



ALIGNMENT

MY INTRODUCTION TO BIOLOGICAL
SEQUENCE ALIGNMENT

About me

- Dr. Benjamin J. Buchfink, computer scientist and mathematician by training
- Started working on sequence alignment in 2012
- Developed the popular DIAMOND tool (12 citations/day) for protein alignment
- Mostly worked in academic research since then, at the University of Tübingen and the Max Planck institute for biology Tübingen
- Read more:
- Buchfink et al., Nature Methods 2014
- Buchfink et al., Nature Methods 2021
- Buchfink et al., Nature Methods 2026 (to appear)

Questions to the audience

1. Have you ever **run** a sequence alignment software?
2. Was it **directly** or as part of a pipeline?
3. Done **multiple** sequence alignment?
4. Know who/what **Smith-Waterman** is?
5. What is the difference between similarity and homology?
6. Why align? Variant calling, RNA-seq quantification, taxonomic classification, etc.
7. Slides adopted from Rayan Chikhi (last year)

Types of alignments

- Pairwise

meiotic recombination protein REC8 homolog [Lepidochelys kempii]

Sequence ID: [XP_073164915.1](#) Length: 573 Number of Matches: 1

Range 1: 340 to 561 [GenPept](#) [Graphics](#)

▼ [Next Match](#) ▲ [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
160 bits(405)	3e-47	Compositional matrix adjust.	115/241(48%)	145/241(60%)	22/241(9%)
Query 35	RRRLLFWDKETQISPEKFQEQ-LQTRAHCWECPMVQPPERTIRGPAELFRTPTLSGWLPP	93			
	RR+L F D+ TQI ++F++Q L RA C +V+ R R PAEL R PT GWLPP				
Sbjct 340	RRQLRFLDEVTQIPRDQFKKQILDVRAQCRPLVVVEVAARQRRTPAELLRAPTY-GWLPP	398			
Query 94	ELLGLWTHCAQPPPKALRRELPEEAAAAEEERRKIEVPSEIEVPREALEPSVPLMVSLEIS	153			
	L LW HCA + R + A EE RK E SE+EV REALEPS+P+MVS EIS				
Sbjct 399	ALHALWQHCA-----ILRPVDYAALWAEERKEEPVSELEVAREALEPSIPVMVSSEIS	452			
Query 154	LEAAAAEKSRLSLIPPEERWAWPEVEAPEAPALPVVPELPEVPMEMPLVLPPELELLSLE	213			
	LE EEE R SL+ PEER PE E ALP+VPELPEV +E LPP+ +L++LE				
Sbjct 453	LETTEEEVPRPSLVTPERRLVPEPEE-----ALPIVPELPEVSLE-----LPPDRDLITLE	504			
Query 214	AVHRAVALELQANREPDFSSLVSPSPRRMAARVFYLLLGECMHVCVCMWGR--DTETRG	271			
	+ R VA EL+ E DF SLV + R +A+R+FYL C+ +C + + TE G				
Sbjct 505	YIRRLVAAELEQVGEMDFRSLVPTTTSRGIASRIFYL-----CLVLCGLQFLQLHQTEPYG	560			
Query 272	P 272				
	P				
Sbjct 561	P 561				

Types of alignments

- Multiple

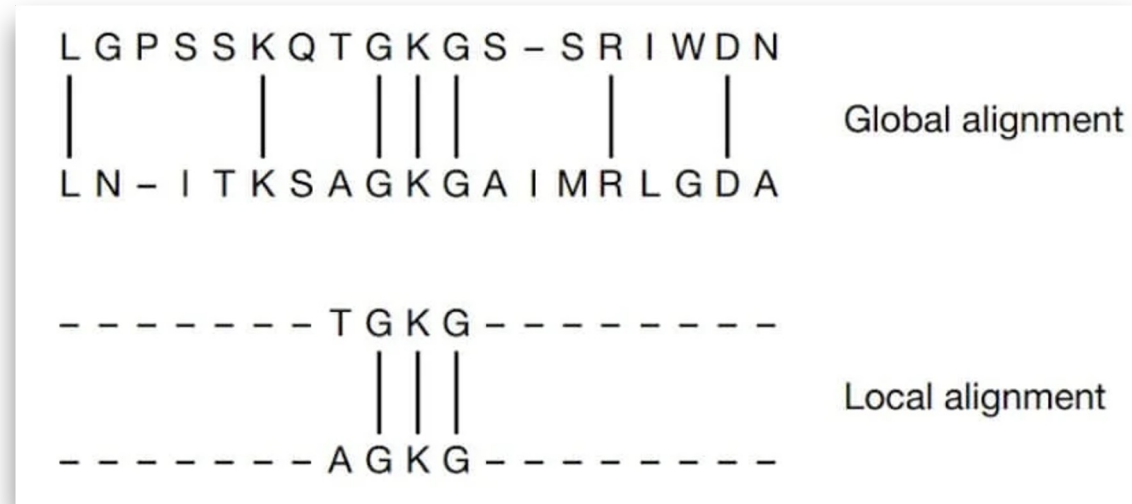
```

      *      .      :      .      *      :      :
Q5E940_BOVIN -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_HUMAN -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_MOUSE -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_RAT -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_CHICK -----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_RANSY -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--SALE 76
Q7ZUG3_BRARE -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_ICTPU -----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PALE 76
RLA0_DROME -----MVRENKAAWKAQYFIKVVLFDFEPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLVGKNTMMRKAIRGHLENN--PQLE 76
RLA0_DICDI -----MSGAG-SKRKKLFIEKATKLTFTYDKMIVAEADFGSSQLQKIRKSIRGI-GAVLMGKKTMRKVIRDLADSK--PELD 75
Q54LP0_DICDI -----MSGAG-SKRKNVFIKATKLTFTYDKMIVAEADFGSSQLQKIRKSIRGI-GAVLMGKKTMRKVIRDLADSK--PELD 75
RLA0_PLAF8 -----MAKLSKQKKQMYIEKLSLIQQYSKILIVHYDNVGSNOMASVRKSLRGK-ATILMGKNTIRITALKKNLQAV--PQIE 76
RLA0_SULAC -----MIGLAVTTTKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFNIALKNAG----YDTK 79
RLA0_SULTO -----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNLTFGIAAKNAG----LDVS 80
RLA0_SULSO -----MKRLALALKQKRVASWKLIEEVKELTELIKNSNTILIGNLEGGFPADKLHEIRKKLRGK-ADIKVTKNLTFKIAAKNAG----IDIE 80
RLA0_AERPE MSVVS LVGQMYKREKPIPEWKTLMRLRELELFSKHRVVFADLTGTPTFVVRVRKKLWKK-YPMVAKKRILRAMKAAGLE---LDDN 86
RLA0_PYRAE MMLAIGKRRYVRTQYYPARKVKIVSEATELLQKYVYVFLFDLHGLSSRLIHEYYRRLRY-GVIKIIKPLFKIAFTKVYGG---IPAE 85
RLA0_METAC -----MAEERHHTHEIPQWKDEIENIKELIQSHKVFQMGVIEGILATKMKIRRDLDV-AVLKVSRTTLTERALNQLG----ETIP 78
RLA0_METMA -----MAEERHHTHEIPQWKDEIENIKELIQSHKVFQMGVIEGILATKMKIRRDLDV-AVLKVSRTTLTERALNQLG----ESIP 78
RLA0_ARCFU -----MAAVRGS---PPEYKVRAVEEIKRMISKPVVAIVSFRNVPAGOMKIRREFRGK-AEIKVVKNLLELALDALG----GDYL 75
RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELELMDEYENVGLVDLEGIPAPOLQEIIRAKLRERDTIIRMSRNTLMRIAEEKLDER--PELE 88
RLA0_METTH -----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPAROLQKMRQLDS-ALIRMSKKTLLISLAEKAGREL--ENVD 74
RLA0_METTL -----MITAESEHKIAPWKIEEVNKLKELKNGQIVALVDMMEVPAVOLQEIIRDKIR-GTMFLKMSRNTLIERAIEVAEETGNPEFA 82
RLA0_METVA -----MIDAKSEHKIAPWKIEEVNALKELKLSANVIALIDMMEVPAVOLQEIIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA 82
RLA0_METJA -----METKVAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPOLQEIIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA 81
RLA0_PYRAB -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRILIRENGGLLRVSRNTLIELAIKKAQELGKPELE 77
RLA0_PYRHO -----MAHVAEWKKKEVEELAKLIKSYPIVIALVDVSSMPAYPLSQMRRILIRENGGLLRVSRNTLIELAIKKAQELGKPELE 77
RLA0_PYRFU -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRILIRENGGLLRVSRNTLIELAIKKAQELGKPELE 77
RLA0_PYRKO -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVAGVPAYPLSKMRDLK-GKALLRVSRNTLIELAIKKAQELGQPELE 76
RLA0_HALMA -----MSAESERKTETIPEWKQEEVDIVMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLELALDDVD----DGLE 79
RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDLVDFIESYESVGVVGAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRLDEVN----DGFE 79
RLA0_HALSA -----MSAEQRTTEEVPEWKQEEVAELVDLLETYDSVGVVNVTCIPSKOLQDMRRGLHCG-AALRMSRNTLLVRALEEAG----DGLD 79
RLA0_THEAC -----MKEVSQKKELVNETTORIKASRSVAIVDTAGIRTRQIDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS 72
RLA0_THEVO -----MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTROMQDIRAKNRDK-VKIKVVKKTLFLKALDSIND----EKLT 72
RLA0_PICTO -----MTEPAQWKIDFVKNELEINSRKVAAIVSKGLRNNFQKIRNSIRDK-ARIKVSRAILLRLAIENTGK----NNIV 72
ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90
```

Terminology

- Query: sequence to align
- Reference (or target, subject): other sequence to align to
- Hit (or match or alignment): part of query aligned to part of reference
- Homology: shared ancestry
- Similarity, identity: mathematical ways to detect homology
- Letter (or residue or monomer): base pair or nucleotide or amino-acid
- CIGAR string (or edit transcript): series of edit operations to transform one sequence into the other

Global vs local



<https://microbenotes.com/local-global-multiple-sequence-alignment/>

Alignment is based on scoring

- What is a *good* alignment?
One that *minimizes* a *penalty* (or *maximizes* a score).
- Example for DNA alignment:
- Match score +1 (M)
- Mismatch penalty -1 (X)
- Gap penalty -2
Two kinds of gaps:
- Deletion from the reference (D)
- Insertion into the reference (I)

Needleman-Wunsch

- Make a dynamic programming matrix

reference

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2					
T	-4					
C	-6					
C	-8					

query

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1				
T	-4					
C	-6					
C	-8					

Each cell is the optimal alignment score of [query up to this row] vs [reference up to this column]

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1			
T	-4					
C	-6					
C	-8					

AG

A-

MD

score: -1

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1			
T	-4	-1				
C	-6					
C	-8					

A-

AT

MI

score: -1

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1			
T	-4	-1	0			
C	-6					
C	-8					

Three possibilities:

MX -> score 0

MDI -> score -3

MID -> score -3

Needleman Wunsch

		-	A	G	T	C	A
	-	0	-2	-4	-6	-8	-10
M	A	-2	1	-1	-3	-5	-7
	T	-4	-1	0			
	C	-6	-3				
MIII	C	-8	-5				

← M ← MIII

→ MX → MDDD

Insight: each filled cell corresponds to the score of an optimal alignment of ref/query so far

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
T	-4	-1	0	?		
C	-6	-3				
C	-8	-5				

Three possibilities:

AGT- or AGT or AGT

A--T A-T AT-

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
T	-4	-1	0	0	-4	-6
C	-6	-3	-2	-1	1	-1
C	-8	-5	-4	-3	0	0

Needleman Wunsch

	-	A	G	T	C	A
-	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
T	-4	-1	0	0	-4	-6
C	-6	-3	-2	-1	1	-1
C	-8	-5	-4	-3	0	0

Then the alignment is
the CIGAR string at the bottom right cell.
It traces back to the top left cell:

←
MDMMX

AGTCA
A-TCC

Smith Waterman

	-	A	G	T	C	A
-	0	0	0	0	0	0
A	0	0	0	0	0	1
T	0	0	0	1	0	0
C	0	0	0	0	2	0
C	0	0	0	0	0	1

1. Cells cannot be negative
2. Find the highest scoring cell
3. Trace it back to a zero

Here: TC aligned to TC (.. how surprising)

$$H_{i,j} = \begin{cases} \max \begin{cases} H_{i-1,j-1} + P[q_i, d_j] \\ E_{i,j} \\ F_{i,j} \\ 0 \end{cases} & \begin{cases} i > 0 \\ \cap \\ j > 0 \end{cases} \\ 0 & \begin{cases} i = 0 \\ \cup \\ j = 0 \end{cases} \end{cases} \quad (1)$$

$$E_{i,j} = \begin{cases} \max \begin{cases} H_{i,j-1} - Q \\ E_{i,j-1} - R \end{cases} & |j > 0 \\ 0 & |j = 0 \end{cases} \quad (2)$$

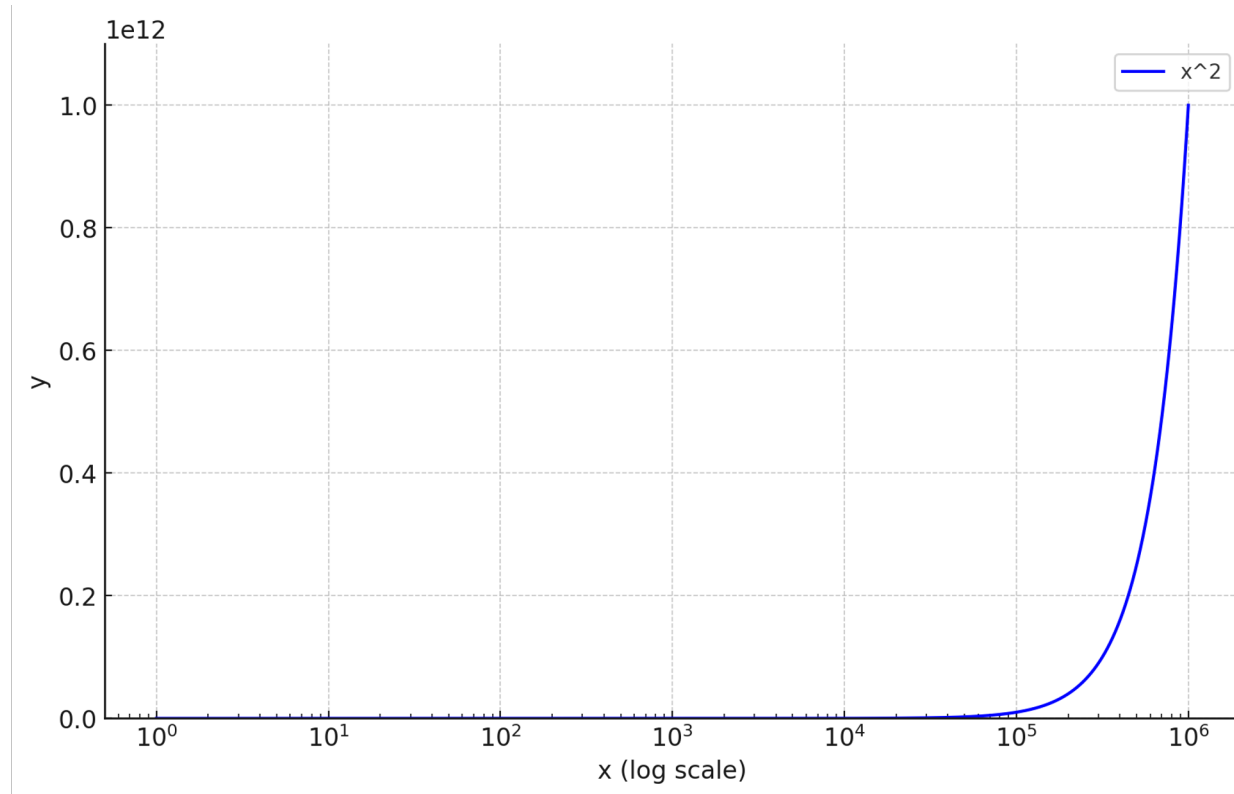
$$F_{i,j} = \begin{cases} \max \begin{cases} H_{i-1,j} - Q \\ F_{i-1,j} - R \end{cases} & |i > 0 \\ 0 & |i = 0 \end{cases} \quad (3)$$

$$S = \max_{1 \leq i \leq m \cap 1 \leq j \leq n} H_{i,j} \quad (4)$$

Why can't we Smith Waterman everything?

It requires $(n*m)$ operations, where n and m are the sequence lengths.

When $n \sim m$, it's n^2 operations:



The BLAST algorithm

```
Query= d5gy3a_ a.102.1.0 (A:) automated matches {Klebsiella pneumoniae [TaxId: 573]}

>d1h12a_ a.102.1.2 (A:) Endo-1,4-beta-xylanase {Pseudoalteromonas haloplanktis [TaxId: 228]}
Length=404

Score = 45.8 bits (107), Expect = 1.3e-05
Identities = 53/195 (27%), Positives = 84/195 (43%), Gaps = 37/195 (18%)

Query    7  YKARFMPDPGRIIDTANGNVSHTEGQGFAMLLAVANNDRPAFDKLWQWTDSTLRD----- 61
          YKA ++    + I+   G+   TEGQ + M  AV  N +   FD LW++  +   ++
Sbjct    59  YKAHYI----KAINPDEGGDDIRTEGQSWGMTAAVMLNKQEEFDNLWRFKAYQKNPDNHP 114

Query    62  --KSNGLFYW--RYNPVAPDPIADKNNASDGDTLIAWALLRAQKQWQDKRRYAIASDAIT 116
          K  G++ W   + N           D+  A DG+   A+ALL A  +W +   +   +DAIT
Sbjct   115  DAKKQGVYAWKLKLNQNGFVYKVDEGPAPDGEEYFAFALLNASARWGNSGEFNYYNDAIT 174

Query   117  A-----SLLKYTVVTFAGRQVMLPGVKGFNLNDHLNLP SYFIFPAWRAFAERTHLTA- 169
          L++  ++ F+       P +   NL D       PSY I   + FA       A
Sbjct   175  MLNTIKNKLMEHQIIRFS-----PYID--NLTD-----PSYHIPAFYDYFANNVTNQAD 221

Query   170  ---WRTLQTDGQALL 181
          WR + T   + LL
Sbjct   222  KNYWRQVATKSRTLL 236
```

The BLAST algorithm

I. Seed: Find two neighborhood 3-mer matches within a window of 40 on the same diagonal

```
Query= d5gy3a_ a.102.1.0 (A:) automated matches {Klebsiella pneumoniae [TaxId: 573]}
```

```
>dlh12a_ a.102.1.2 (A:) Endo-1,4-beta-xylanase {Pseudoalteromonas haloplanktis [TaxId: 228]}
Length=404
```

```
Score = 45.8 bits (107), Expect = 1.3e-05
Identities = 53/195 (27%), Positives = 84/195 (43%), Gaps = 37/195 (18%)
```

```
Query    7  YKARFMPDGRIIDTANGNVSHTEGQGFAMLLAVANNDRPAFDKLWQWTDSTLRD----- 61
      YKA ++    + I+    G+    TEGQ + M  AV  N +  FD LW++  +  ++
Sbjct   59  YKAHYI----KAINPDEGDDIRTEGQSWGMTAAVMLNKQEEFDNLWRFAKAYQKNPDNHP 114

Query   62  --KSNGLFYW--RYNPVAPDPIADKNNASDGDTLIAWALLRAQKQWQDKRRYAIASDAIT 116
      K  G++ W  + N          D+  A DG+  A+ALL A  +W +  +  +DAIT
Sbjct  115  DAKKQGVYAWKLKLNQNGFVYKVDEGPAPDGEEYFAFALLNASARWGNSGEFNYYNDAIT 174

Query  117  A-----SLLKYTVVTFAGRQVMLPGVKGFNLNDHLNLPSTYFIFPAWRAFAERTHLTA- 169
      L++  ++ F+          P +   NL D      PSY I   +  FA      A
Sbjct  175  MLNTIKNKLMEHQIIRFS-----PYID--NLTD-----PSYHIPAFYDYFANNVTNQAD 221

Query  170  ---WRTLQTDGQALL 181
      WR + T  + LL
Sbjct  222  KNYWRQVATKSRTLL 236
```

The BLAST algorithm

2. Extend using gapped X-drop

```
Query= d5gy3a_ a.102.1.0 (A:) automated matches {Klebsiella pneumoniae [TaxId: 573]}

>dlh12a_ a.102.1.2 (A:) Endo-1,4-beta-xylanase {Pseudoalteromonas haloplanktis [TaxId: 228]}
Length=404

Score = 45.8 bits (107), Expect = 1.3e-05
Identities = 53/195 (27%), Positives = 84/195 (43%), Gaps = 37/195 (18%)

Query      7  YKARFMPDGRIIDTANGNVSHTEGQGFAMLLAVANNDRPAFDKLWQWTDSTLRD----- 61
           YKA ++    + I+    G+    TEGQ + M AV  N +  FD LW++  +  ++
Sbjct     59  YKAHYI----KAINPDEGDDIRTEGQSWGMTAAVMLNKQEEFDNLWRFAKAYQKNPDNHP 114

Query     62  --KSNGLFYW--RYNPVAPDPIADKNNASDGDTLIAWALLRAQKQWQDKRRYAIASDAIT 116
           K  G++ W  + N          D+  A DG+ A+ALL A  +W +  +  +DAIT
Sbjct    115  DAKKQGVYAWKLKLNQNGFVYKVDEGPAPDGEEYFAFALLNASARWGNSGEFNYYNDAIT 174

Query    117  A-----SLLKYTVVTFAGRQVMLPGVKGFNLNDHLNLPSTYFIFPAWRAFAERTHLTA- 169
           L++  ++ F+          P +  NL D      PSY I  +  FA      A
Sbjct    175  MLNTIKNKLMEHQIIRFS-----PYID--NLTD-----PSYHIPAFYDYFANNVTNQAD 221

Query    170  ---WRTLQTDGQALL 181
           WR + T  + LL
Sbjct    222  KNYWRQVATKSRTL 236
```

Anything can align to anything

Two random DNA sequences:

ATTTTAGGGGGG-GAAGGTTG-

| | | | | | | |

GCG--AGGGCCGTGTTGCCGGT

- Be careful of “coercing” alignments. Sometimes there is just no homology. Those alignments are meaningless.

BLAST's E-value

E-value = number of hits one can “expect” to see by chance on a database this size.

Always raise an eyebrow if your E-value is ≥ 0.01 .

Common thresholds: < 0.01 , or $< 1e-5$

Short read mapping, in principle

ACAAC**T**GTCTGCTTCAGGAGTTAAATCTTACA-GGATGA **reference**

ACAAC**T**GTCTGCTT

read1

TCTG-TTCAGGAGTT

read2

CTGCTTCAGGAGTT

read3

GGGAGTTAAATCTT

read4

GAGTTAAAT

read5

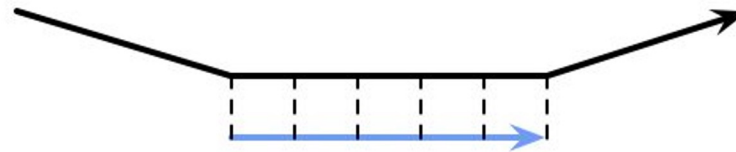
Wait.. is this local alignment or global alignment?

Types of pairwise alignment

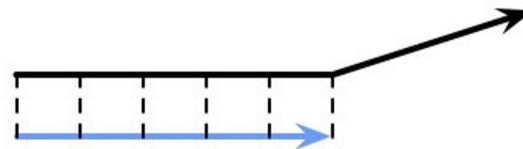
Neither. It's glocal.



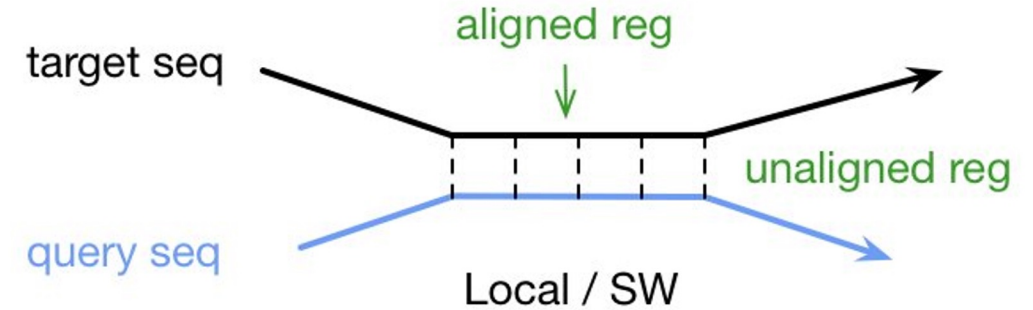
Global / NW



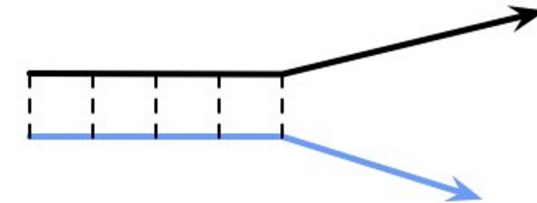
Semi-global / glocal / infix / HW



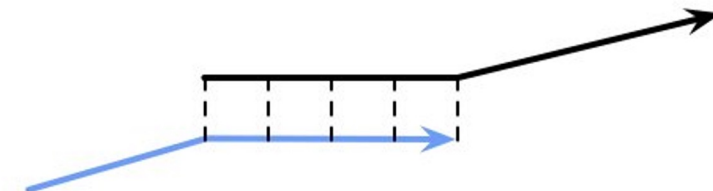
Global-extension / prefix / SHW



Local / SW



Extension



Overlap / dovetail

Why is it difficult? Need to find a home for every read

Problem: Half of the human genome is comprised of repeats

```
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
accctaaccctaaccctaaccctaaccctaaccctaacccta  
cctaaccctaaccctaaccctaaccctaaccctaacccta  
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
ccccctaaccctaaccctaaccctaaccctaaccctaacccta  
ccctaaccctaaccctaaccctaaccctaaccctaacccta  
cccaacccaacccaacccaacccaacccaacccaacccaac  
ctaccctaaccctaaccctaaccctaaccctaaccctaacccta  
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
aacctaaccctaaccctcgcggtaccctcagccggcccgccggg  
tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg  
tgagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct  
gaggagaacgcaactccgccggcgagggcgagagaggcgccgcgccg  
gcgcaggcgagacacatgctagcgcgtcggggtggaggcggtggcgagg  
cgagagaggcgccgcgccggcgagggcgagagacacatgctaccgc  
gtccaggggtggaggcggtggcgagggcgagagaggcgaccgcgccggc  
gcaggcgagagacacatgctagcgcgtccaggggtggaggcggtggcgca  
ggcgagagacgcaagcctacgggcgggggtggggggcggtgtgttgca  
ggagcaaagtcgcacggcgccgggctggggcggggggagggtggcgccgt  
gcacgcgcagaaactcacgtcacggtggcgcgccgcagagacgggtagaa
```

Slide: A. Quinlan

(first bit of human chromosome 1)



Tools

- Bowtie2
- BWA-MEM
- Strobealign
- Minimap2
- ...

Which one to choose? *Not going to answer that!*

Bowtie2, BWA-MEM: battle-tested, well-documented

minimap2: faster

Strobealign: ultra fast, newer

PacBio CLR / ONT:

- Minimap2
- Variants of minimap2 for ~ 2-5x speed gain (mm2-fast, BLEND, ..)

PacBio HiFi:

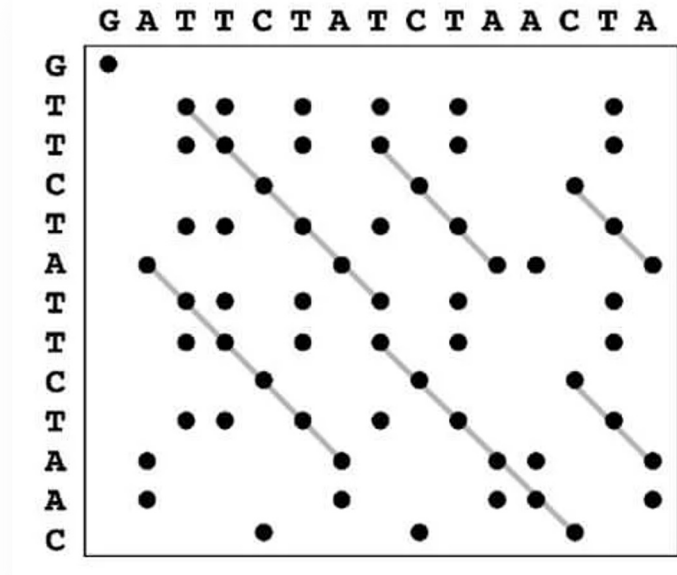
- Minimap2
- Winnowmap2 (better accuracy)
- Mapquik (30x faster mapping, but no alignment)

Tools (long vs long)

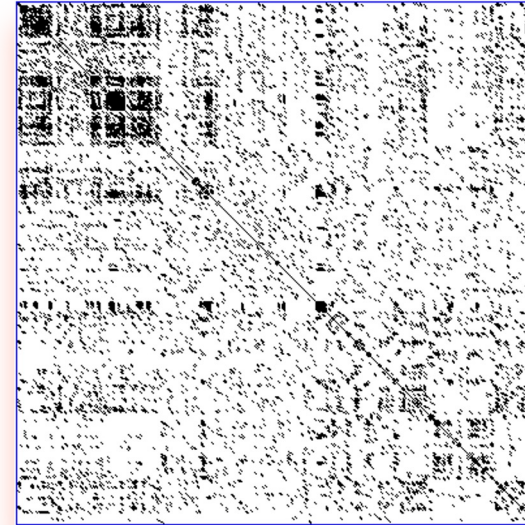
- BLAT
- Exonerate
- LASTZ
- MUMmer
- minimap2
- wfmash
- FASTGA

Dot plots

- Tools: LASTZ, D-Genies, yass, MUMmer, ModDotPlot



<https://microbenotes.com/local-global-multiple-sequence-alignment/>



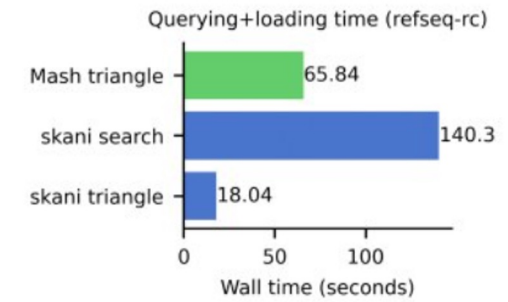
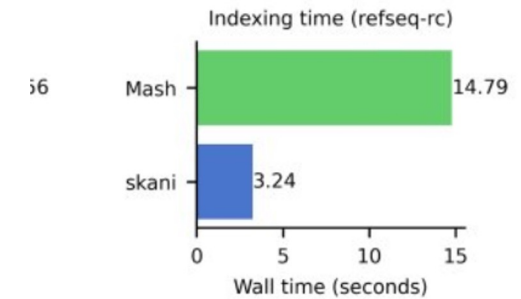
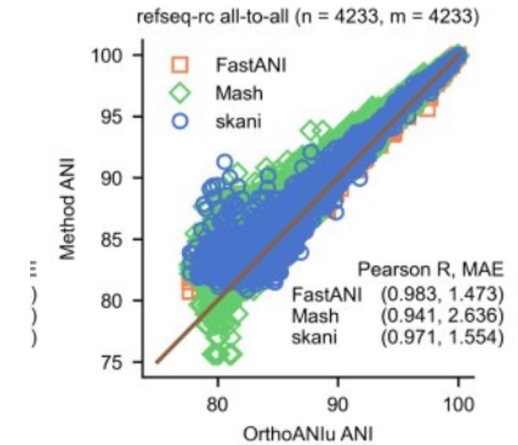
Wikipedia

ANI (average nucleotide identity)

A strange identity metric, used to compare two bacterial genomes:

1. Extract many 1 Kbp fragments from query
2. ANI = mean identity of the reciprocal best hits

(from FastANI: <https://www.nature.com/articles/s41467-018-07641-9>)



How does Bowtie2 work?

Specializes in aligning Illumina reads to genomes.

- 1) Find seeds using FM-index, typically 20 nt length, up to 1 mismatch
- 2) Prioritizes seeds to further align
- 3) Extend seeds using SW-like algorithm

Minimizers

Minimap2 and strobealign use minimizers as seeds, then SW extension.

Minimizers: select only *some* k-mers as seeds

reference: CTAAAAAGGTCA..

2nd window: TAAAAAGG

TAAAA

seed: AAAAA

AAAAG

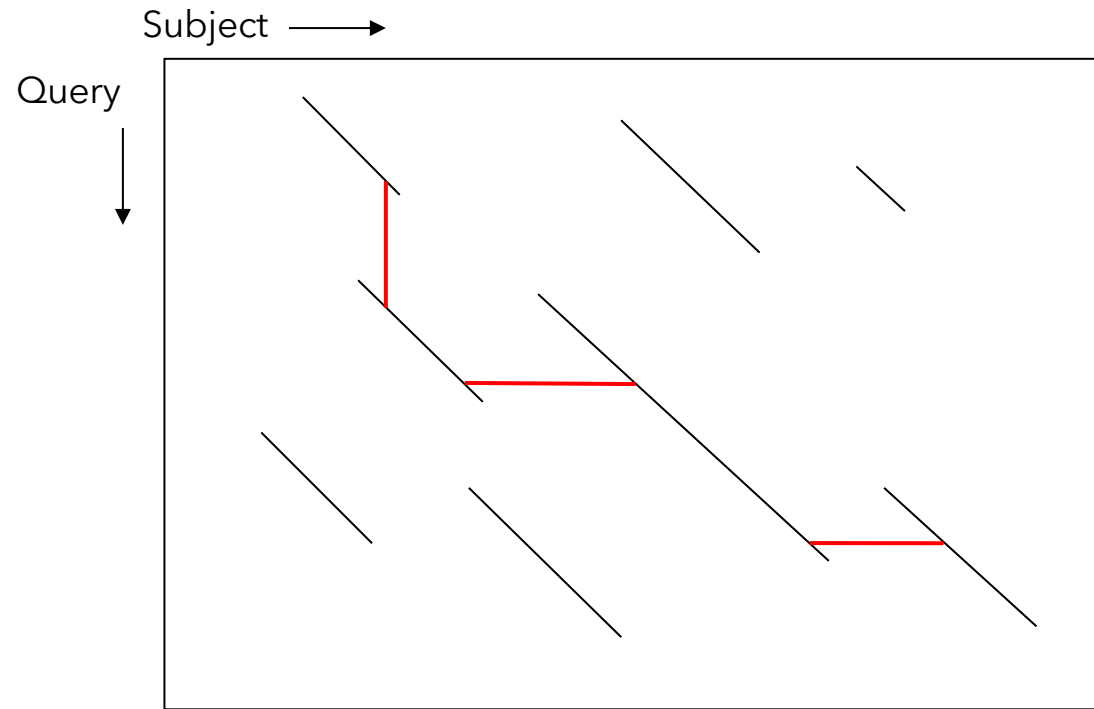
AAAGG

all reference k-mers	Found at position(s)
AAAAA	10, 65, 147, ...
AAAAC	80
....	
AAAGG	none
....	
TAAAA	49, 101

Which “some”? Slide a window over the reference, and pick the (lexicographically) smallest seed within that window. Do that for all windows

Chaining

DP on the ungapped extensions at seed hits



1. Provides estimate of alignment score to limit the search space.
2. Infer the trajectory of the best alignment to guide banded DP.

Paired reads

In some cases, Illumina sequencers output pairs of reads.



Aligners consider both reads jointly to improve precision. Need to specify:

- Orientation (forward-reverse is most common)
- Format: interleaved in one file, or two separate files

Mapping quality

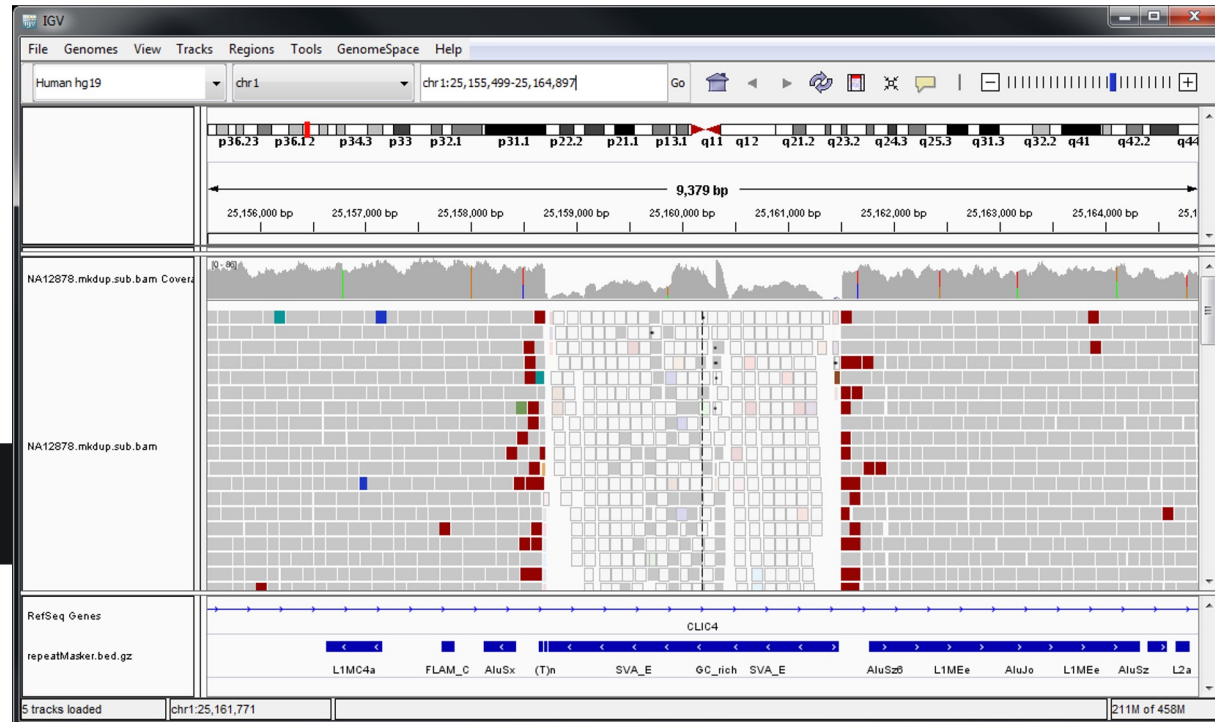
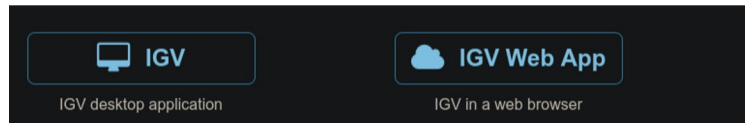
...is your best friend, to avoid errors downstreams.

Mapq: how confidently each read is mapped (in log probability).

Grab only highly-confident alignments: `samtools view -q 60 [file.bam]`

Grab all alignments except trash ones: `samtools view -q 1 [file.bam]`

Visualization of alignments



RNA

RNA read alignment is very similar to DNA, except:

- Split mapping (on genomes) due to splicing
- Ambiguity (on transcriptomes) due to many isoforms

Tools:

- Kallisto, Salmon
- STAR, HiSAT2



Proteins: What changes compared to pairwise DNA?

- Different alphabet, shorter sequences
- Some substitutions are more likely than others
- Distant homology search (high evolutionary distances)

The BLOSUM62 substitution matrix

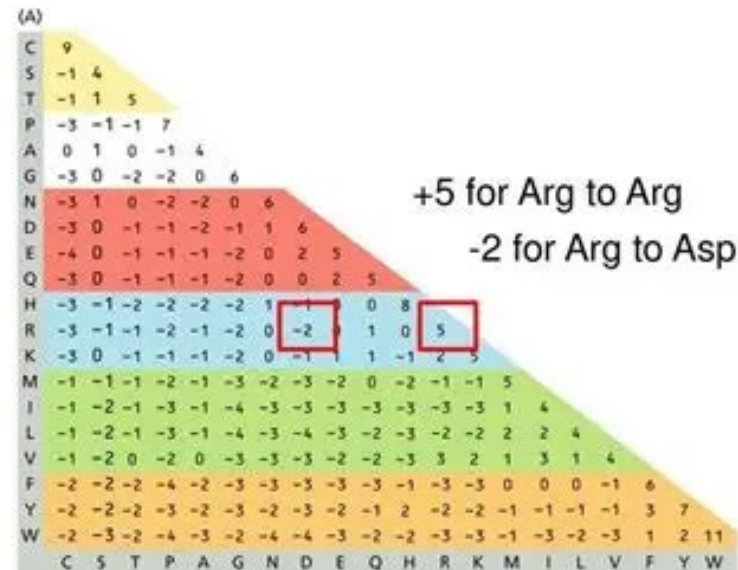


Fig 4.4

Slides: Mike Hallett, David Walsh

Algorithms

- To correctly interpret output from bioinformatics tools, you need to (at least roughly) understand its algorithm.
- Do not treat a tool as some magic black box!
- The tool cannot read your mind. Always adjust its parameters for your application.
- Risk of missing information, or reach wrong conclusions.
- Default settings are most likely not optimal for your application.

The BLAST algorithm

I. Seed: Find two neighborhood 3-mer matches within a window of 40 on the same diagonal

```
Query= d5gy3a_ a.102.1.0 (A:) automated matches {Klebsiella pneumoniae [TaxId: 573]}

>dlh12a_ a.102.1.2 (A:) Endo-1,4-beta-xylanase {Pseudoalteromonas haloplanktis [TaxId: 228]}
Length=404

Score = 45.8 bits (107), Expect = 1.3e-05
Identities = 53/195 (27%), Positives = 84/195 (43%), Gaps = 37/195 (18%)

Query      7  YKARFMPDGRIIDTANGNVSHTEGQGFAMLLAVANNDRPAFDKLWQWTDSTLRD----- 61
          YKA ++      + I+      G+      TEGQ + M  AV  N +  FD LW++  +  ++
Sbjct     59  YKAHYI----KAINPDEGDDIRTEGQSWGMTAAVMLNKQEEFDNLWRFAKAYQKNPDNHP 114

Query     62  --KSNGLFYW--RYNPVAPDPIADKNNASDGDTLIAWALLRAQKQWQDKRRYAIASDAIT 116
          K  G++ W  + N          D+  A DG+  A+ALL A  +W +  +  +DAIT
Sbjct    115  DAKKQGVYAWKLKLNQNGFVYKVDEGPAPDGEEYFAFALLNASARWGNSGEFNYYNDAIT 174

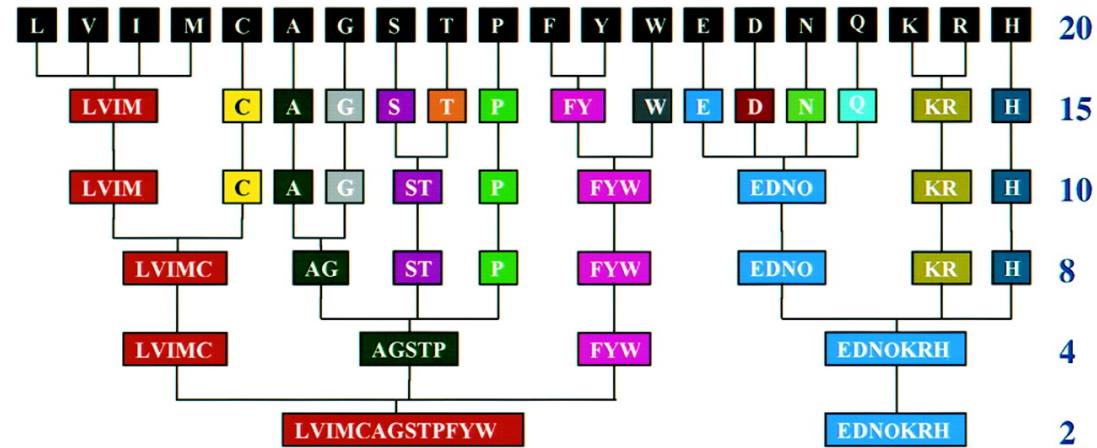
Query    117  A-----SLLKYTVVTFAGRQVMLPGVKGFNLNDHLNLPSTYFIFPAWRAFAERTHLTA- 169
          L++  ++ F+          P +  NL D      PSY I  +  FA      A
Sbjct    175  MLNTIKNKLMEHQIIRFS-----PYID--NLTD-----PSYHIPAFYDYFANNVTNQAD 221

Query    170  ---WRTLQTDGQALL 181
          WR + T  + LL
Sbjct    222  KNYWRQVATKSRTLL 236
```

Expected value of initial two-hit extensions between two random proteins of length 300: ~7.4

Alphabet reductions

Group amino acids based on substitution frequencies:
A [KR] [EDNQ] C G H [ILVM] [FYW] P [ST]



Murphy et al. Simplified amino acid alphabets for protein fold recognition and implications for folding. (2000)

Spaced seeds

Pattern of match and “don’t care” positions to extract seeds from sequence:

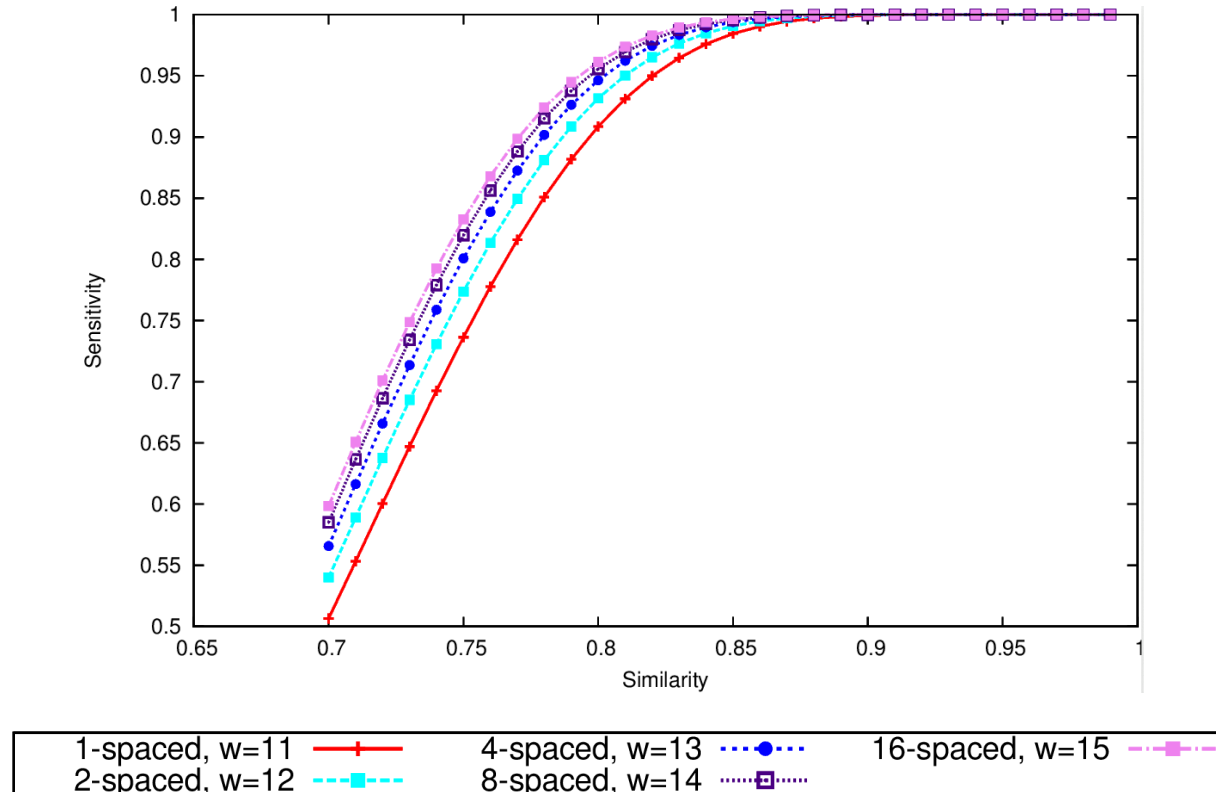
Reference	SL WAKKR TV DGQPKWLP LVAHLVDASNVSRMLFNQWLSD
Spaced Seed	111101101110111
Query	F WAKKR TV NGQPLWLP LTQHLEDASNVSR

Choose set of seed shapes that complement each other:

1011110111	10101001101011
1100101111	110100100010111
101110001111	111001001001011
111101010011	110011000110011
1111010010101	1111000010000111
1100011011011	1101010000011011
11011010001101	1110001010101001
11011101100001	1101001100010011

Multiple spaced seeds

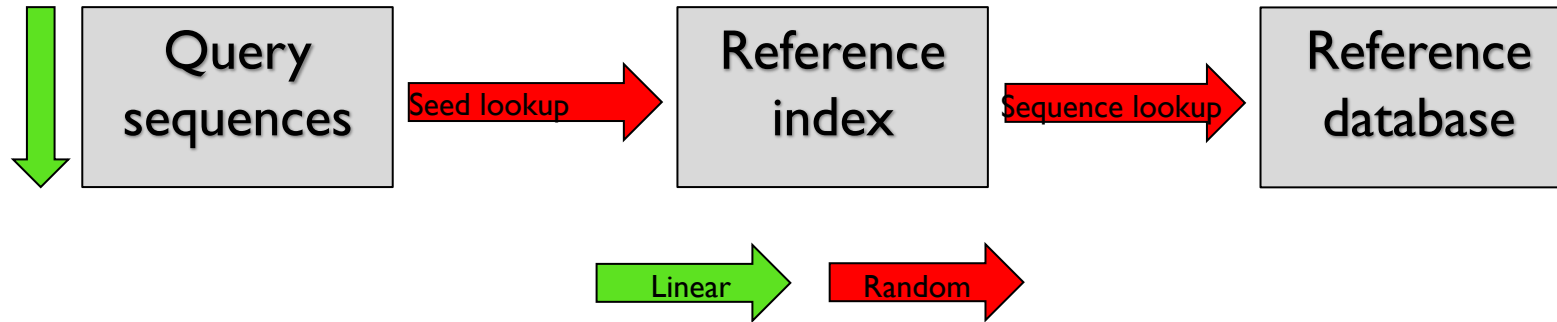
Sensitivity for homologous region of length 70 (Ilie et al., 2011)



Technical challenge: High memory consumption and many seed lookups

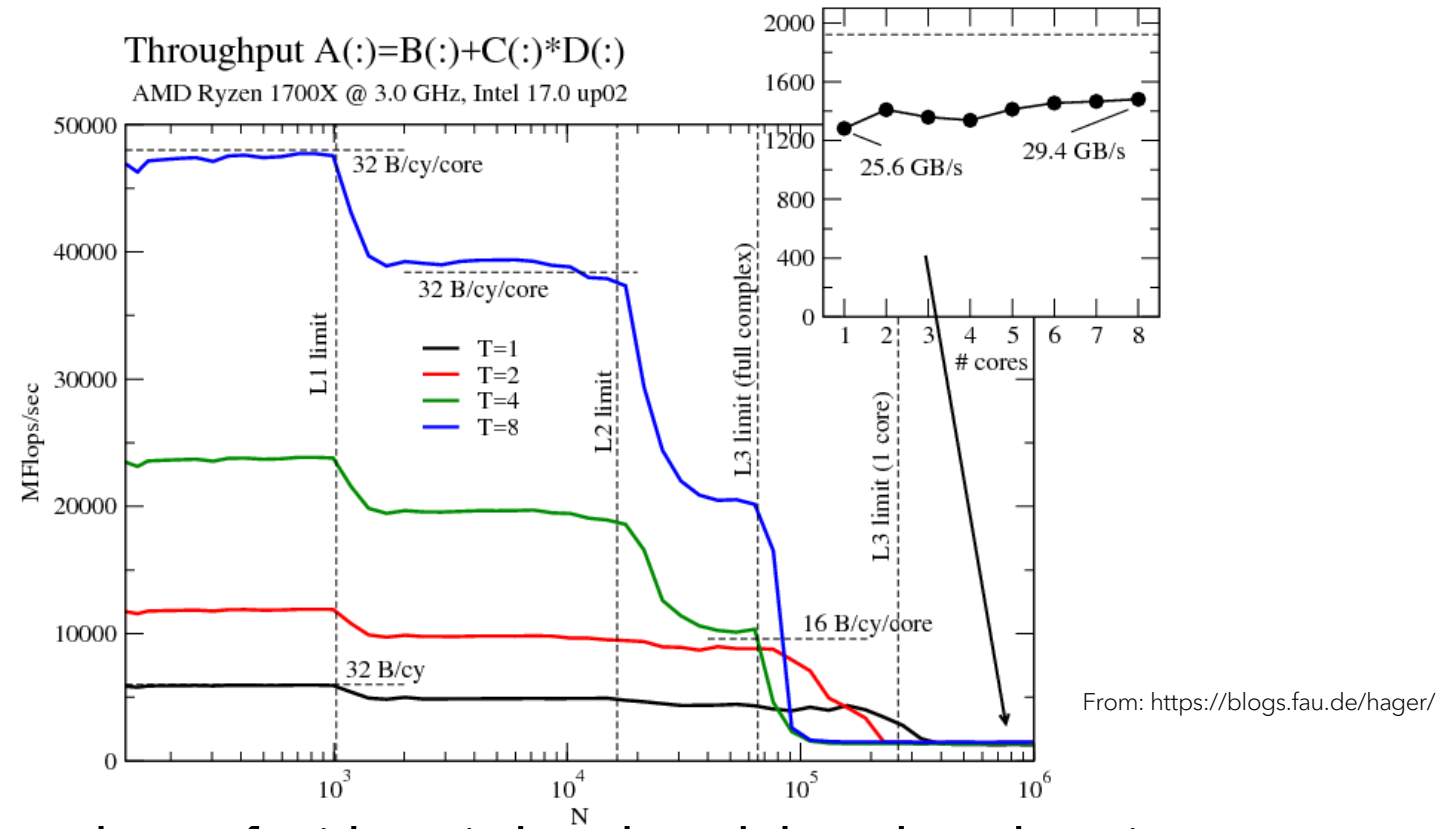
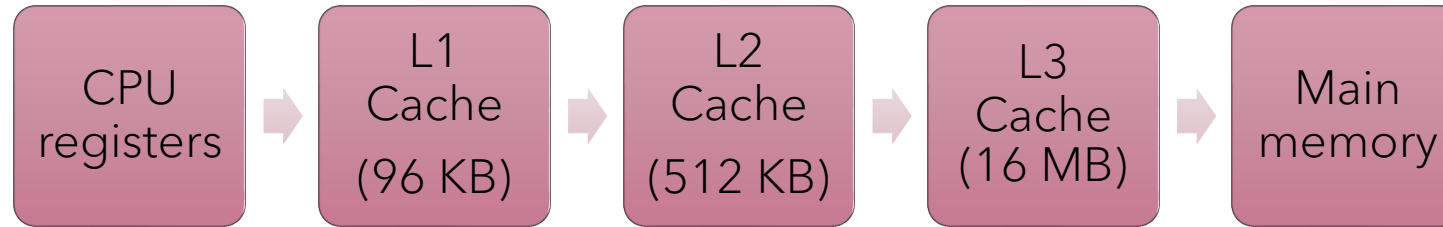
Index based alignment

Linear processing of the queries:



- Random memory access pattern to data structures that far exceed cache sizes.
- Bottleneck: latency of main memory.
- Compounded by using multiple spaced seeds.

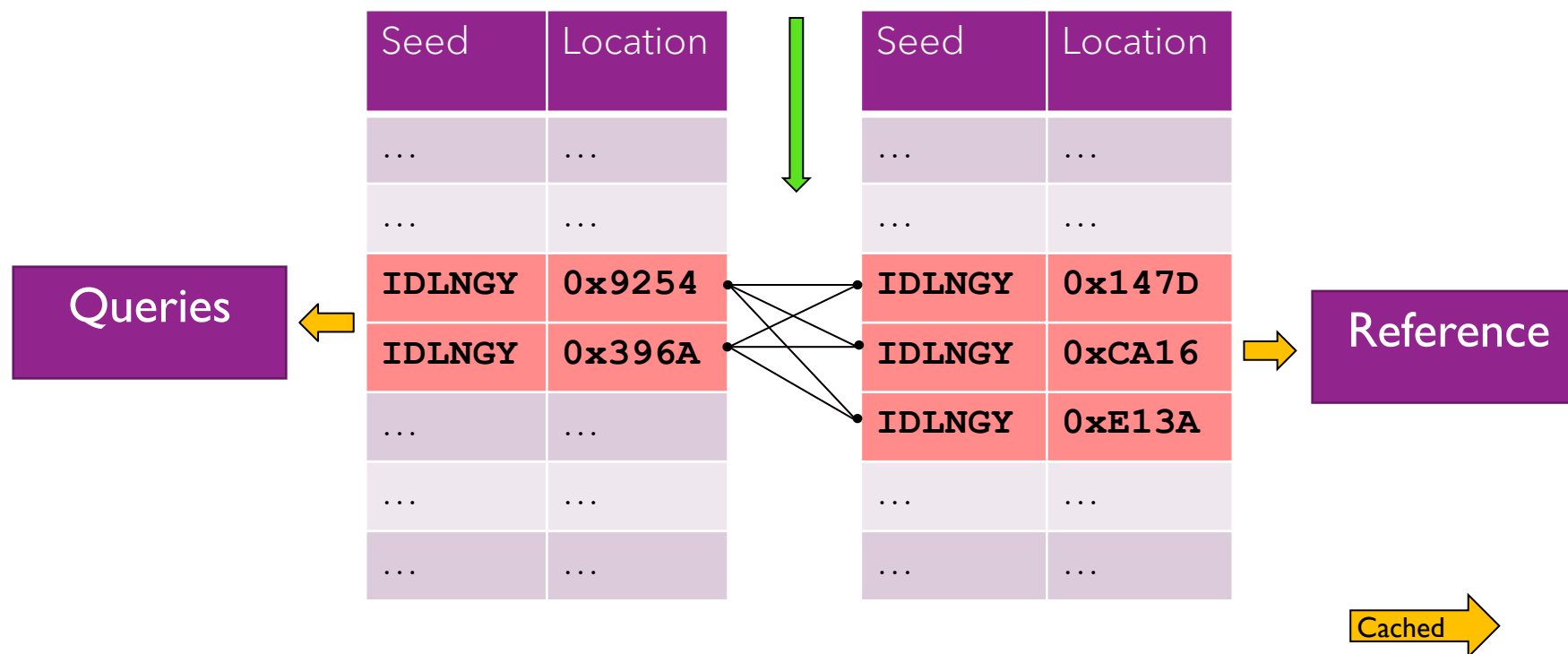
The memory hierarchy



Throughput of arithmetic benchmark based on data size

Double indexing

Sort-merge join technique with on-the-fly computed double indexes



Compare operations for each seed are streamed through the CPU efficiently by cache-aware partitioning.

Example run (nr database, 514 million sequence)

>tr|A0A0D6XPW3|A0A0D6XPW3_9STAP MerR family transcriptional regulator
OS=Staphylococcus microti OX=569857 GN=tipA PE=4 SV=1

MTTYAIGDIAKMFDISTRTLQYYDEKGVLPAYVDHNNYRVYTEHEVEKLQLIVLMKTLG
MQLKEIQALLTAEGTLDTVRLLLDQKSKALEQTIQKQQAQYQQIQAMQRMINETSQSSIT
KLQEMDRMLMKNEHLKPLRVKSLGLASIGTSLWSGIWVGTKLGTPTYPLLGLGTSVGIA
YYLTHDYYHQVKYMCPCNHVFPVPTMRQFILARHTPSTRHLTCPSCEFTGYCVEVYEPH

Mode	Shapes	Seed hits	Alignments	Ratio
Default	2x10	1,083,599	10,524	100
Sensitive	16x8	248,372,076	49,204	5,000
Very-sensitive	14x7	1,055,031,248	51,464	20,000
Ultra-sensitive	64x7	4,687,907,463	51,684	90,000

Hamming distance filter

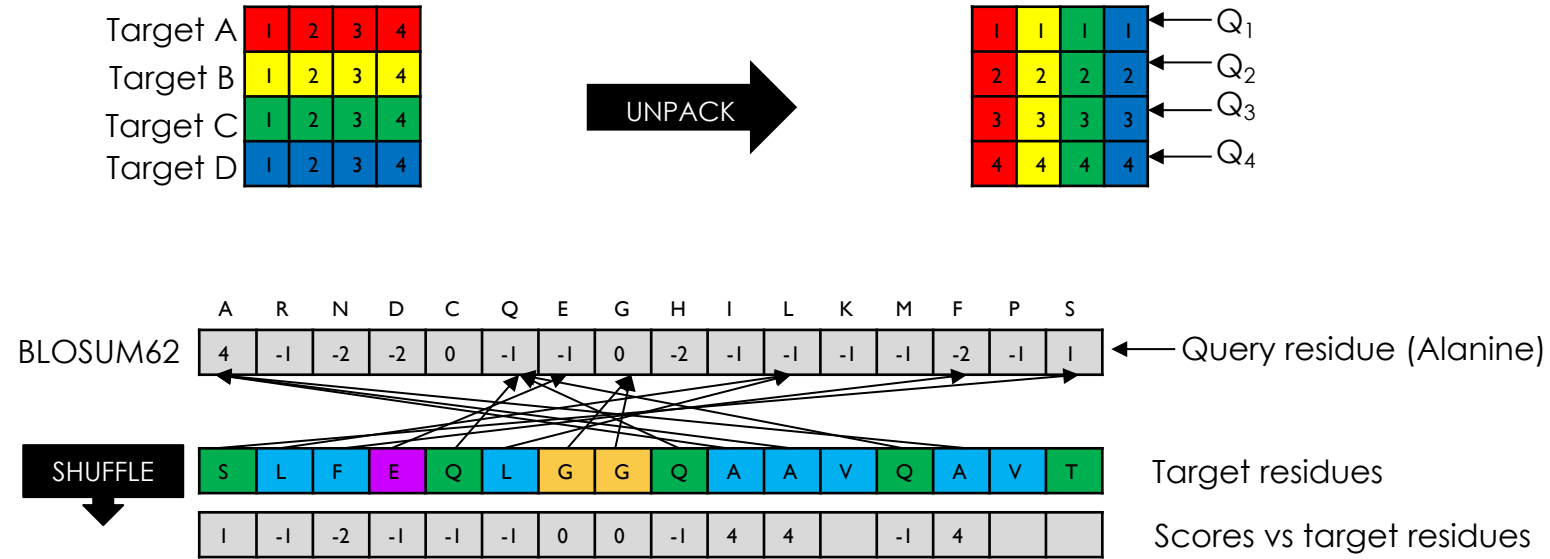
KKLNHPCIIKIKNF~~F~~DAE~~D~~YYAIV~~L~~E~~L~~ME~~G~~HE~~L~~FDKVVGNKRLK~~E~~ATC
+ P + + F D + +L+LM G++L + + E
STGDCPFAVCMSYA~~F~~H~~T~~P~~D~~KLSFI~~L~~D~~L~~M~~N~~G~~N~~D~~L~~HYHLSQHGVS~~E~~NDM

- Compute the hamming distance of 2 16-letter strings:

```
_popcnt32(_mm_movemask_epi8(_mm_cmpeq_epi8(x, y)))
```

- Throughput: 16 ps/Letter, compared to 170 ps/Letter for ungapped extension with BLOSUM62 scores.
- Preload all query and subject windows into linear buffers.
- Use loop tiling to optimise cache use.
- Filtration: ~92%

Ungapped extension filter



```

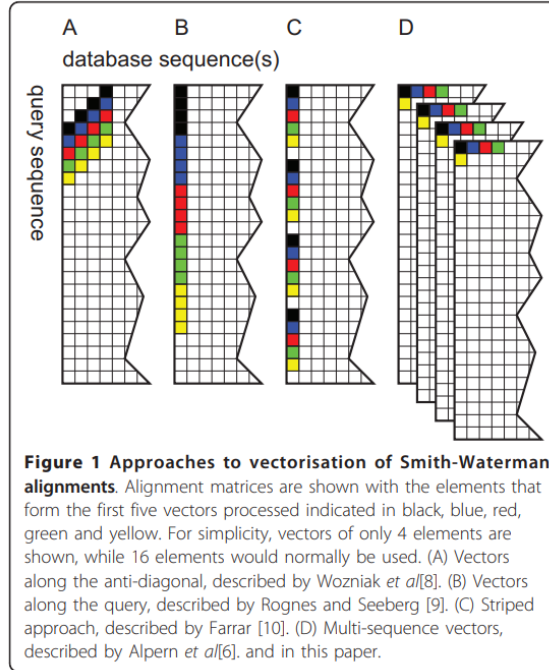
s1,...,s16 ← 16x16 byte subject sequences
t1,...,t16 ← transpose(s1,...,s16)
for i in [0,16[
    match_scores ← BLOSUM62[query[i], ti]
    score += match_scores
    max_score ← max(max_score, score)

```

Filtration: ~83%

Data-parallel Smith Waterman

- Technical challenges: Intra-register data dependencies, loading match scores into registers.
- SWIPE: Inter-sequence SIMD parallelization
- Does not easily afford banded DP.

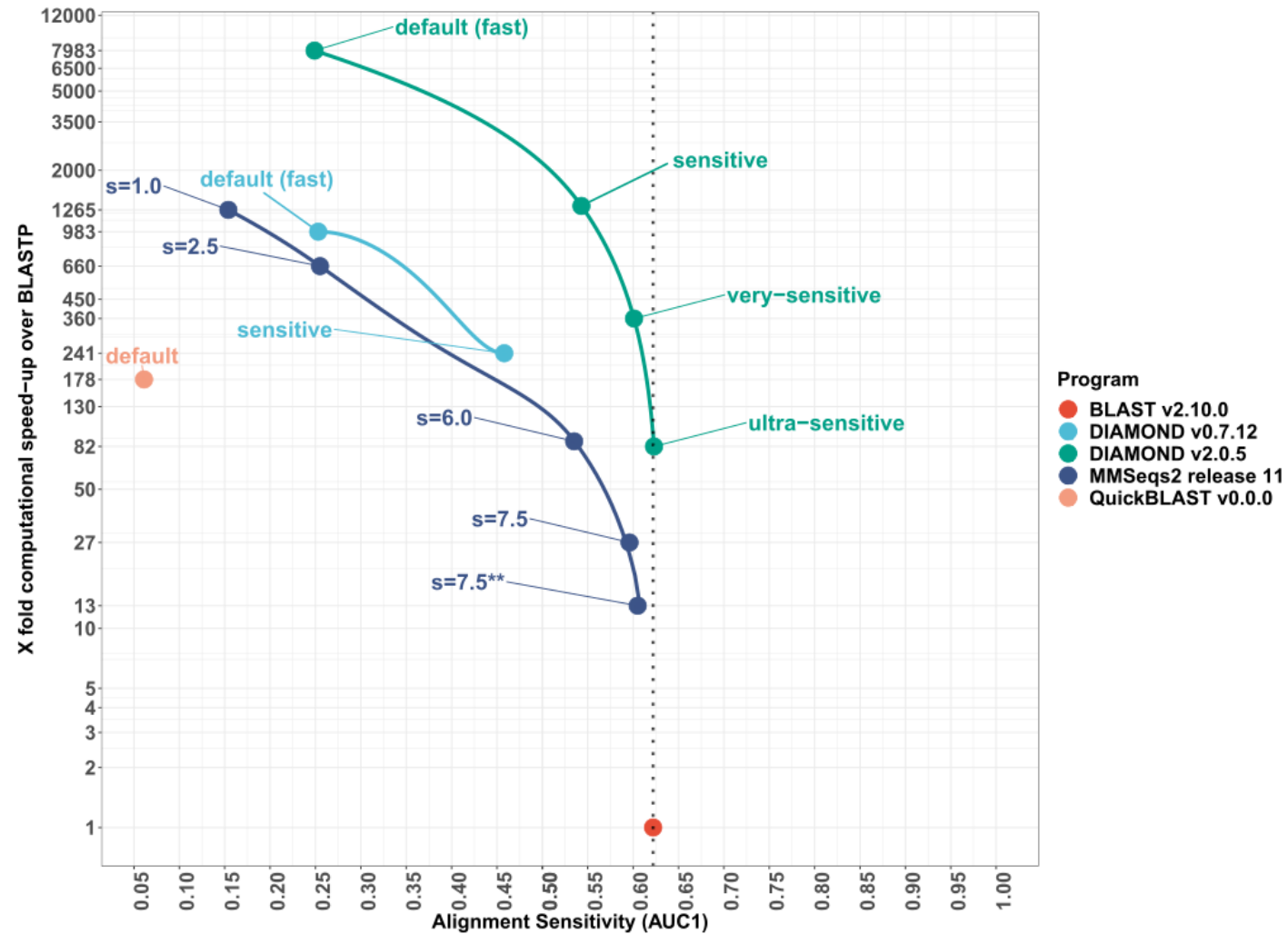


From: Rognes, Faster Smith-Waterman database searches with inter-sequence SIMD parallelization (2011)

Benchmark

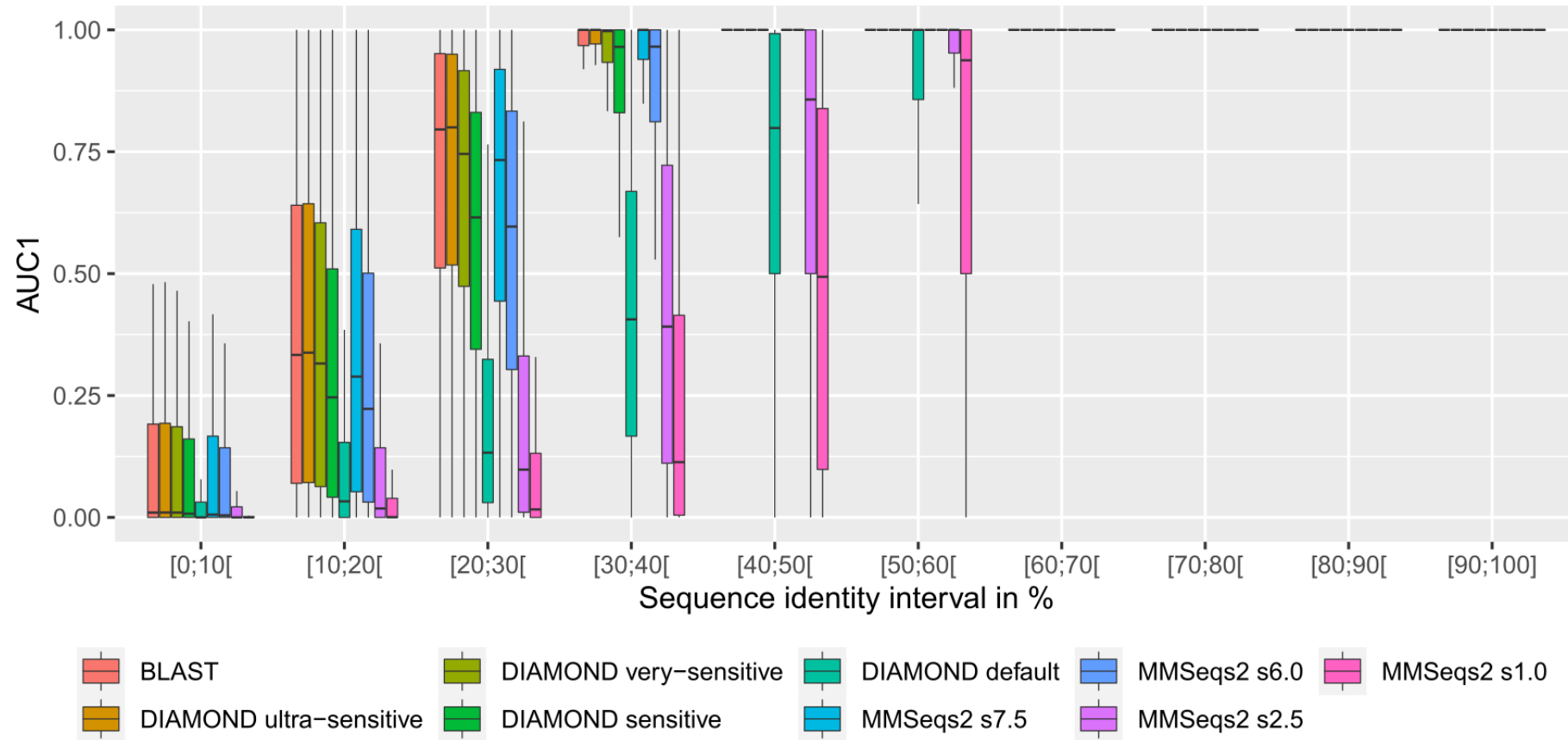
- Annotate a query file and sequence database using SCOP domains.
- SCOP (Structural Classification of Proteins): Proteins are classified into classes, folds, superfamilies, and families. Contains ~2,000 superfamilies and ~4,500 families.
- For each query, compute AUCI: fraction of sequences from the same family covered until the first false positive.
- Alignments to targets annotated with a different SCOP fold are false positives; shuffle non-annotated ranges.
- In our case: 1.7 million queries vs. 8 million targets.

DIAMOND (2021)



From: Buchfink et al., Nature Methods (2021)

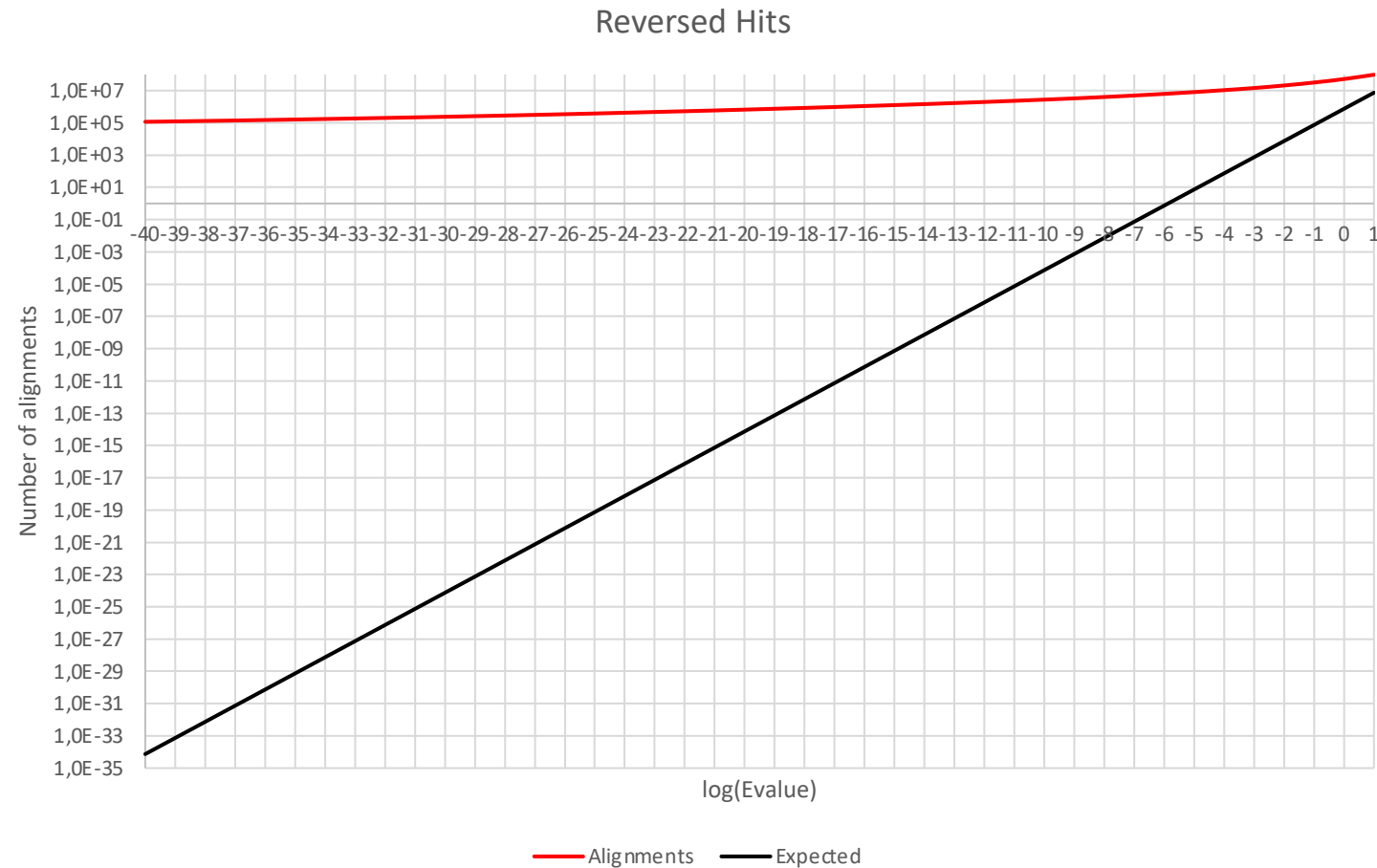
Sensitivity by sequence identity



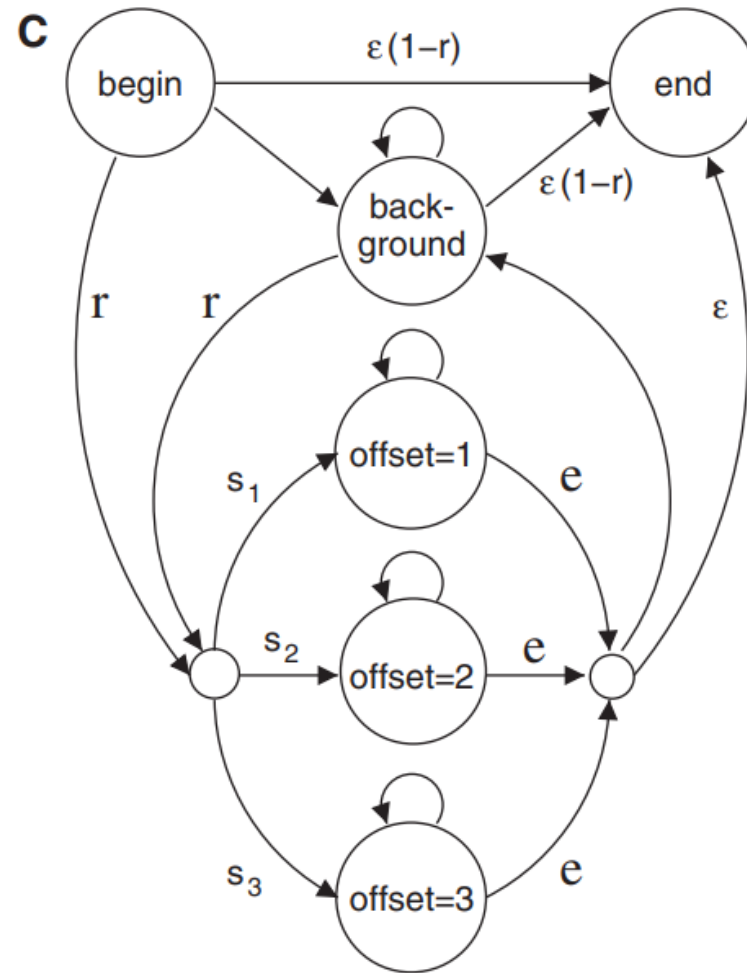
From: Buchfink et al., Nature Methods (2021)

Repeat masking

Alignment of 740,000 queries against the NR database (220 million REVERSED proteins) using BLAST 2.10.0:



tantan masking



(Frith, 2016)

Examples of reversed alignments

```
Query= VUS24436.1 hypothetical protein SB6410_00906 [Klebsiella grimontii]
Length=1828

>\RSK04010.1 hypothetical protein D8809_02240 [Streptococcus gordonii]
Length=1896

Score = 1144 bits (2958), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 537/1621 (33%), Positives = 1361/1621 (84%), Gaps = 20/1621 (1%)

Query 10      SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 69
          +++LS++E L+E+ +LS++E L+++ +LS++E L+++ +LS++E L+++++L E++ L+E
Sbjct 262      TDALSDAEVLAETLALSDAEVLADTLALSDAEVLADTLALSDAEVLADTDALVEADVLA 321

Query 70      SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 119
          +E+LS++E L+++ +LS+++ L+          L+++ +LS+++ L+++ +L ++E L++
Sbjct 322      TEALSDAEVLADTLALSDADVLADTLALSDADVLADTLALSDADVLADTLALVDAEVLAD 381

Query 120     SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 179
          +++L E++ L+++ +L E+E L+++ +L ++E L+++++LS++E L+++ +LS+++ L+E
Sbjct 382      TDALVEADVADTLALVEAEVLADTLALVDAEVLADTDALSDAEVLADTLALSDADVLAE 441

Query 180     SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 239
          +++L +++ L+++E+LS+++ L+++ +LS+++ L+ S++LS++E L+++ +LS+++ L+E
Sbjct 442      TDALVDADVLADTEALSDADVLADTLALSDADVLADSDALSDAEVLADTLALSDADVLAE 501

Query 240     SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 299
          +E+LS+ E L+++ +LS+ E L+E+ +LS++E L+E+E+LS++E L++S++LS++E L++
Sbjct 502      TEALSDDEVADTLALSDDEVLAETLALSDAEVLAETLALSDAEVLADSDALSDAEVLAD 561

          . . .

Query 1490    SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 1549
          +E L+++ +LS+++ L+++++LS++E L+++++LS+++ L+++E+L +++ L+E+ + S
Sbjct 1762    AEVLADTLALSDADVLADTDALSDAEVLADTDALSDADVLADTEALVDADVLAEATLAFSD 1821

Query 1550    SESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSESESLSE 1609
          +E L+++ +LS++E L+++++LS+++ L+++++LS++E L+++++LS+++ L+++E+L +
Sbjct 1822    AEVLADTLALSDAEVLADTDALSDADVLADTDALSDAEVLADTDALSDADVLADTEALVD 1881

Query 1610    S 1610
          +
Sbjct 1882    A 1882
```

Examples of reversed alignments

Query= BAG82822.1 polyprotein [Enterovirus A71]

Length=2438

>\WP_124134815.1 hypothetical protein [Campylobacter hepaticus]

Length=113

Score = 185 bits (470), Expect = 1e-50, Method: Composition-based stats.

Identities = 89/96 (93%), Positives = 90/96 (94%), Gaps = 2/96 (2%)

Query 84 H D F F K S A M P E G Y V Q E R T I F F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E D G N I L G H K 143

D F F K S A M P E G Y V Q E R T I F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E D G N I L H K

Sbjct 20 Q D F F K S A M P E G Y V Q E R T I S F K D D G N Y K T R A E V K F E G D T L V N R I E L K G I D F K E D G N I L - H K 78

Query 144 L E Y N Y N S H N V Y I M A D K Q K N G I K V N F K I R H N I E D G S V 179

L E Y N Y N S H V Y I A D K Q K N G I K N F K I R H N I E D G S +

Sbjct 79 L E Y N Y N S H - V Y I T A D K Q K N G I K A N F K I R H N I E D G S M 113

Examples of reversed alignments

```
Query= XP_018736024.1 DNA-binding transcription factor YAP1 [Sugiyamaella  
lignohabitans]
```

```
Length=561
```

```
>\XP_027639320.1 plectin [Falco peregrinus]  
Length=6846
```

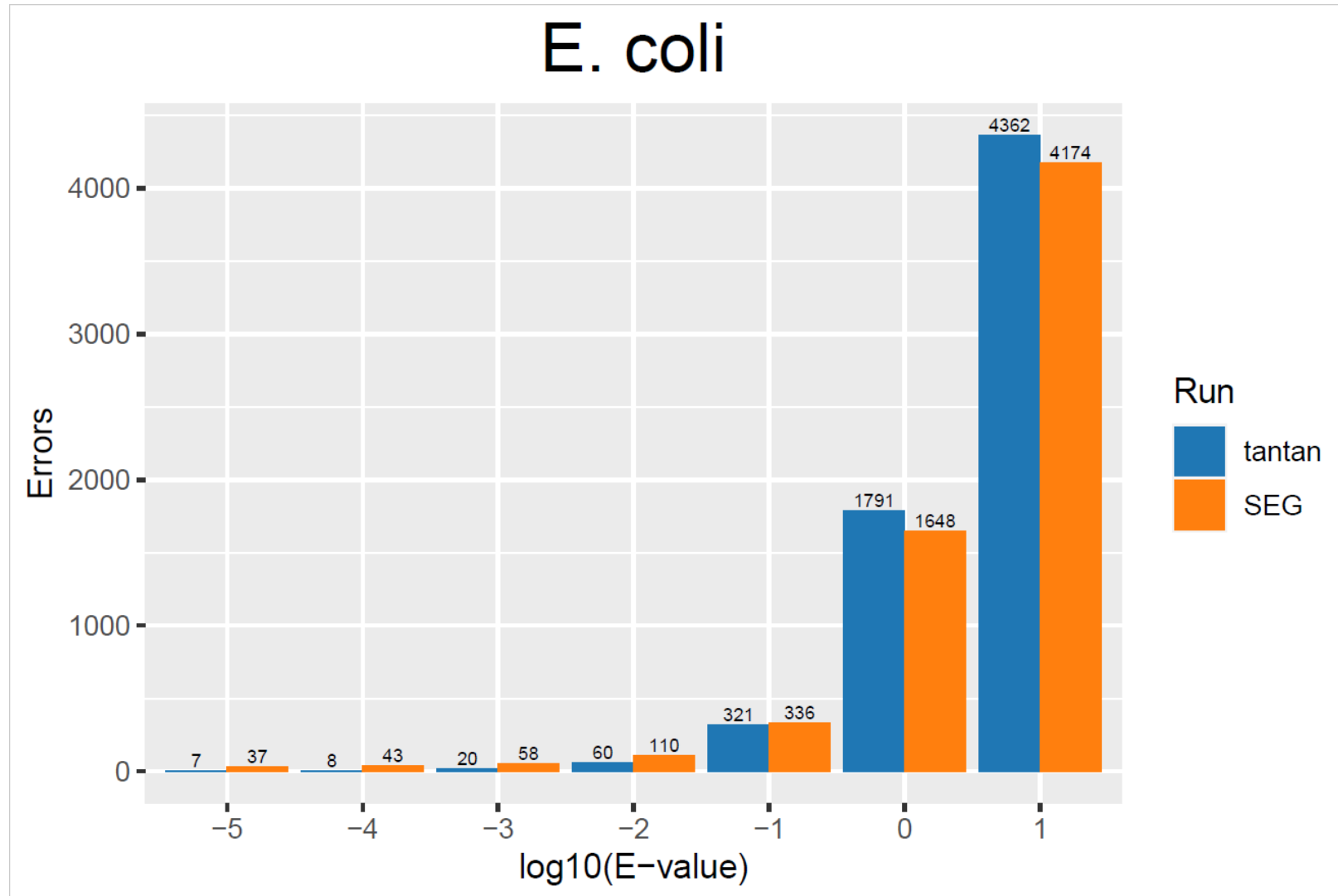
```
Score = 64.7 bits (156), Expect = 7e-07, Method: Composition-based stats.  
Identities = 43/146 (29%), Positives = 71/146 (49%), Gaps = 6/146 (4%)
```

```
Query   66      ETPSPNNTFSLPEKRKIASPDRENSEEPVQVKSEANTESDNIKSKAAKADASSTVSSSKV   125  
          E  +   FSL + +  A   R+  E+ VQ++ E   E+  ++ +A +A           +  
Sbjct   5209  ELQATKEAFSLRKSQLEAEASRQAVEQAVQIQREKEVEAQRMREAEAAQLRLKELDAVA  5268  
  
Query   126      AAAAAAAKKAGRKPEEKEPENKRKAQNRAAQRAFRE-----RKERHLKELEDRLVQLQLE  179  
          AA ++A  K + +E  ++K      QR  RE      RK+R  E   R+  QLE  
Sbjct   5269  KQKERAADREAQVKKLEEEAERKKQSEKVVQRRLREAEQAQRKQREADEARARLEQLE  5328  
  
Query   180      NEATATNDENNFLRLQVERLQEEELKK   205  
          EA+A   E N L+L++ R++EE+KK  
Sbjct   5329  TEASARQKETNELQLRIVRIEEEVKK   5354
```

Error benchmark

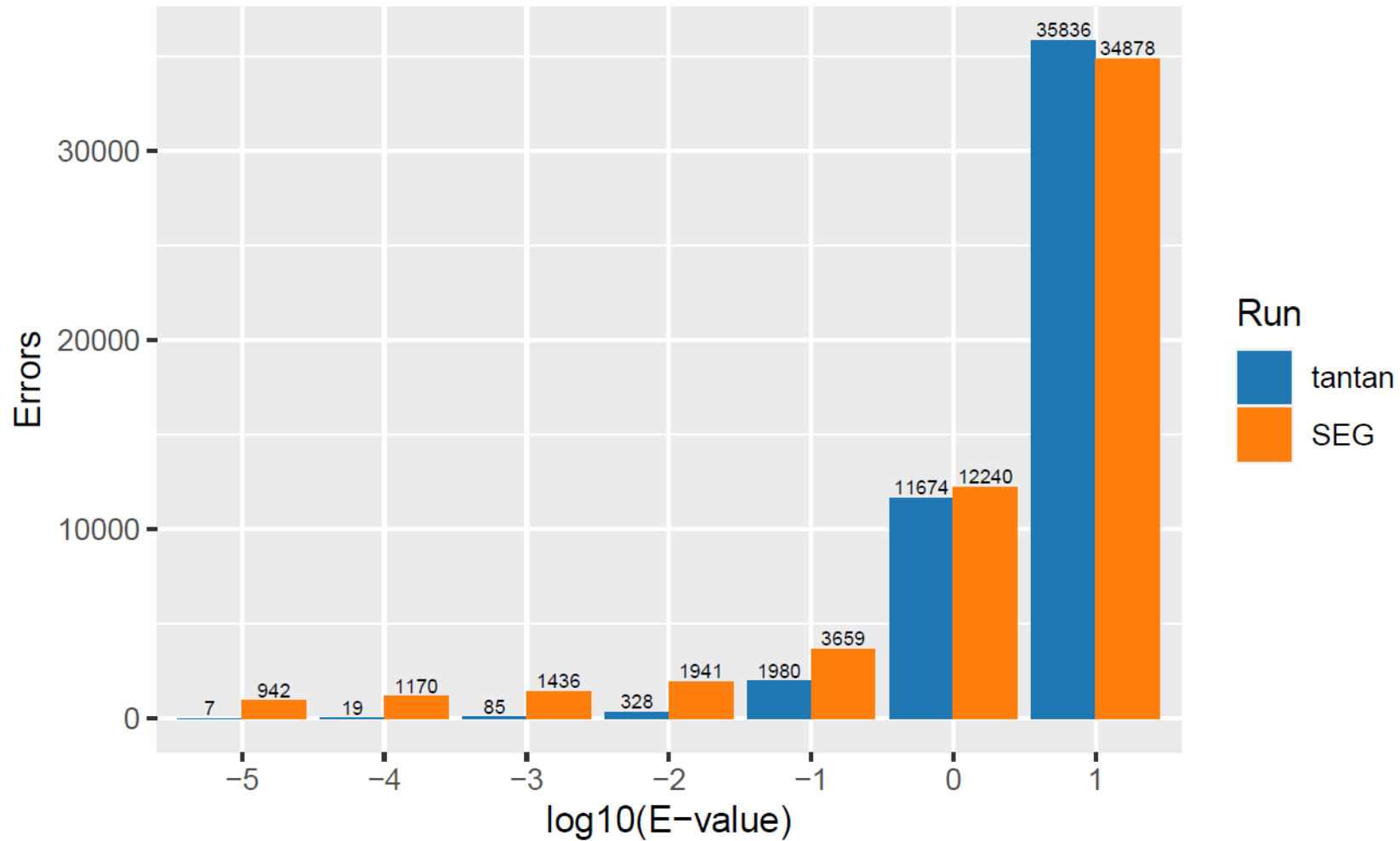
- Database: UniRef50 processed by DecoyPyrat (Wright, 2016).
- Query: *E. coli*, *A. thaliana* (TAIR10) reference assemblies
- Diamond in very-sensitive mode

Queries with errors

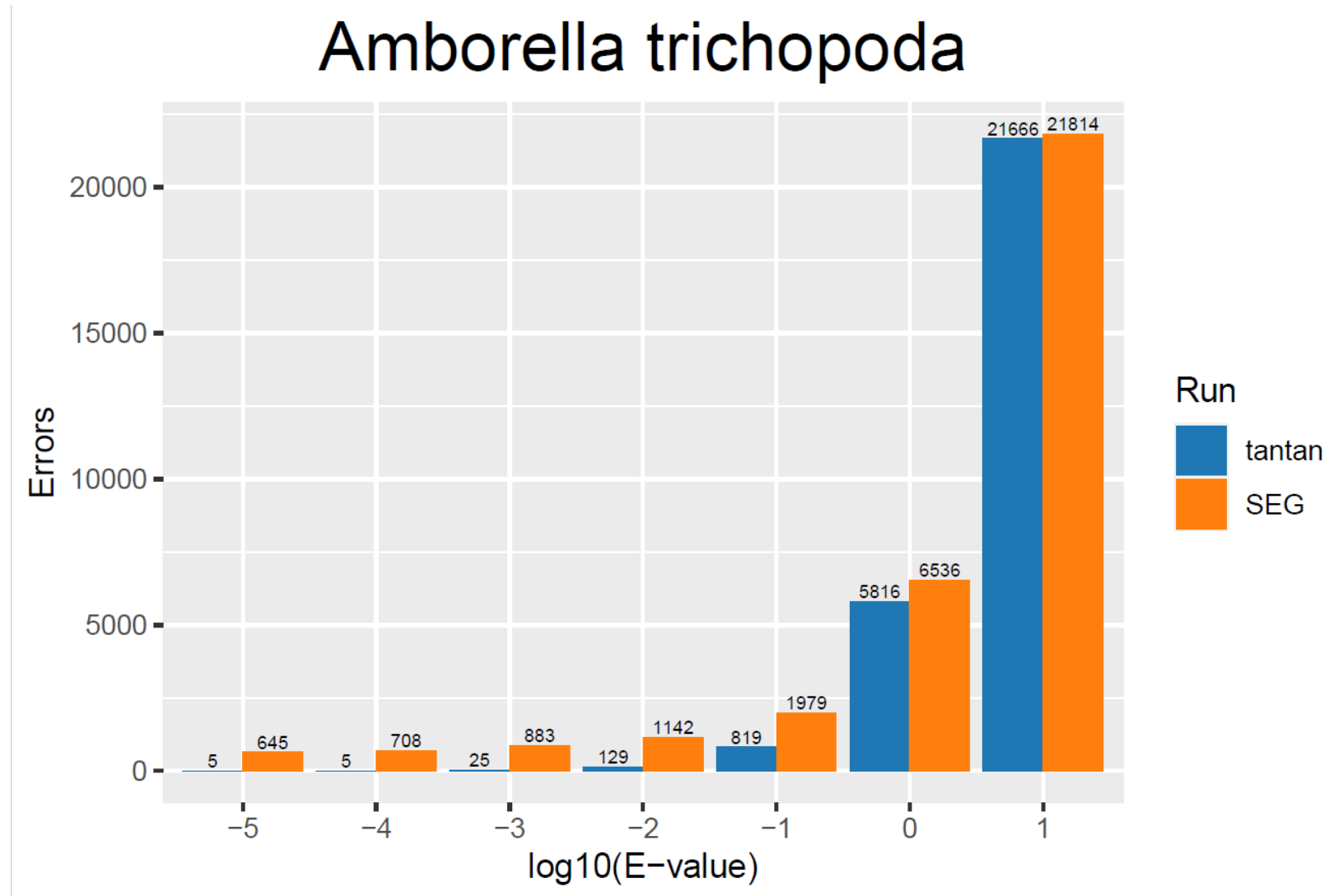


Queries with errors

A. thaliana

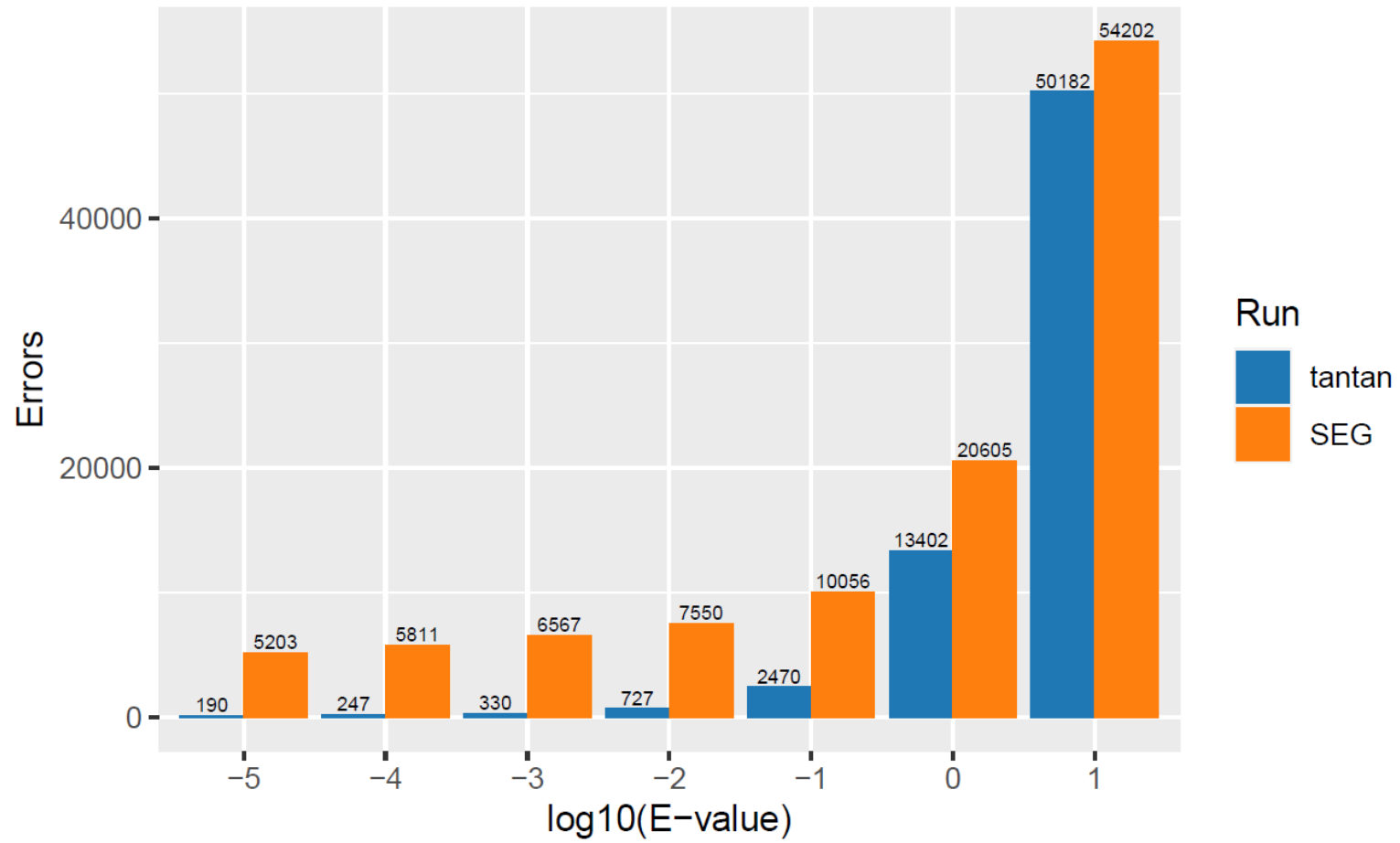


Queries with errors



Queries with errors

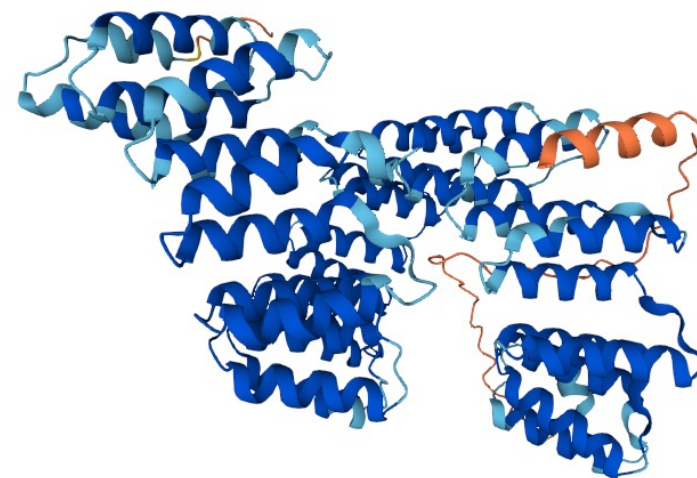
Pan troglodytes



Examples

Pentatricopeptide repeat-containing protein AtIgl2775, mitochondrial

```
MVRMMIRRLSSQASRFVQPRLLGTGLRIALINCPNELLFCCERGFSTFS
DRNLSYRDKLSSGLVGIKADDAVDLFRDMIQSRPLPTVIDFNRLFSAIAK
TKQYELVLALCKQMESKGIAHSIYTLSIMINCFRCRKLSTYAFSTMGKIM
KLGYPEPDTVIFNTLLNglclecrvsealelvdrmvemghkptlitlntlv
nglclngkvsvdavlldrmvetgfgpnevtygpvlmvmcksgqtalamel
lrkmeernikldavkysiidglckdgsldnafnlfnemeikgfkadiit
yntliggfcnagrwdgakkllrdmikrkispnvvtfsvlidsfvkegklr
eadqllkemmqrgiapntitynslidgfckenrleeaiqmvdlmiskgcd
pdimtnilingyckanriddglelfrmslrgviantvtvntlvqgfcq
sgklevakklfgemvsrrvrpdivsykilldglcdngelekaleifgkie
kskmeldigiymiihgmcnaskvddawdlfcsplpgkvkldaraynimi
selcrkdslskadilfrkmteeghapdeltynilirahlgddattaael
ieemkssgfpadvSTVKMVINMLSSGELDKSFLDMLSTTRASLK
```



PFAM: PPR repeat family

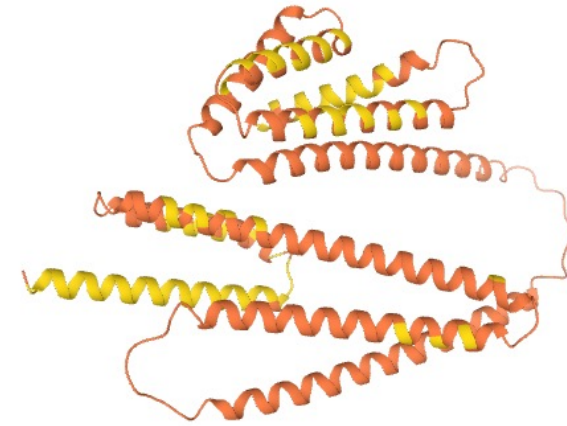
This repeat has no known function. It is about 35 amino acids long and is found in up to 18 copies in some proteins. The family appears to be greatly expanded in plants and fungi.

```
B9SD26_RICCO/490-539
B9SD26_RICCO/630-679
B9S1N6_RICCO/475-524
B9SD26_RICCO/560-609
PP325_ARATH/551-600
PP432_ARATH/611-660
CSXQ26_SORBI/466-515
C4IZC9_MAIZE/290-339
PP318_ARATH/291-340
B9T3G5_RICCO/315-364
B9S9V6_RICCO/201-250
PP432_ARATH/226-275
A8N2J3_COPC7/429-478
B9HLF4_POPTR/253-302
PPR40_ARATH/253-302
B9H106_POPTR/474-523
PP101_ARATH/253-302
PPR37_ARATH/315-364
B9T1X9_RICCO/197-247
PPR79_ARATH/150-199
PP187_ARATH/226-275
B9RA74_RICCO/824-873
B9SPV8_RICCO/115-164
AOA0E0QDE3_ORYRU/551-600
VDEVSYGILLMGY.F.....K.DGK.SVERMKLWDEMKE.K.EI...IPSIITTYNMIIGLCH
IDAVTYNTIIISGL.C.....K.EDR.FEEAFDLAEMEE.K.KL...SPDCYTYNAILSALAD
LDSISYNTLILAC.C.....K.EGK.VEEFPLKEEMVR.R.GI...QPDMYTYNMLLHGLCN
PDEITYNTIILGY.C.....R.EGQ.VEKAFQFHNKMKV.K.SF...KPDLTTCNILLRGLCT
MDRWSYNTLISGL.C.....G.KKK.LDEAFMFLDEMVK.R.GI...KPDNVTYSILICGLFN
VDIWMYNTLLIAM.C.....K.SGN.LAKAVSLFSEMVO.R.SI...LPDSYTYTSLISGLCR
PNEYTYNTLMKGY.I.....K.KGA.FDKAVALLNEMHQ.N.GV...SGEYTYNTILLINGLCM
LDNACYNTLIHLR.C.....Q.EGK.LDAFAELINMEE.G.GI...ESDEYTFESILVNGLC
HDIYSYNTILLNLY.F.....K.DGN.LDAVDDLIETIER.R.GM...KDEYTHITIVNGLLR
HDLASYNTLINLY.C.....K.QGN.LEAAYQLLDEIEK.E.GI...QODEYTHITILINMCK
PSVITYNTIVLHWY.C.....K.KGR.YKAELELDQMS.K.GI...EADACTYNMLVDLCK
PTIVTYNTIVLHWY.C.....K.KGR.FKAELELDHMK.S.K.GV...DADVCTYNMLIHDLCR
PDVITYNTIFLYE.H.....R.RGD.FAEIGKMLSQMDK.E.RV...VEDVYTESILLIALLR
PDLYTYNTVLLNYC.C.....N.EFM.FEENVKLLKMEC.S.AI...EPDVYSYNQLLKAHCK
FDLYTYNTVLLNYY.Y.....D.NNM.LKRAEFVMAEMVR.S.GI...QLDAYSYNQLLRHCR
RNVITYNTLIDGH.C.....K.NKR.LAEAEELDQHEL.Q.GV...SRNSVTYNILIDGLCK
PNLVTFNALIDAF.V.....K.EGK.FVEAEKLHDDMIK.R.SI...DPDIFTYNSLINGFCM
PDVVAESALIDCF.V.....K.EGK.LREAEELHKEMIQ.R.GI...SPDITVYTSILIDGFCCK
PNTQIENILVKHH.C.....K.SGD.LESALEVMHEMKSR.RS..YFNVTYTSILIDGLCK
PDACTYNILIHGC.S.....Q.SGC.FDDALKLDEMVK.K.KW...KRTGTFETLHGLCK
PDVCTENILINGY.C.....R.SSK.FDLALDLFREMKE.K.GC..EPNVYSNTILIRGLS
PNHVTYTYLIEYH.C.....T.VGN.IKAEQQLFEMEMQK.R.NW..MPNVLYTSLIHGYNR
PNQVTHIILIEAH.C.....R.AGE.LDHATELENLMN.N.AY...SPDKVTYNILLRSLCK
PNEYTYNTIMINGY.C.....K.NGR.LEDAFSLERQMAS.K.GV...NPGIVTYSTHICGLFQ
```

Examples

Late embryogenesis abundant domain-containing protein

```
MMERRRTALVLFV VVVVLTWQEGVLGKWLESTAKEKTGSWAGWVSDKITT  
GFGTKKEETGIYQKSKDEARKAAQAAENYAYDKANYvkdsaydnagyakd  
faenkaeyakdfaydkagdakmayekaghakdfaydkagnakdmayeta  
gyakdfaydkggyakdvaydkagnakdmayekagnakdmayekaehikdf  
tydkVGSAYgsaqsmmdsgydkagdakmayekaGIVkdMAYdkagdakd  
vayekagiakdmaydkagnakdmaydkvGSAYGSAQkakdsgyekageak  
dyaykkagnakdiayekaqdakdfaydkaGYGydkagdvirmatdksgea  
yegakeksakdTAGEAMDDSIDYMKdkshnakdgatrgfeeamekvge  
kygvakestkyayetakkkaSQVAGEIRDYAEEL
```



PFAM: Late embryogenesis abundant protein

Different types of LEA proteins are expressed at different stages of late embryogenesis in higher plant seed embryos and under conditions of dehydration stress. The function of these proteins is unknown.

```
Q06431_BETPN/202-245  
Q06431_BETPN/246-289  
LEAD8_DAUCA/104-147  
LEAD8_DAUCA/155-198  
Q39871_SOYBN/297-340  
Q39871_SOYBN/348-391  
Q1XI26_POLVA/252-295  
Q1XI26_POLVA/413-456  
Q9ZFW6_ARATH/322-365  
Q9ZFW6_ARATH/238-281  
Q9ZFW6_ARATH/176-219  
Q9ZFW6_ARATH/77-120  
LEAD8_DAUCA/326-369  
LEAD8_DAUCA/238-281  
Q7P280_FUSNV/15-58  
ECP63_ARATH/99-142  
ECP63_ARATH/143-186  
Q2N1E0_PHAVU/119-162  
Q39871_SOYBN/158-201  
Y3304_ARATH/135-178  
Q06431_BETPN/78-121  
Y3304_ARATH/354-397  
ECP63_ARATH/220-263  
Q2N1E0_PHAVU/163-206  
Q9FFW5_PHAVU/180-222  
Q9ZTY1_SOYBN/170-213
```

```
KDYAAEKAKKAKDKTVGKASEYKDYAAEKAKETKDSALSKAEY  
KDYTAEMEKEVKDKIVGKAGEYKDYAAEKAKETKDYTAETIEG  
AEEAKEKAKAKDKITMGKAGEYKDYTAQKAEAKAKAKAEET  
KNVTAQKAGEAKDITLGKAGEYKDYAAQKAEAKDITTAQKAE  
AETANKAGEMKEATKKKTAETAENAKNKASEIKDRAAETAEAA  
AEVTKNKALEMKDAAKDRTAETTDAAKQKTAQAKENTKENVSGA  
KEAAKDKIGHAVDVITDKLGQAKDATAEKLVAQKDATAEKLGYA  
KDKTADKLGQADKTEKELVEAKDVTADKLGAKDVTADKLGRA  
KDYAYKKAGNAKDIAEKAQDAKDFAYDKAGYGYDKAGDVIRMA  
MDSGYDKAGDAKDMAYEKAQIVKDMAYDKAGDAKDVAYEKAQIA  
KDFAYDKGGYAKDVAYDKAGNAKDMAYEKAQNAKDMAYEKAHEH  
ENYAYDKANYVKDSAYDNAQYAKDFAENKAQYAKDFAYDKAGDA  
KDYTAQKAAETKDATMEKAKAKDITVQKTEYKDYAAEKAKEG  
KDYAAQKAAEKDATMQKTEYKDYAAQKTAETKDATMEKAKAY  
KDTVADKAKELKDEAVTKVGELKDKATEKAGELKDKVVDKAKEL  
KDAIVEKAKETADYTAEKVGEYKDYTVDKAKAKDITAEKAKET  
ANYTADKAVEAKDKTAETKGEYKDYAVDKAVEAKDKTAEKAKET  
ADDCASQKAKEAKDAAMAKAGDYADYASQKAKEAKDKTLEKGGY  
TDYASQKAKEAKKTIEMKGGYKDYSAEKAKERKDATVNMKGGEY  
ADYAAEKAREAKDRTADKTKETAETAEKAREAKDITADKLG  
TEETAESTRESSENAAEKARKTKDSAAEKTRETKCAAEKAKAY  
DEDASRKTTQSTESAADKAEHETKDSVAQRGEEGKSIMGALGNM  
RDYTAEKATIEAKDKTAETKGEYKDYTVVEKATEGKDYTVSKLGEL  
KDYAAEKAKEGKDATVNLKGEYKDYTAETKTEGKDTAMEKLGEL  
KDY.AERIKAEKDATLQKAAEYKDYTAEKAKEEKDSATSKLGEL  
KDYTAEKAKNAKDTIVQKAGEYKDYTAEKATIEGKDSAVGKLGEL
```

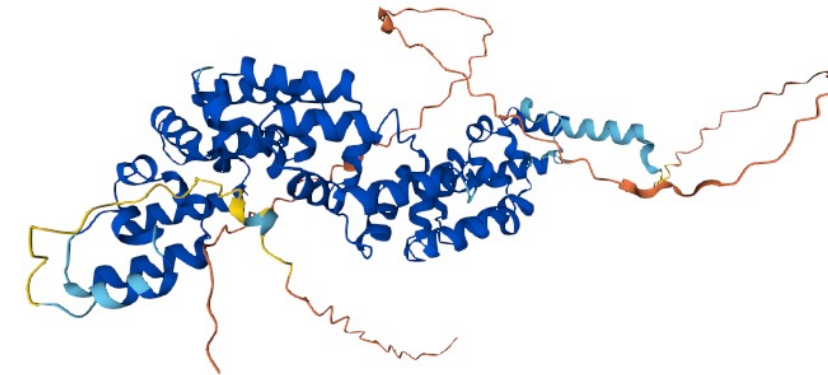
Examples

Transcription termination factor MTERF4, chloroplastic

```
MKIRFCNGFTKPGFLLVHFEPSPFFAVRSRSLSDSTYGNLCNHKKRPGTG
IGLTVQCAIANRRFSSRSLDSPRRERSSSRssssssgrdrdrdkdkGRDSKS
LYSRPSLLDMNKEKAANRAKVYEFRLGIGIVPDELDGLELPVTADVMKER
VEFLHKLGLTIEDINNYPLVLGCSVKKNMVPVLDYLGLKLGVRKSTFTEFL
Rrypqlvhssvvidlapvvkylqgldikpsdvprvlerypevlvgfklegt
mstsvaylvlgigvarreiggiltrypeilgmrvariikplveylevlgip
rlaaarliekrphilgfelddtvkpnvqilqdfnvretslpsiaaypei
igidLKPCLDTQRKLLCSAHLNPEDLGSliermpQFVSLSESPMLKHID
FLTKCGFSIDQTREMVIGCPQVLALNLGIMKLSFEYFQKEMKRPLQDLVD
FPAFFTYGLESTVKPRHKKI IKKGIKCSLAWMLNCSDEKFEQRMSYDTID
IEEVETDPSSFDMNTLMQPEREeesdseyeeeddddeEFA
```

PFAM: mTERF

This family contains one sequence of known function Human mitochondrial transcription termination factor (mTERF) the rest of the family consists of hypothetical proteins none of which have any functional information. mTERF is a multizipper protein possessing three putative leucine zippers one of which is bipartite. The protein binds DNA as a monomer. The leucine zippers are not implicated in a dimerisation role as in other leucine zippers [1].



```
M4F6E8_BRARP/81-426      LSLRSHSTDSHSSITVINPRLITADAFSLSPMLKTKSRSSSSESETISKVPRILVYKKEKSTIYDFVRDWI
D7KUT9_ARALL/459-797    LSLRSYAFIDSQSIIVTDYFQLITADAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
Q8GY68_ARATH/88-435     LSLRSYAFINSQSSIIITDYFQLITADAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
Q80701_ARATH/88-433     LSLITSYQFTKSSSIITTYFRLALDAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
Q9STG7_ARATH/88-426     LNLRSNPFDSQSRIIRAYFRLITDAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
M4DN04_BRARP/80-417     LNLRSNPFDSQSRIIRAYFRLITDAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
M4FEA5_BRARP/52-398     LNLITSYQFTSTQSKIIITTYFRLITDAKSLAPLQILSRSSSSEIANVSTVEKILKKKDKTSTIYVDIVKEII
W1PW38_AMBIC/164-472    LDYLSKLGVRKSNTEFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
M0LSV7_SOLIU/122-430    LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
MTEF4_ARATH/183-491     LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
I1JH93_SOYEN/166-474    LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
B7ZY02_MAIZE/130-439    LSYLEKLGVRTRALAAFVRAVRAELHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
W1NFR5_AMBIC/294-602    LAYLEKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
O80572_ARATH/148-456    LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
D8RFQ1_SEIML/131-440    LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
D8RWN7_SEIML/115-422    LDYLSKLGVRKSTITDFIRRYFQVLHSSVVIDLAFVKKYLQSLDKNDIFSVLERYPEVLGFKLEGMTSTSVAYLVSI
M0SIE4_MUSAM/283-589    VEFLELLGVFEATATILLSEPPPIIFDIEKIKKPLCAFSKSTEEKDIAMLMKKEWILSTIQENYEKILAFNEK
F6GYK9_VITVI/267-574    MKEFEDIGVQKSRMNVLLYPPPIIFDYDEKIKPELLAFEKIHAADKDIAGRMVKEWILSTIQENYEKILSFFYRE
B9RH14_RICCO/263-569    VEFESLGVEKEHMDSTIFLLFPVLIQDIKVIRKVLALKEKVAVDDEFGKMTFKWILSTIQDNKYEILSFCDAE
B9HG78_POPTR/272-578    VEFESLGVEKEHMDSTIFLLFPVLIQDIKVIRKVLALKEKVAVDDEFGKMTFKWILSTIQDNKYEILSFCDAE
Q0NRY2_ARATH/268-575    VEFESLGVEKEHMDSTIFLLFPVLIQDIKVIRKVLALKEKVAVDDEFGKMTFKWILSTIQDNKYEILSFCDAE
M1AI20_SOLIU/279-586    MYFLDDIGVVECKRQIILLFPPIIFDYDEKIKPELLAFEKIHAADKDIAGRMVKEWILSTIQENYEKILSFFYRE
W1PWJ1_AMBIC/270-577    IEFKLIGVSESSRSVLSFPPIILDYDERMKPRMNLKELGDKNDKIGMIRKVEWILSSSIQANYKILAFSTDE
J3M035_ORYER/173-480    IDFYQYIGIKARTASVLSFPPIILDYDERMKPRMNLKELGDKNDKIGMIRKVEWILSSSIQANYKILAFSTDE
M0YQP7_HORVV/81-388    IDFYQYIGIKARTASVLSFPPIILDYDERMKPRMNLKELGDKNDKIGMIRKVEWILSSSIQANYKILAFSTDE
O64531_ARATH/84-357    LQVLRWCCDDDESKLFTRRBALQRANVACLEFHLSLKELSTISSDILVKILNCRFFFSCLV.....
HOV117_CAVPO/84-397    LNNFTMGV...LDIDAKKQKQGVFNRTVINEQVQMFLSKKASKEVLASISRYERAITRTTESLSKRNWLNRKVI
G3I220_CRIGR/35-348    LSNLLTMGV...LQDMARRRQKGVFNRTVINEQVQMFLSKKASKEVLASISRYERAITRTTESLSKRNWLNRKVI
```

Rost, 1999

Twilight zone of protein sequence alignments

Burkhard Rost^{1,2,3}

¹EMBL, 69 012 Heidelberg, ²LION Bioscience AG, Im Neuenheimer Feld 517, 69 120 Heidelberg, Germany and ³Columbia University, Department of Biochemistry and Molecular Biophysics, 650 West 168 Street, New York, NY 10032, USA

Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). The signal gets blurred in the twilight zone of 20–35% sequence identity. Here, more than a million sequence alignments were analysed between protein pairs of known structures to re-define a line distinguishing between true and false positives for low levels of similarity. Four results stood out. (i) The transition from the safe zone of sequence alignment into the twilight zone is described by an explosion of false negatives. **More than 95% of all pairs detected in the twilight zone had different structures.** More precisely, above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were. (ii) Whether or not sequence homology implied structural identity depended crucially on the alignment length. For example, if 10 residues were similar in an alignment of length 16 (>60%), structural similarity could not be inferred. (iii) The 'more similar than identical' rule (discarding all pairs for which percentage similarity was lower than percentage identity) reduced false positives significantly. (iv) Using intermediate sequences for finding links between more distant families was almost as successful: pairs were predicted to be homologous when the respective sequence families had proteins in common. All findings are applicable to automatic database searches.

pe
(F
be
ar
id
T
m
(I
ai
(V
l
Fi
ti
m
to
id
fc
ol
di
tv
L
P
be
zc
pe
a
T
pl
al
se
Ir
,

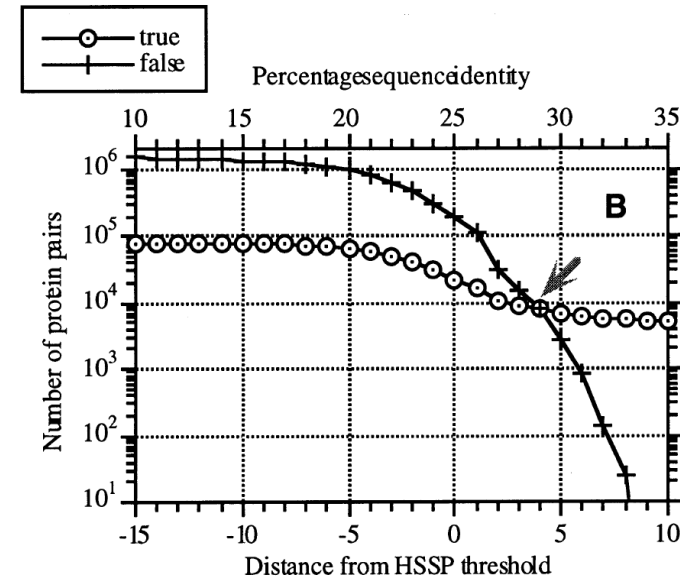


Fig. 2. Explosion of structurally dissimilar pairs in the twilight zone.

Distance to HSSP threshold

The HSSP-curve was originally defined by (Sander and Schneider, 1991):

$$p^l(n) = n + \begin{cases} 290.15 \cdot L^{-0.562}, & \text{for } L < 80 \\ 25, & \text{for } L \geq 80 \end{cases} \quad (1)$$

where L gave the number of residues aligned between two proteins; p^l the cut-off percentage of identical residues over the L aligned residues; and n described the distance in percentage points from the curve ($n = 0$ corresponds to the original HSSP-curve; $n = 5$ to the official HSSP database releases; curve plotted in Figure 3). Once Schneider and Sander

Rost, 1999

```
Query= A0A011

Length=281

>A0A023GPI5
Length=231

Score = 22.7 bits (47), Expect = 95.7
Identities = 31/124 (25%), Positives = 39/124 (31%), Gaps = 25/124 (20%)

Query 109 VRGADALLLPALLGSGDDYFVWKSFLETLAAFPGRIPREEWPPELLLTVALT-----FG 161
          VRGA  + AL  + D+  ++  F+ T      G   R E P L              FG
Sbjct  39 VRGATLAMPGALDRADDETLIYPLFMAT-----GWFTRSELPRRLALAGAPKARILPPFG 93

Query 162 EDPRTGDLLGTVFPVSTASTEEID---RYLHVARAFGFHVMVLYSRNEHVPPEVVRHFRK 217
          DP   L   +   A T+      R L A   G              P E R
Sbjct  94 SDPGLPALCLALIAQAETQGWPLAGTRLLVAAHGSG-----RSRAPSEAARRIAA 144

Query 218 GLGP 221
          GL P
Sbjct 145 GLAP 148
```

CATH annotation of PDB sequences

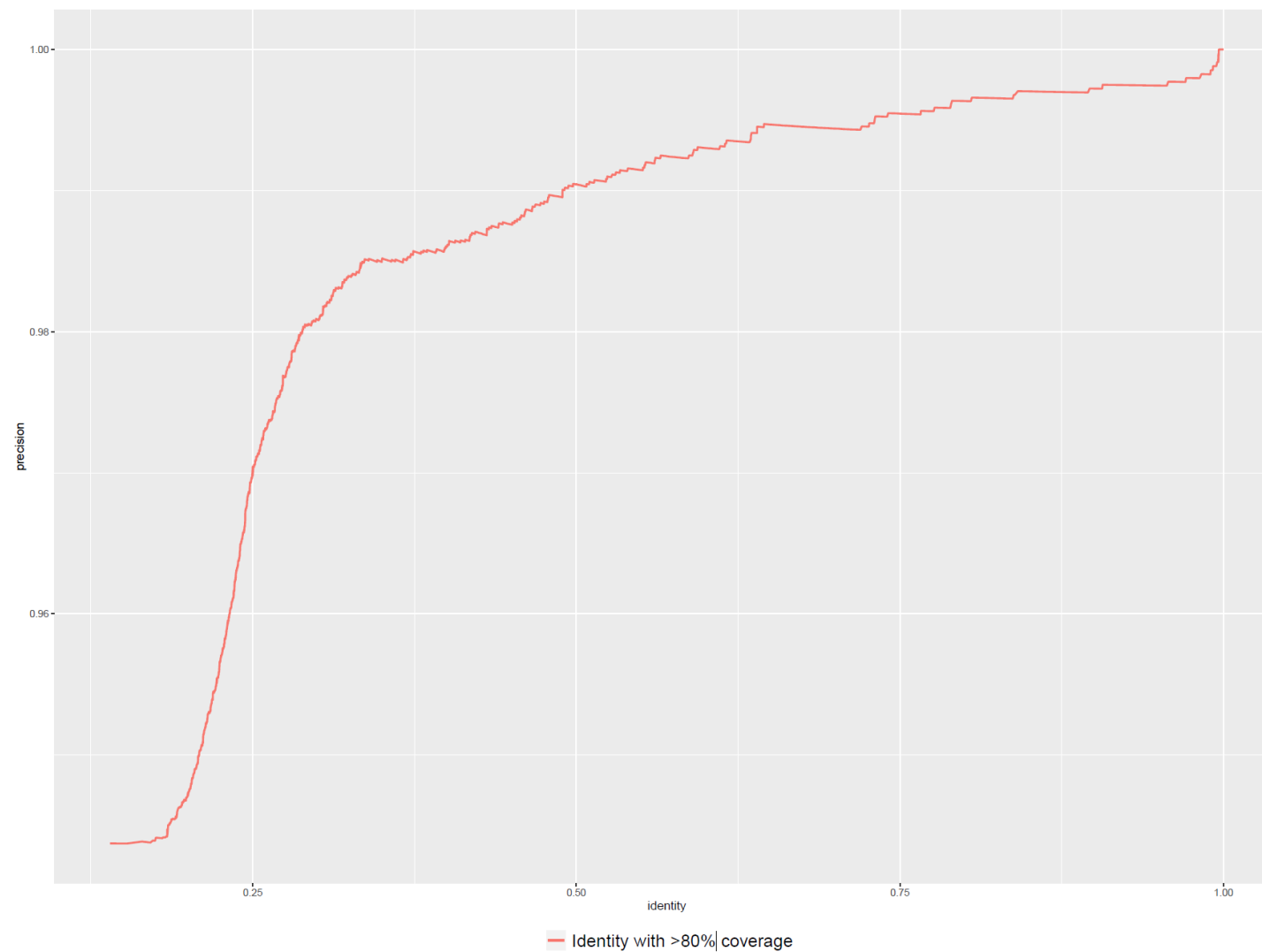
187,170 structures in the PDB, 128,000 with CATH fold annotation, corresponding to 35,713 primary sequences from UniProt.

14,987 of these sequences are annotated with CATH domains over >70% of their range.

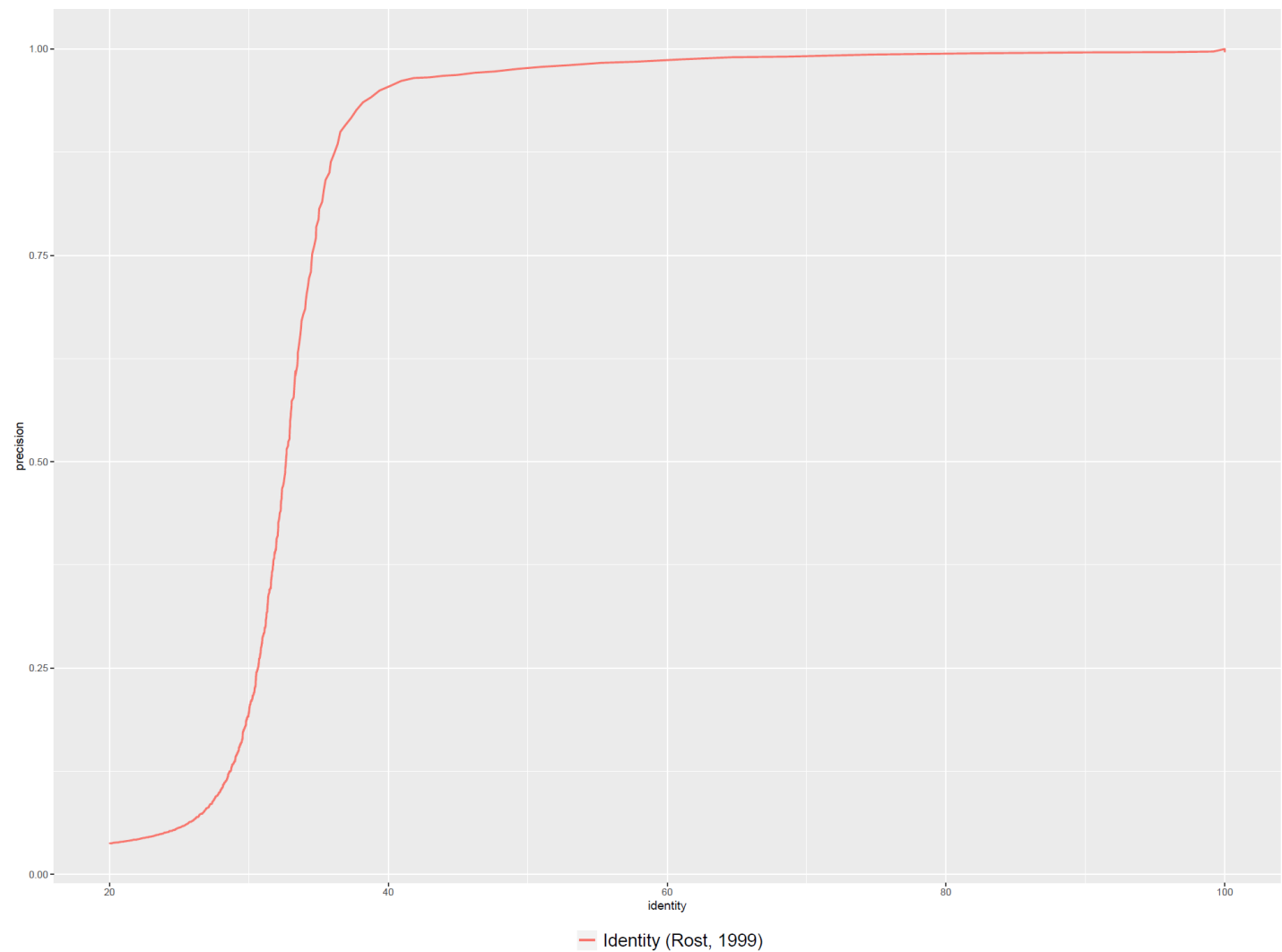
Clustering at 25% identity results in a query set of 3,138 sequences.

Positives := Sequence pairs with identical CATH domain architecture

CATH annotation of PDB sequences



Rost, 1999



Vector instructions

Count the matching amino acids:

...ITV**T**AGATEALYA**A**I**T**ALVRN**G**DEVICFD**P**S**Y**DSYAPAIALS**G**GIVKR...
...VVI**T**P**G**SSGGFLL**A**F**T**ALFDS**G**DRV**G**IGAP**G****Y**PSYRQILRAL**G**LVPVD...

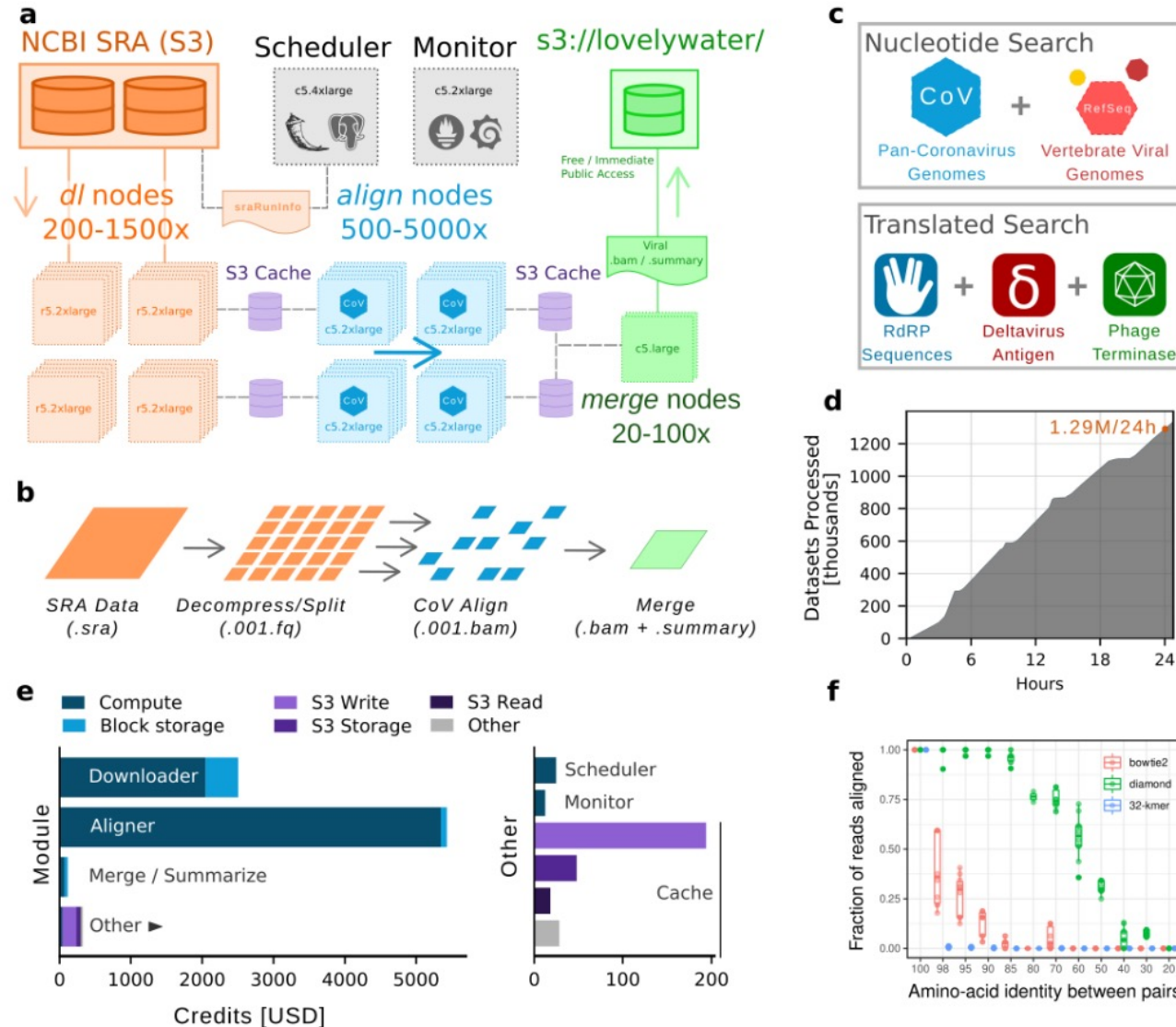
Naïve solution (~15ns):

```
int n = 0;
for(int i = 0; i < 48; ++i)
    if(a[i] == b[i]) ++n;
```

Using AVX2 instructions (~0.58ns)

```
_popcnt32(_mm256_movemask_epi8(_mm256_cmpeq_epi8(a, b)))
```

Project Serratus (<http://serratus.io>) (Edgar et al., Nature 2021)



Project Serratus (<http://serratus.io>)

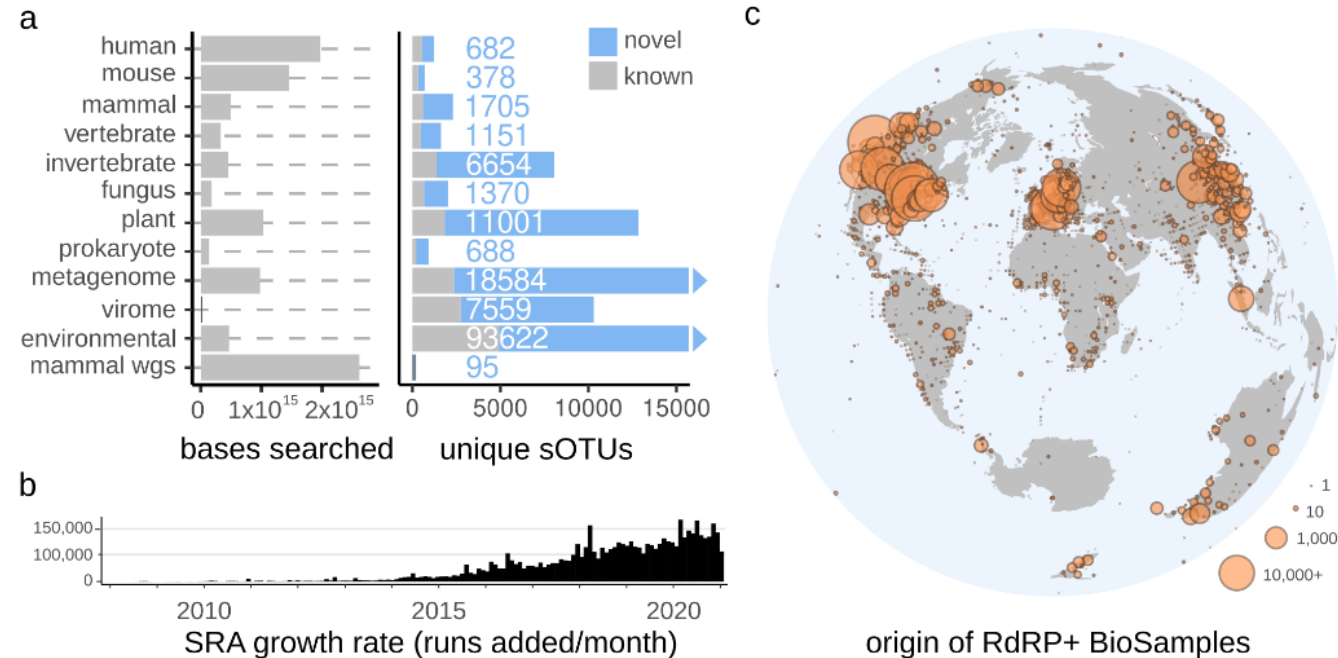


Figure 1: Searching the planetary virome

a Total bases searched from the 5,686,715 SRA sequencing runs analysed in the viral RdRP search grouped by sample taxonomy, where available (see Extended Figures 1 and 3, and Extended Table 1). 8,871/15,016 (59%) of known RdRP species-like operational taxonomic units (sOTUs) were observed in the SRA, and 131,957 unique and novel RdRP sOTUs were identified (see Extended Figure 2). sOTUs identified in multiple taxonomic groups are counted in each group separately, numbers shown indicate the number of novel sOTUs in each group. **b** Release dates of the runs included in the analysis reflecting the growth rate of available data. **c** Sample locations for 635,656 RdRP-containing contigs (27.8% of samples lacked geographic metadata). The high density of RdRP seen in North America, Western Europe and Eastern Asia reflects the substantial acquisition bias for samples originating from these regions. Interactive map is available at <https://serratus.io/geo>.

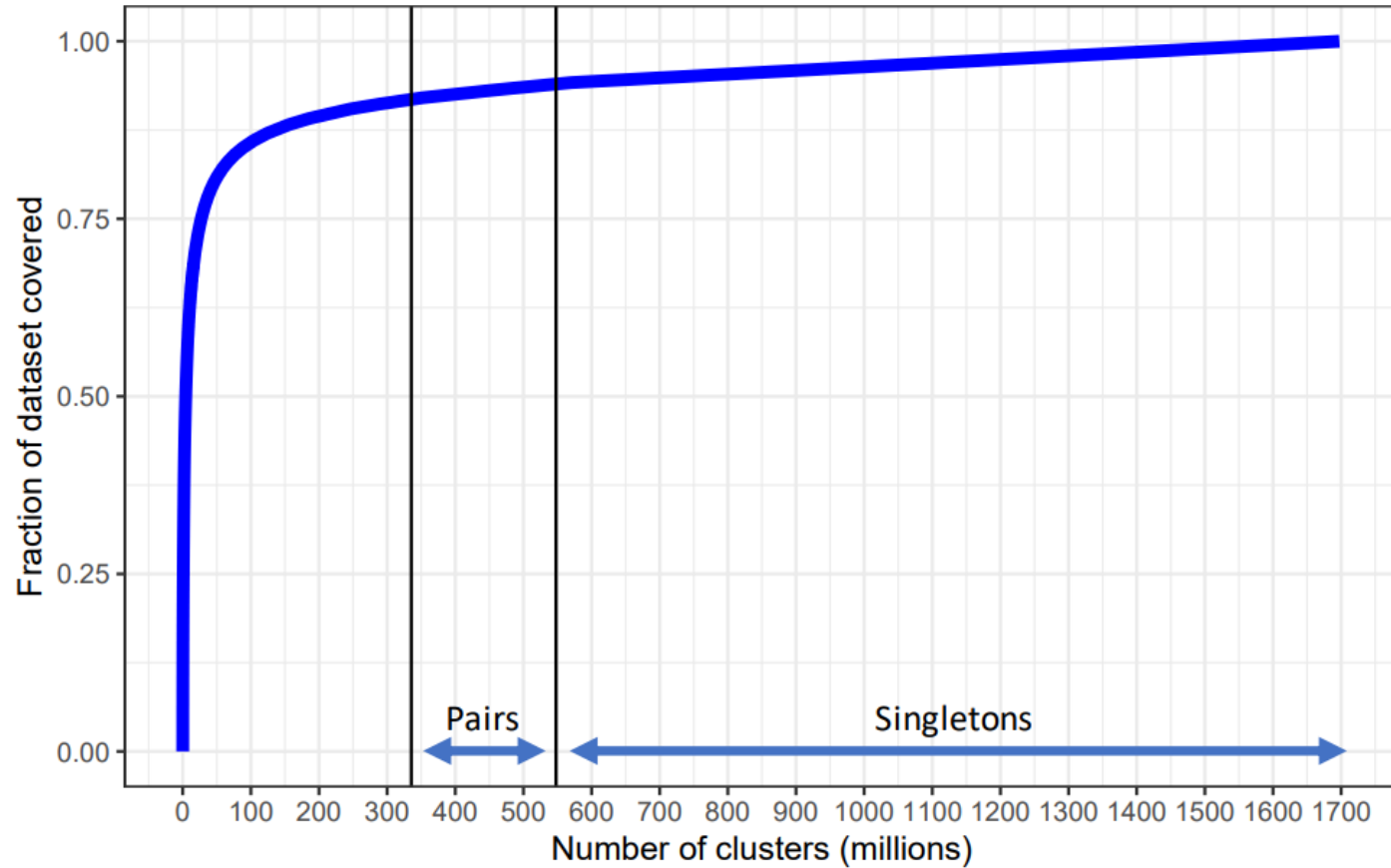
DIAMOND DeepClust

Database name	Number of sequences	Download date/release
IMG Environmental Aquatic Metagenomes	6,232,188,951	March 2022
IMG Environmental Non-Aquatic Metagenomes	6,851,118,226	March 2022
IMG Host-Associated Metagenomes	1,905,657,909	March 2022
IMG Engineered Metagenomes	801,972,930	March 2022
SRC	2,022,891,389	March 2022
MGnify	1,977,479,951	2022_05
metacrust	1,757,323,526	March 2022
NCBI NR	465,406,186	April 2022
AGNOSTOS	427,306,945	Ver 5
MERC	292,137,902	March 2022
MetaEuk	12,111,301	2019_11
SMAGs	10,207,435	v1
TOPAZ	8,405,914	v1
GPD	7,581,807	April 2022
NovelFams	4,587,583	March 2022
MGV	11,837,198	v1.0
Total	22,788,215,153	

	Alphafold2 BFD	DIAMOND run
Input sequences	2,423,213,294	22,788,215,153
Deduplicated sequences	2,423,213,294	19,387,935,704
Clusters	345,159,030	1,697,446,279
Singleton clusters		1,149,664,058
Clusters ≥ 3 members	61,083,719	335,433,899

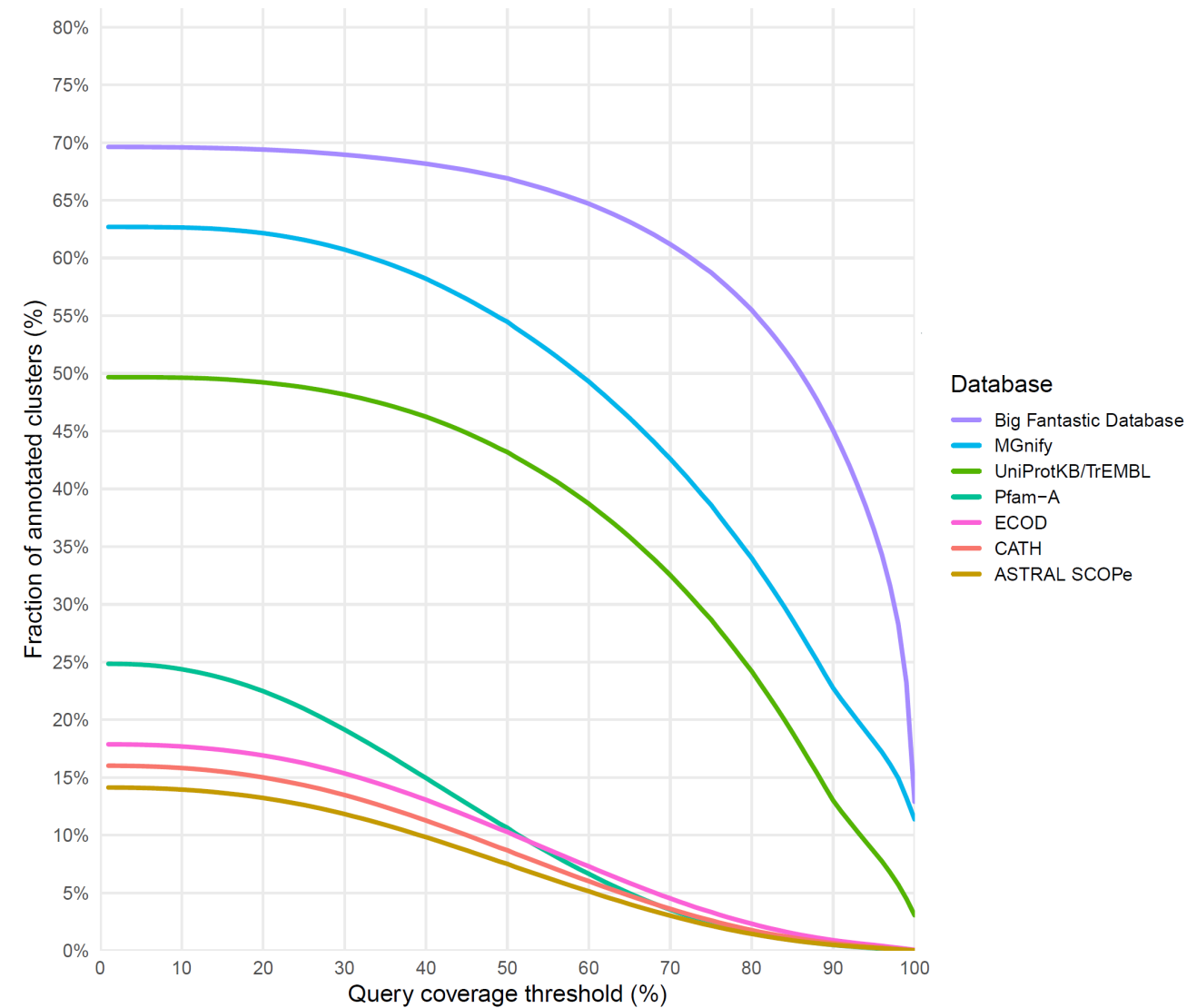
Time: ~250,000 CPU hours

DIAMOND DeepClust

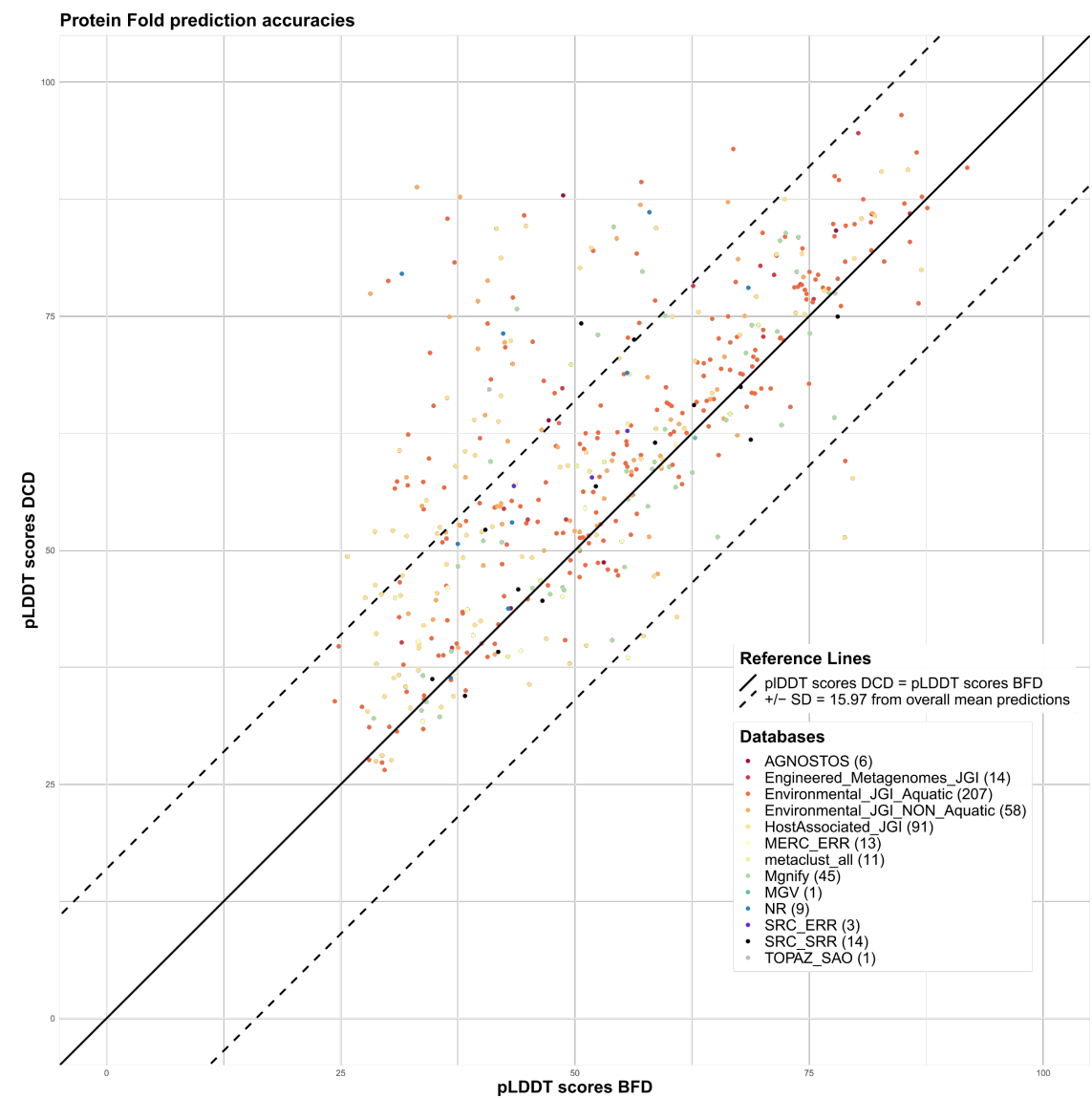


Supplementary fig. 10 | Cluster representation for the experimental study. Shown is the number of clusters generated with DIAMOND DeepClust against the fraction of the 19.4 billion input proteins that are covered by the clustering. The vertical lines indicate the start of clusters of size two and one (from left to right). The result illustrates that ~335 million representatives can capture 92% (~17.48 billion sequences) of the protein universe that comprises 19 billion sequences.

DIAMOND DeepClust



DIAMOND DeepClust



Multiple Sequence Alignment

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLLDNDYPKCFIVGADNVGSKQMQQIRMSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAAWKAQYFIKVVELFDEFPPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PQLE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSQLOKIRKSIRGI-GAVLMGKNTMMRKAIRGHLENN--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSQLOKIRKSIRGI-GAVLMGKNTMMRKAIRGHLENN--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKLSSLIQQYSKILIVHVDNVGSKQMASVRKSLRGK-ATILMGKNTMRIRALKKNLQAV--PQIE	76
RLA0_SULAC	----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIITANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFNIALKNAG----YDTK	79
RLA0_SULTO	---MRIMAVITQERKIAKWKIEEVKELEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNLTFGIAAKNAG----LDVS	80
RLA0_SULSO	---MKRLALALKQRKVASWKLEEVKELTELIKSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNLTFKIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSILVGQMYKREKPIPEWKTLMLELEELFSKRRVVFADLTGTPTFVVQVRKKLWKK-YPMMAVAKKRIILRAMKAAGLE---LDDN	86
RLA0_PYRAE	-MMLAIGKRRYVRTROYPAKVKIVSEATELLQKYPYVFLFDLHGLSSRIILHEIYRRLRY-GVIKIIKPTLFFKIAFTKVYGG---IPAE	85
RLA0_METAC	-----MAEERHHTEHIPQWKKDEIENIKELIQSHKVFQMGVIEGILATKMQKIRRDLDV-AVLKVSNTLTERALNQLG----ETIP	78
RLA0_METMA	-----MAEERHHTEHIPQWKKDEIENIKELIQSHKVFQMGVIEGILATKMQKIRRDLDV-AVLKVSNTLTERALNQLG----ESIP	78
RLA0_ARCFU	-----MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNLTLERALDALG----GDYL	75
RLA0_METKA	MAVKAKGQPPSGYE PKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLOKMRQTLRDS-ALIRMSKKTLLISLALEKAGREL--ENVD	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAQLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLOEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA	82
RLA0_METJA	-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAQLOEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA	81
RLA0_PYRAB	-----MAHVAEWKKKEVEELANLIKSPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKKEVEELAKLIKSPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEELANLIKSPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEELANLIKSPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAQELGQPELE	76
RLA0_HALMA	-----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD----DGLE	79
RLA0_HALVO	-----MSESEVRQTEVIPQWKREEVDELVDIESYESVGVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN----DGFE	79
RLA0_HALSA	-----MSAEEQRTTEEVPEWKQEEVDELVDIESYESVGVGVAGIPSKQLQDMRRGLHGG-AALRMSRNTLLVRALEEAG----DGLD	79
RLA0_THEAC	-----MKEVSQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFFKALENLGD----EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAQDITKSKAVIVDIKGVTRQMODIRAKNRDK-VKIKVVKKTLLFFKALDSIND----EKLT	72

Multiple Sequence Alignment

- Comparative genomics
- Phylogeny
- Protein structure prediction
- RNA structure and function
- ...

Multiple Sequence Alignment

- MUSCLE
- ClustalW
- T-Coffee
- MAFFT

Challenging alignment

FLVRESQARNPQG-FVLSLC	HLQ---KVKHY
FIIRFSERNPGQ-FGIAYI	GVEMPARIKHY
FLLRFSESSREGAITFTWV	--E---RSQNG
FLVRDASTKMHGDYTLTLR	--K---GGNNK

Alternative MSAs
of same sequences

Which one is
correct / better?

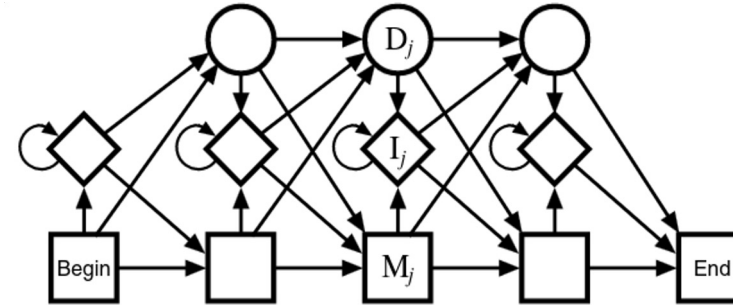
FLVRESQARNPQG-FVLSLC	HLQ----KVKHY
FIIRFSERNPG-QFGIAYI	GVEMP-ARIKHY
FLLRFSESSREGAITFTWV	ERSQNGGEPD-F
FLVRDASTKMHGDYTLTLR	--K---GGNN-K

Hard / impossible
to decide, even
with structures

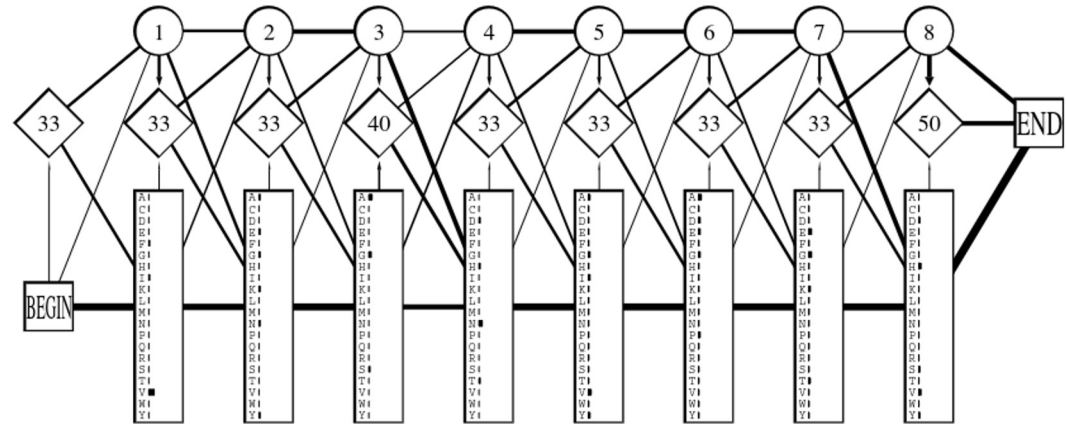
<https://www.youtube.com/watch?v=2HmjHStpu7I>

HMM profiles

Tools: HMMER, HHblits

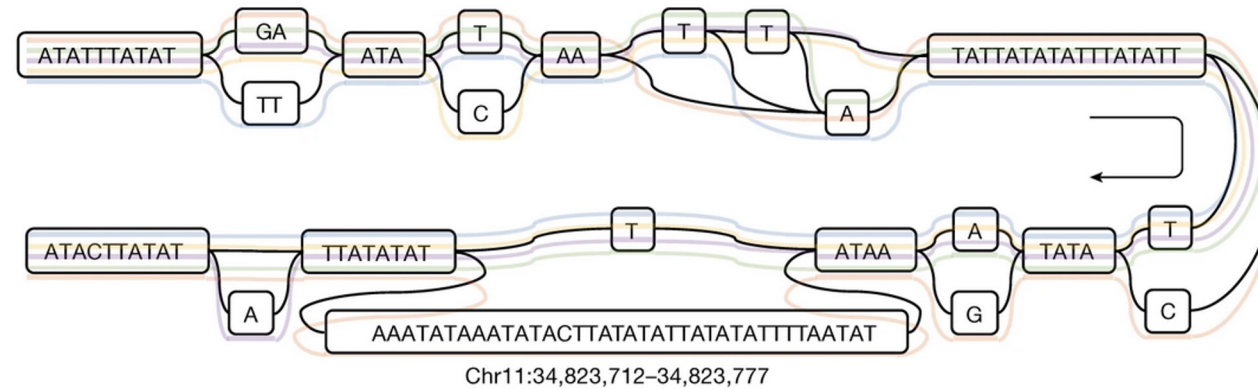


HBA_HUMAN	...VGA--HAGEY...
HBB_HUMAN	...V---NVDEV...
MYG_PHYCA	...VEA--DVAGH...
GLB3_CHITP	...VKG-----D...
GLB5_PETMA	...VYS--TYETS...
LGB2_LUPLU	...FNA--NIPKH...
GLB1_GLYDI	...IAGADNGAGV...
	*** *****



Multiple DNA

- Wayyy longer sequences
- Duplications, inversions, and translocations wreck linearity
- Tools: SibeliaZ, Cactus
- State of the art: human genome graphs, look for pangenomics papers.



AI

- More and more important.
- Main issue in biology: overfitting, limited training data
- Very easy to game any given benchmark.
- Something looks good on paper does not mean it will perform in the real world.
- Be careful and critical.

*Und es rief der Herr von Sachsen,
der von Bayern, der vom Rhein:
"Graf im Bart, Ihr seid der Reichste!
Euer Land trägt Edelstein!„
Lied der Württemberger, Justinus Kerner (1818)*

Thank you!

- Contact me: buchfink@gmail.com
- <https://github.com/bbuchfink/diamond>
- Join my discord