Subquadratic optimal alignment using A*

Pesho Ivanov and Ragnar Groot Koerkamp

ETH Zurich

SRILAB
https://sri.inf.ethz.ch/

BIOMEDICAL INFORMATICS
https://bmi.inf.ethz.ch/

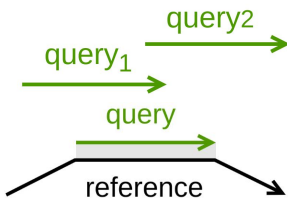
eth-sri/astarix

RagnarGrootKoerkamp/astar-pairwise-aligner

ETH zürich

Pairwise alignment variants

Semi-global / Mapping



now

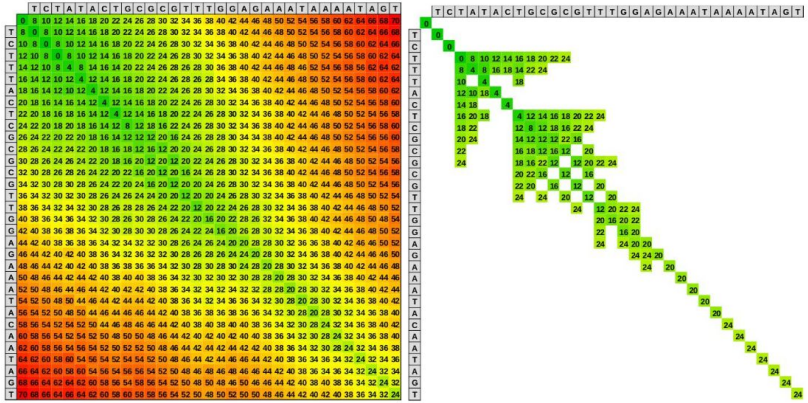
Global

in 15min

Local

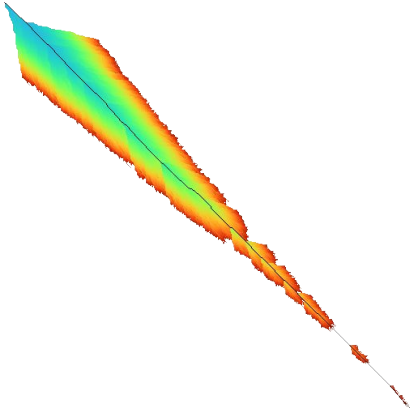
some day?

Optimal alignment algorithms

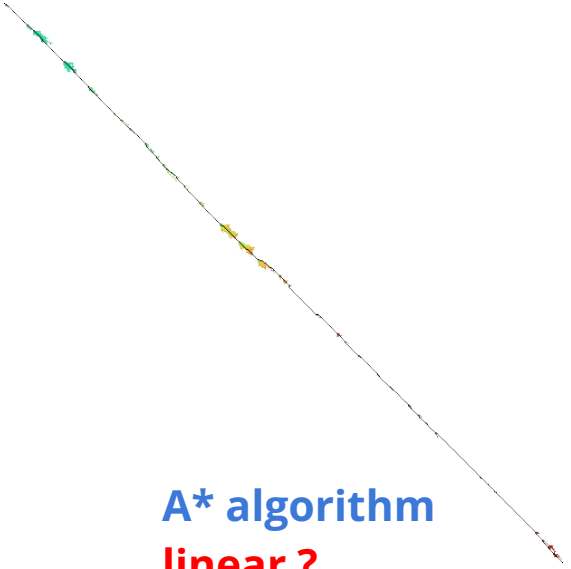


NW
$O(n^2)$

WFA
$O(sn) \equiv O(en^2)$



Dijkstra
$O(n^2)$

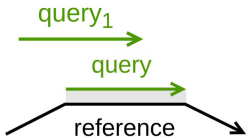


A* algorithm
linear ?

Uninformed search

Informed search

Results

	Core approach	Empirical runtime	Tool
Sequence-to-graph 	**trie / suffix tree** Ivanov et al. (2020) **seed heuristic** Ivanov et al. (2022)	$O(mN) \rightarrow m^{1.2}N^{0.4}$?	AStarix
Global aligning 	**seed heuristic pruning** Groot Koerkamp & Ivanov (draft)	$O(en^2) \rightarrow n$?	A*PA

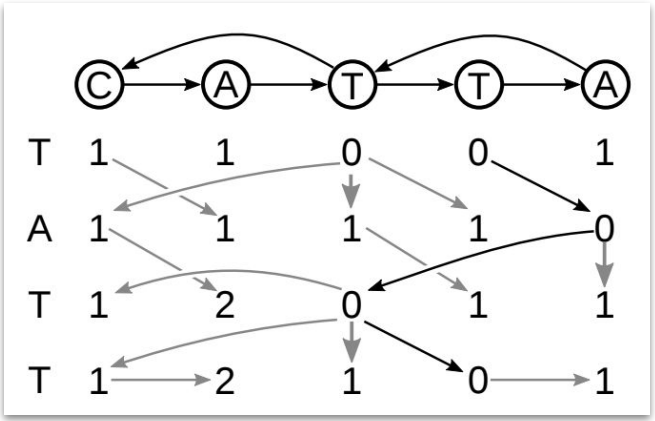
m – query length
N – reference size
n – sequences length
e – error rate

A* algorithm and admissible heuristic

A* heuristic:

- **Admissible**
- **Precise**
- **Efficient to compute**

Sequence-to-graph mappers



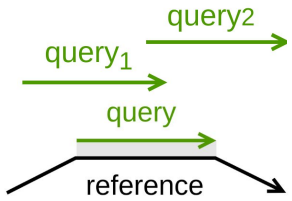
Edit graph

dp(u,i) — Minimal cost to align **q[1..i]**
from node **u** in the reference

Mikko Rautiainen and Tobias Marschall (2017) —
Aligning sequences to general graphs in O(V+mE) time

Except of **AStarix** other **optimal** mappers are **quadratic**:
GraphAligner, Vargas, PaSGAL

A* for sequence-to-graph



Scaling with

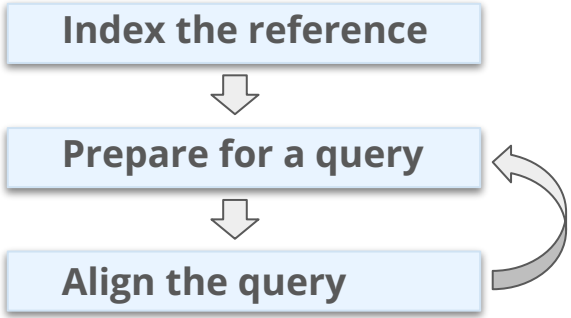
Reference size	**A* + trie**	$N^{0.11\text{-}0.46}$
Query length	**seed heuristic**	$m^{1.21}$

Ivanov P, Bichsel B, Mustafa H, Kahles A, Rätsch G, Vechev M (RECOMB 2020) – AStarix: Fast and Optimal Sequence-to-Graph Alignment

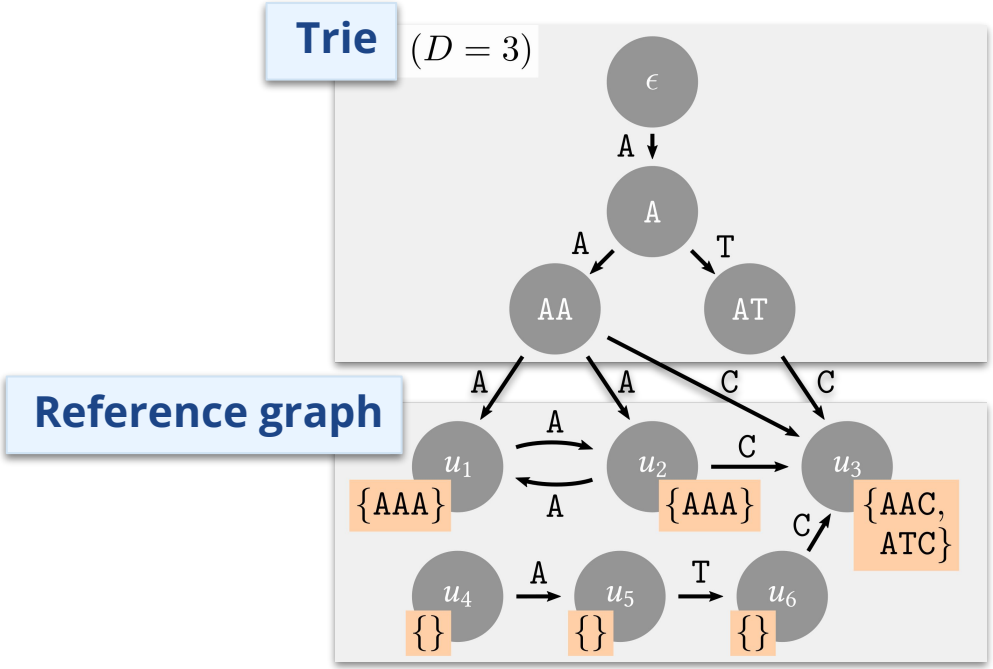
Ivanov P, Bichsel B, Vechev M (RECOMB 2022) – Fast and Optimal Sequence-to-Graph Alignment Guided by Seeds

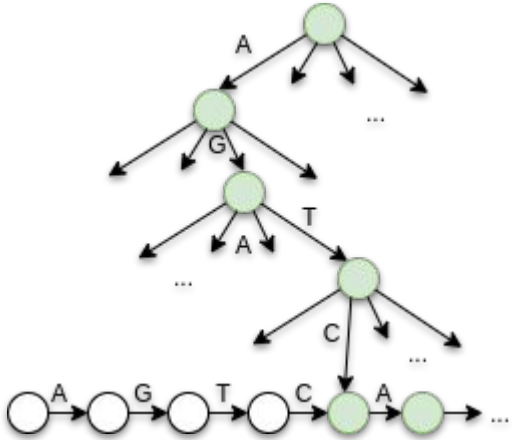
Fastest (consistently >60x) for:

- Illumina reads (200bp, 2-4% error rate)
- HiFi reads (up to 25kbp, 0.3% error rate)



1. Index the reference



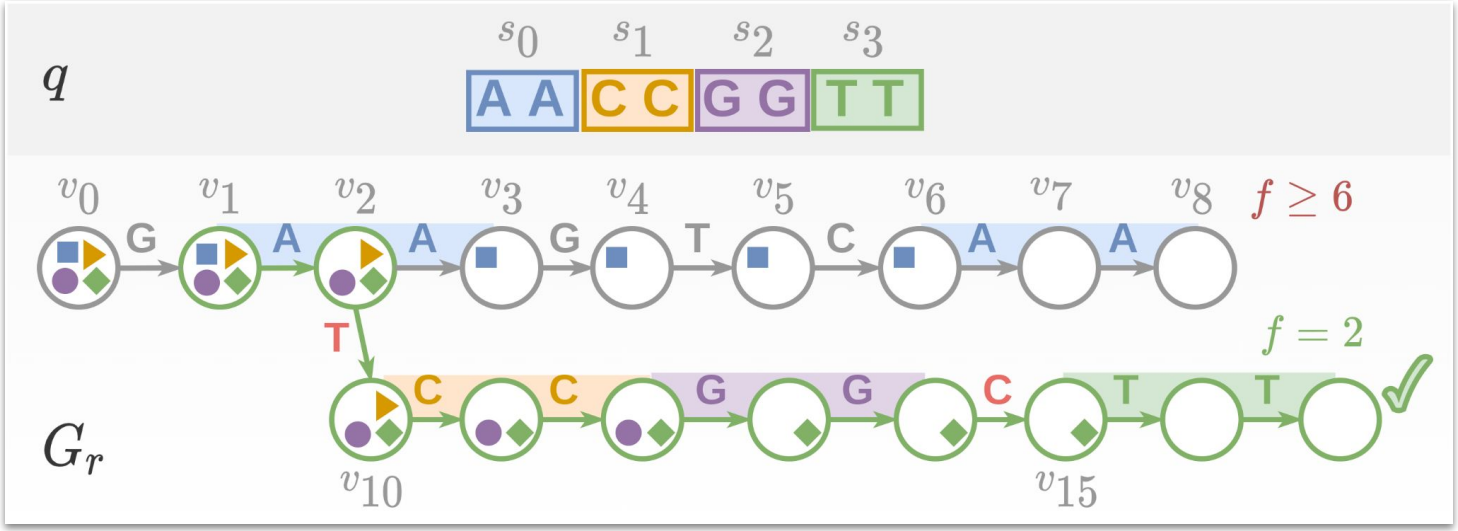


Aligning with a good heuristic

Dox and Fostier (Master's thesis, 2018), Efficient algorithms for pairwise sequence alignment on graphs

Ivanov et al. (RECOMB 2020), AStarix: Fast and Optimal Sequence-to-Graph Alignment

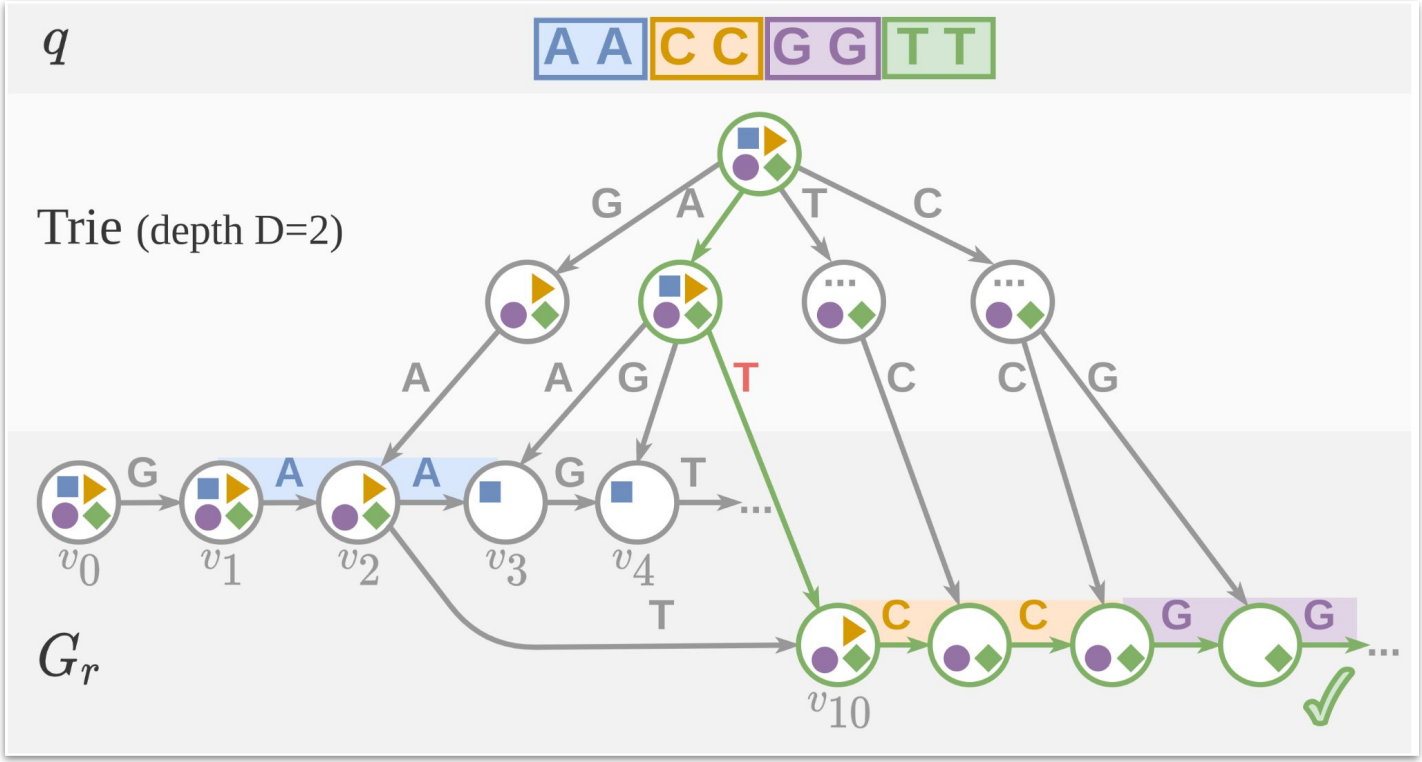
Seed heuristic: preparation



Query→Seeds→Matches→Crumbs

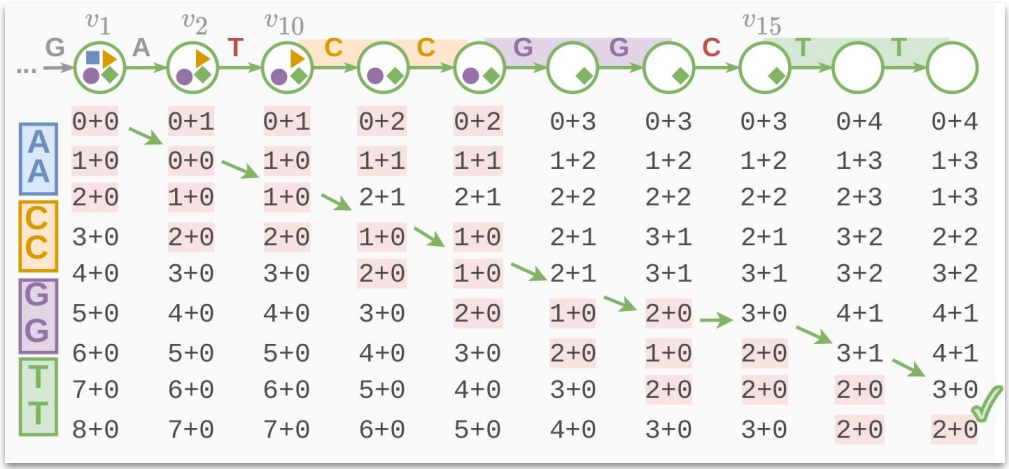
Ivanov, Bichsel and Vechev (RECOMB 2022) –
Fast and Optimal Sequence-to-Graph Alignment Guided by Seeds

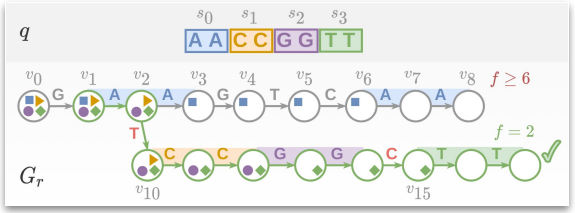
Seed heuristic on trie



Query→Seeds→Matches→Crumbs

Seed heuristic: query





$$h\langle v, i\rangle = misses\langle v, i\rangle \cdot \delta_{min}$$

$$\left|\{s \in Seeds_{\geq i} \mid s \text{ has no crumb in } v\}\right| \qquad \min(\Delta_{\text{subst}}, \Delta_{\text{del}}, \Delta_{\text{ins}})$$

Speedup results

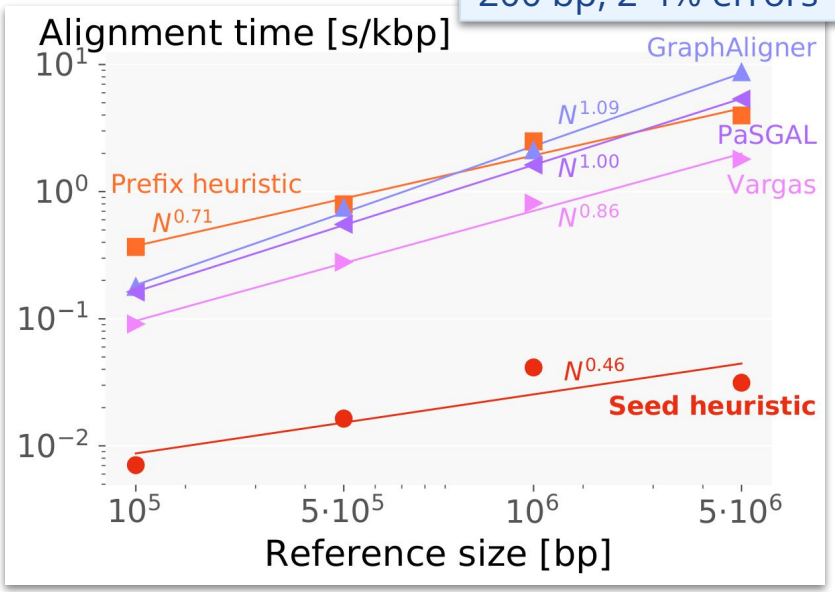
Tool	Illumina *E. coli*	Illumina MHC	HiFi *E. coli*	HiFi MHC	
Seed heuristic	0.019	0.041	0.001	0.002	s/kbp
Prefix heuristic	269x	180x	n/a	n/a	x slowdown
GraphAligner	424x	212x	118x	64x	
Vargas	133x	67x	1 413x	705x	
PaSGAL	263x	130x	1 367x	736x	

Seed heuristic skips >99.99% of the table cells

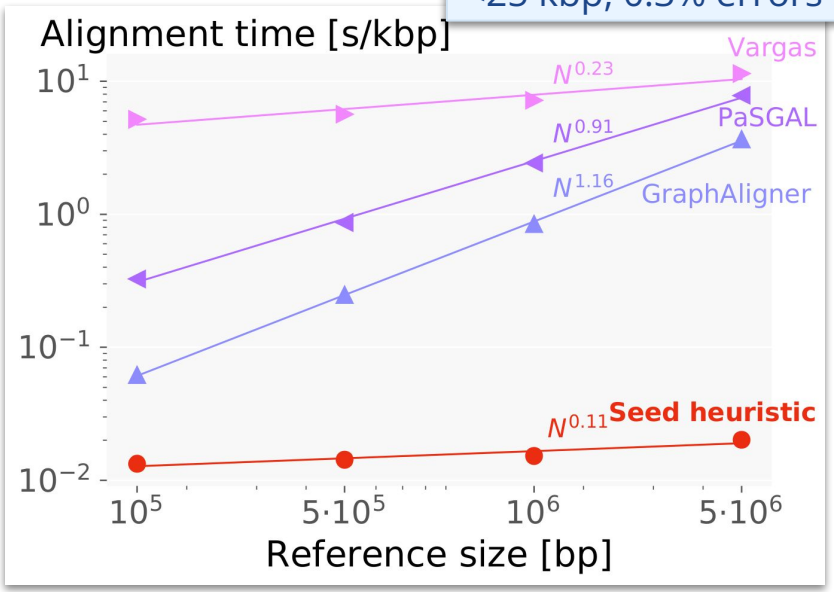
References: *E. coli* – linear 4.6M, Major Histocompatibility Complex (MHC) – 5.3M
Simulated queries: 200bp Illumina with Δ=(0,1,5,5); HiFi: 5–25kbp, e=0.3%, Δ=(0,1,1,1)
Prefix heuristic parameters: length cap d=5, cost cap c=5, trie depth D=log(N)
Seed heuristic parameters: D=14 ≈ $\log_4 N$; k=25 for Illumina, k=150 for HiFi reads

Scaling: reference size

Illumina reads:
200 bp, 2-4% errors



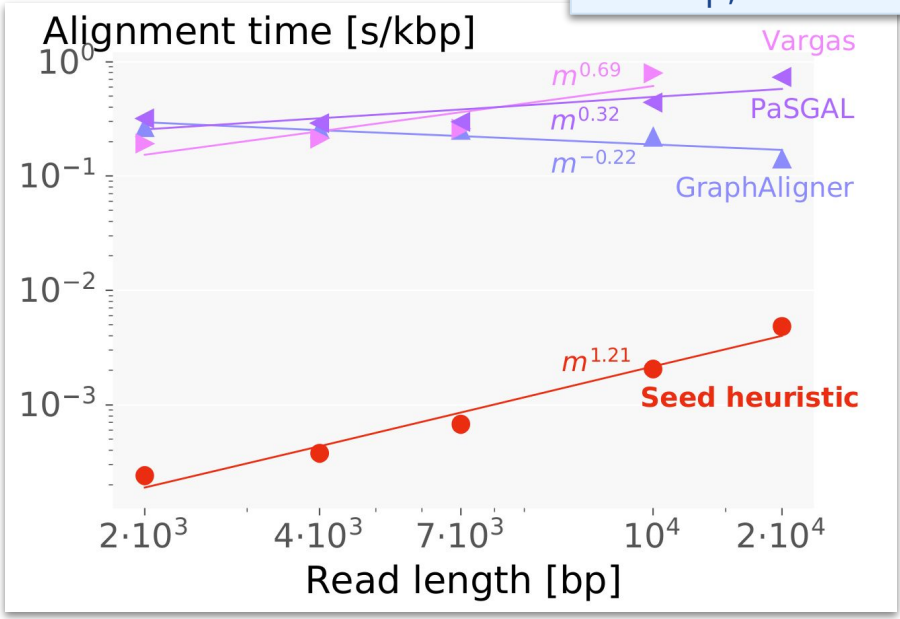
HiFi reads:
<25 kbp, 0.3% errors



References: *E. coli* – linear 4.6M, Major Histocompatibility Complex (MHC) – 5.3M
Simulated queries: 200bp Illumina with Δ=(0,1,5,5); HiFi: 5–25kbp, e=0.3%, Δ=(0,1,1,1)
Prefix heuristic parameters: length cap d=5, cost cap c=5, trie depth D=log(N)
Seed heuristic parameters: D=14 ≈ $\log_4 N$; k=25 for Illumina, k=150 for HiFi reads

Scaling: query length

HiFi reads:
<25 kbp, 0.3% errors



References: Major Histocompatibility Complex (MHC) – 5.3M
Simulated queries: HiFi: 2–25kbp, e=0.3%, Δ=(0,1,1,1)
Prefix heuristic parameters: length cap d=5, cost cap c=5, trie depth D=log(N)
Seed heuristic parameters: D=14 ≈ $\log_4 N$; k=150 for HiFi reads

Future work

Applicability:
- longer indels
- complex references

Performance:
- memory-efficiency
- parallelization
- optimize for linear references

Extensions:
- local alignment
- affine costs

Theoretical analysis:
- scaling proofs

Fast and Optimal Sequence-to-Graph Alignment Guided by Seeds

Pesho Ivanov, Benjamin Bichsel and Martin Vechev

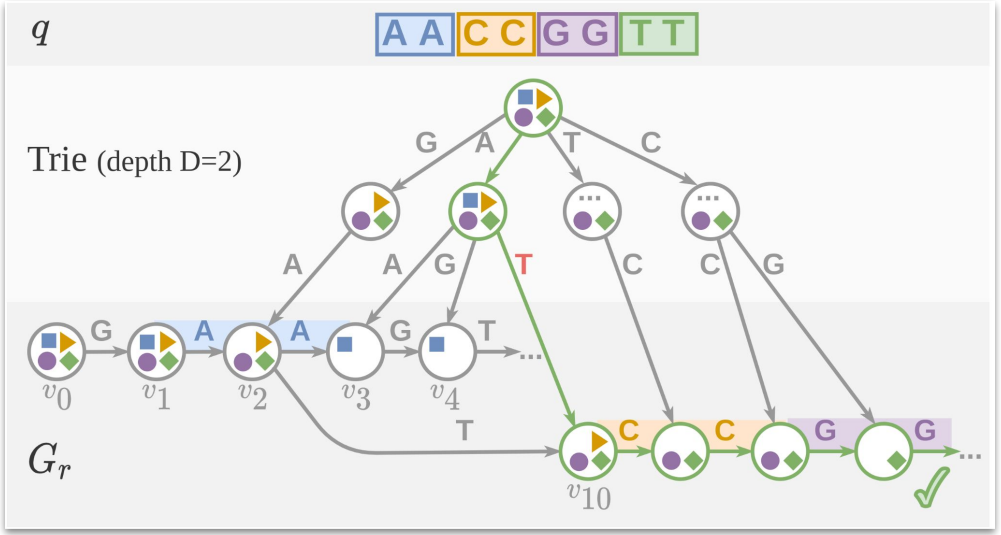
ETH Zurich

SRILAB

https://sri.inf.ethz.ch/

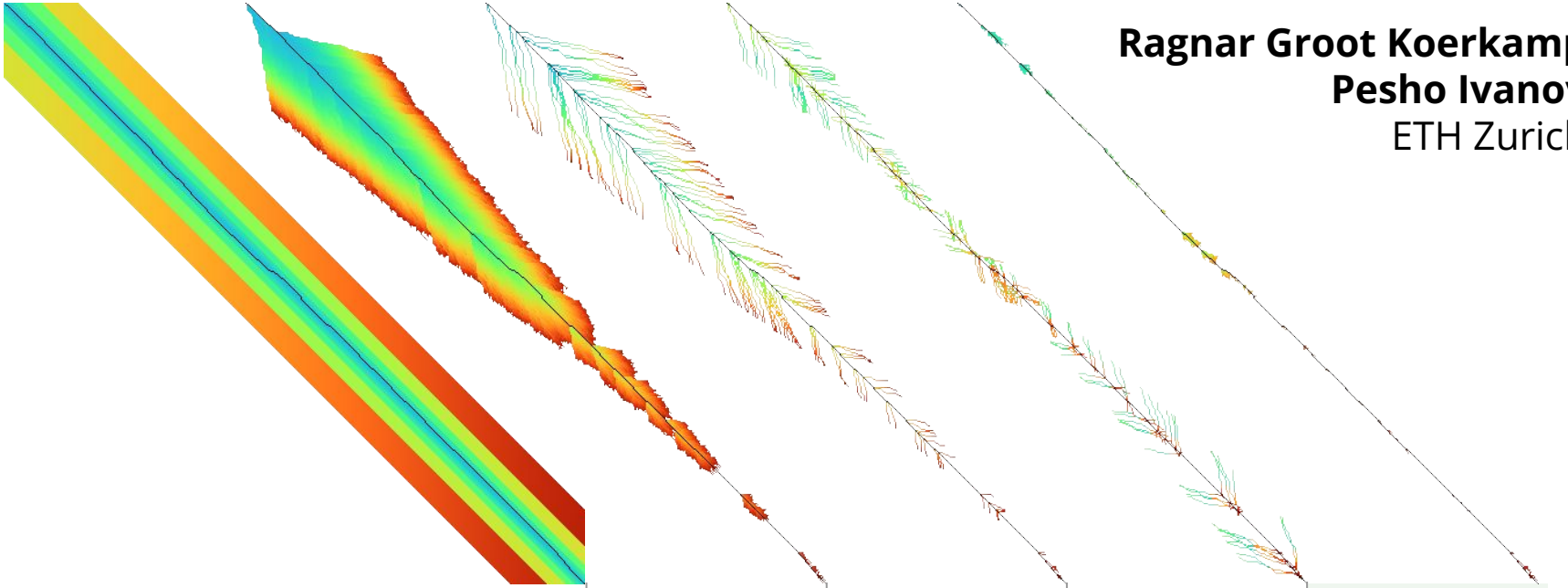
eth-sri/astarix

ETH zürich



Exact global pairwise alignment using A*

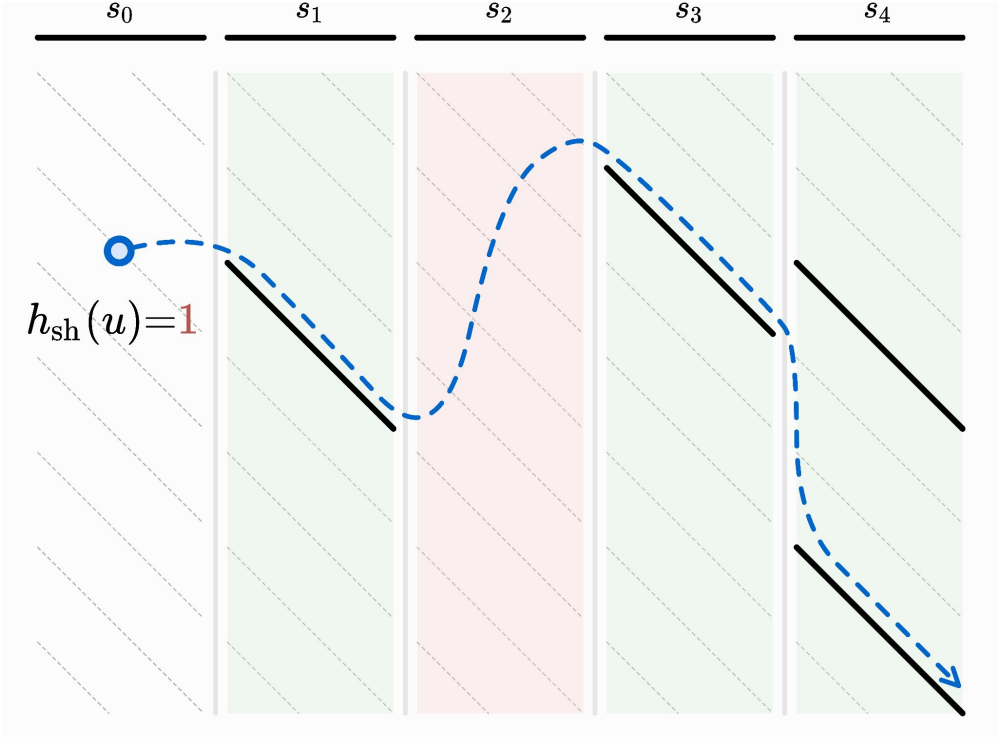
Ragnar Groot Koerkamp
Pesho Ivanov
ETH Zurich



Exponential band Ukkonen'85	**Dijkstra** Ukkonen'85	**Diagonal Transition** Ukkonen'85, Myers'86	**+ Divide & Conquer** Hirschberg'75, Myers'86	**A* + SH + pruning** Groot Koerkamp, Ivanov'22
Edlib Šošić, Šikić'17	**-**	**WFA** Marco-Sola et al'20	**BiWFA** Marco-Sola et al'22	**A*PA**
O(ns), O(n)	**O(ns)**	**O(s^2)**	**O(s^2), O(s)**	**O(n)?**

n=500, e=20%
Visualization by Mykola Akulov

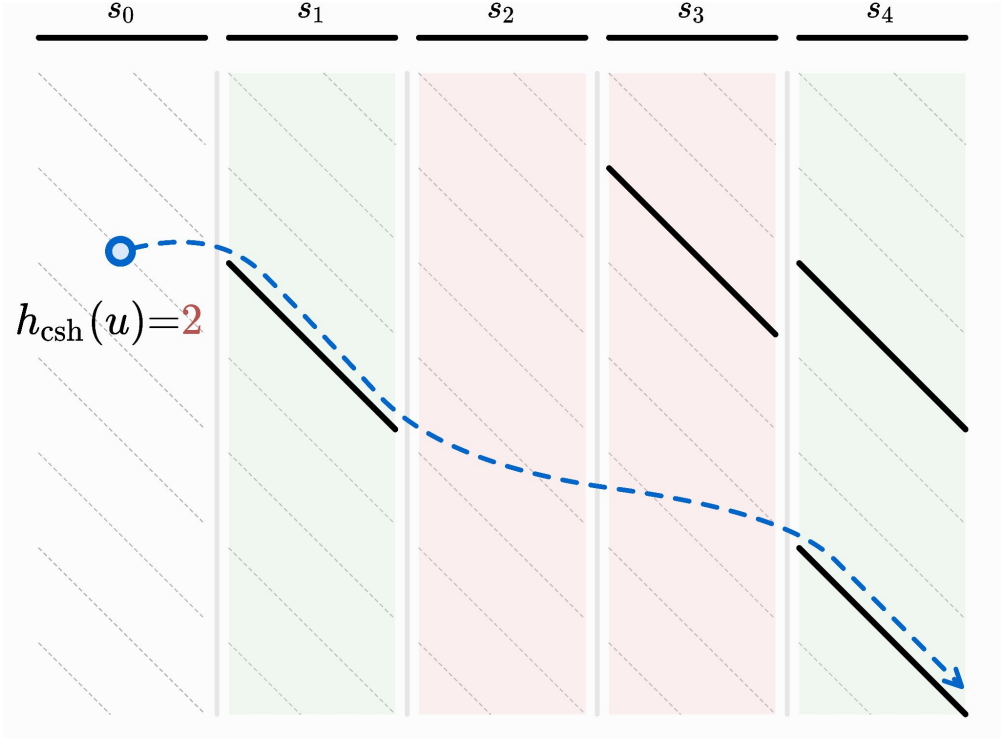
Seed Heuristic (SH)



Number of seeds without a match:

h_{sh} = #{upcoming seeds} – #{matching upcoming seeds}

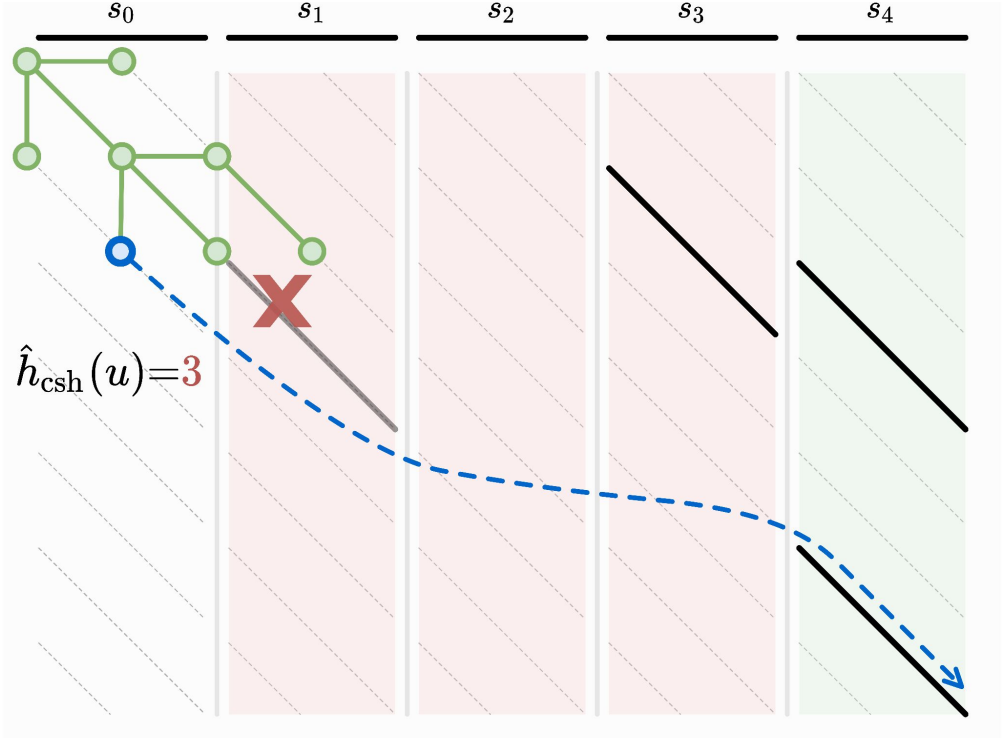
Chained Seed Heuristic (CSH)



Require matches to form a chain:

h_{csh} = #{upcoming seeds} – #{longest **chain** of matches}

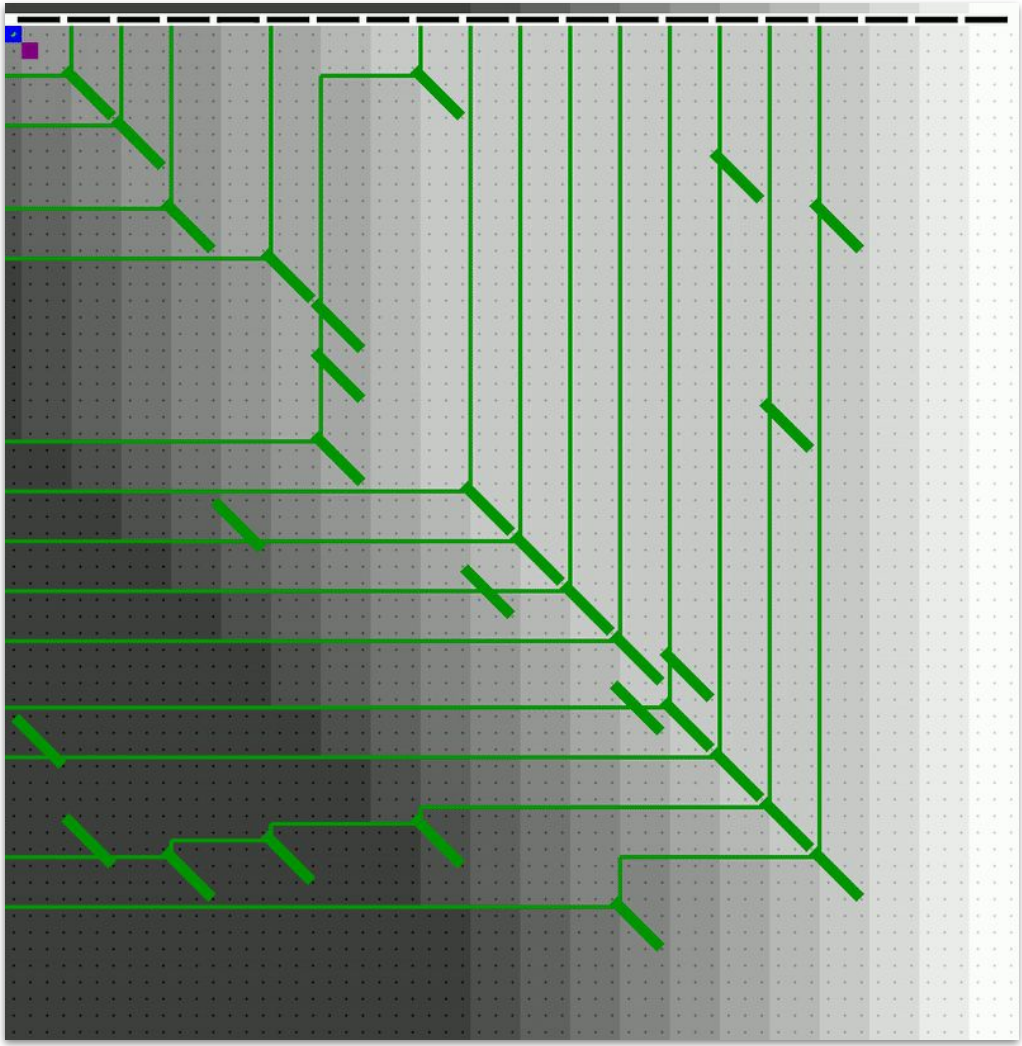
Match pruning



Exclude paths containing visited states => Prune visited matches.

$\hat{h}_{csh}$ = #{upcoming seeds} – #{longest chain of **unpruned** matches}

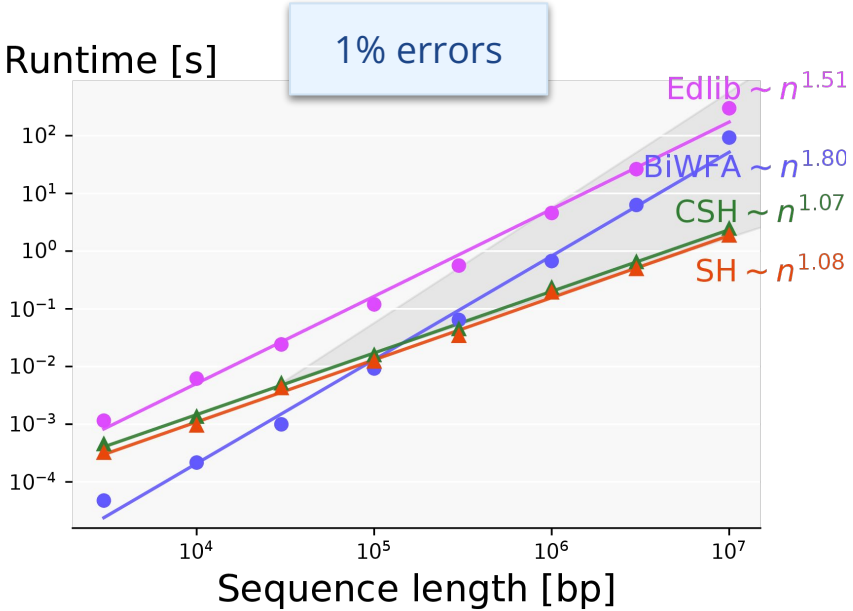
Demo

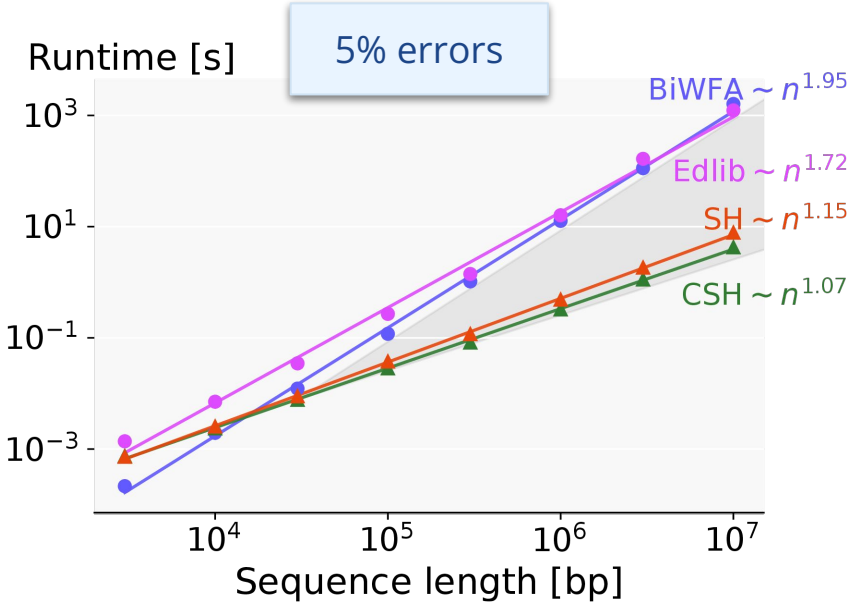


Visualization by
Mykola Akulov

- Blue: expanded
- Purple: explored
- Lime path: current state
- Background: heuristic
- Black: seeds
- Green: seed matches
- Green lines: contours
- Lime matches: pruned

Comparison on random data

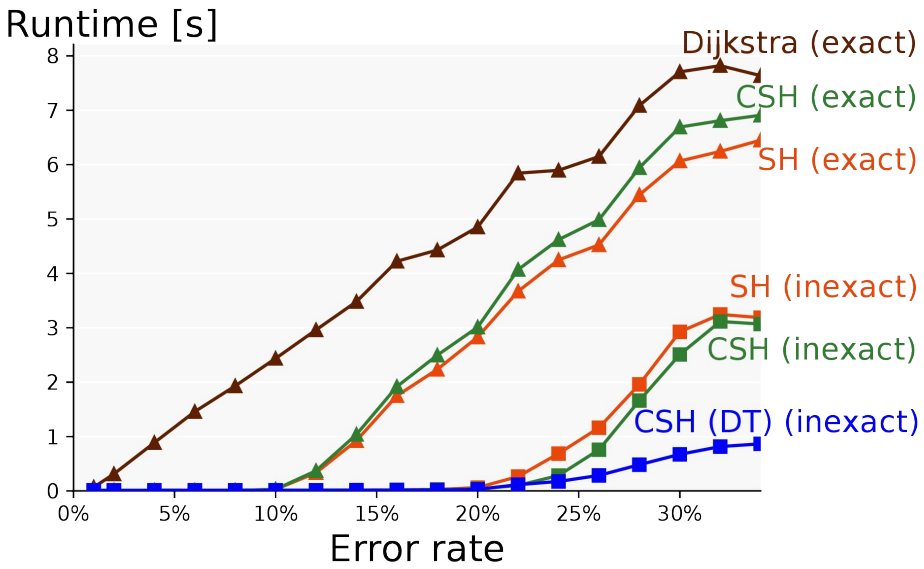




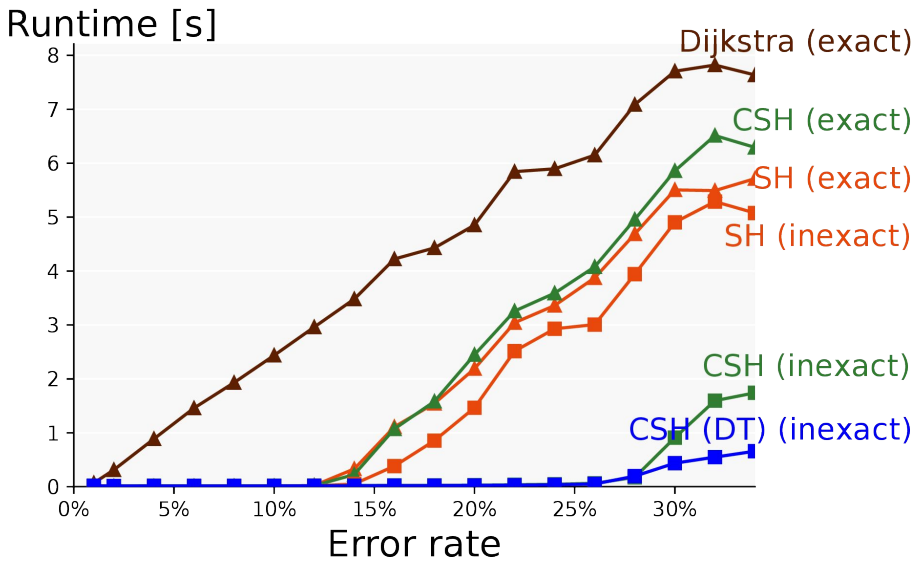
Seed length k=15
Exact matches

Randomness is important!
On human data:
~2x faster to ~100x slower

Scaling with error rate



n=10^4, seed length k=11



n=10^4, seed length k=9

Limitations & future work

Limitations

- Sequences must not be repetitive
- Mutations must be uniform random: No long indels!

Performance

- Use diagonal transition (work in progress)
- Variable seed length
- Gap cost for joining chains

Scope

- More cost models (affine costs)
- Semi-global alignment
- Extend to sequence-to-graph alignment

Prove expected linear time on random input

Q&A:
A* for Optimal Sequence Alignment

Pesho Ivanov and Ragnar Groot Koerkamp

ETH Zurich

eth-sri/astarix

@peshotrie

RagnarGrootKoerkamp/astar-pairwise-aligner/

@curious_coding

research.curiouscoding.nl

ETH zürich

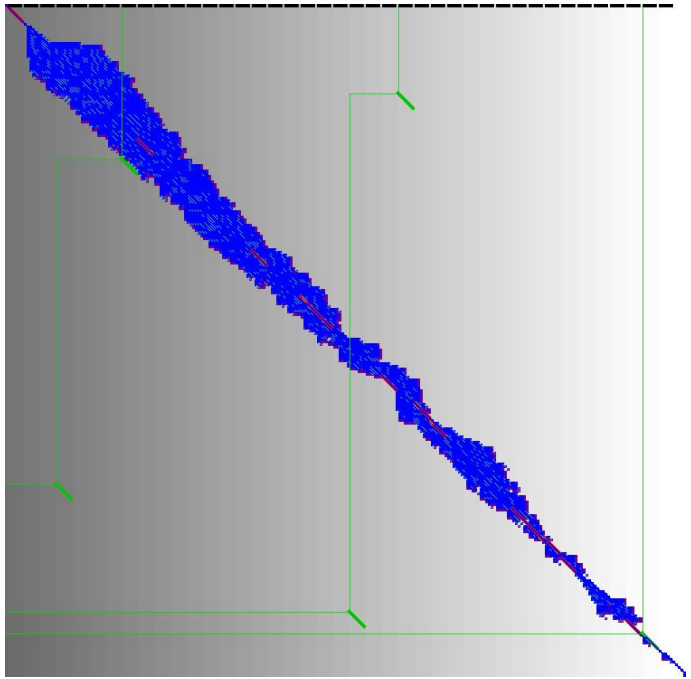
Q&A

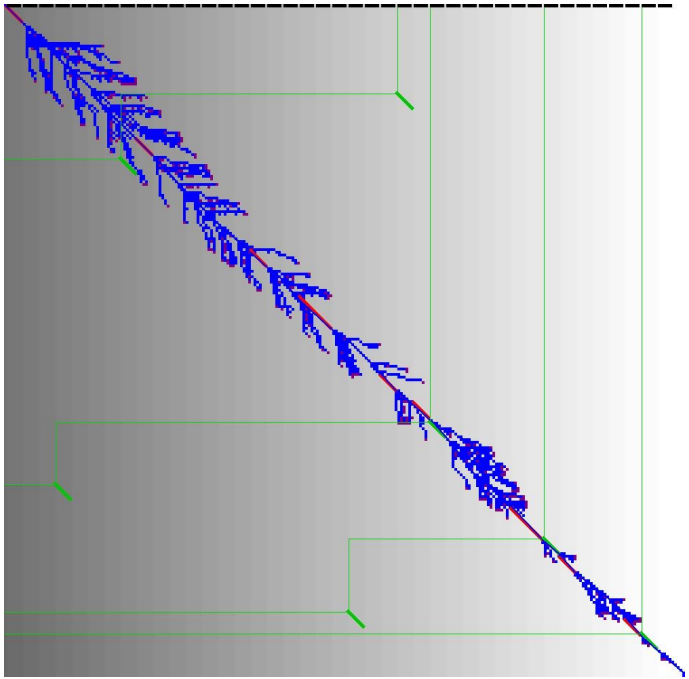
	DP Needleman, Wunsch'69	**Exponential band** Ukkonen'85	**Dijkstra** Ukkonen'85	**Diagonal Transition** Ukkonen'85, Myers'86	**+ Divide & Conquer** Hirschberg'75, Myers'86	**A* + SH + pruning** Groot Koerkamp, Ivanov'22
		Edlib Šošić, Šikić'17		**WFA** Marco-Sola et al'20	**BiWFA** Marco-Sola et al'22	**A*PA**
Expected[#] runtime	**O(n^2)**	**O(ns)**	**O(ns)**	**O(n+s^2)**	**O(n+s^2)**	**O(n)[*]?**
Expected[#] memory	**O(n^2)**	**O(ns) / O(n)**	**O(ns)**	**O(s^2)**	**O(s)**	**O(n)[*]?**

[#]: random string with random mutations

*: For n and e small enough

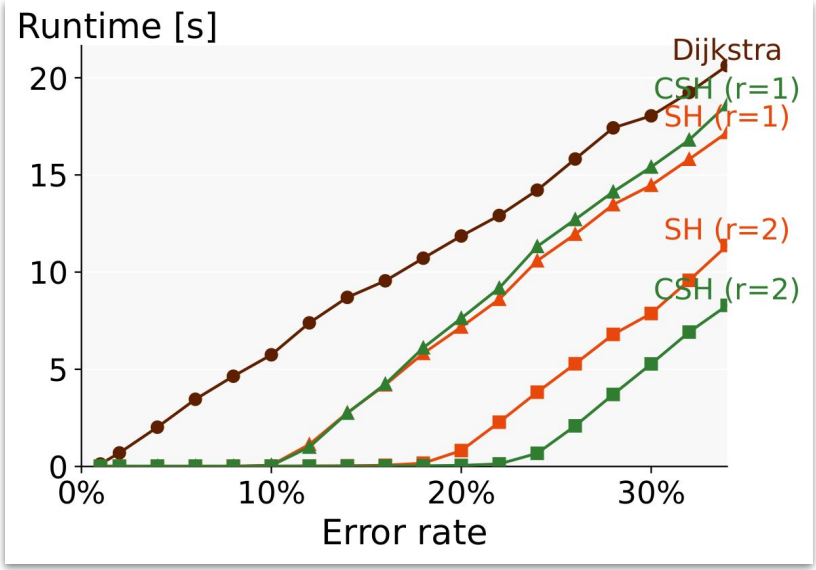
A* + Diagonal transition

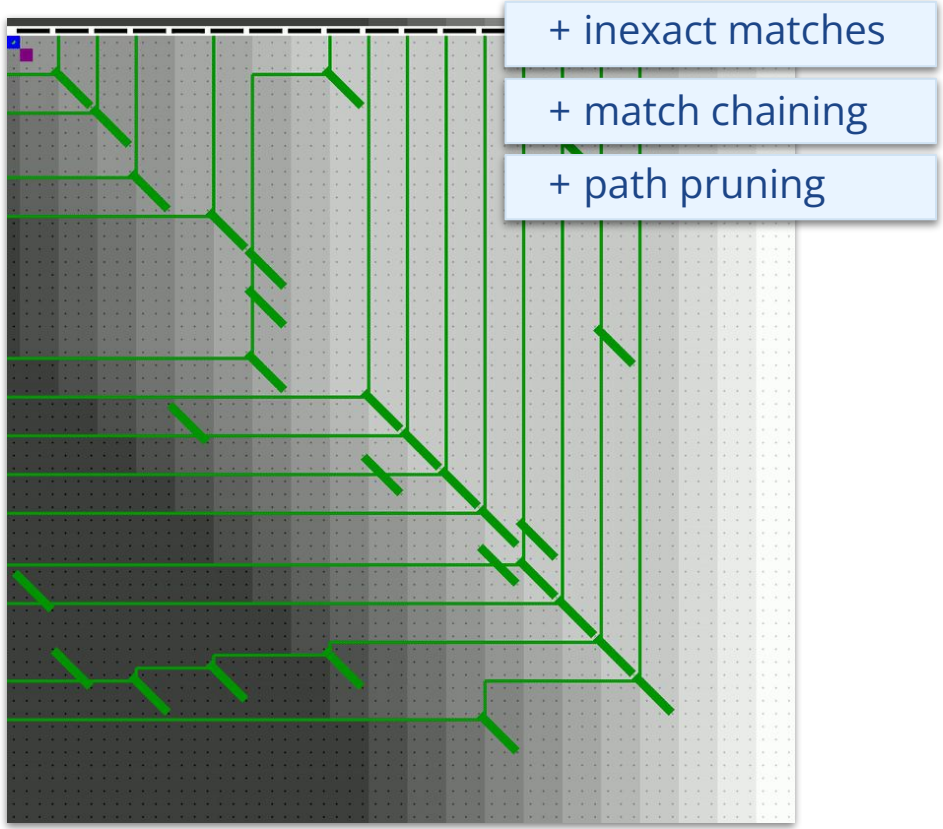




without/with DT
(n=250, e=30%, k=6)

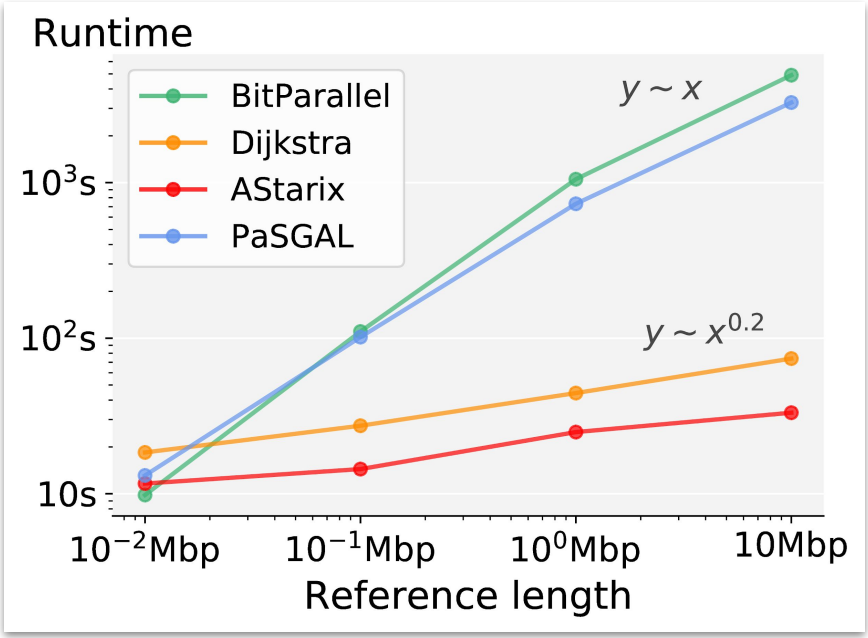
Scaling: error rate *(in progress)*

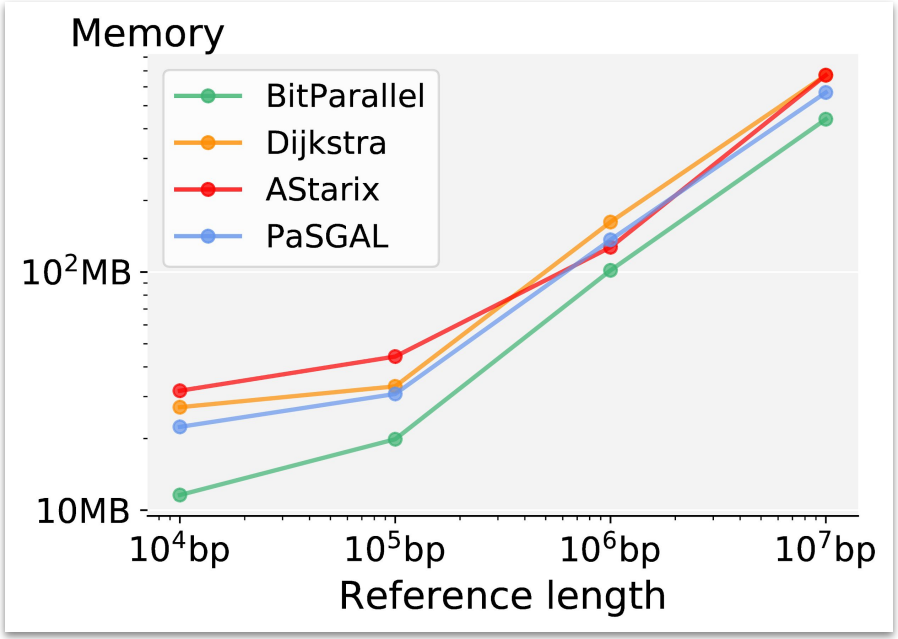




Joint work with **Ragnar Groot Koerkamp**
Visualization by Mykola Akulov
ETH Zurich

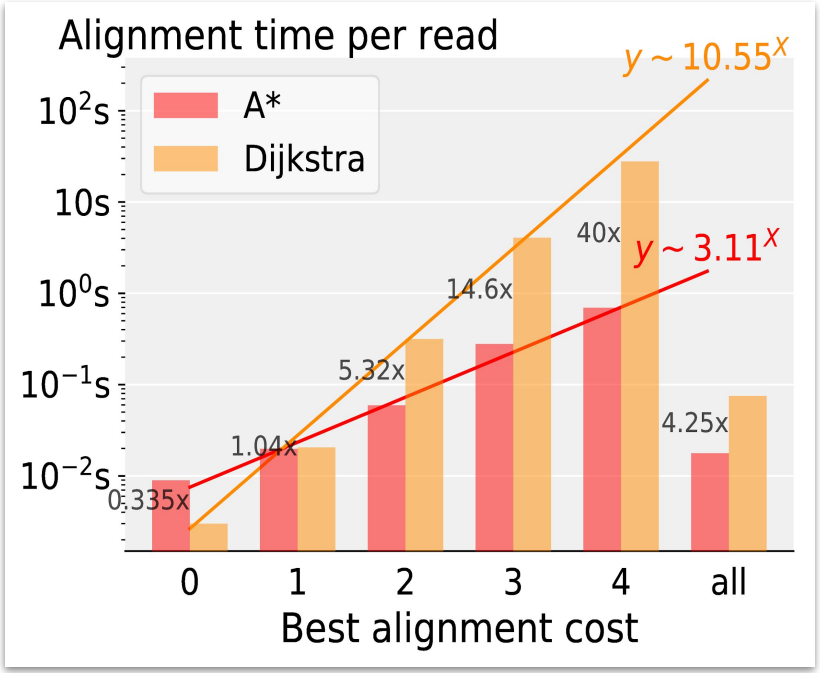
Prefix heuristic: Scaling with reference size

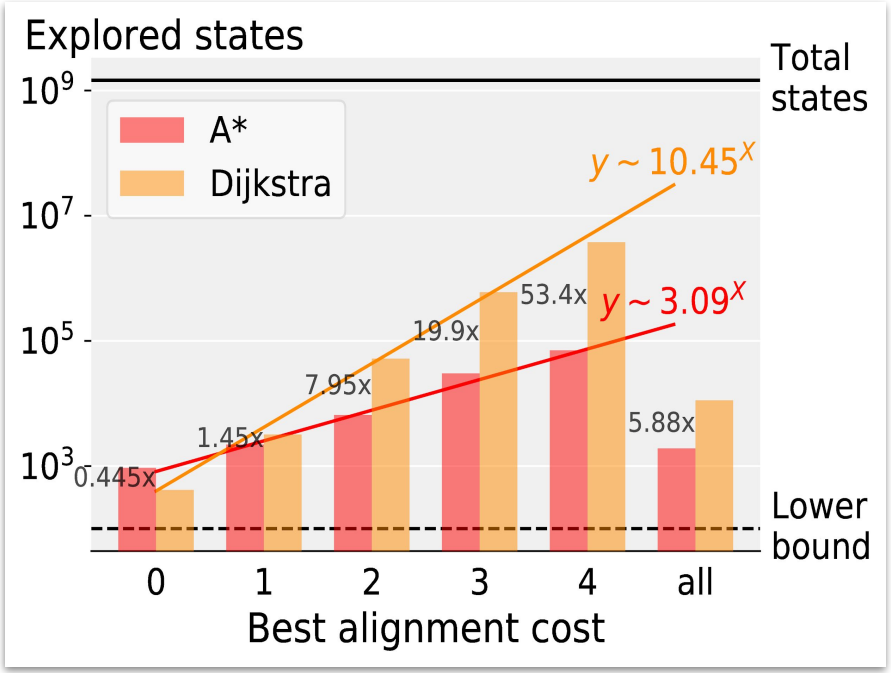




The reference length is the prefix of the linear E.coli used as a reference reference

Prefix heuristic: Alignment cost↗





`Dijkstra` reuses all codebase, optimizations and parameters for `A*`, except for the `A*`-specific.

Abstract. We present a novel A* *seed heuristic* enabling fast and optimal sequence-to-graph alignment, guaranteed to minimize the edit distance of the alignment assuming non-negative edit costs.

We phrase optimal alignment as a shortest path problem and solve it by instantiating the A* algorithm with our novel *seed heuristic*. The key idea of the *seed heuristic* is to extract *seeds* from the read, locate them in the reference, mark preceding reference positions by *crumbs*, and use the crumbs to direct the A* search. We prove admissibility of the *seed heuristic*, thus guaranteeing alignment optimality.

Our implementation extends the free and open source AStarix aligner and demonstrates that the *seed heuristic* outperforms all state-of-the-art optimal aligners including GraphAligner, Vargas, PaSGAL, and the *prefix heuristic* previously employed by AStarix. Specifically, we achieve a consistent speedup of $>60\times$ on both short Illumina reads and long HiFi reads (up to 25kbp), on both the *E. coli* linear reference genome (1Mbp) and the MHC variant graph (5Mbp). Our speedup is enabled by the *seed heuristic* consistently skipping $>99.99\%$ of the table cells that optimal aligners based on dynamic programming compute.

AStarix 2.0:
Scaling optimal sequence-to-graph alignment to long reads

presented by Pesho Ivanov
ETH Zurich

ETH zürich

1 Chop the read into seeds

Long read with unknown errors

Read chopped into seeds

$seed_1$ $seed_2$ $seed_3$ $seed_4$

2 Generate all seeds at distance ≤ 2

Generate all seed modifications

no modifications

1 modification

All modified seeds
(only substitutions shown)

2 modifications

Throw away most seeds as they do not appear in the reference

/dev/null
Default penalty ≥ 3

Filter only the seeds that appear in the reference

3 Map the modified seeds exactly

Modified seeds that appear in the reference

exact match
$\Rightarrow$ no penalty

1 modification
$\Rightarrow$ penalty ≥ 1

2 modifications
$\Rightarrow$ penalty ≥ 2

Exact mapping of modified seeds to the reference

Parameters

```
Match cost: 0                    Seeds: 4
Mismatch cost: 1          Max. #corrections: 2
Indel cost: +∞ (no indels)   ⇒ Max. penalty: 3
```

```
Potential = Seeds * Max.Penalty * Max.edit cost
```

This heuristic can "pay" for $\leq$**12 errors**

Trie
with Seeds heuristic values for the specific read

4 Update the heuristic values while climbing up (independently for each seed)

Links to the reference

Seed "votes" for a start

Misleading matches resulting in noise

12 12 12 12 12 12 12 12 12 12 12 12 12 12

...CCGAT...

Rerefence
rawn linear for simplicity)

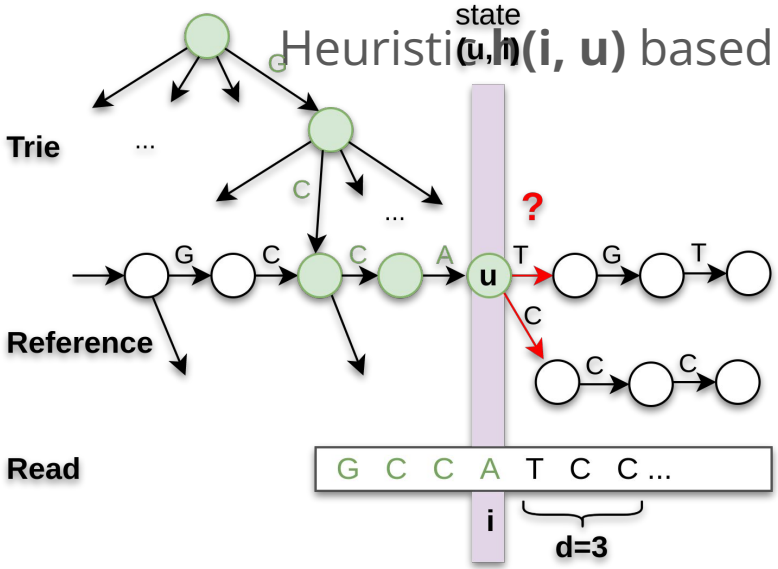
Legend

- 4, 12 Heuristic value from 0 to the potential
- Read errors not known in advance (only substitutions shown for simplicity)
- Incorrect correction in a seed
- Correct correction in a seed
- Corrected seed seen in the reference
- Corrected seed not seen in the reference

A*-ix

https://github.com/eth-sri/astarix

AStarix: A* *prefix* heuristic

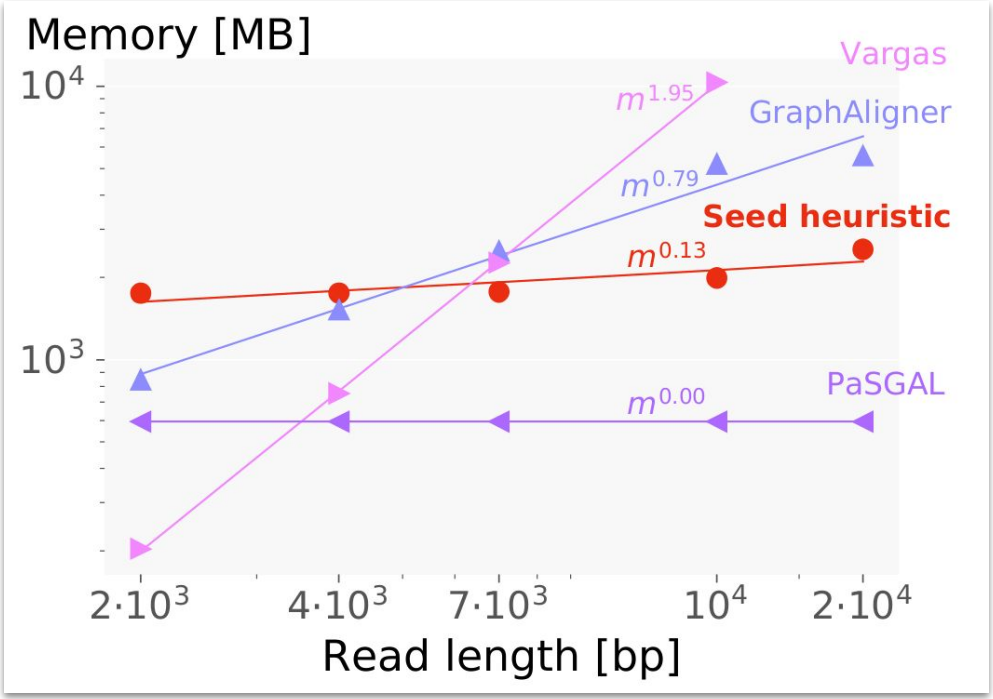


Heuristic $h(i, u)$ based on the ***d*** following read nucleotides

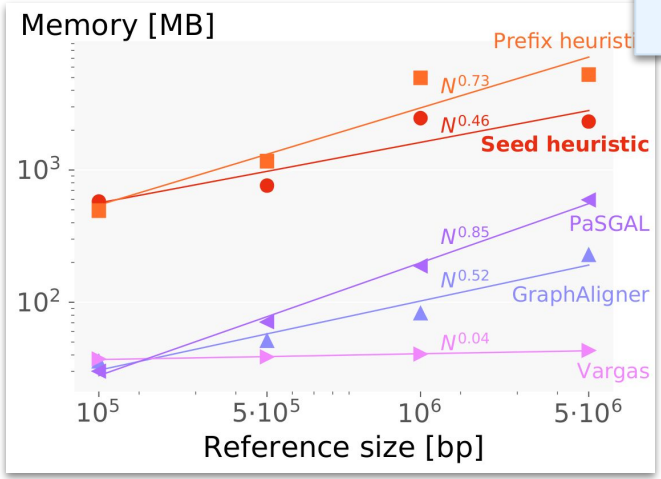
Algorithm 3 Recursive alignment used by Heuristic in Algorithm 1.

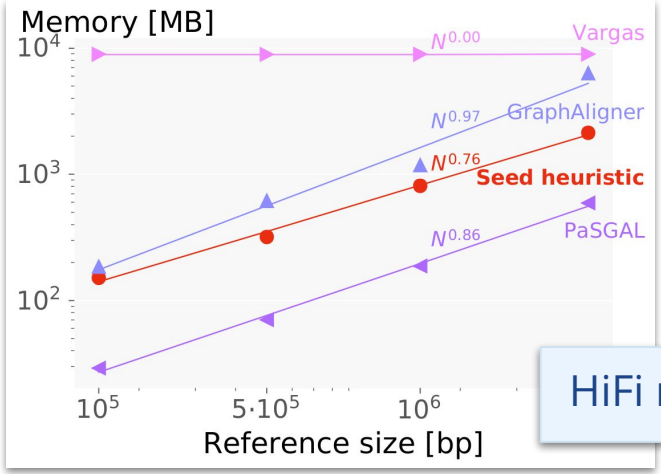
```
1: function RecursiveAlign(u, s, curr, best)          ▷ Return value is ≤ best
2:     if curr ≥ best then
3:         return best                                 ▷ Branch and bound: bounding
4:     if s = ϵ then                                   ▷ Reached a target
5:         return curr
6:     for all (u, v, ℓ, w) ∈ E_e where ℓ ∈ {s[0], ϵ} do
7:         suff = s[1 :] if ℓ ≠ ϵ else s
8:         best = RecursiveAlign(u, suff, curr + w, best)
9:     return best
```

Memory



Illumina reads





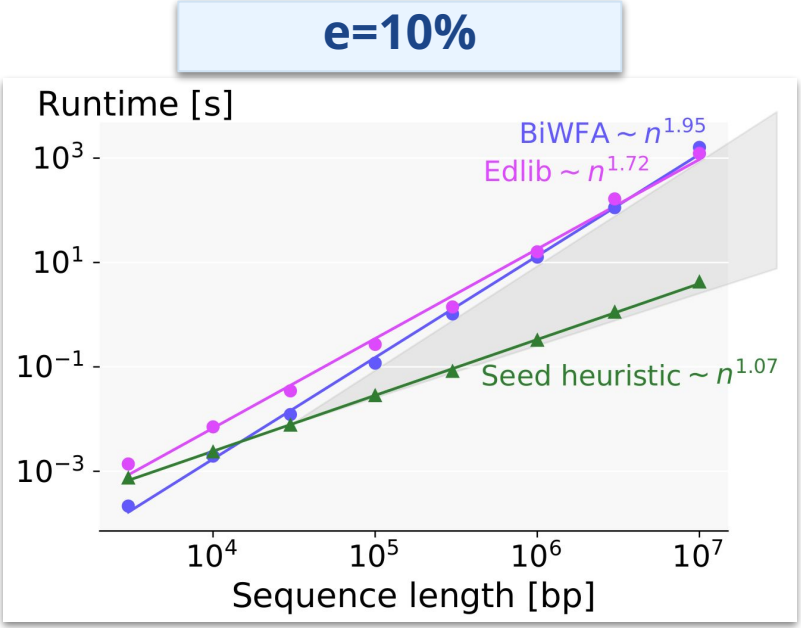
HiFi reads

Results

Tool	Illumina		HiFi		
	E. coli	MHC	*E. coli*	MHC	
Seeds heuristic (this work)	0.019	0.041	0.001	0.002	s/kbp
	2.4	2.6	2.4	1.7	GB
	99.9996	99.9981	99.9989	99.9984	% skipped states
Prefix heuristic	269x	180x	n/a	n/a	x slowdown
	7.7	9.6	>20	>20	
	99.9501	99.9501	n/a	n/a	
GraphAligner	424x	212x	118x	64x	
	0.2	0.2	3.6	3.4	
Vargas	133x	67x	1 413x	705x	
	<0.1	<0.1	7.3	7.3	
PaSGAL	263x	130x	1 367x	736x	
	0.6	0.6	0.6	0.6	

References: *E. coli* – linear 4.6M, Major Histocompatibility Complex (MHC) – 5.3M
Simulated queries: 200bp Illumina with Δ=(0,1,5,5); HiFi: 5–25kbp, e=0.3%, Δ=(0,1,1,1)
Prefix heuristic parameters: length cap d=5, cost cap c=5, trie depth D=log(N)
Seed heuristic parameters: D=14 ≈ $\log_4 N$; k=25 for Illumina, k=150 for HiFi reads

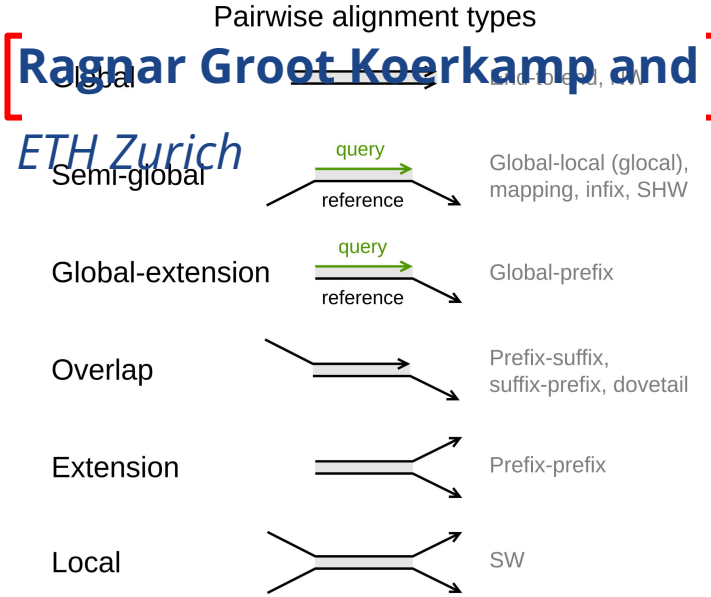
Near-linear scaling

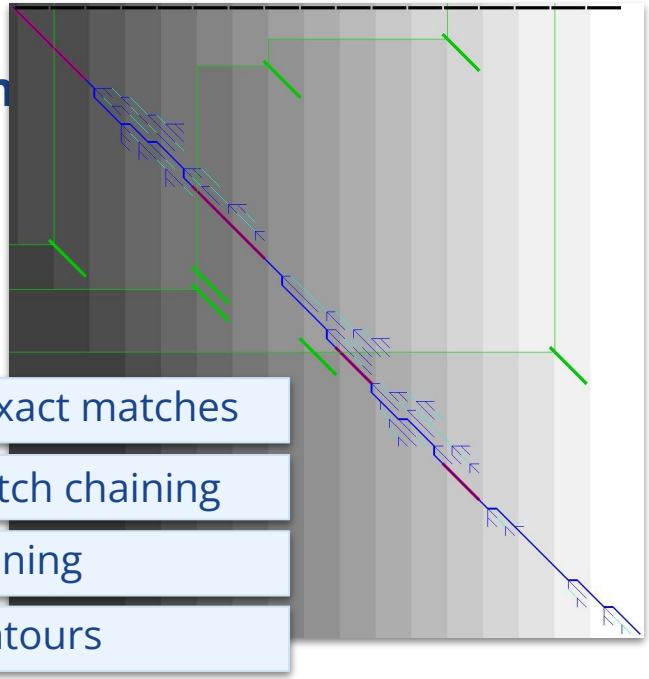


Seed heuristic for global alignment (unpublished)

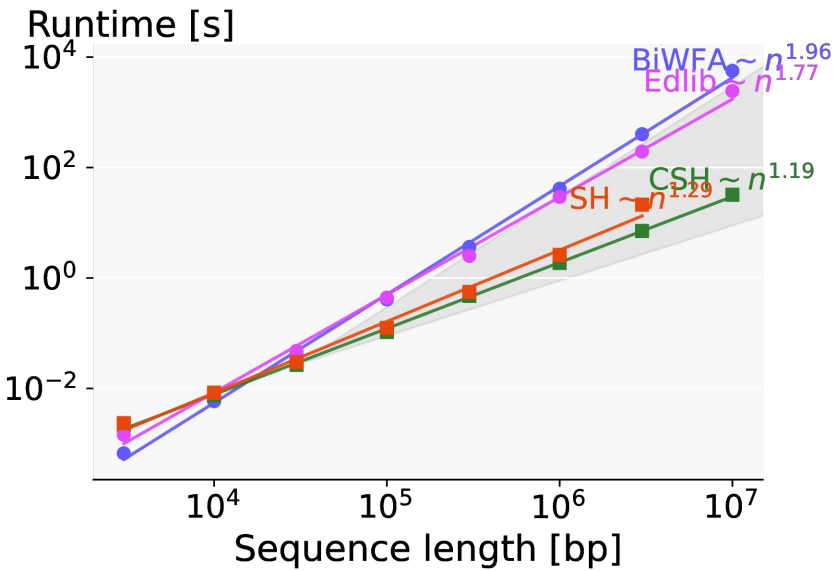
Ragnar Groot Koerkamp and Pesho Ivan

ETH Zurich

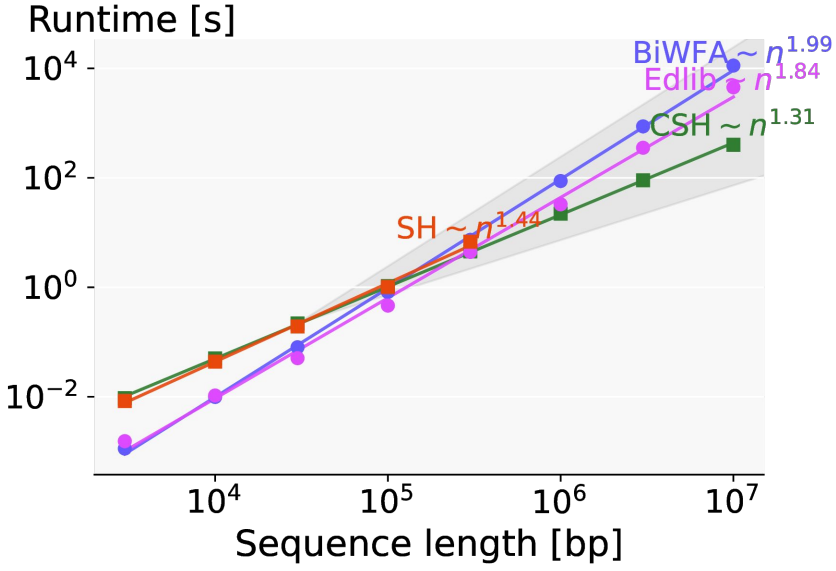




Comparison for high error rate

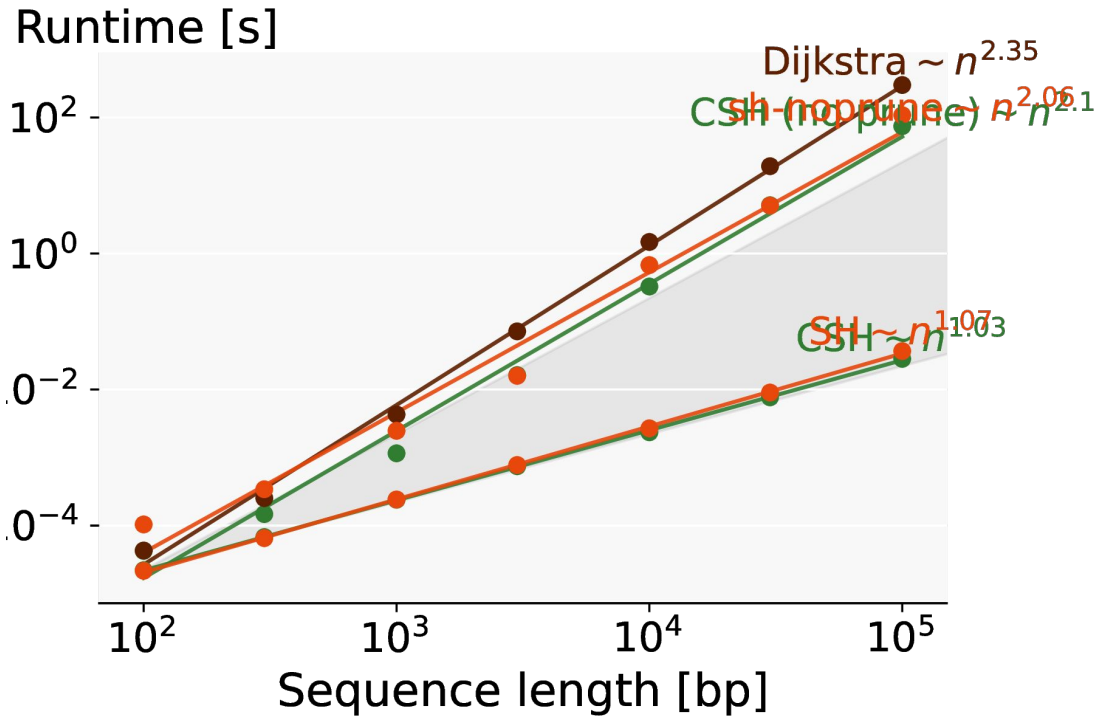


Error rate: 10%
Seed length: 15
Inexact matches



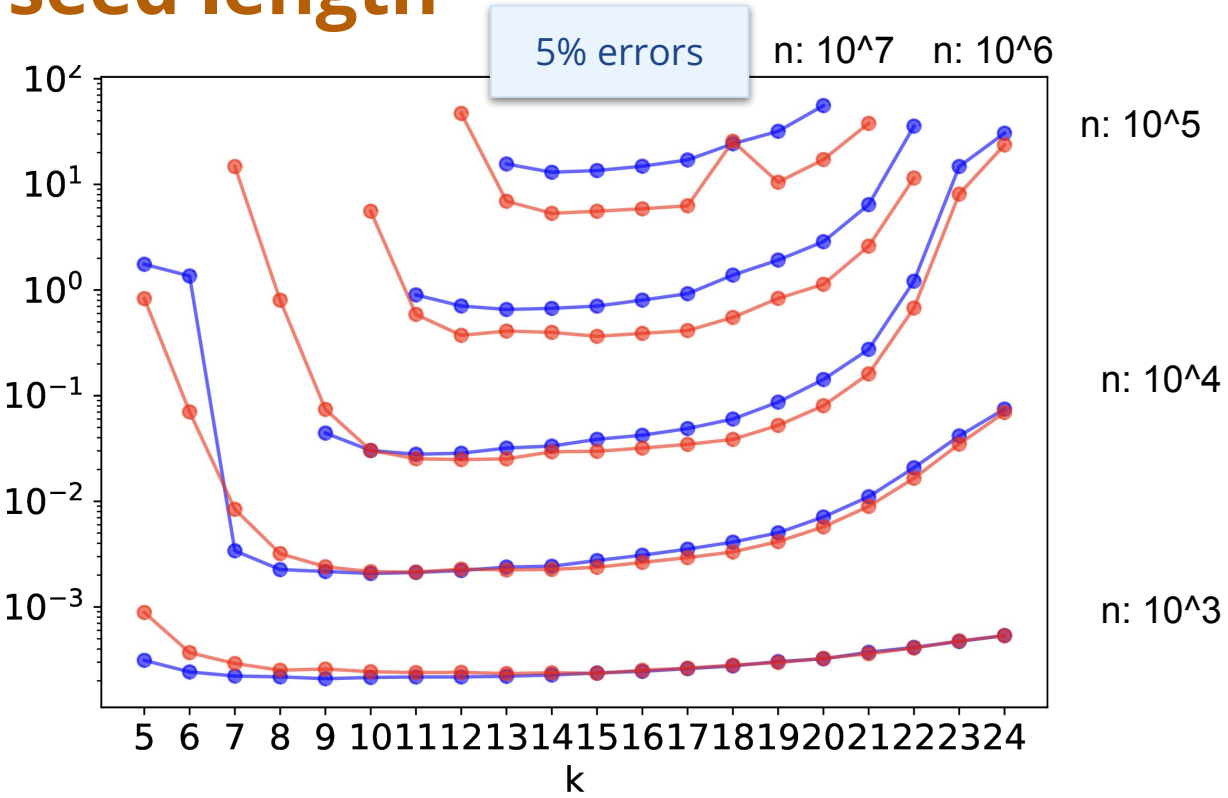
Error rate: 15%
Seed length: 15
Inexact matches

Effect of pruning



Error rate: 5%, seed length: 15

Choosing seed length



log(n) < k < 1/e