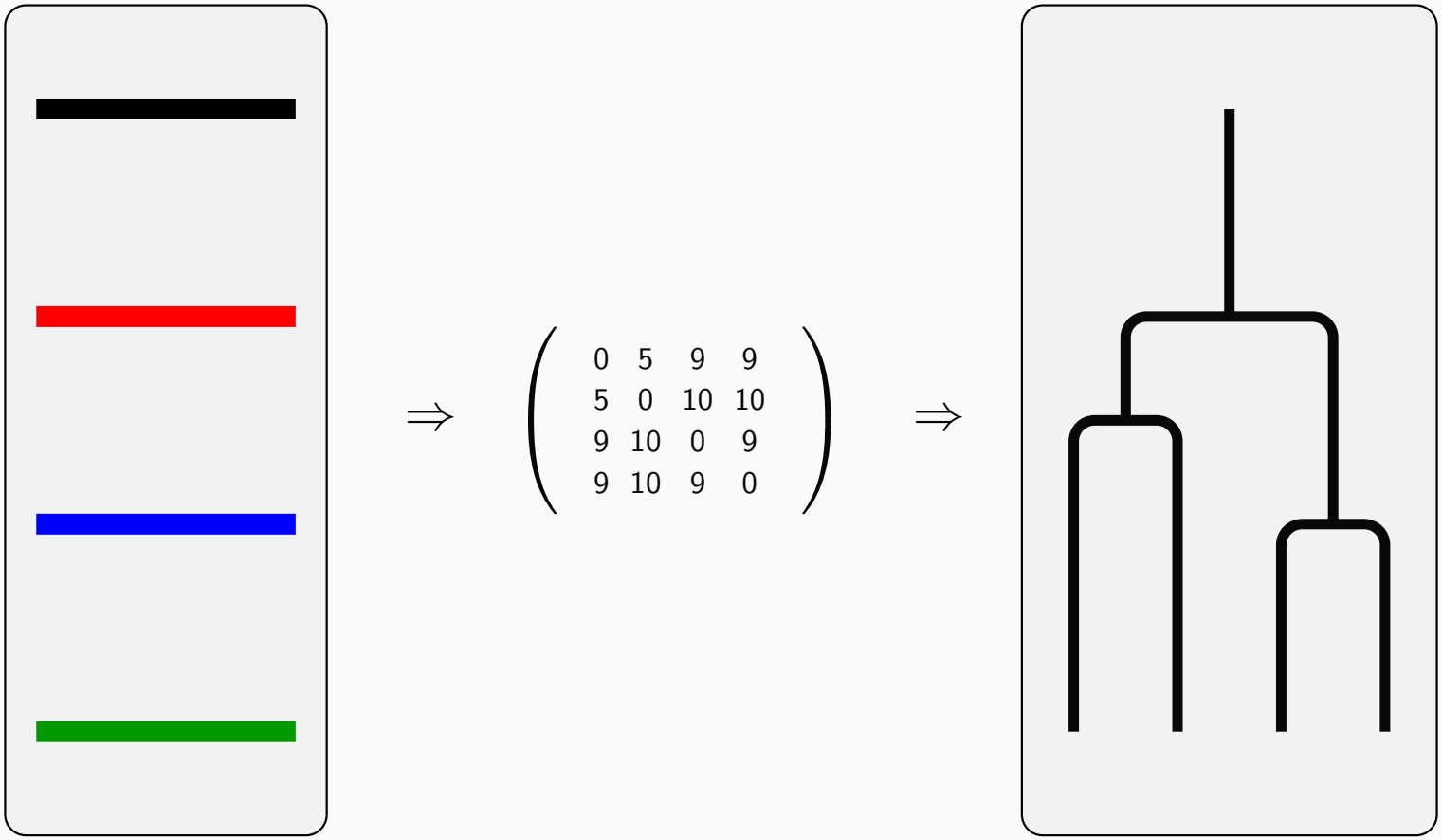
Stephen K. Gire,[1,2*] Augustine Goba,[3*†] Kristian G. Andersen,[1,2*†] Rachel S. G. Sealfon,[2,4*] Daniel J. Park,[2*] Lansana Kanneh,[3] Simbirie Jalloh,[3] Mambu Momoh,[3,5] Mohamed Fullah,[3,5‡] Gytis Dudas,[6] Shirlee Wohl,[1,2,7] Lina M. Moses,[8] Nathan L. Yozwiak,[1,2] Sarah Winnicki,[1,2] Christian 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).]

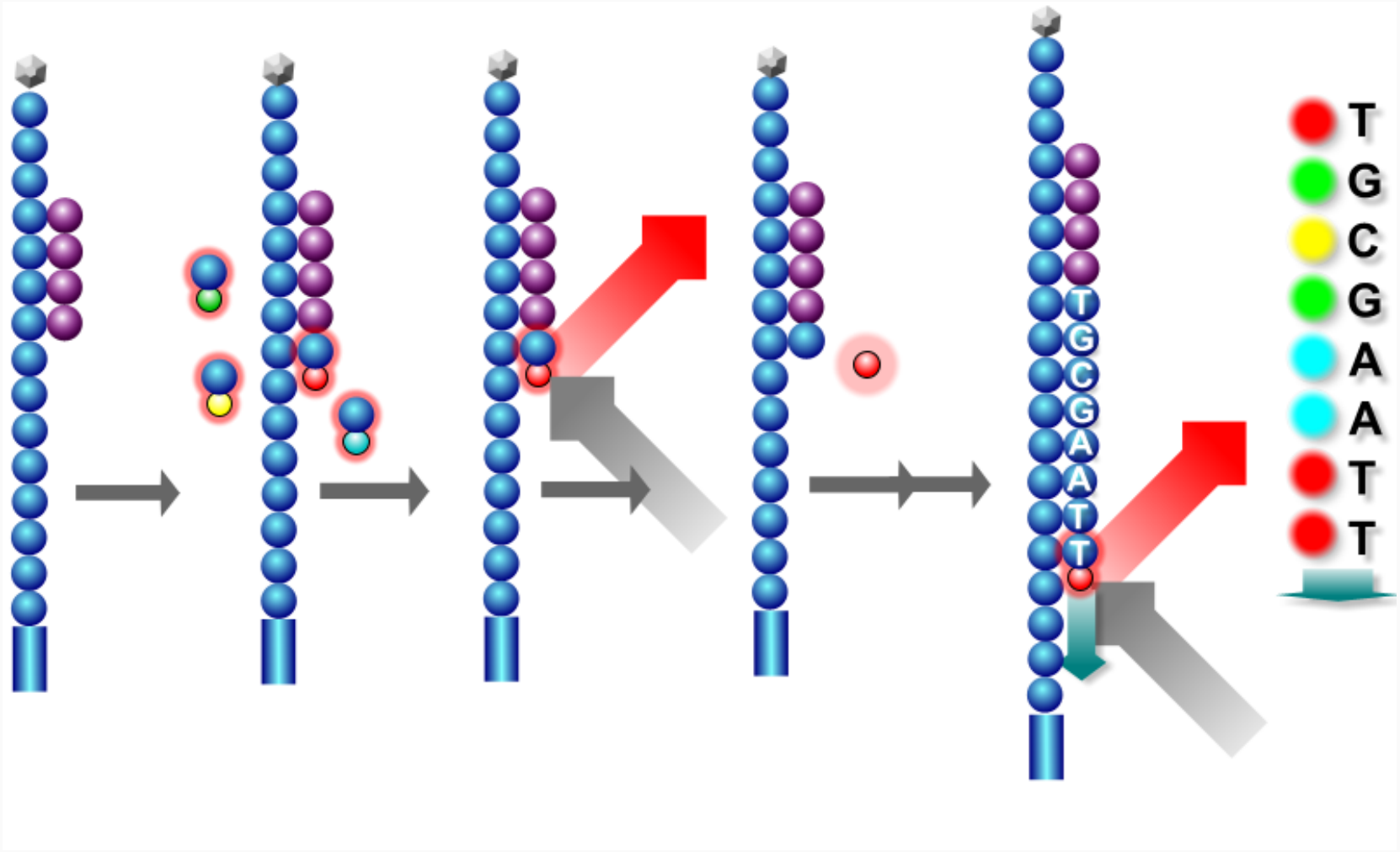
Hierarchical clustering algorithms

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



BACKGROUND - IS IT THAT SIMPLE?

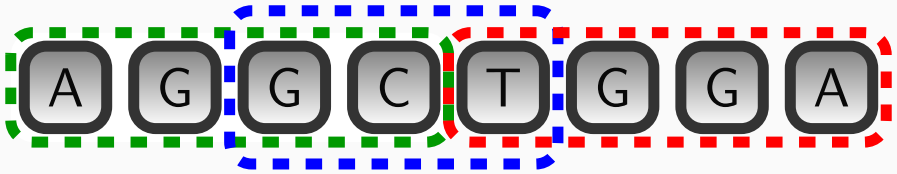
Sequencing by synthesis

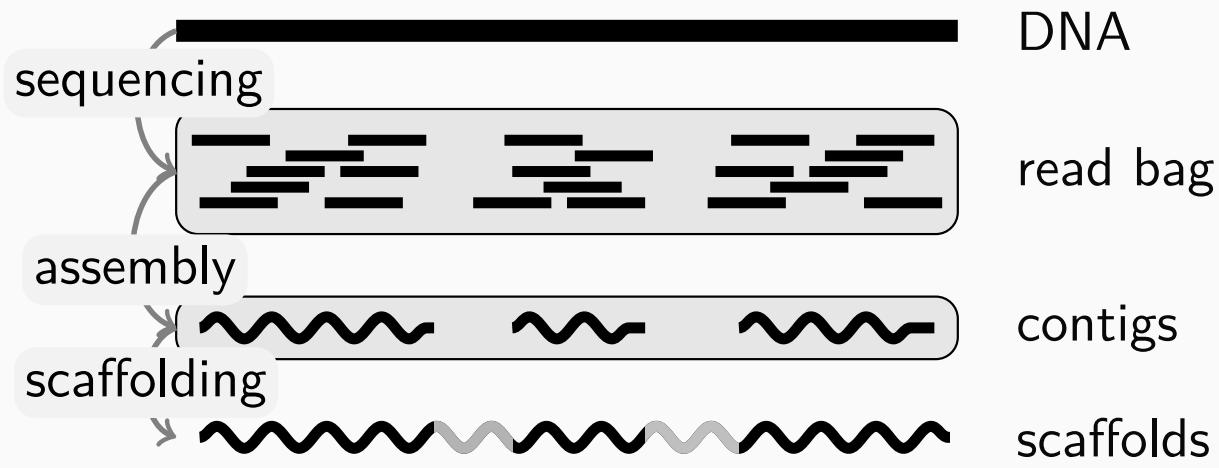


[By Abizar Lakdawalla, CC BY-SA 3.0, https://en.wikipedia.org/wiki/File:Sequencing_by_synthesis_Reversible_terminators.png]

Sequencing, read-sets

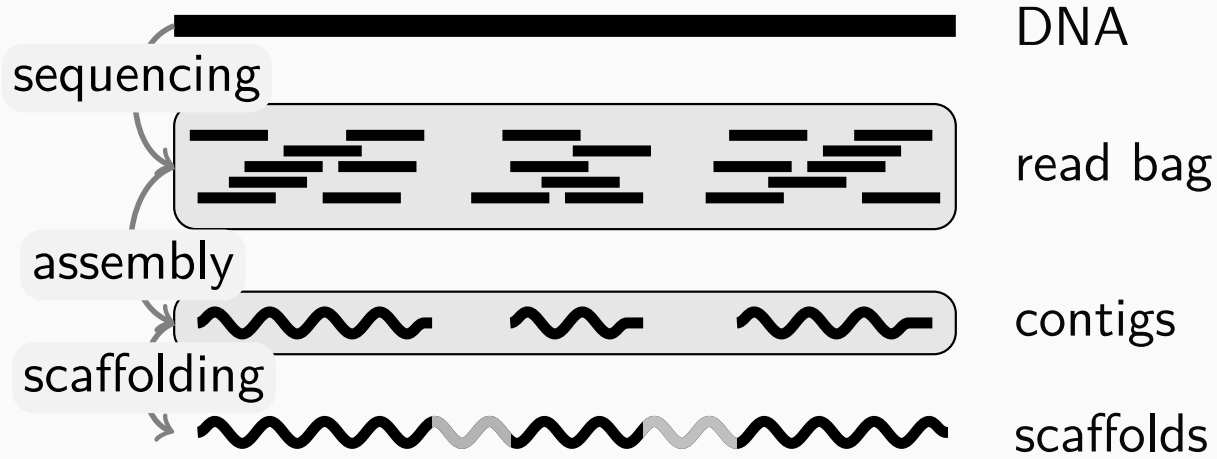
- Product of sequencing is not a long sequence, but short substrings called reads
- Reads have length of 10s to 100s of symbols
- Sequence AGGCTGGA is represented by set $\{\mathrm{AGGC}, \mathrm{TGGA}, \mathrm{GCT}\}$.





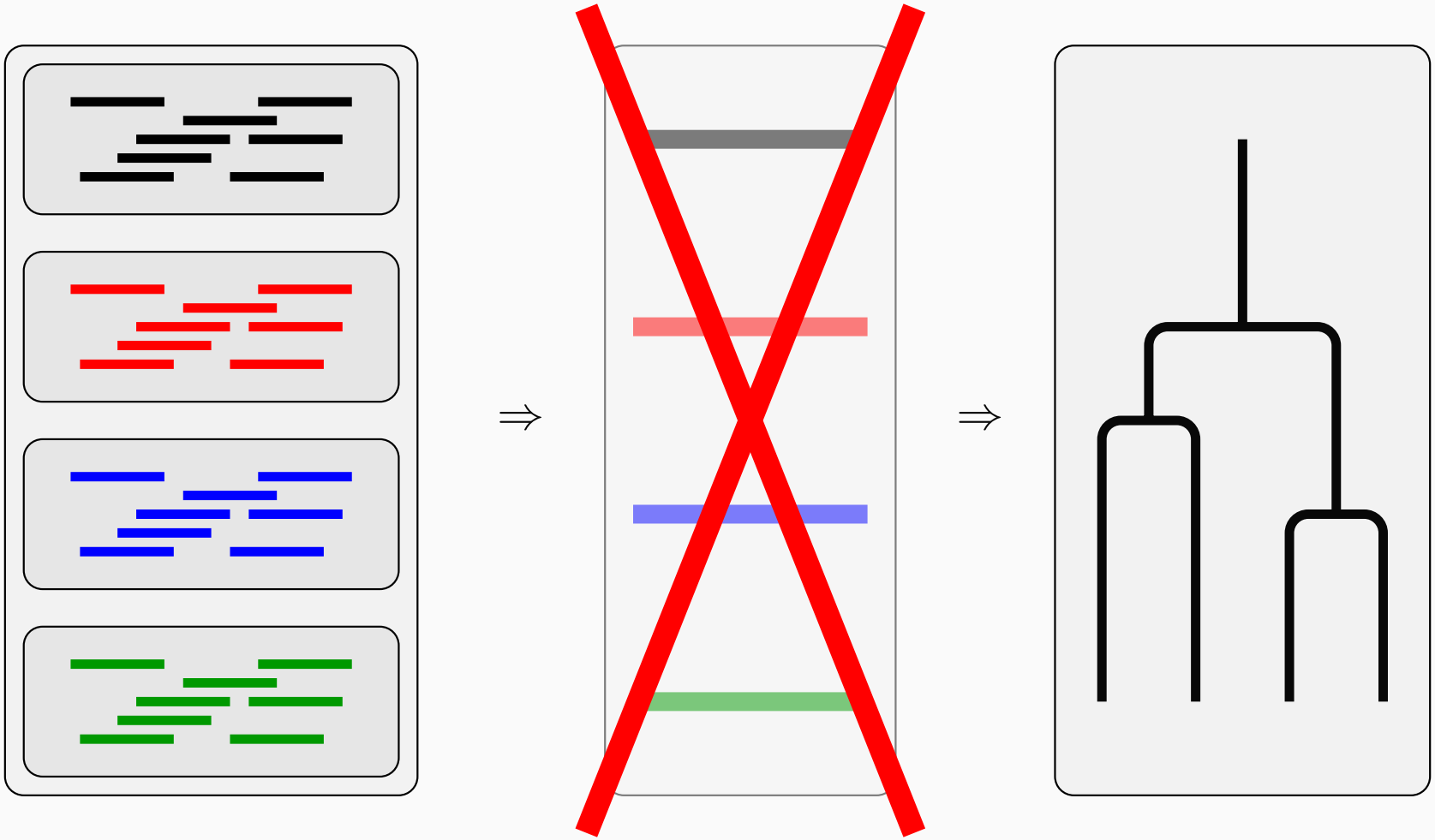
Contigs

- 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 locations and orientation



Our approach - skip assembly.

- The goal is to build a dendrogram directly from the read sets



- Assumption: no reference sequence known

Alignment-free approaches

- Originally designed to 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
Advance Access published on 23 September 2013
doi:10.1093/bib/bbt067

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

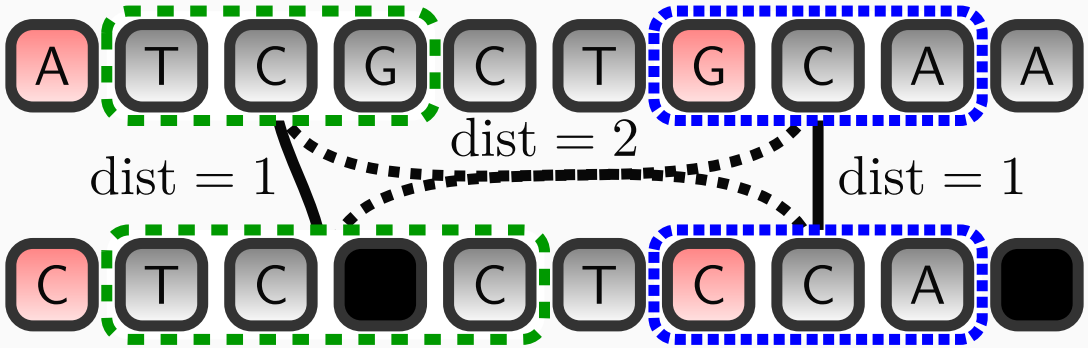
Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S. Waterman and Fengzhu Sun

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

DISTANCE FUNCTION DESIGN

Monge-Elkan distance

- Our approach is based on the 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

Estimating Sequence Similarity from Read Sets for Clustering Sequencing Data

Petr Ryšavý & Filip Železný

Conference paper | First Online: 21 September 2016

1631 Accesses | **4** Citations

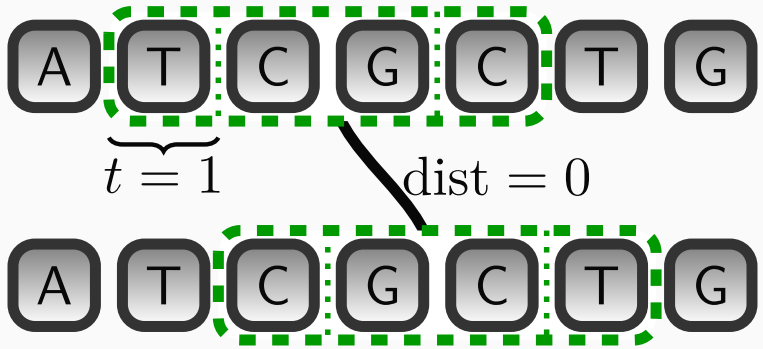
Part of the Lecture Notes in Computer Science book series (LNISA,volume 9897)

Distance scale and symmetry

- Our measure should be symmetric
- The 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

IMPROVEMENTS

Too slow ? - sample

- Coverage c around 2 provides results that are good enough.
- For high coverage data, downsample to $c = 2$.

Published: 04 August 2018

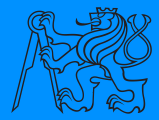
Estimating sequence similarity from read sets for clustering next-generation sequencing data

Petr Ryšavý & Filip Železný

Data Mining and Knowledge Discovery **33**, 1–23 (2019) | Cite this article

734 Accesses | **3** Citations | **2** Altmetric | Metrics

Too slow ? - use embedding

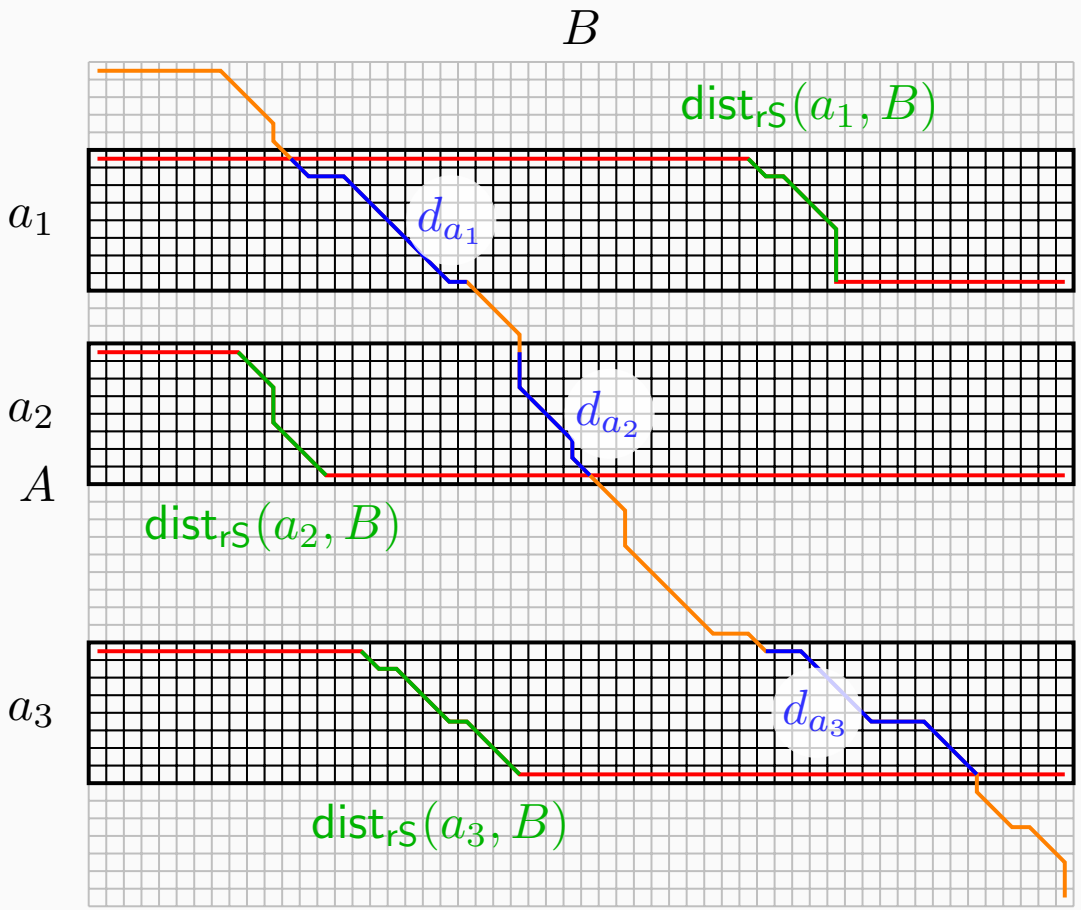


- We do not need an exact minimum in the Monge-Elkan distance.
- We use embedding to identify good candidates.
- q-gram profile is vector of counts of all possible q-grams, i.e. strings from Σ^q.
- q-gram distance of two strings is the Manhattan distance of their q-gram profiles.
- Inspiration by BLAST 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.

THEORETICAL ANALYSES

Works well - why? (unpublished)

- If we use a read bag and a sequence, a Monge-Elkan-alike distance serves as a lower-bound



Practical Applicability - p-value

- We developed a p-value algorithm for the Monge-Elkan distance
- Based on generating polynomials, combinatorics, and improved with FFT

An Algorithm to Calculate the p-value of the Monge-Elkan Distance *

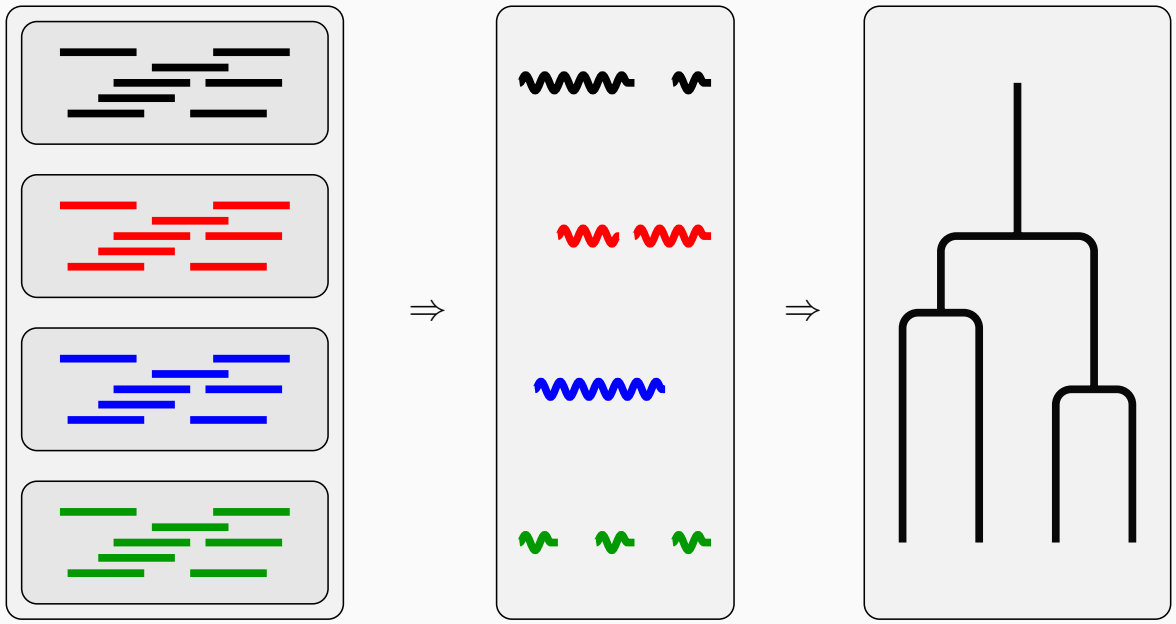
Petr Ryšavý[0000−0002−6597−6616] and Filip Železný[0000−0001−9780−3376]

Department of Computer Science,
Faculty of Electrical Engineering, Czech Technical University in Prague,
Prague, Czech Republic
{petr.rysavy,zelezny}@fel.cvut.cz

USING CONTIG-SETS

Our approach - skip assembly.

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



Estimating Sequence Similarity from Contig Sets

Petr Ryšavý ✉ & Filip Železný

Conference paper | First Online: 04 October 2017

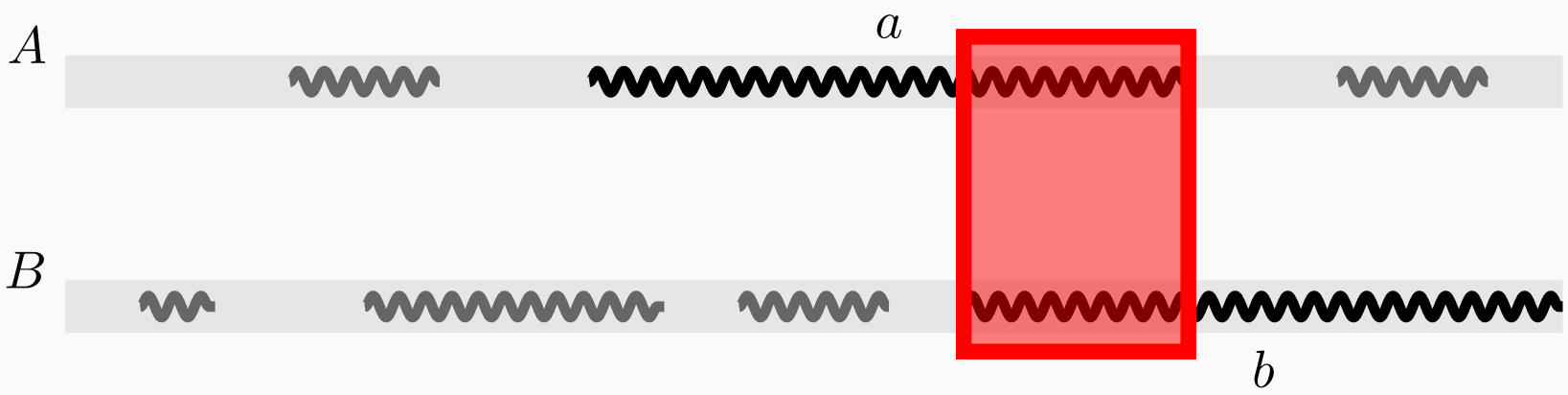
Part of the Lecture Notes in Computer Science book series (LNISA,volume 10584)

1) Estimating overlaps for contig pairs

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

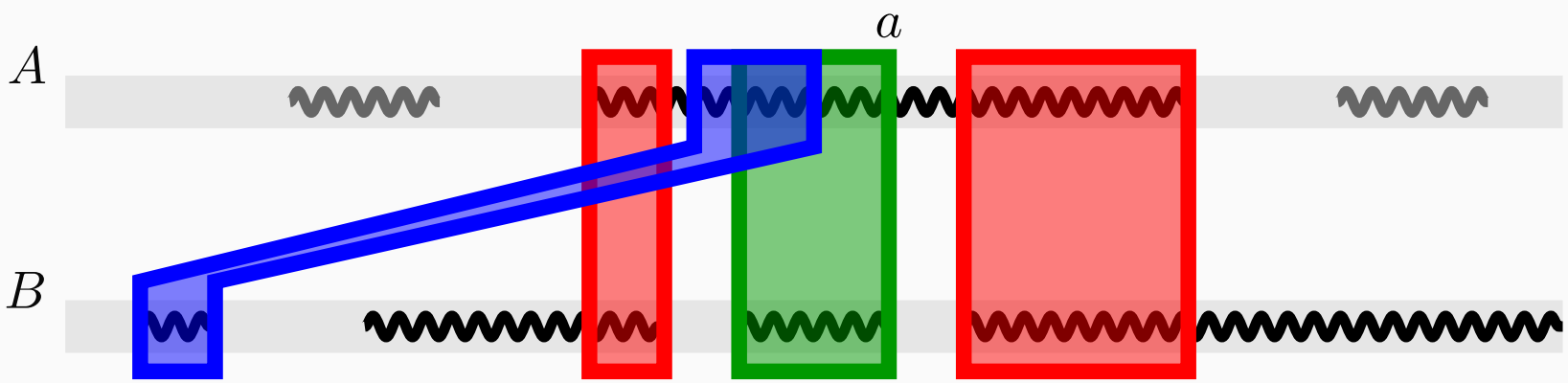
$$\overline{\mathsf{dist}}(\alpha, \beta) = \frac{\mathsf{dist}(\alpha, \beta)}{\max\{|\alpha|, |\beta|\}}. \tag{1}$$

- 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 scheduling problem*

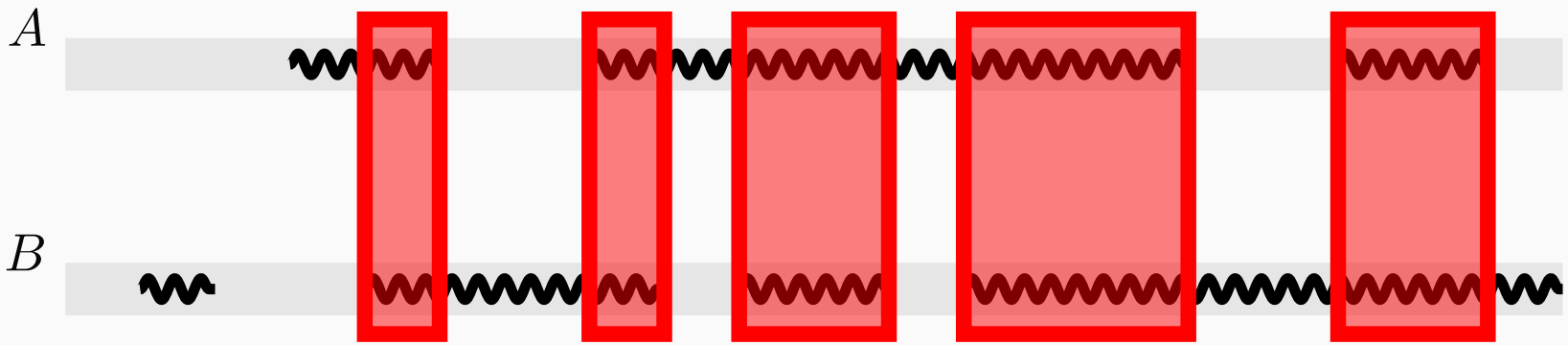


3) Combining the Results

- Sum distances of overlap pairs

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

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

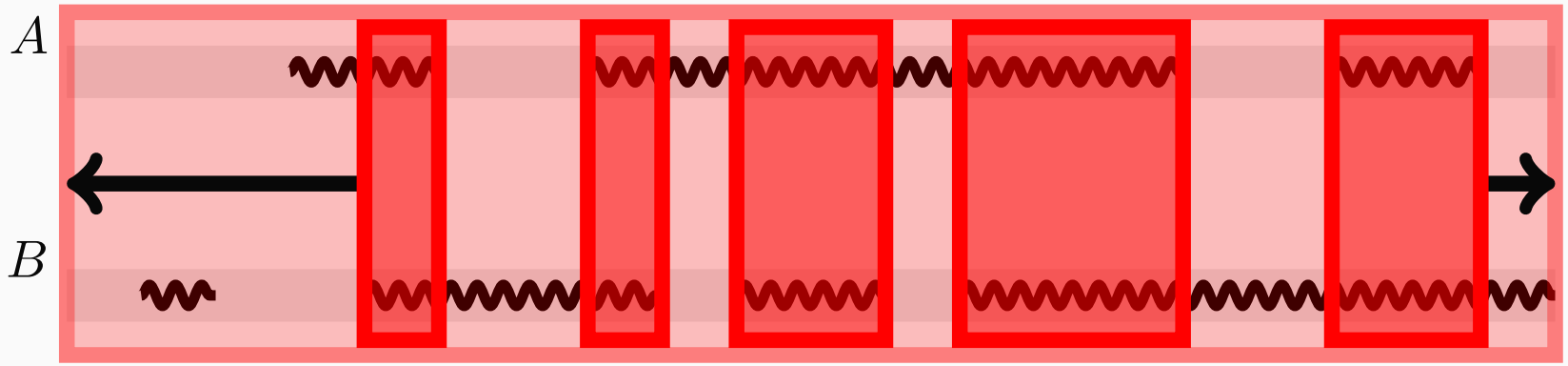


3) Combining the Results

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

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

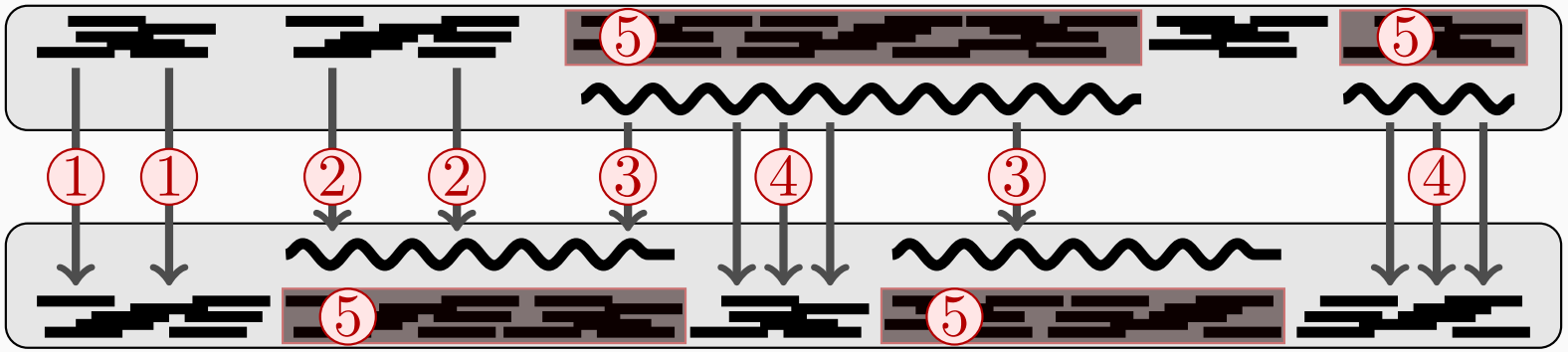
- Finally, make the resulting measure symmetric ...



COMBINATION OF THE MEASURES

Combine the Read-Based and Contig-Based Methods

- Requires dealing with many read/contig combination pairs



Research | Open Access | Published: 27 March 2023

Reference-free phylogeny from sequencing data

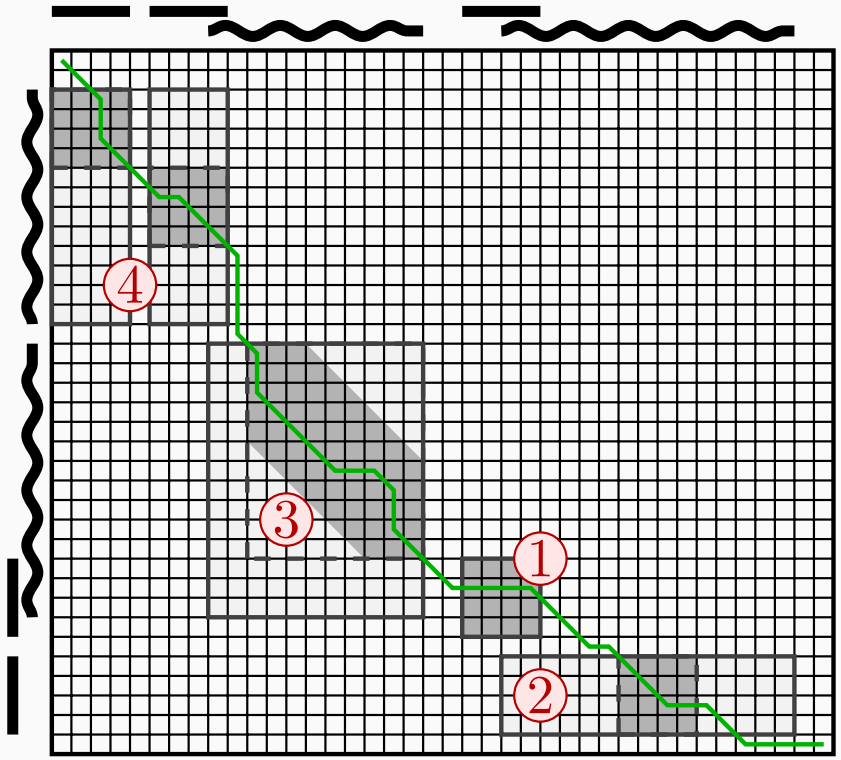
Petr Ryšavý & Filip Železný

BioData Mining **16**, Article number: 13 (2023) | Cite this article

1108 Accesses | **5** Altmetric | Metrics

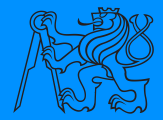
Efficiency Improvements

- Ideally, the algorithm should calculate an alignment only around the optimal alignment path



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 using 5 common de novo assemblers (ABySS, edena, SSAKE, SPADes, velvet)
- Comparison with alignment-free measures (8 d-type statistics, Mash, co-Phylog)

Measured characteristics:

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

- Exact evaluation of the 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
	$\max(\|R_A\|, \|R_B\|)$	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
	$\max(\|R_A\|, \|R_B\|)$	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
	$\max(\|R_A\|, \|R_B\|)$	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
	$\max(\|R_A\|, \|R_B\|)$	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)
$\text{Dist}_{\text{MESSG(M)q}\alpha}$	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

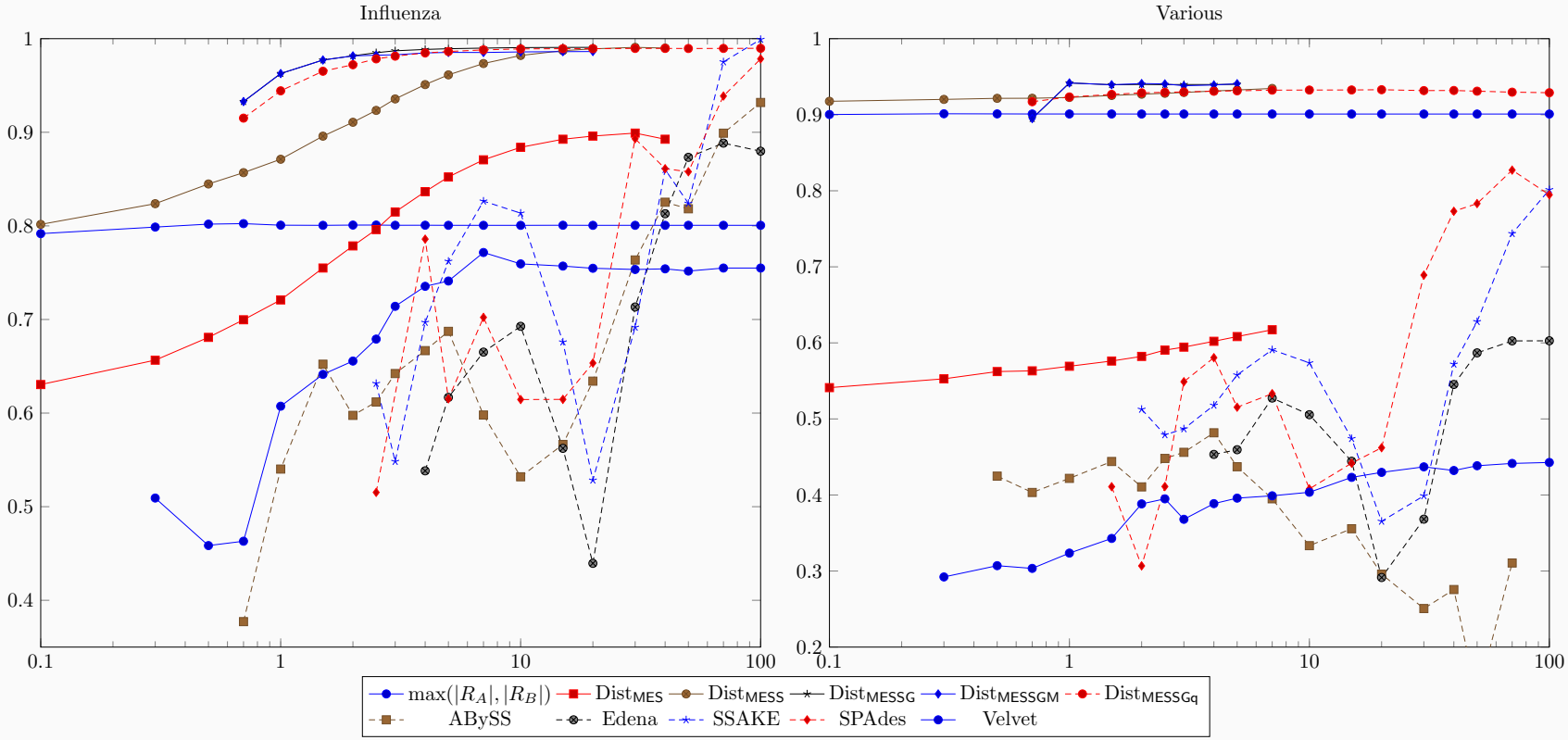


Figure 2: Plot of average Pearson's correlation coefficient for several choices of coverage values.

Results

- Our approach works better for short reads than assembly

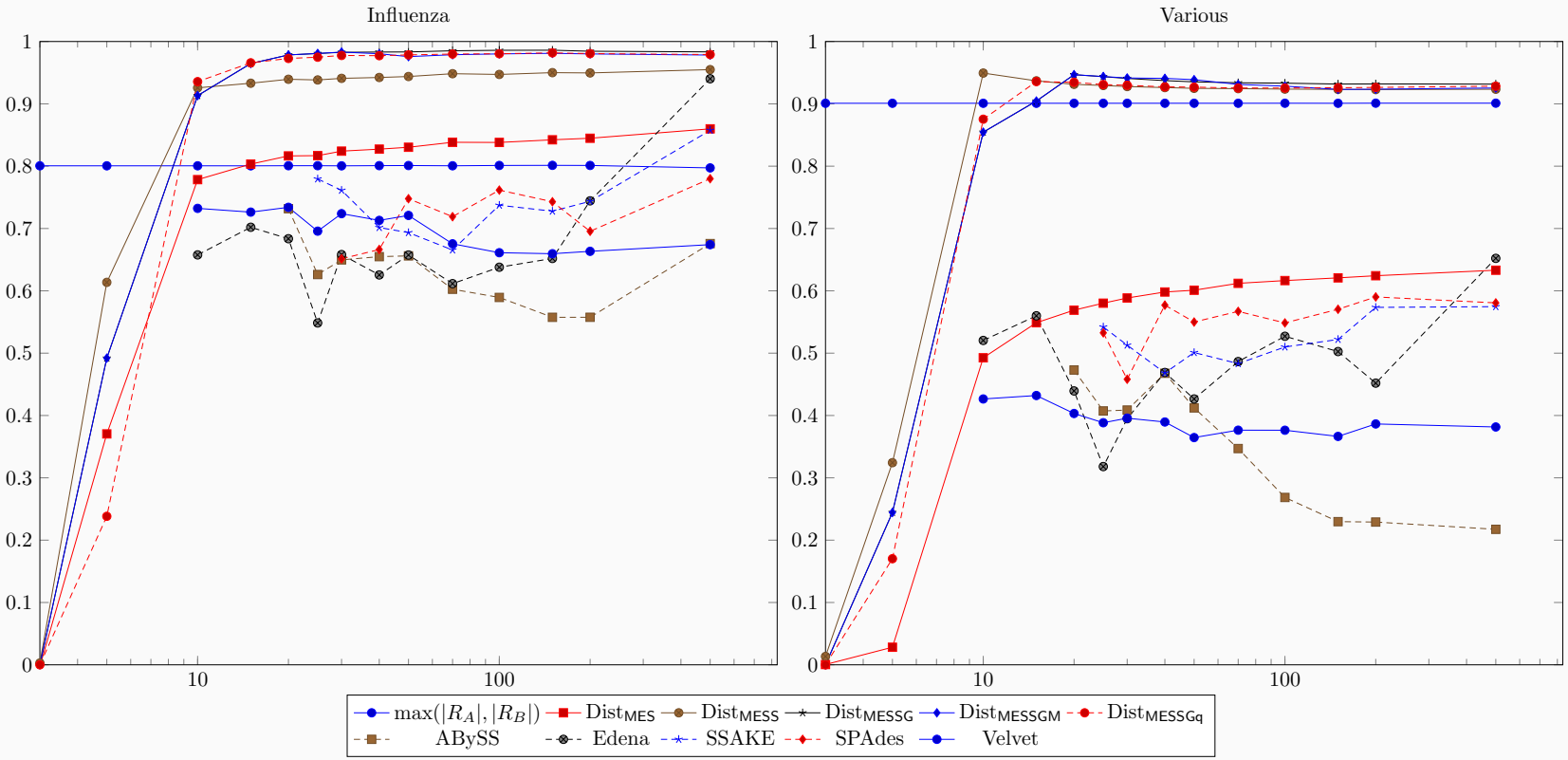


Figure 3: Plot of average Pearson's correlation coefficient for several choices of read length.

Conclusion

- We have seen three methods for estimating sequence similarity from read/contig sets or both
- Only single approximation step
- Adapts advantages of both alignment-free approaches and alignment similarity
- Due to low coverage requirements and small read length requirements, possible applications might include MiSeq, or as part of supertree methods
- Applicable to other similarity-based learning methods, as k-NNs

Circular RNAs

Research | Open Access | Published: 27 September 2022

circGPA: circRNA functional annotation based on probability-generating functions

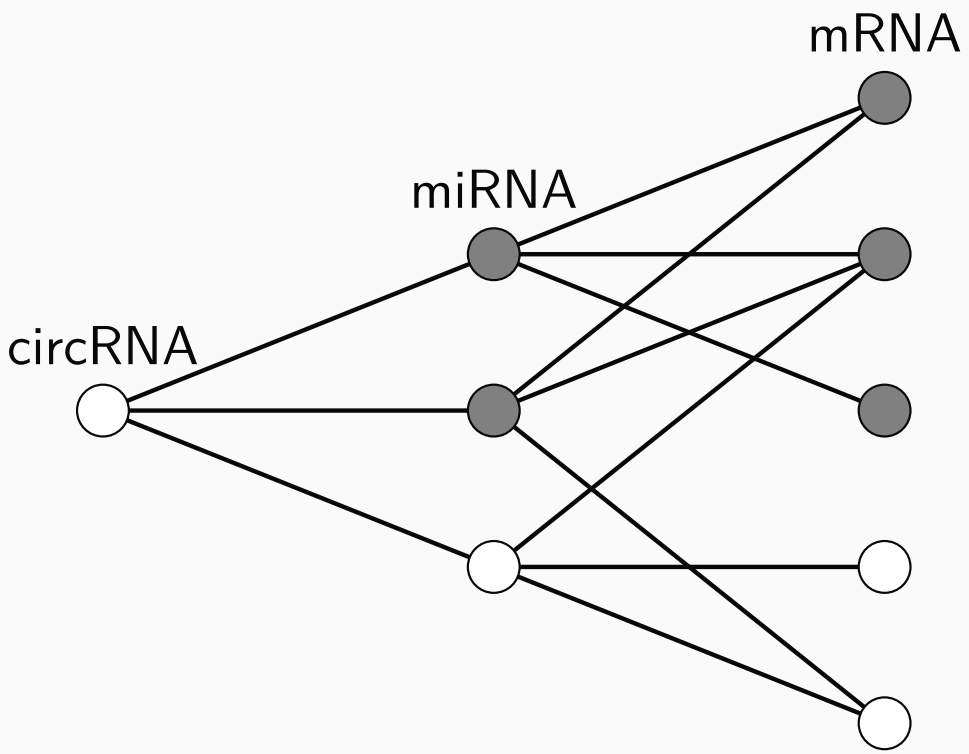
Petr Ryšavý ✉, Jiří Kléma & Michaela Dostálová Merkerová

BMC Bioinformatics **23**, Article number: 392 (2022) | Cite this article

1101 Accesses | **1** Citations | **4** Altmetric | Metrics

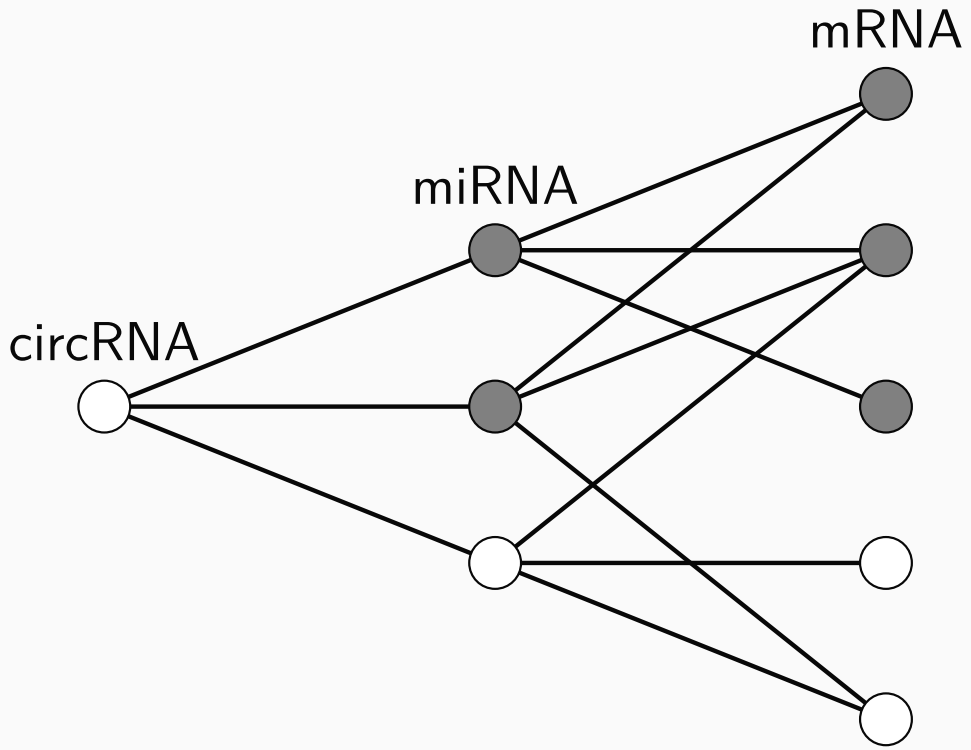
Circular RNA-interaction Graph

- In a second project, we focused on annotation of circular RNAs with annotation terms (as gene ontology terms)
- Today, the function of many circRNAs remains unknown
- An automatic tool to annotate circRNAs needed
- Annotation of miRNAs and mRNAs available
- Interaction graph is known (miRNA silencing and circRNA sponging)



Vector notion

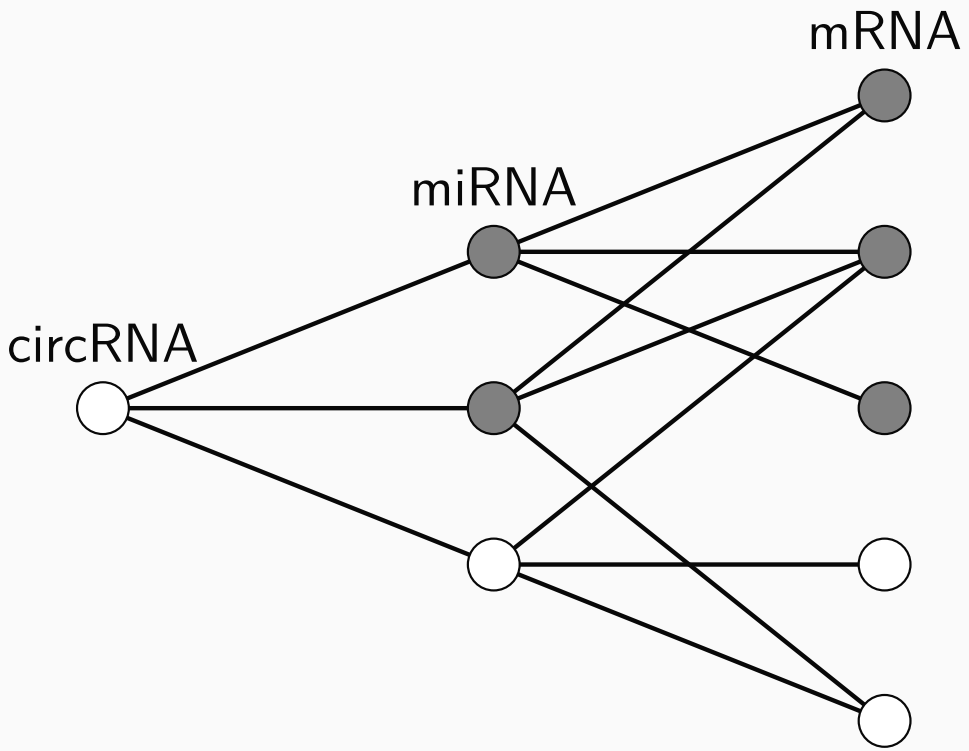
- $\vec{a}^{\mu,c} = (1,1,1)$
- $\mathbf{A}^{m,\mu} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$
- $\vec{g}^{\mu} = (0,1,1)$
- $\vec{g}^{m} = (1,1,1,0,0)$



ANNOTATION ALGORITHM

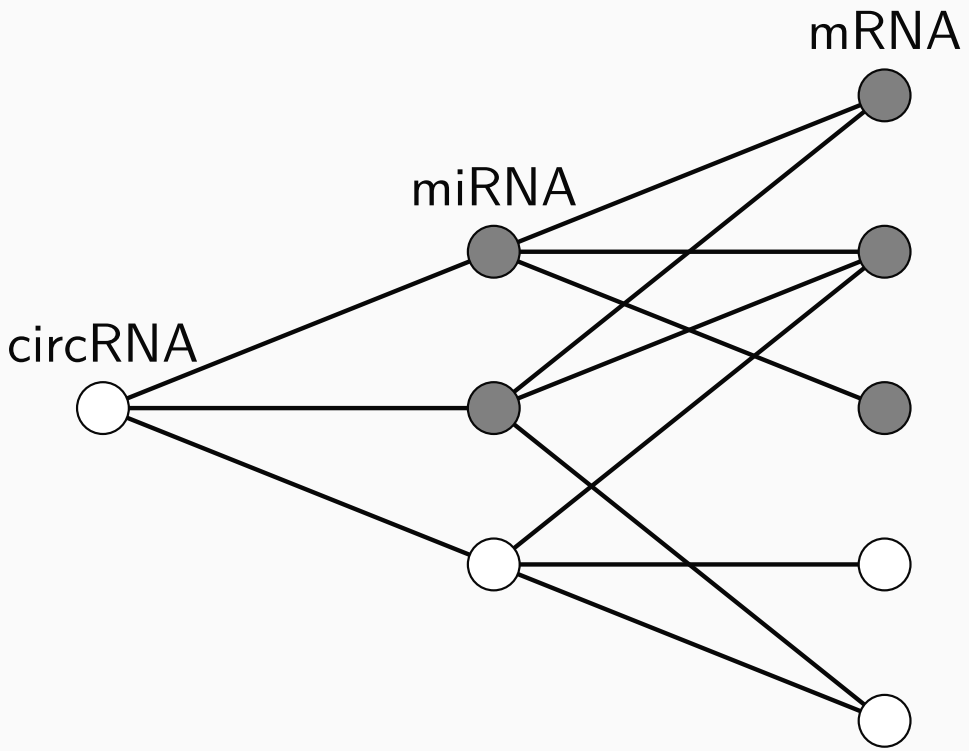
Use the number of paths from the circRNA

- Problems are independent for different circRNAs and different terms
- Guilt by association principle
- We can count the number of paths from the circRNA that end in a miRNA/mRNA annotated with the term

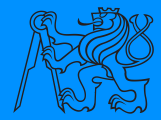


... in matrix form

$$\vec{a}^{\mu,c} \cdot \vec{g}^{\mu} + \mathbf{A}^{m,\mu} \vec{a}^{\mu,c} \cdot \vec{g}^{m}$$



The statistic is not comparable among terms



- More frequent terms have a higher statistic
- We can solve this bias by using the p-value
- Calculate the probability that we get a higher statistic for the same size term by chance
- Traditionally solved by generating random subsets and calculating the statistic (Barnard's Monte-Carlo sampling)

How to calculate the p-value

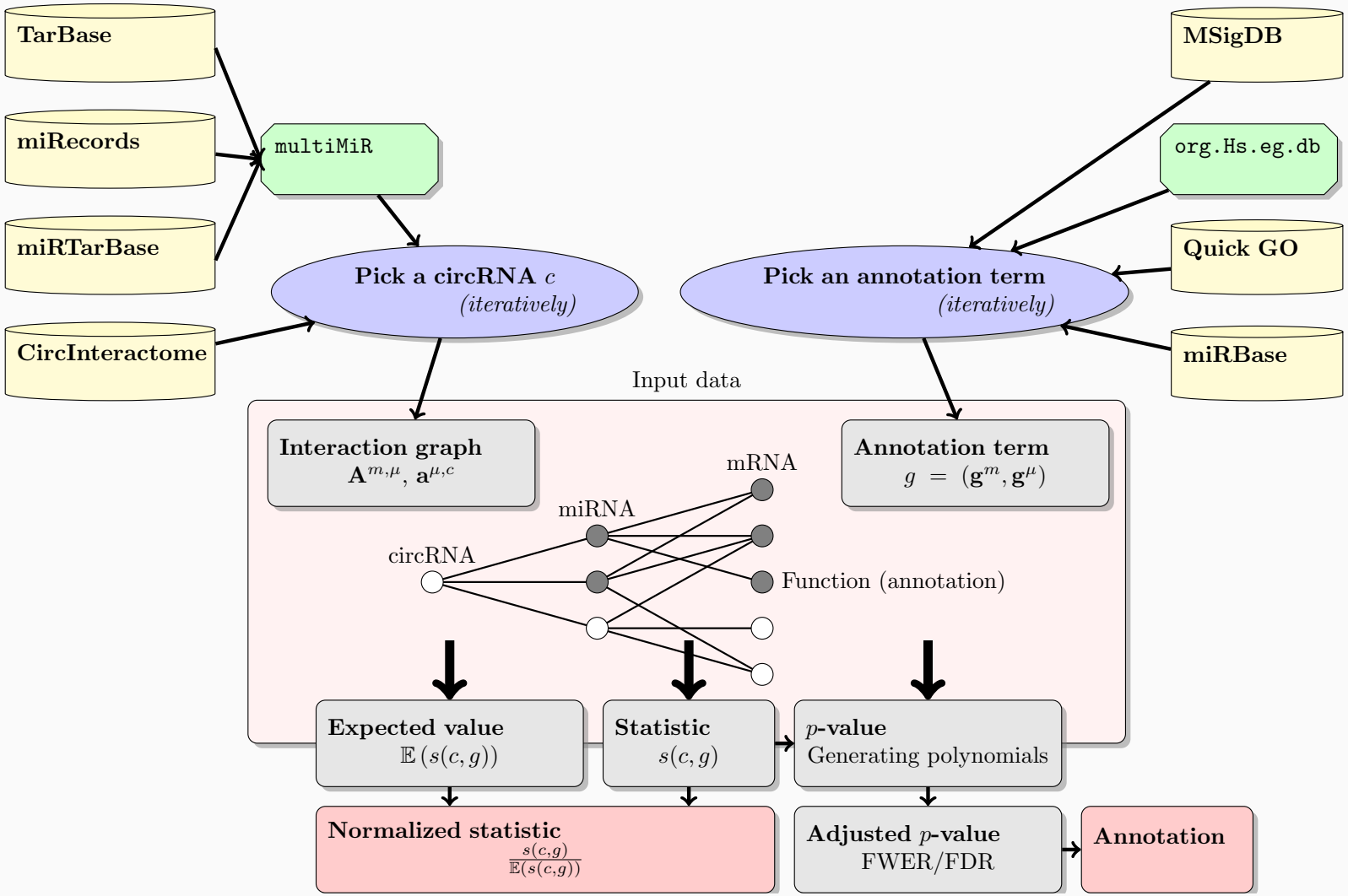
- Using a generating polynomial
- Imagine tossing a dice with sides $(1, 2, 3, 3, 3, 4)$, then the generating polynomial is

$$x + x^2 + 3x^3 + x^4$$

- Tossing this dice twice gives

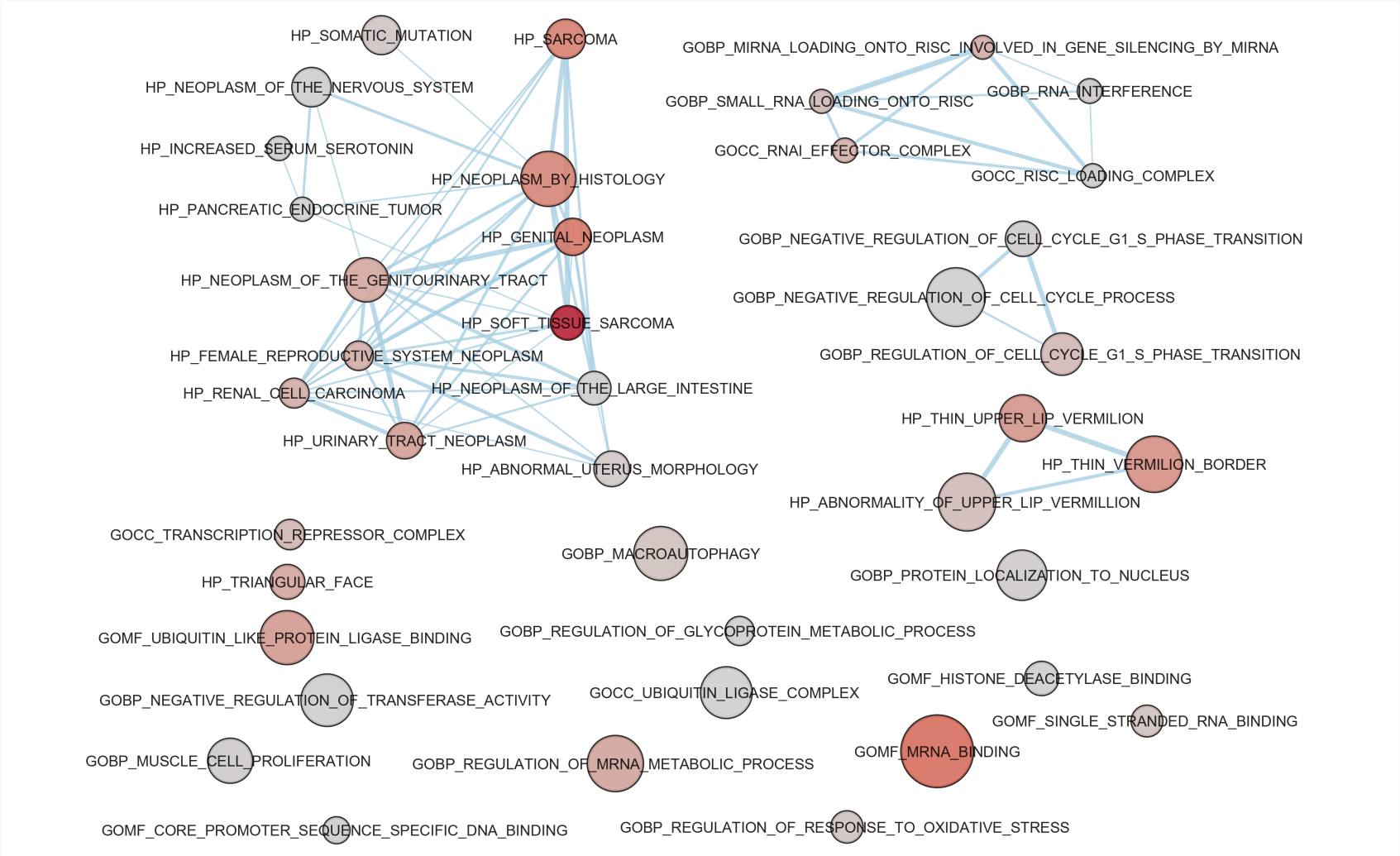
$$\left(x + x^2 + 3x^3 + x^4\right)^2 = x^2 + 2x^3 + 7x^4 + 8x^5 + 11x^6 + 6x^7 + x^8$$

The complete pipeline

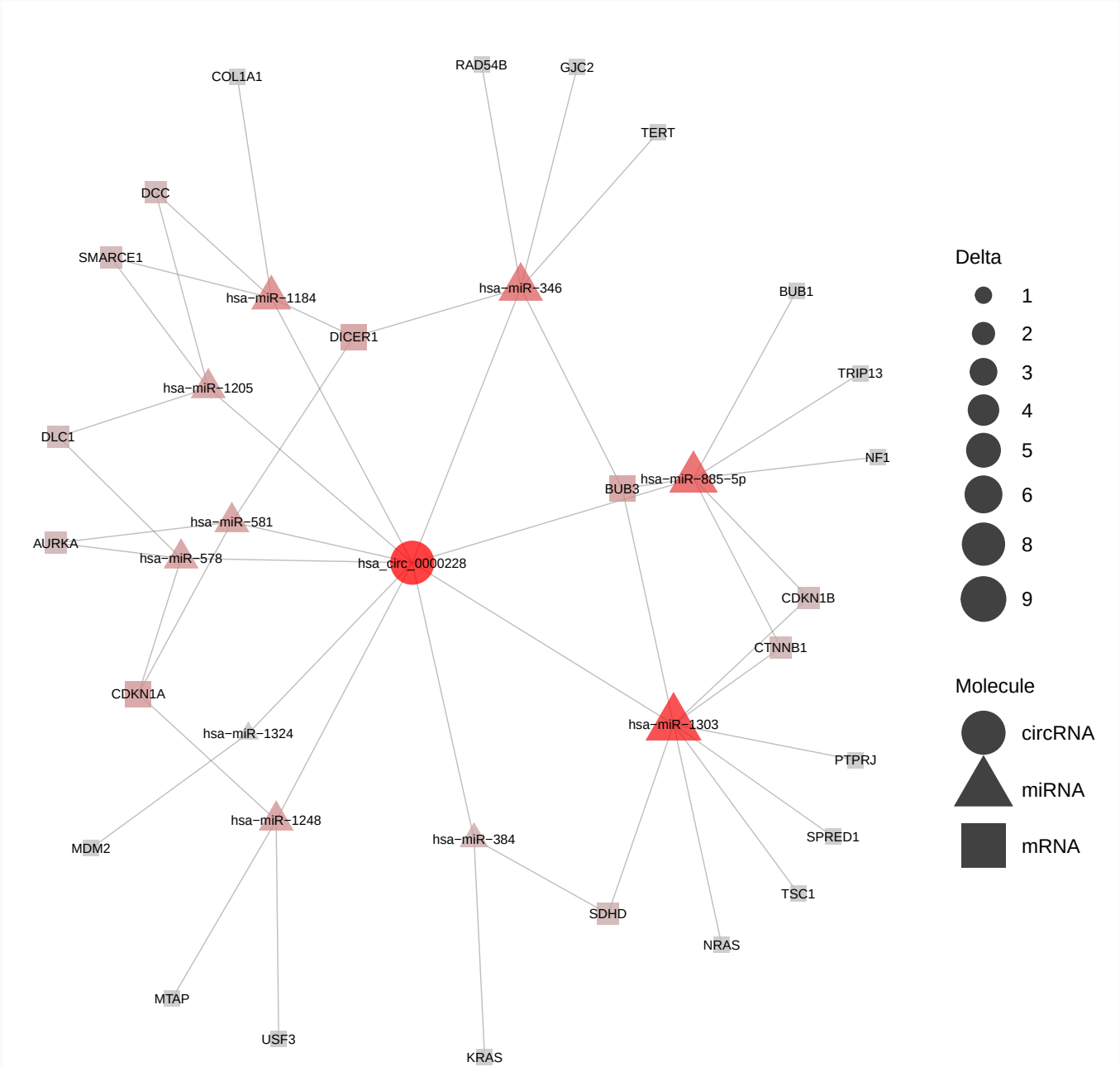


RESULTS

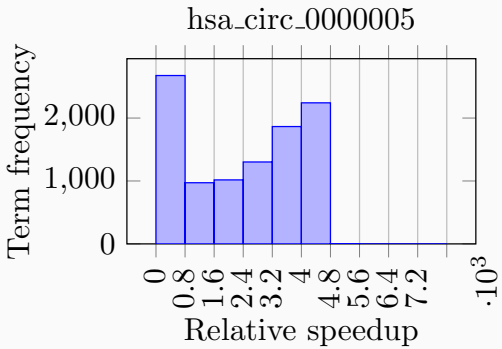
A sample output - annotation of hsa_circ_0000228

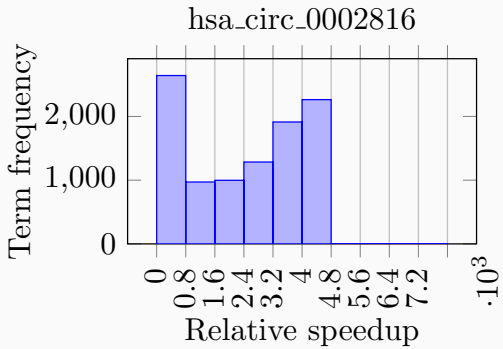


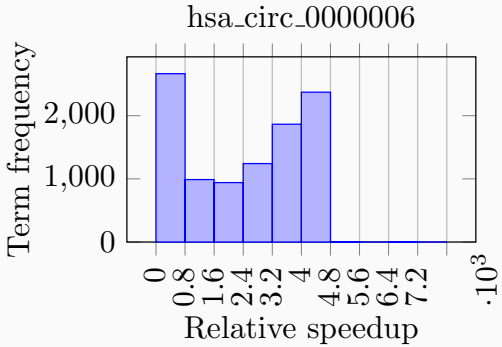
A sample output - influence of interacting RNAs

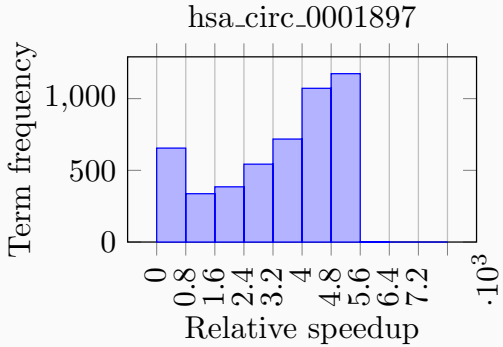


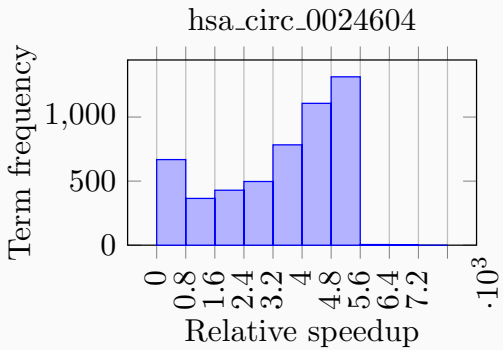
Evaluation - faster than the Monte-Carlo sampling

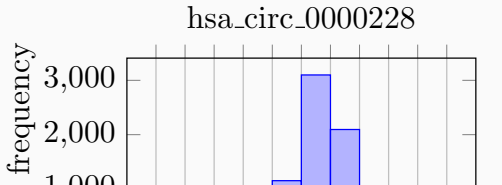


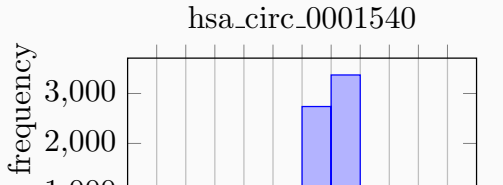












CONCLUSION - WORK IN PROGRESS

Conclusion - Work in Progress

- The edges can be weighted by co-expression
- The vertices can be weighted by the log-fold changes
- This way, we incorporate the expression matrix

Another problem:

- The tool can be used to mine circRNA-disease associations

Estimating Sequence Similarity from Read Sets for Clustering Sequencing Data

Petr Ryšavý & Filip Železný

First Online: 21 September 2016

Citations

Notes in Computer Science book series (LNISA,volume 9897)

Estimating Sequence Similarity from Contig Sets

Petr Ryšavý & Filip Železný

Conference paper | First Online: 04 October 2017

Part of the Lecture Notes in Computer Science book series (LNISA,volume 1

Published: 04 August 2018

Estimating sequence similarity from read sets for clustering next-generation sequencing data

Petr Ryšavý & Filip Železný

owledge Discovery 33, 1–23 (2019) | Cite this article

Research | Open Access | Published: 27 March 2023

Reference-free phylogeny from sequencing data

Petr Ryšavý & Filip Železný

BioData Mining 16, Article number: 13 (2023) | Cite this article

An Algorithm to Calculate the p-value of the Monge-Elkan Distance *

Petr Ryšavý[0000−0002−6597−6616] and Filip Železný[0000−0001−9780−3376]

Department of Computer Science,
lectrical Engineering, Czech Technical University in Prague,
Prague, Czech Republic
{petr.rysavy,zelezny}@fel.cvut.cz

Research | Open Access | Published: 27 September 2022

circGPA: circRNA functional annotation based on probability-generating functions

Petr Ryšavý, Jiří Kléma & Michaela Dostálová Merkerová

BMC Bioinformatics 23, Article number: 392 (2022) | Cite this article

Thank you for your attention.
Time for questions!

The support of the following grants is highly appreciated:

CZ.02.1.01/0.0/0.0/16 019/0000765 "Research Center for Informatics", CESNET LM2015042, GA CR 14-21421S, GA CR 20-19162S, SGS17/189/OHK3/3T/13, and SGS20/178/OHK3/3T/13.