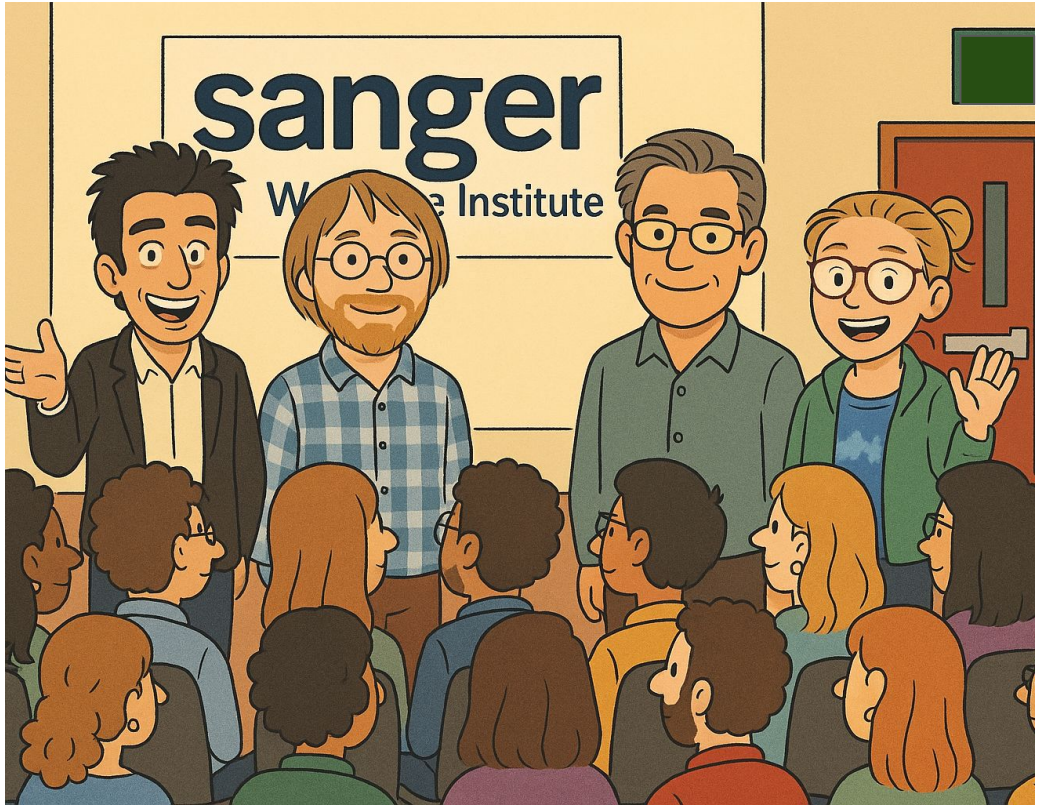
History and overview of k-mers in bioinformatics

Rayan Chikhi, Kamil Jaron, Katie Jenike, Gene Myers

K-mer Workshop for
Biodiversity Genomics
2025

Introductions



Scope of the talk

- Sequencing and k-mers
- Focus: historical aspects
- How we got to today
- Technical depth is later in the course

Genome sequencing

assembly

Genome sequencing

assembly

genome profiling

13,000,000 individual plants in the harvest;
on average each weighted 843g

How to look at the “bag of reads”?!

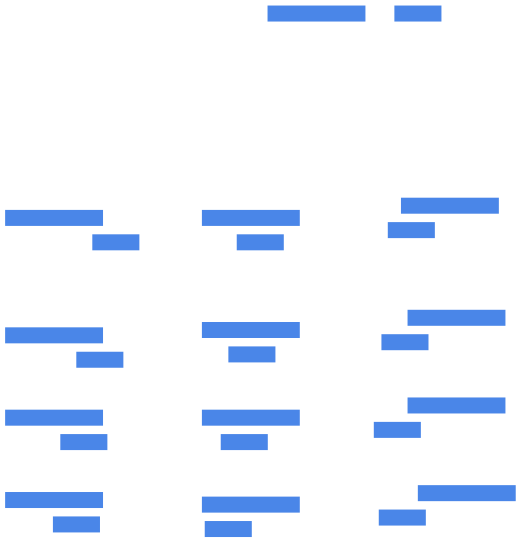
To get coverage, we need to identify what sequences belong together!

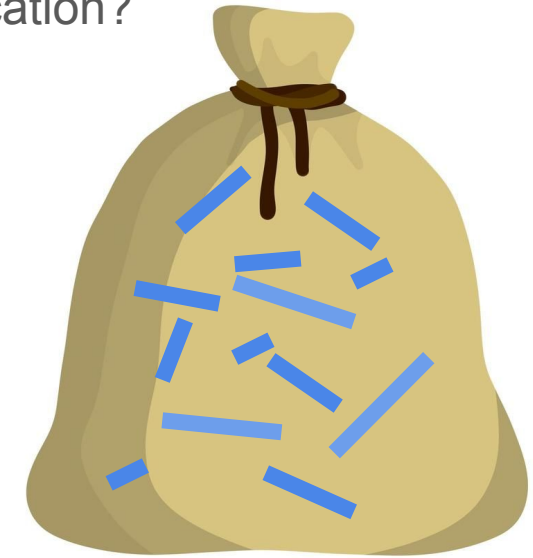
How to look at the “bag of reads”?!

How AI sees sequencing…

To get coverage, we need to identify what sequences belong together!

Imagine just two reads… Are they from the same location?

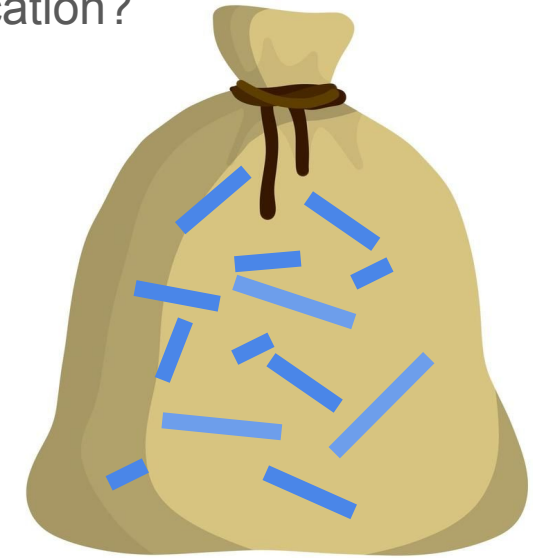




Try all possible overlaps?

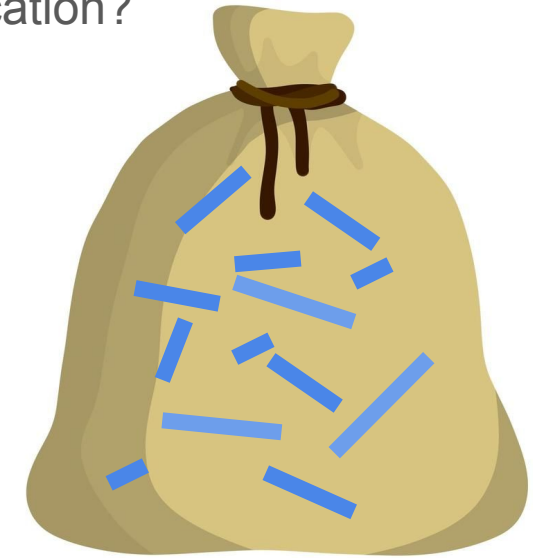
Imagine just two reads… Are they from the same location?

What if one has a sequencing error?
Do we discard all the reads that have an error?



Imagine just two reads… Are they from the same location?

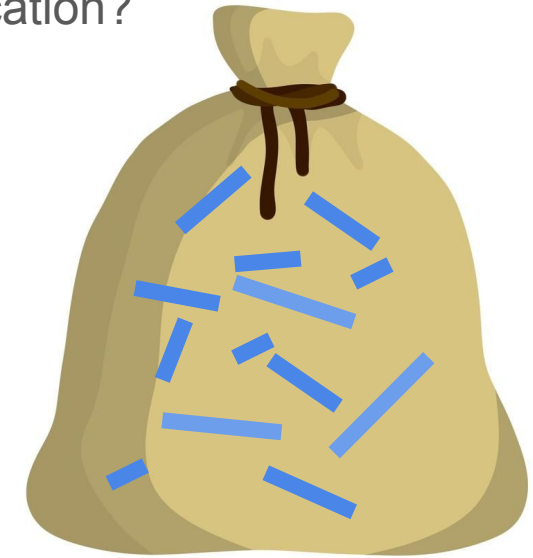
What if one has a sequencing error?
Do we discard all the reads that have an error?



This is not a problem with one good solution

Imagine just two reads… Are they from the same location?

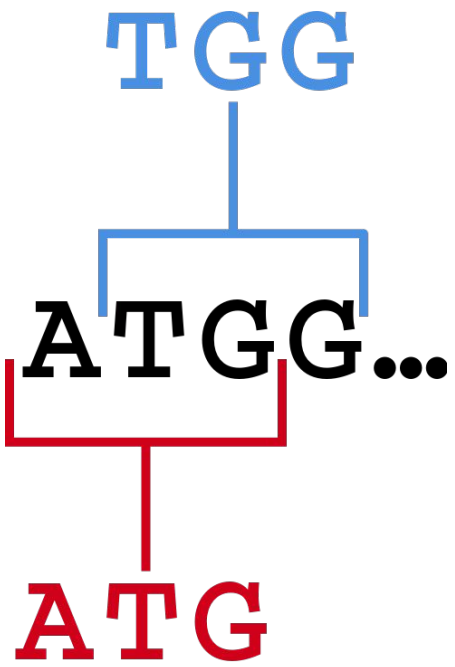
What if one has a sequencing error?
Do we discard all the reads that have an error?



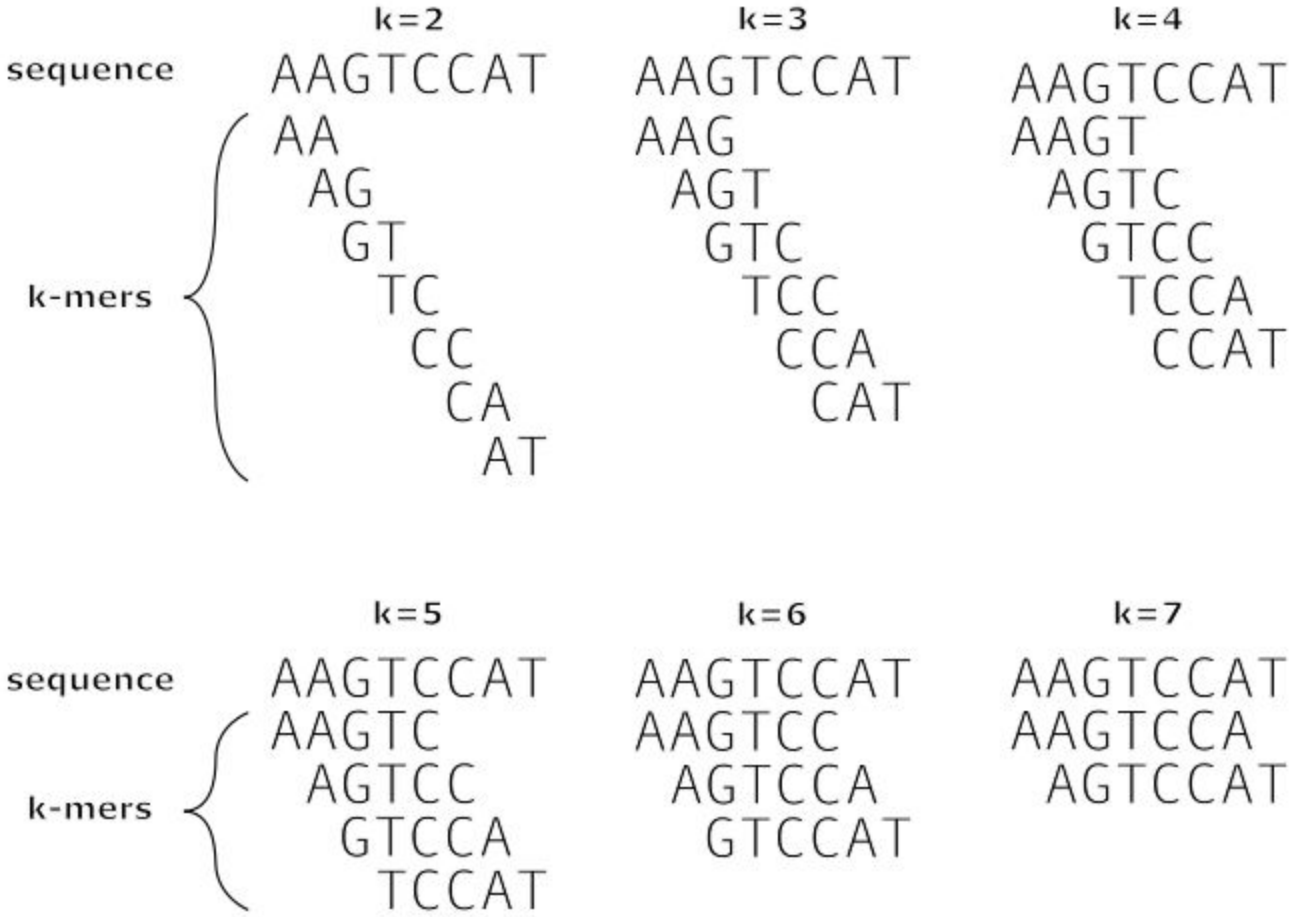
This is not a problem with one good solution

…but *k*-mers are part of nearly all existing ones!

What is a k-mer?



Sequence ATGG has two 3-mers: ATG and TGG.



https://github.com/KamilSJaron/k-mer-approaches-for-biodiversity-genomics/wiki/Concept-of-k-mers

History of the term

etymology slides

- **N-gram** (Shannon, 1948)
- **k-tuple** (Drmanac et al. 1991; Idury and Waterman 1995)
- **ktup** (Lipman and Pearson 1985)
- **L-tuple** (Idury and Waterman 1995)
- **ℓ-tuple** (Li and Waterman 2003)
- **k-word** (Lippert et al. 2002; Li and Waterman 2003)
- **word** (Reinert et al. 2000)
- **11-mers** (Drmanac et al. 1989)
- **w-mers** (BLAST, 1990)
- **k-mer** (MUMmer, 1999)

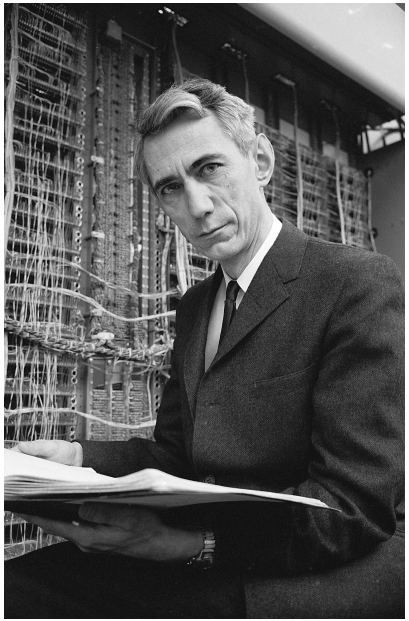
Jenike et al, 2025

Origins of k-mer concepts in early bioinformatics

Historical Roots

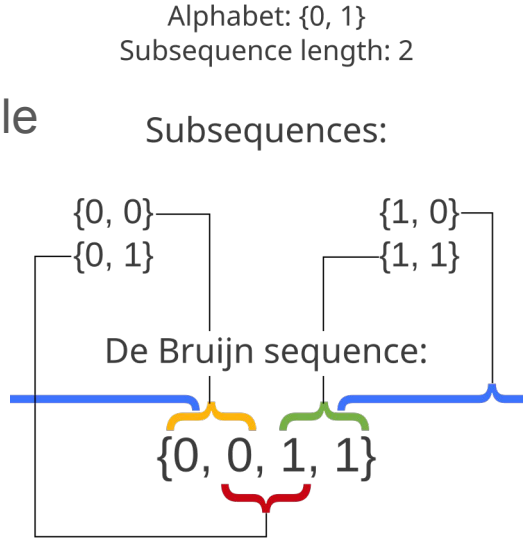
N.G. de Bruijn (1946)
de Bruijn sequences
Construct a short text that contains all k-mers

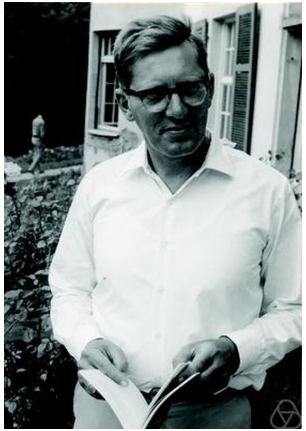
C. Shannon (1948)
Information theory
Predict what character is likely to follow a given k-mer



de Bruijn sequence

- A sequence where every possible k-tuple appears
- Its length must be: |alphabet|^k
- Not a lot of practical uses

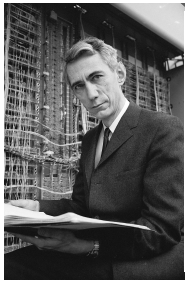




Math interests:

- How to construct it
- Counting how many possible sequences

Shannon's theory



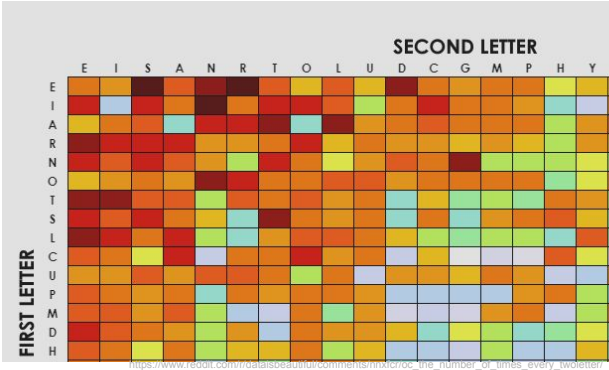
Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

- Given **frequencies** of English letters
- Generate a "plausible" text, drawing at random:
  - OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI …
- Now knowing the **2-mer** frequencies in English texts:
  - ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY …
- The **3-mer** frequencies:
  - IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID …
- Take-home: **k-mer frequencies encode information**
- n consecutive **words,** not letters:
  - n=2: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT …



How did Shannon call *k*-mers? N-grams

gram means “small weight”

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

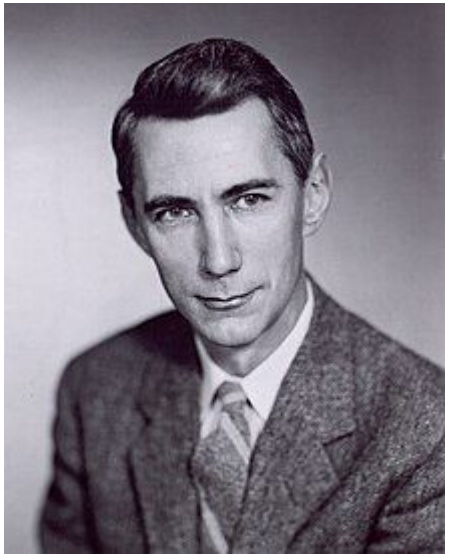
   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further struc-ture is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.



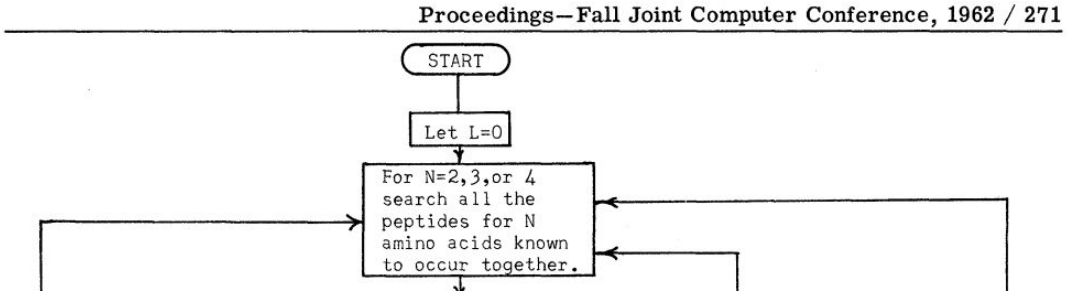
Claude Shannon

–A Mathematical Theory of Communication, 1948

Who introduced k-mers to bioinformatics?

Margaret Dayhoff:

- k-mer search in “COMPROTEIN: A Computer Program to Aid Primary Protein Structure Determination.” 1962

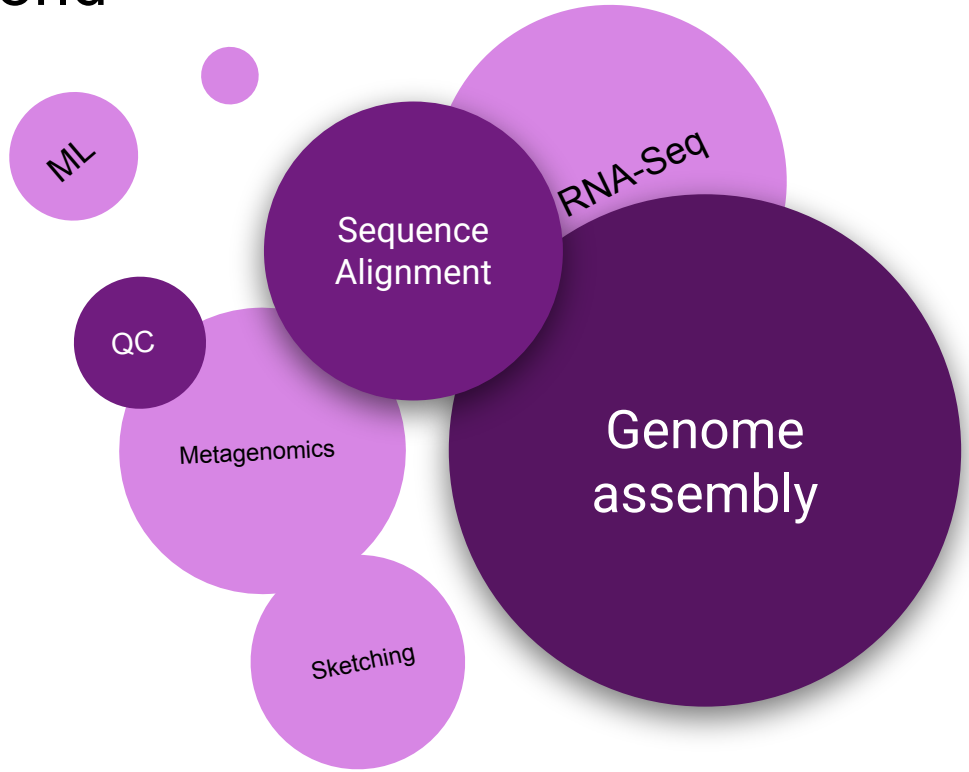


technologies to support advances in biology and medicine, most notably the creation of protein and nucleic acid databases and tools to interrogate the databases. She originated one of the first substitution matrices, point accepted mutations (*PAM*). The one-letter code used for amino acids was developed by her, reflecting an attempt to reduce the size of the data files used to describe amino acid sequences in an era of punch-card computing.

Margaret Oakley Dayhoff

Born	Margaret Belle Oakley March 11, 1925 Philadelphia, Pennsylvania
Died	February 5, 1983

It’s a k-mer world



Early Applications: Similarity searches in the 1980s

- **Smith-Waterman** (1981): local alignment (no k-mers)
- **FASTA** (1985): heuristic for faster search
  - used short exact matches as seeds — conceptually close to k-mers.
- **GenBank** launched (1979): early sequence databases
- Conceptual precursors to k-mers : “motifs” & “short matches”

FASTA and the use of 2-mers as seeds

- **FASTA** (Lipman, Pearson, 1985)
  - k=2 for proteins
  - k=4,6 for nucleotides

1) Identifies regions with high density of k-mer matches
2) Run Smith-Waterman alignment on them

The k-mer matches do not serve as anchors for the alignment.
(They serve to “light up” the right region(s))

“-tuples” or “words”? Why don’t you pick one?

“ordered list” or “finite sequence”

Michael Waterman

“k-tuple”
“L-tuple”
“ktup”

Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. J Comput Biol. 1995;2: 291–306.

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227: 1435–1441.

“ℓ-tuple”

Li X, Waterman MS. Estimating the Repeat Structure and Length of DNA Sequences Using ℓ-Tuples. Genome Res. 2003;13: 1916–1922.

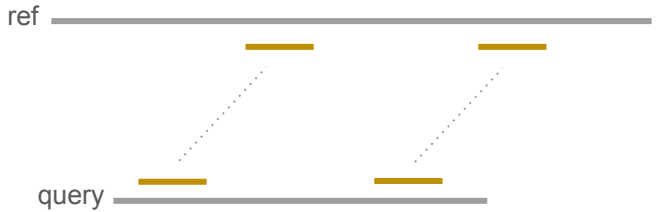
“k-word”

Lippert RA, Huang H, Waterman MS. Distributional regimes for the number of k-word matches between two random sequences. Proc Natl Acad Sci U S A. 2002;99: 13980–13989.

“word”

Reinert G, Schbath S, Waterman MS. Probabilistic and statistical properties of words: an overview. J Comput Biol. 2000;7: 1–46.

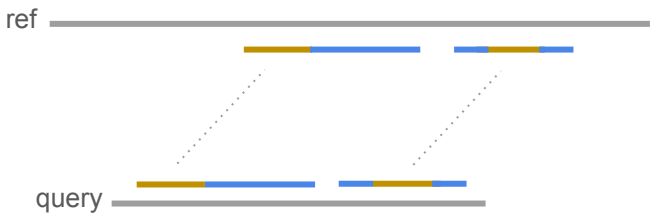
k-mers as Seeds in Alignment

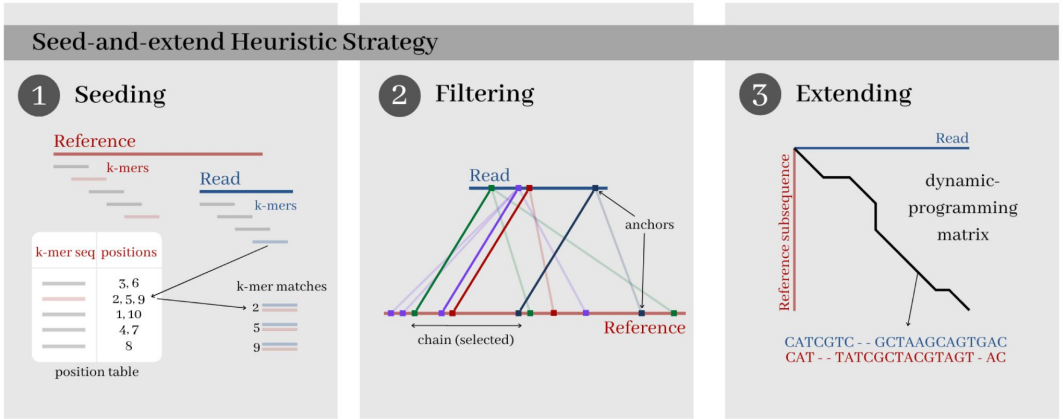


- **Query**
- **Reference**
- **Seed**: short sequence (possibly a k-mer) found in both the query and reference
- **Minimizers**: covered later in the course

Seed-and-extend approaches for alignment

1. Find seeds
2. Filter (or chain) seeds
3. Extend alignment to the left/right of seeds using exact algorithm (Smith-Waterman)





Teng et al, 2023

BLAST: longer seeds

- **BLAST** (Altschul, Gish, Miller, Myers, Lipman, 1990)
- Improves upon FASTA
- Uses k-mers as seeds
  - k=11,12 for blastn
- Then runs Smith-Waterman around the seeds

Compare to FASTA which has k=6.

“-mer” comes from “meros” which means “a share”

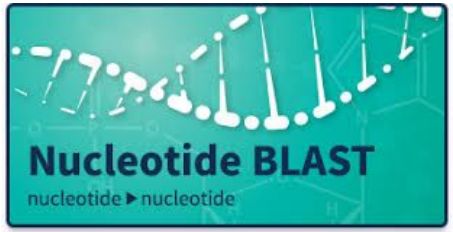
1968: “Unique Sequences in Tobacco Mosaic Virus”

oligomers
“Ψ-mer” and “Ω-mer”

Is Stanley Mandeles father of *k*-mers?!

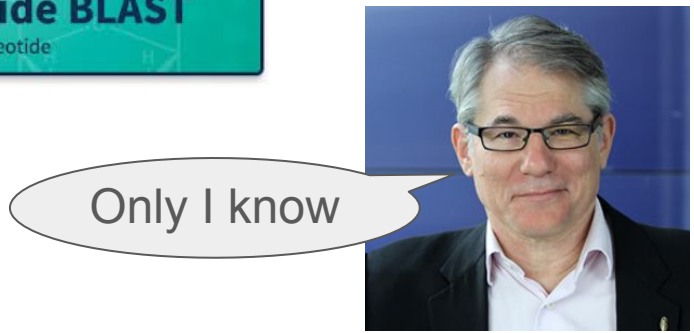
1990: BLAST uses “w-mers”

- Is the “w” for “words” here?



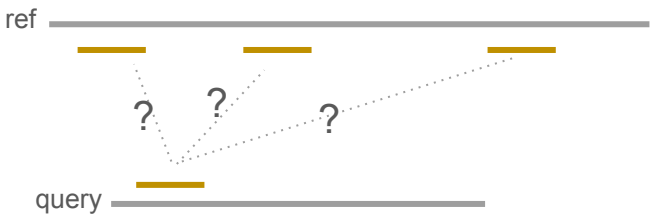
1999: MUMmer

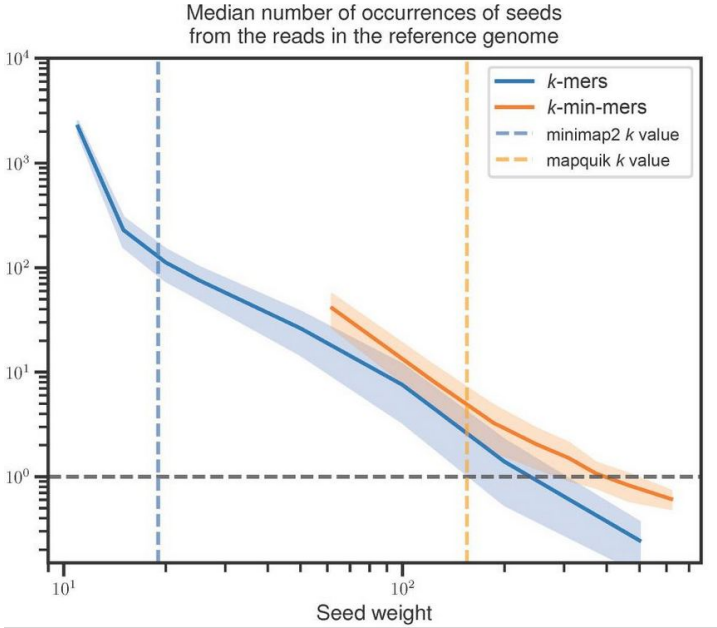
2010: Jellyfish; *k*-mers reach fixation



Balancing k-mer size for alignment speed vs. sensitivity

- **Long** k-mer seeds:
  - high specificity (= few candidate alignment locations)
  - low sensitivity (= less likely to find the right location)
- **Short** k-mer seeds:
  - low specificity (= many potentially wrong options)
  - high sensitivity (= more likely to find the right location)





Ekim et al, 2023 (mapquik, human genome)

k-mers in various alignment settings

- Reads-vs-genome
  - **Bowtie1** (Langmead et al, 2009)
    - k=28 as seed, with 2 mismatches

- Genome-vs-genome
  - **LASTZ** (Bob Harris)
    - Uses spaced seeds, k=19 with 8 mismatches
  - **YASS** (Noe, Kucherov, 2005)
    - Uses spaced seeds too

- Protein-vs-protein
  - **DIAMOND** (Bunchfink, Huson, 2015)
    - Uses a set of spaced seeds, k=??

Methods with even shorter k-mers (k=4, 5, 6, ..)

- Metagenomic binning tools
  - They look at nucleotide composition, to group contigs into genomes
  - **TETRA** (Teeling et al, 2004): **k=4**
  - Generalizes over GC content (k=1)
  - Many others too: MetaBAT2, MaxBin 2
- Metagenomics taxonomic classification
  - **QIIME2**'s q2-feature-classifier (Bukolich et al, 2018): **k>=4**
  - Naive Bayes classifier on k-mer counts
- Virus binning
  - **VirFinder** (Ren et al, 2017): **k=4**
- Genome masking
  - Komplexity (Clark 2019): **k=4**
  - Removes low-complexity regions

k-mers for determining genome structure

- Genome size and repeats (Li and Waterman, 2003)
  - Models k-mer histograms (without calling them that) as a mixture of Poisson distributions
  - Reconstructs repeat consensus sequences(!)
- Many follow-up works in modeling genome histograms
  - (Chor et al 2009, Liu et al 2013, Chikhi and Medvedev 2014, SPAdes assembler, Sun et al 2018, etc..)
- **CRT** (Bland et al, 2007)
  - Finds location of CRISPRs using exact k-mer matching
- **Tallymer** (Kurtz et al, 2009)
  - Find Transposible Elements in plant genomes using k-mers

k-mers in Genome Assembly

IDURY AND WATERMAN

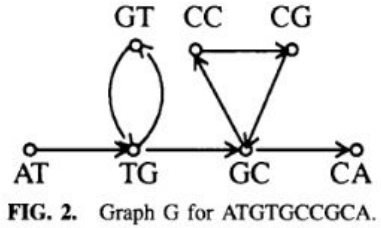


FIG. 2. Graph G for ATGTGCCGCA.

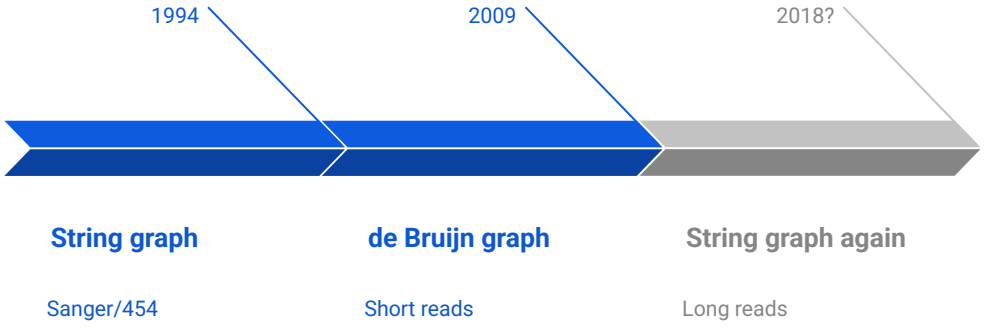
- **Sequencing by Hybridization** (1987)
  - Basically a microarray of k-mers
  - k=8
- String graphs and **de Bruijn graphs**, both introduced at the same conference (DIMACS 1994) A History of DNA Sequence Assembly, G. Myers, 2016
  - **Idury-Waterman** (1995 journal version)
    - de Bruijn graph theory for assembly (without naming it)
- **EULER** (Pevzner, Tang, Waterman 2001)
  - first assembly tool using a dBG
- Many many more tools: EULER-SR, Velvet, SOAPdenovo, SPAdes,

De Bruijn graphs, **unitigs**, **contigs**: will be covered later in the course (Logan lecture)

Choice of **best k:** also covered later (“In-depth understanding” lecture)

k-mers vs string graphs

- Early-day question of which formalism is most adequate for assembly:
  - k-mers?
  - or full reads?
- (Still today)



it – Information Technology 2016; 58(3): 126–132 DE GRUYTER OLDENBOURG

Special Issue Open Access

Eugene W. Myers Jr*

A history of DNA sequence assembly

DOI 10.1515/itit-2015-0047
Received October 26, 2015; accepted March 9, 2016

Abstract: DNA sequence assembly is a rich combinatorial problem that arose with the first DNA sequencing projects in the early 80's. Here we give a short history of the progression of algorithmic ideas used to solve the *de novo* problem of inferring a genome given a large sampling of substrings covering it. This classic inverse problem is compounded by a variety of experimental features and artifacts that must be considered in any realistic solution. While current methods produce very good results, the perfect assembler has yet to be built.

Keywords: Shotgun sequencing, shortest common superstring, string graph, de Bruijn graph.

ACM CCS: Applied computing → Life and medical sci-

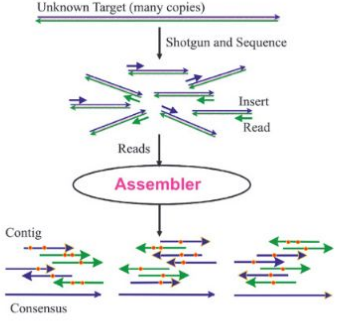
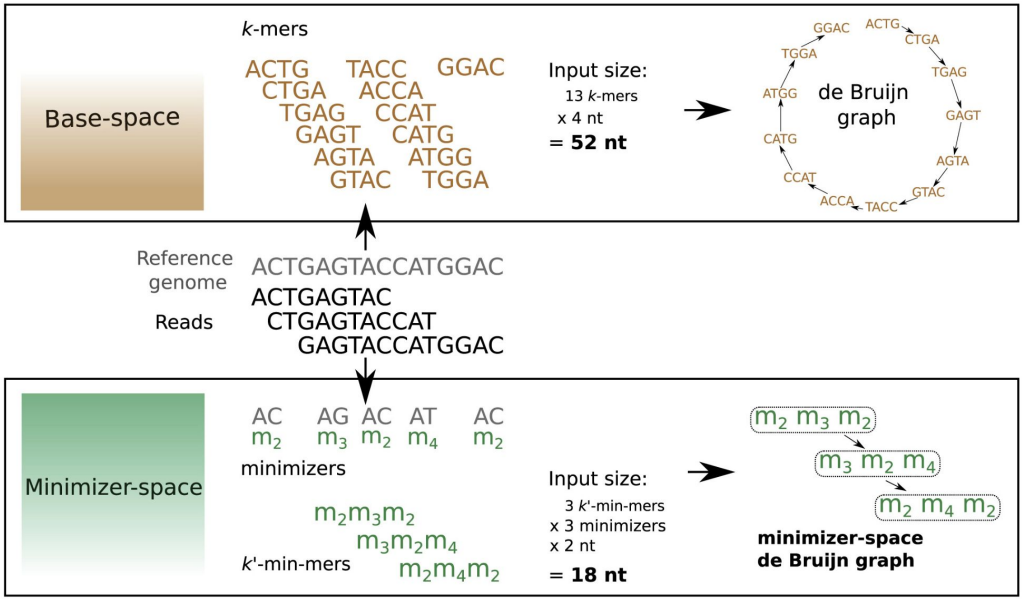


Figure 1: DNA Shotgun Sequencing.

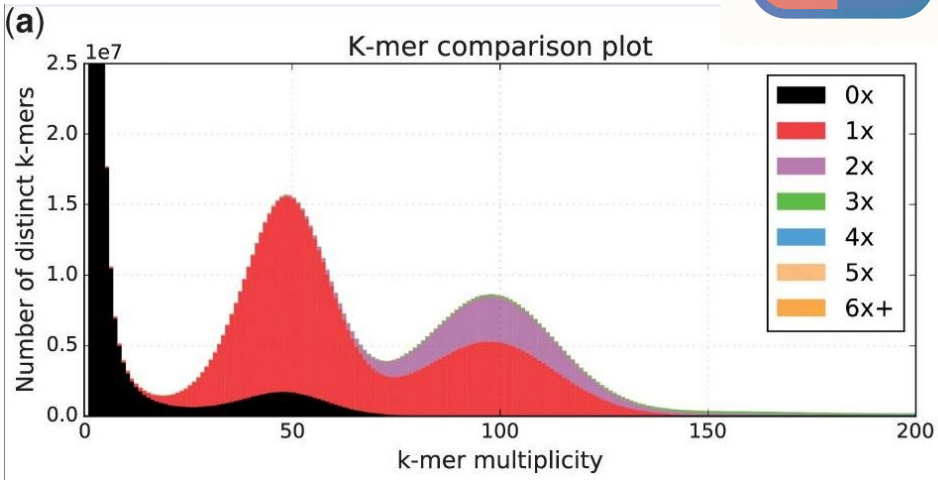
k-mers in more recent genome assemblers

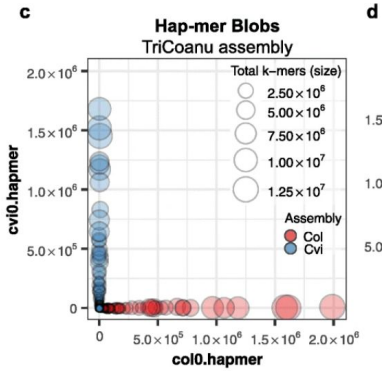
- **Verkko, LJA**
  - Apply long k-mers to long read assembly
- **rust-mdBG**
  - Uses k-mers over a different alphabet (minimizers)
- **metaMDBG**
  - State-of-the-art metagenomics assembly based on mdBG



k-mers for checking genome assembly quality

- **KAT** (Mapleson et al, 2016)
  - Compares k-mers in short reads vs k-mers in assembly
  - Or GC-content vs k-mer frequeny
- **QUAST-LG** (Mikheenko et al, 2018)
  - Uses k-mers to estimate completeness
- **Mercury** (Rhie et al, 2020)
  - Improves upon KAT
  - Extended to phased diploid genomes
  - Generates a consensus quality QV value

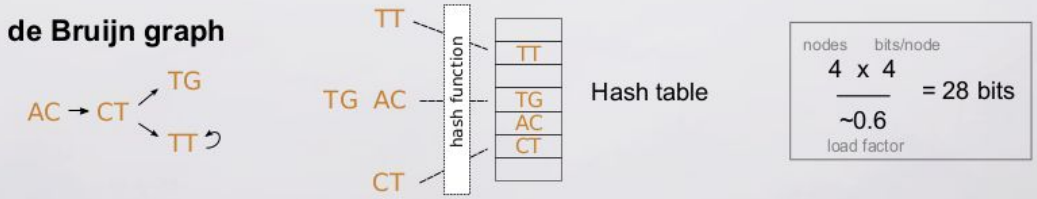




You'll see a lot of this during next week

k-mer storage optimization

Exact encoding of the de Bruijn graph using a hash table



RESEARCH-ARTICLE OPEN ACCESS

Data Structures to Represent a Set of *k*-long DNA Sequences

Authors: Rayan Chikhi, Jan Holub, Paul Medvedev Authors Info & Affiliations

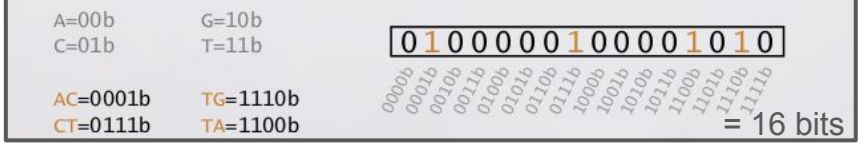
ACM Computing Surveys, Volume 54, Issue 1 • April 2021 • Article No.: 17, pp 1–22 • https://doi.org/10.1145/3445967

Published: 08 March 2021

The birth of a line of research (2011-2012)

- Conway-Bromage (2011) proposed to encode dBG as bit vector
- Bit vector is of size 4^k, **1**s at positions of *k*-mers
- Efficient succinct encoding (Okanohara *et al* 2006): $O(nk)$
- Info-theoretically optimal

Exact encoding of the de Bruijn graph using a bit vector

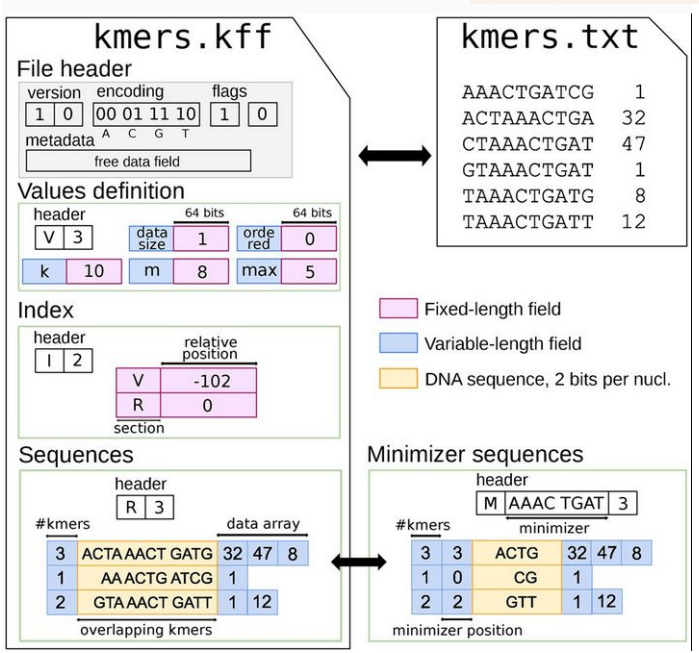


k-mer storage, on disk

- No dedicated file format, until recently
- **KFF** (Kmer File Format) (Dufresne et al, 2022)
  - Implemented in KMC, DSK, some other tools

Otherwise:

- .fasta.txt
- HDF5 (not recommended)



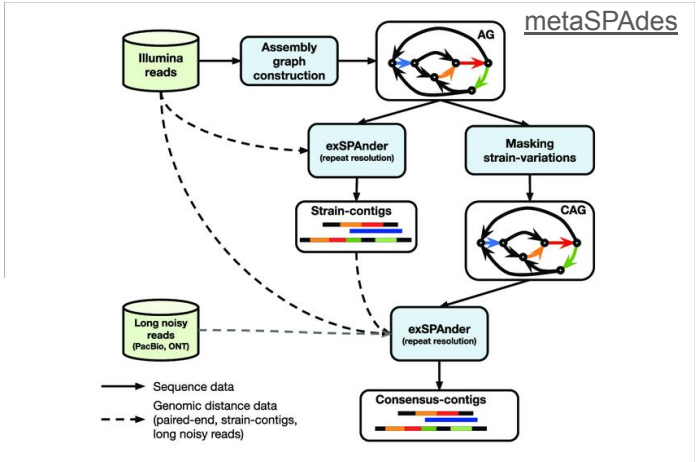
Other Applications of k-mers

- **Metagenomics** and **error correction**
- **Transcriptomics** (isoform quantification)

k-mers in metagenomics

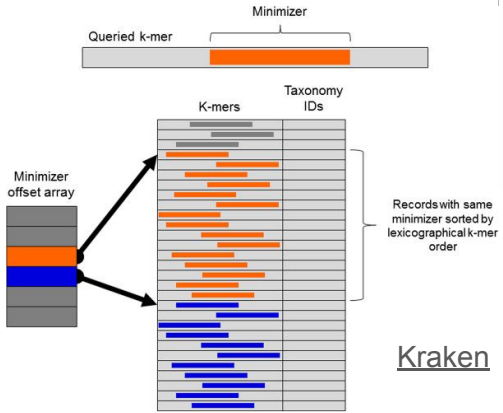
Metagenomic short read assembly: dominated by k-mers

- **metaSPAdes**, **MEGAHIT**, **minia3**



Taxonomic classification: same

- **Kraken, Centrifuge**



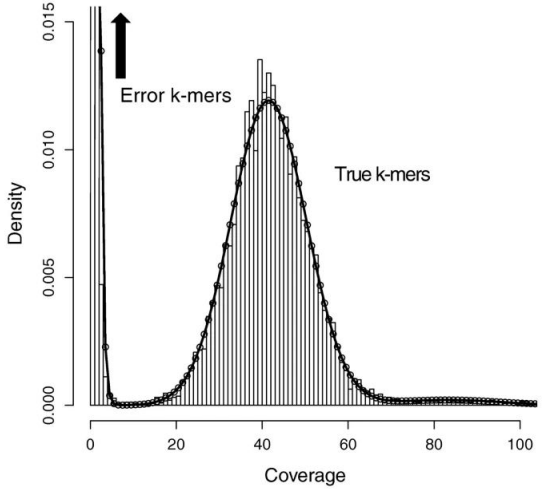
k-mers in read error correction (for genome assembly)

Sanger reads error correction

- **EULER**'s error correction module (Pevzner, 2001)
  - Mutates reads so that all k-mers are solid

Short reads:

- **Quake** (Kelley, Schatz, Salzberg, 2010)
  - First use of k-mer histogram to determine error threshold
  - Frames EULER's method as a Maximum Likelihood
- **BFC**, **Lighter**, Musket, Bloocoo, (Bayes)HAMMER, ..
- Nowadays: none used anymore



Long reads:

- **LoRDEC**, **LoRMA**
  - de Bruijn graph to correct long reads using short reads (or long reads)

k-mers in RNA-seq analysis

Alignment? k-mers.
Abundance estimation? k-mers
Differential analysis?
believe it or not, also k-mers
(sometimes)

Gene quantification:

- **Salmon** (Patro et al, 2017)
- **Kallisto** (Bray et al, 2016)

RNAseq specific mapping:

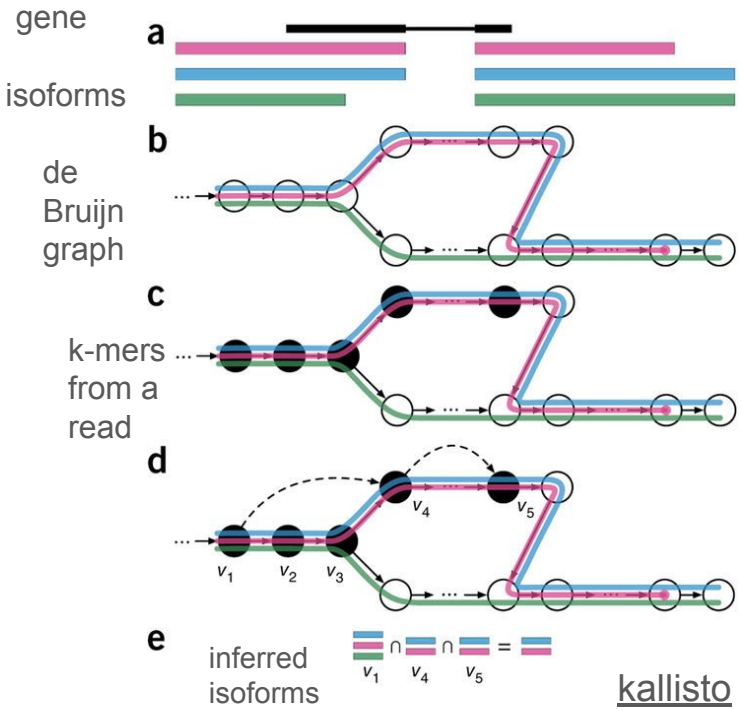
- **STAR** (Dobin et al, 2012)

RNAseq assembler:

- **Trinity** (Grabherr et al, 2011)

Reference-free diff analysis:

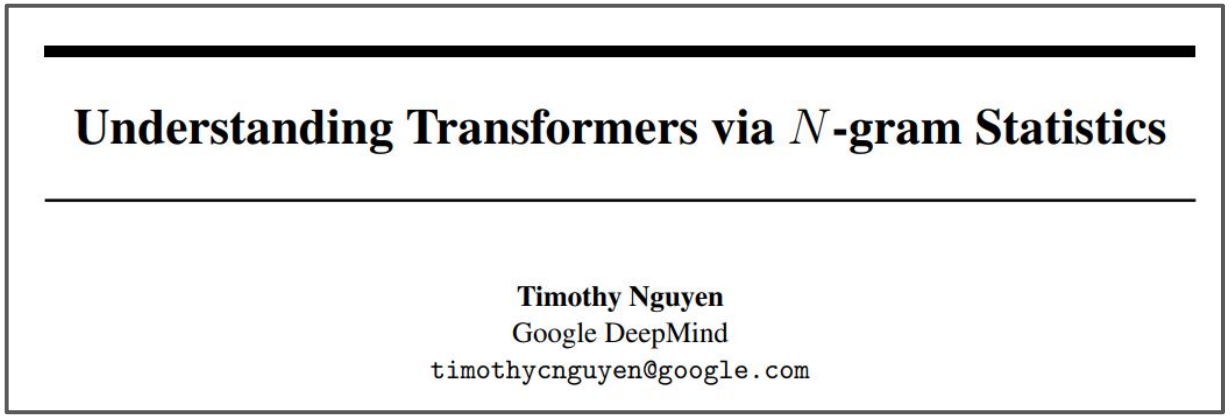
- **Transipedia.org** (Bessière et al, 2011)



Future Directions for k-mers

- Integration with machine learning
- Evolving role of k-mers in long-read technologies
- Misc: sketching, database search, phylogeny
- Omitted: pangenomics

k-mers in ML



Understanding Transformers via N-gram Statistics

Timothy Nguyen
Google DeepMind
timothycnguyen@google.com

How does a transformer-based large language model (LLM) make use of its context when predicting the next token?

Recall Shannon's theory: probably of next gram following a N-gram

k-mers in sketching

- **D_2** (Torney et al, 1990)
  - number of shared k-mers as a proxy of sequence similarity
- **MinHash** (Broder, 1997)
  - Measuring similarity between two documents by random sampling of words
- **Winnowing/minimizers** (Roberts, 2004)
  - Sampled k-mers
- **Bloom filters** (Bloom, 1970) applied to k-mers (2004)
  - First applied to accelerate BLAST
  - Later to k-mer counting, error correction, assembly, etc..
- **Mash/sourmash/Dashing**
  - Modern k-mer sketching for sequence similarity estimation

k-mers in sequence database search

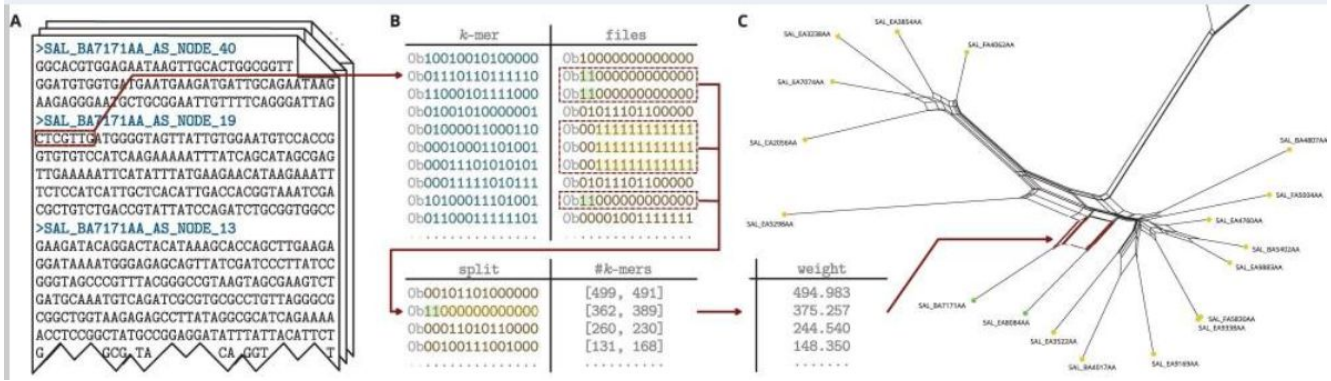
- **BLAST** (1990)
  - Clustered GenBank, assembled sequences, 200 GB
- **BigSI/COBS** (Bradley et al, 2019)
  - Isolate genomes, 170 TB (.3% of SRA)
- **MetaGraph** (Karasikov et al, 2024 preprint)
  - Animals, Plants, Bacteria, cancer data, 3.3 Pbp (8.9% of SRA)
- And many more

k-mers in phylogeny

- Construct phylogenetic trees without multiple sequence alignment
  - “Is Multiple-Sequence Alignment Required for Accurate Inference of Phylogeny” (Hölh, Ragan, 2007)
  - “*we encode K-mers as character states and estimate posterior probabilities of bipartitions using Mr-bayes*”

Recent works close to pangenomics:

- **SANS, SANS serif** (Rempel, Wittler, 2021), **KINN** (Tang et al, 2023),



What we will learn in the course

Sun: Intro to k-mers

Mon: K-mer ops & histograms (Kamil), Fundamentals (Katie)

Tue: Deep dive: k-choice, errors (Katie), Proteins & OMArk (Yannis), FastK counting (Gene)

Wed: Graphs & Logan data (Rayan), Polyploidy & modeling (Kamil), Pangenomes & syncmers (Katie)

Thu: Assembly QC with Merqury (Gene), Sketching: Sourmash & minimisers (Cassandra/Tessa)

Fri: Hackathon & wrap-up

Lex Nederbragt
@lexnederbragt

En réponse à @ctitusbrown

"Finding your way in life is like finding the genome in a De Bruijn graph: it is very easy to find *a* path, very hard to find *the* path".

Thank you for your attention!