

Challenging algorithms in bioinformatics

IN3130, 30 September 2020

Torbjørn Rognes

Department of Informatics, UiO

torognes@ifi.uio.no



UiO : **University of Oslo**

What is bioinformatics?

Definition:

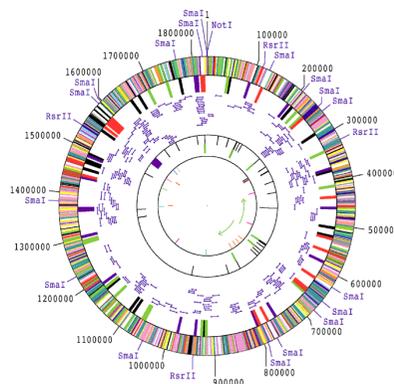
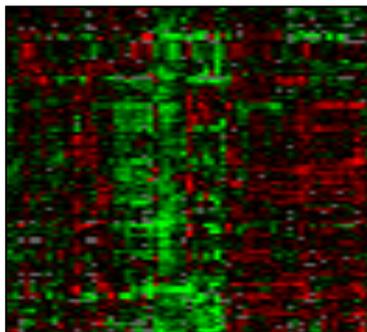
Bioinformatics is the development and use of computational and mathematical methods to gather, process and interpret molecular biological data.

Aim of research:

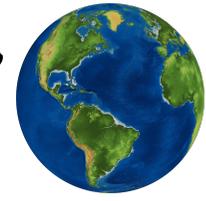
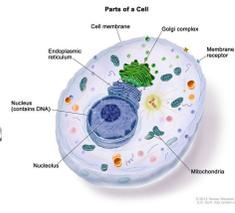
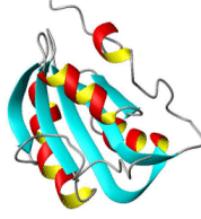
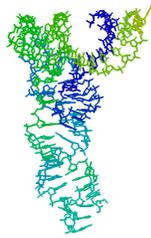
To increase our understanding of the connections between biological processes at different levels while developing better theories and methods in computer science and statistics.

An interdisciplinary subject:

Computer science/statistics/mathematics + biology/medicine



Bioinformatics at many levels



DNA

RNA

Protein

Cell

Organ

Individual

Population

Biosphere

Genomics

Transkript-omics

Proteomics

System biology

Neuro-informatics

Precision medicine

Population genetics

Metagenomics

Genome assembly

RNomics

Structural biology

Cell simulation

Organ modelling/simulation

Variant detection

Epidemiology

Evolutionary biology

Genefinding

Microarrays

Drug design

Metabolism studies

Meta-genomics

Variant detection

Epidemiology

Phylo-genomics

Annotation

RNA-folding

MS analysis

Binding site analysis

Cancer genomics

ChIP-Seq

RNA-seq

Structural biology

Interaction networks

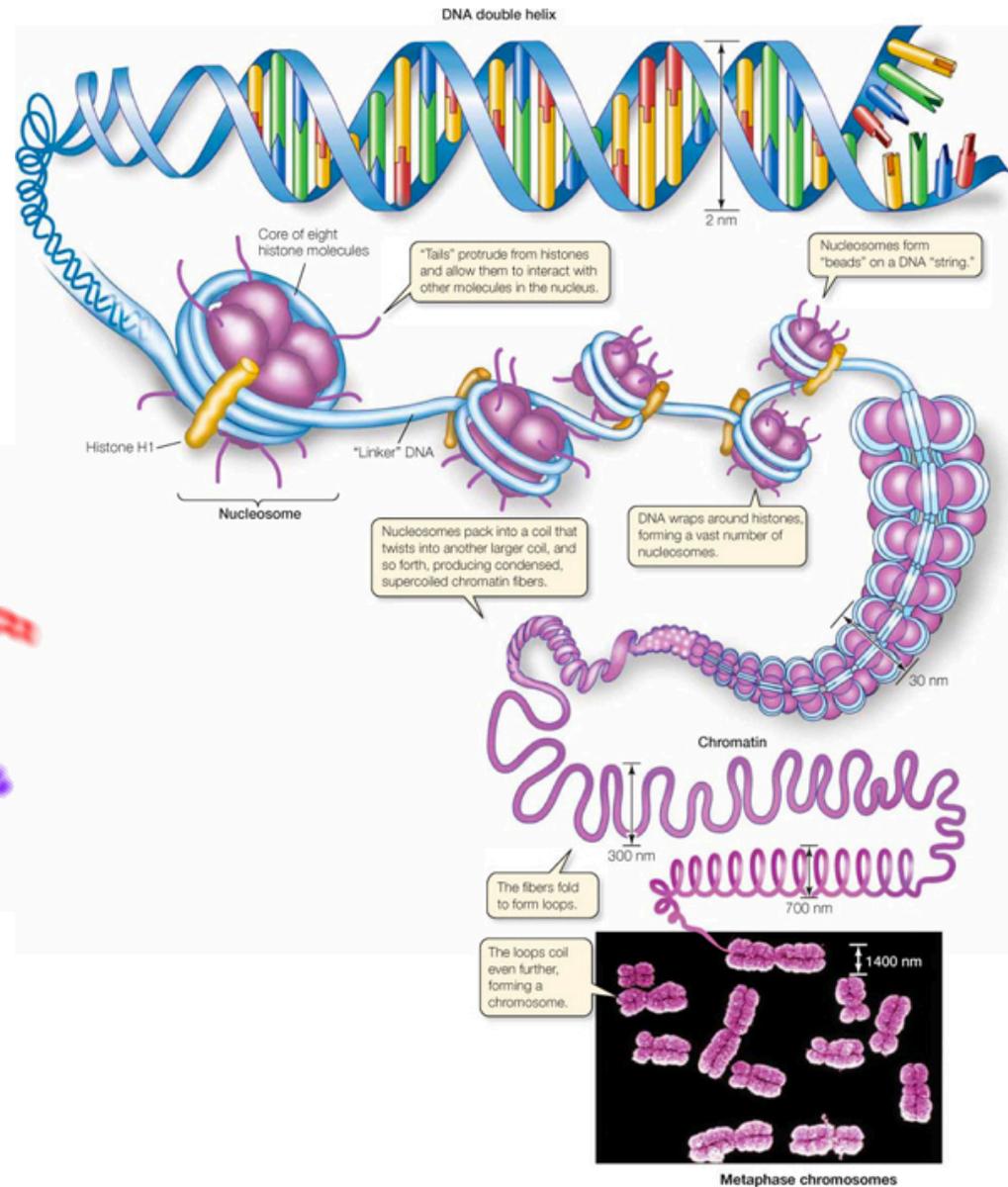
Genomes and chromosomes

The genome is our genetic material. It consists of DNA. From ~ 2 to $\sim 150\,000$ million nucleotides (base pairs).



Human genome with 23 pairs of chromosomes (22 + XY)

ca 3 000 000 000 bp

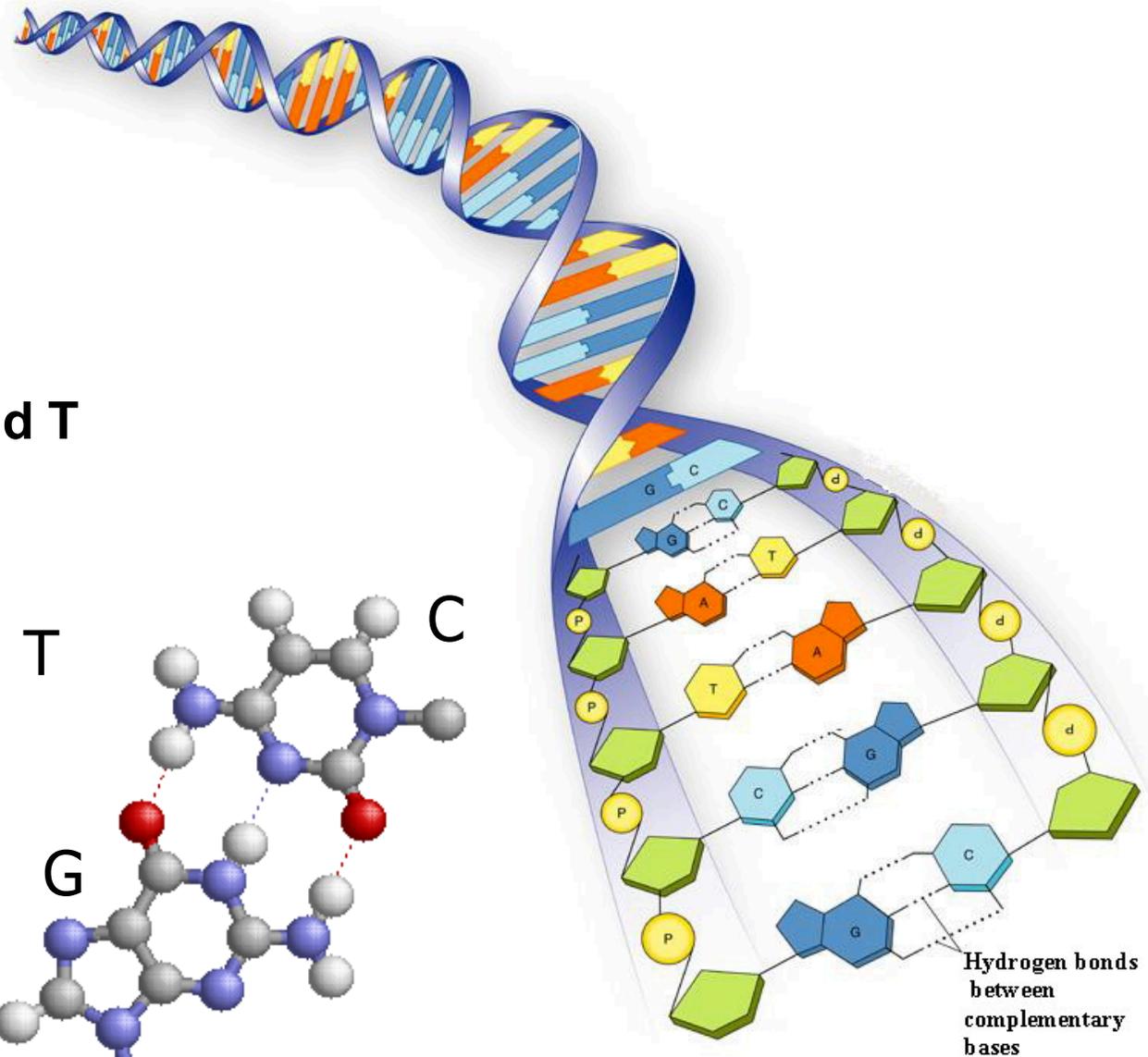
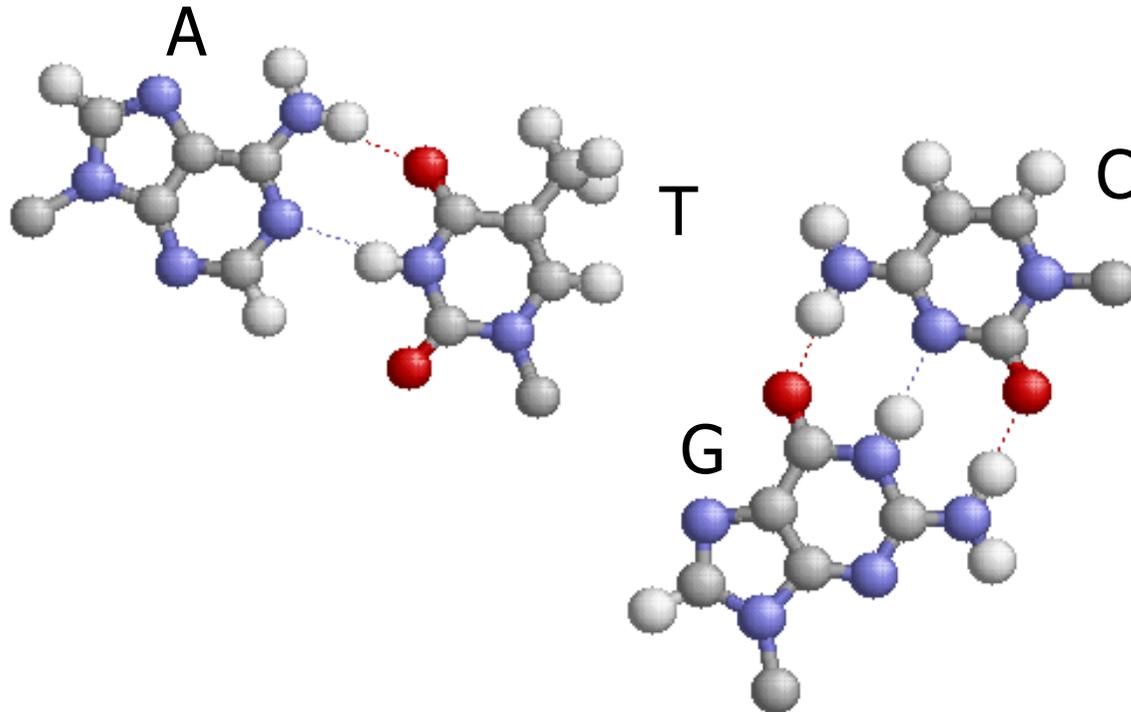


Four nucleotides form 2 pairs

Complementary bases:

- A with T (2 H-bonds)
- C with G (3 H-bonds)

Four bases: A, C, G and T

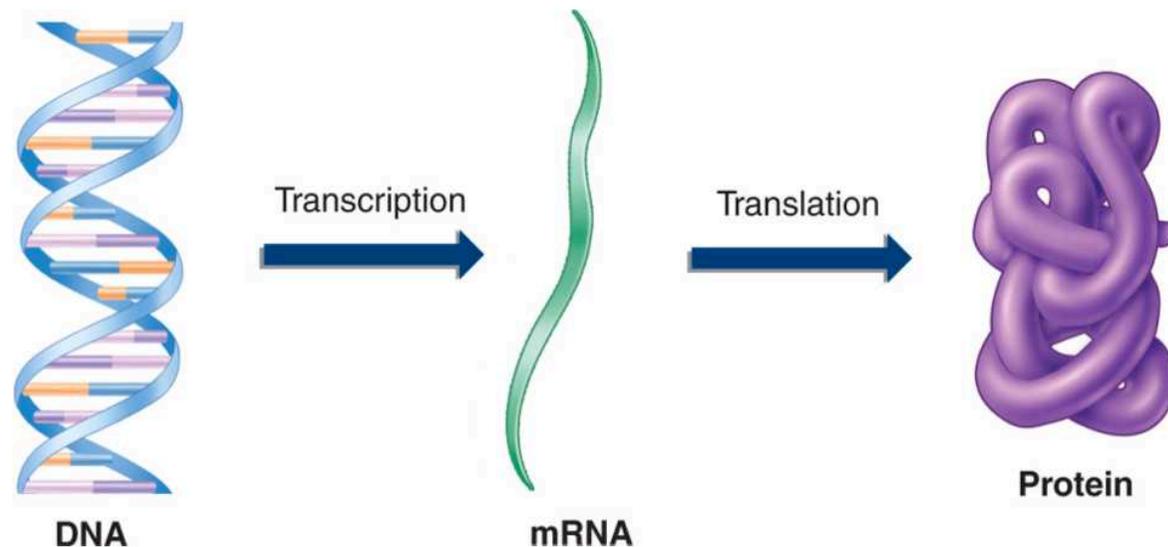


DNA -> mRNA -> Protein

Genes can be turned on and *expressed* (produced) at certain times and places.

The expression of gene consists of at least two steps

- **Transcription:** DNA → mRNA
- **Translation:** mRNA → Protein



The universal genetic code

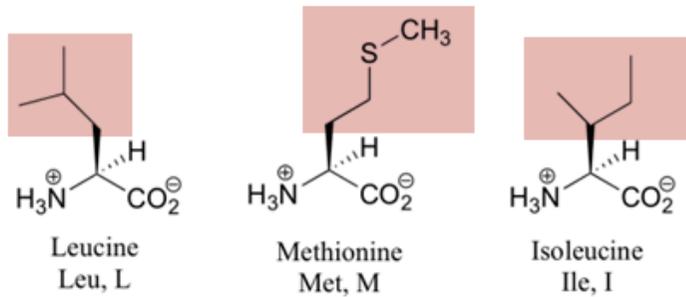
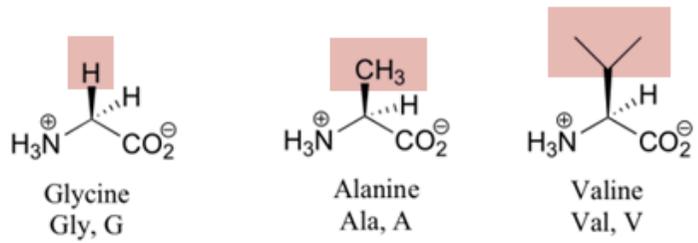
During translation, groups of 3 nucleotides are read from the mRNA. These *codons* select new amino acids to be added to the protein chain.

Start codon:
AUG

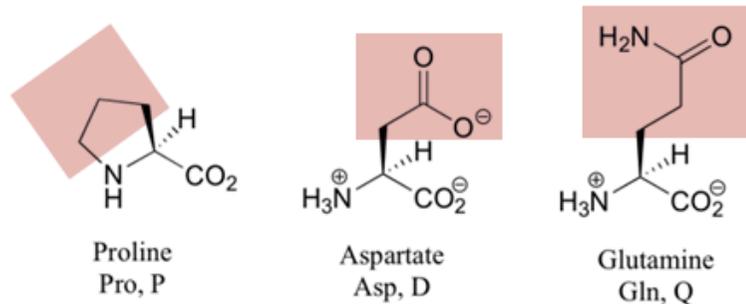
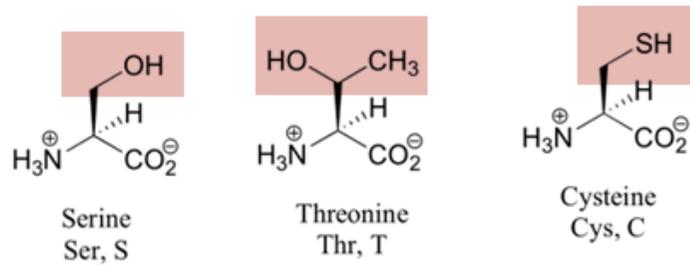
Stop codons:
UAA,
UAG,
UGA

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G
						Third letter

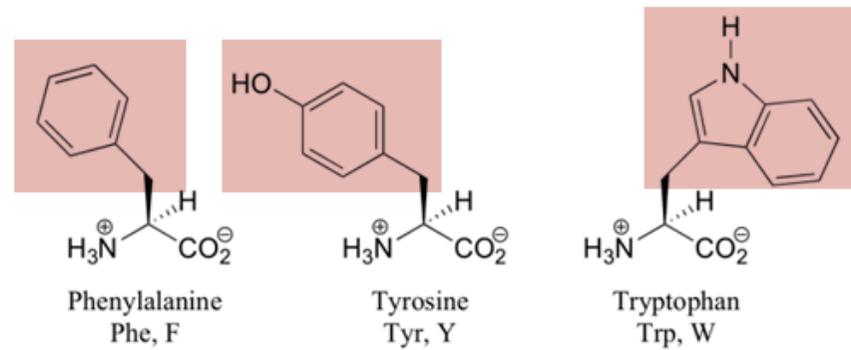
Nonpolar, aliphatic side groups



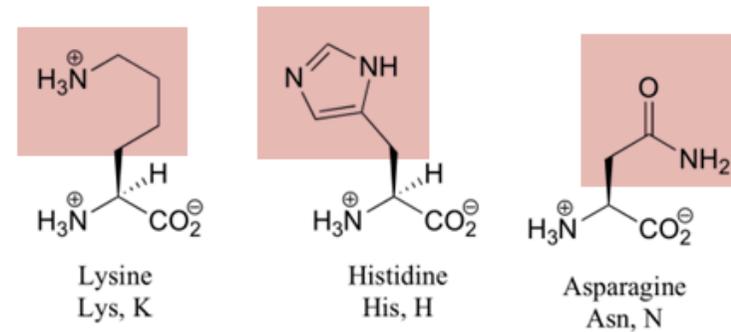
Polar, uncharged side groups



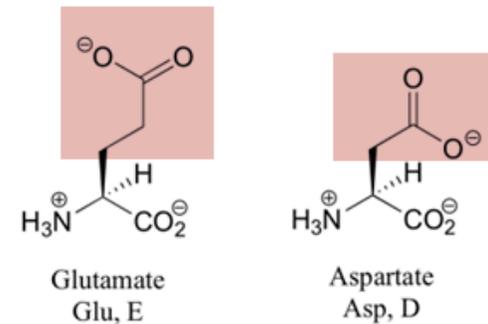
Aromatic side groups



Positively charged side groups



Negatively charged side groups



Computational challenges

Examples of classic and important computational challenges in bioinformatics (hardest problems first):

- Protein structure prediction and design
- Whole-genome *de novo* sequence assembly
- Pairwise and multiple sequence alignment

PROTEIN STRUCTURE PREDICTION AND DESIGN

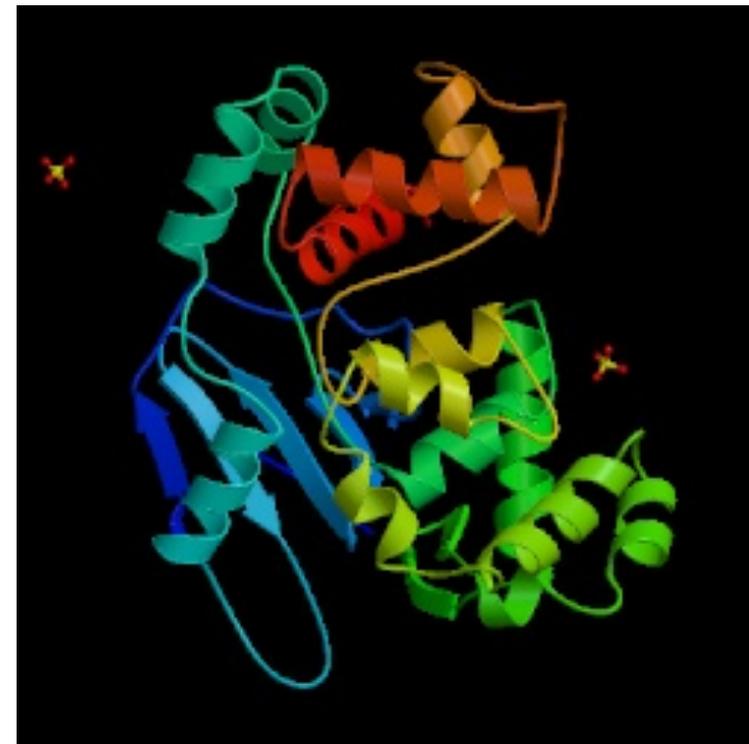
Protein 3D structure and design

MPARALLPRRMGHRT
LASTPALWASIPCPR
SELRLDLVLPSGQSF
RWREQSPAHWGVLA
DQVWTLTQTTEEQLHC
TVYRGDKSQASRPTP
DELEAVRKYFQLDVT
LAQLYHHWGSVD . . .

Structure prediction

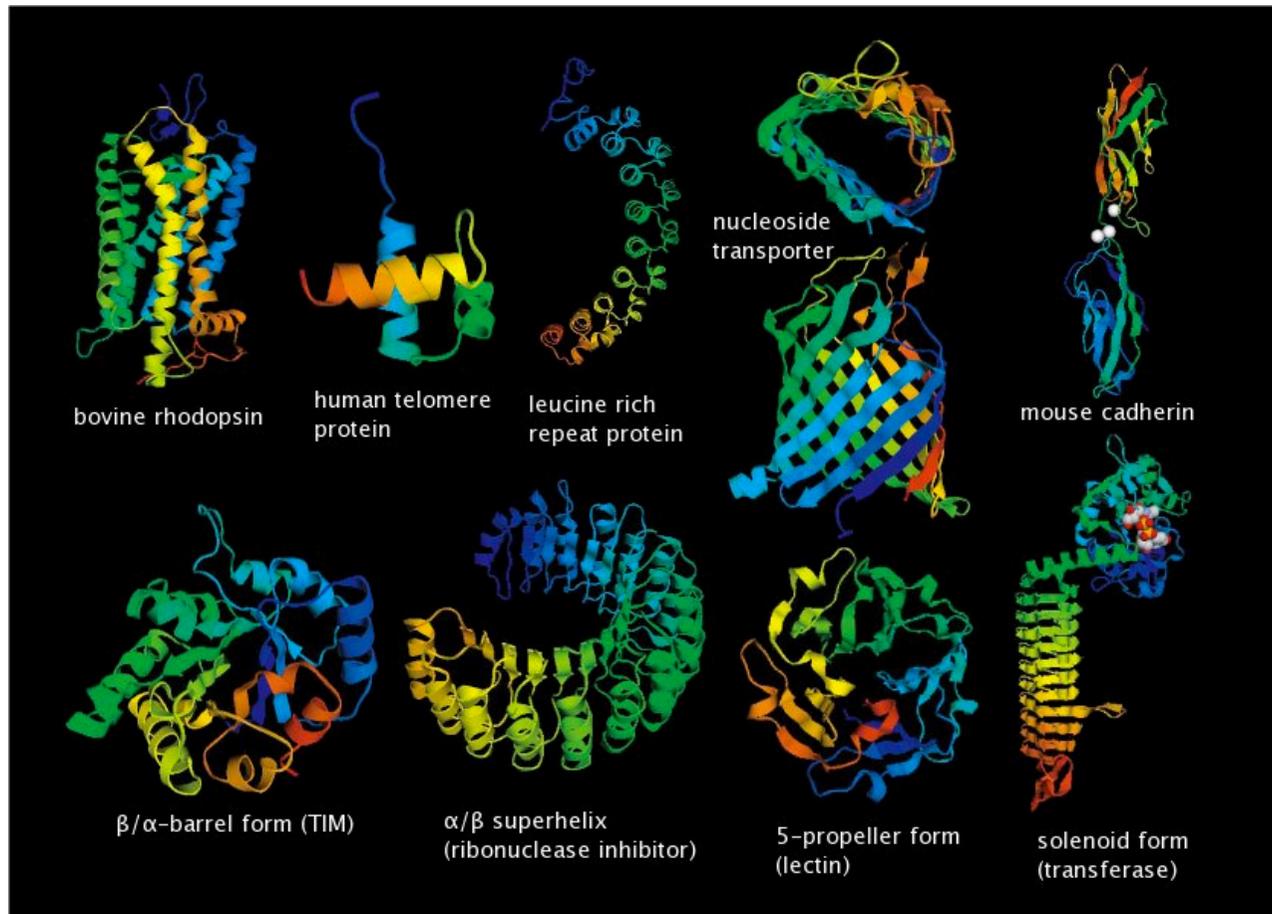


Protein design



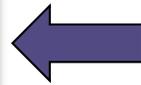
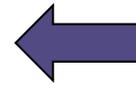
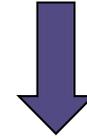
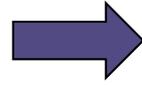
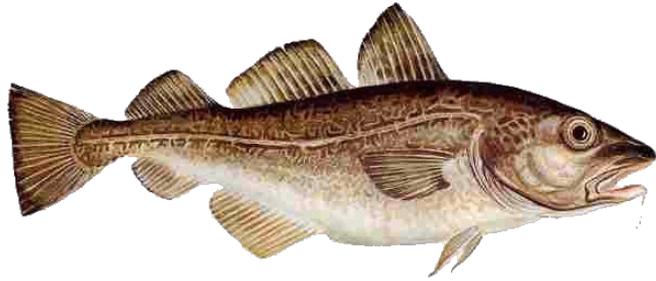
Proteins fold into beautiful structures

- Proteins consist of chains of amino acids (on average 350)
- Proteins form 3D structures
- They act as molecular machines or as structural building blocks

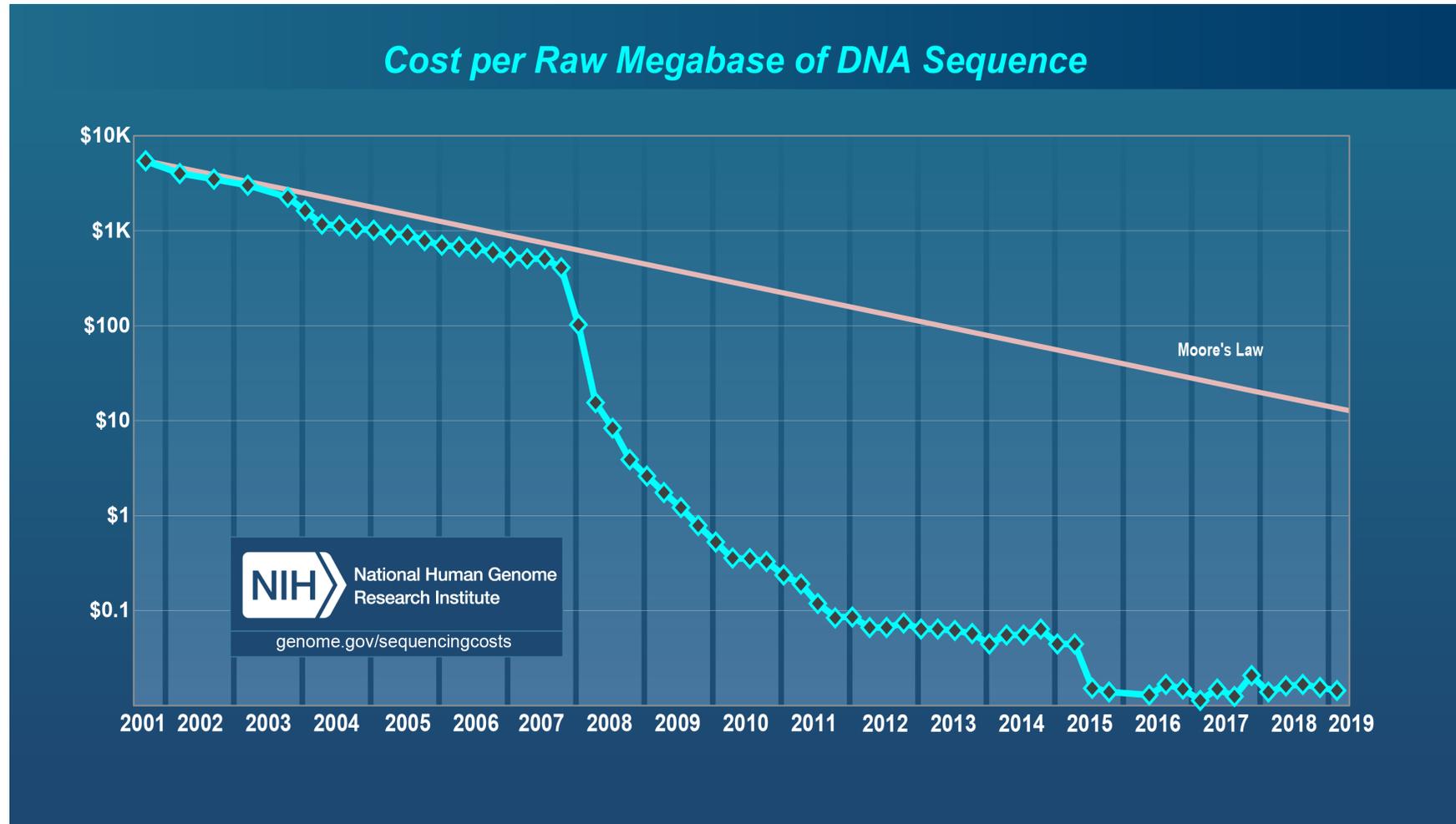


**WHOLE-GENOME *DE*
NOVO SEQUENCE
ASSEMBLY**

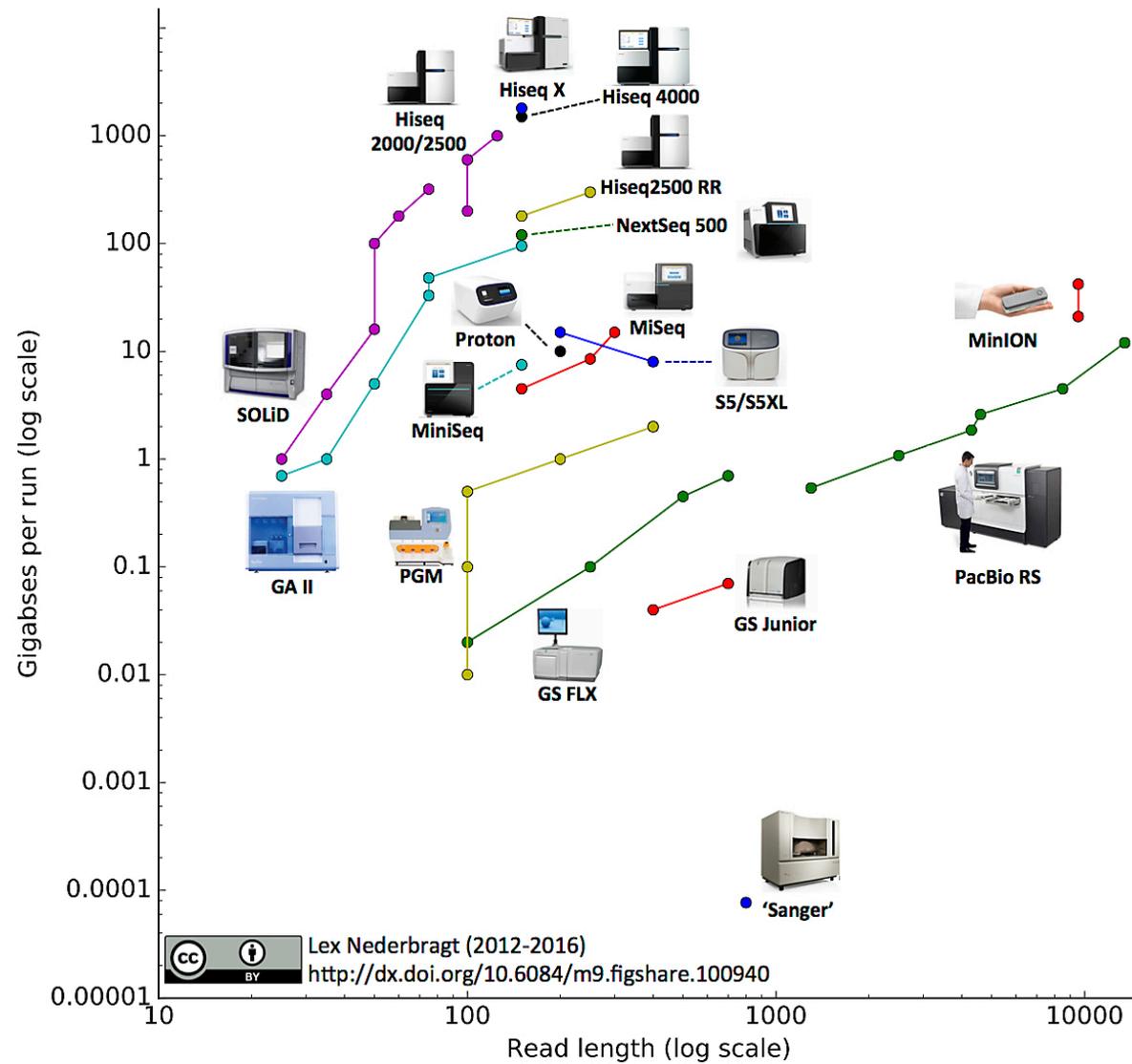
Whole genome sequence assembly



The cost of sequencing



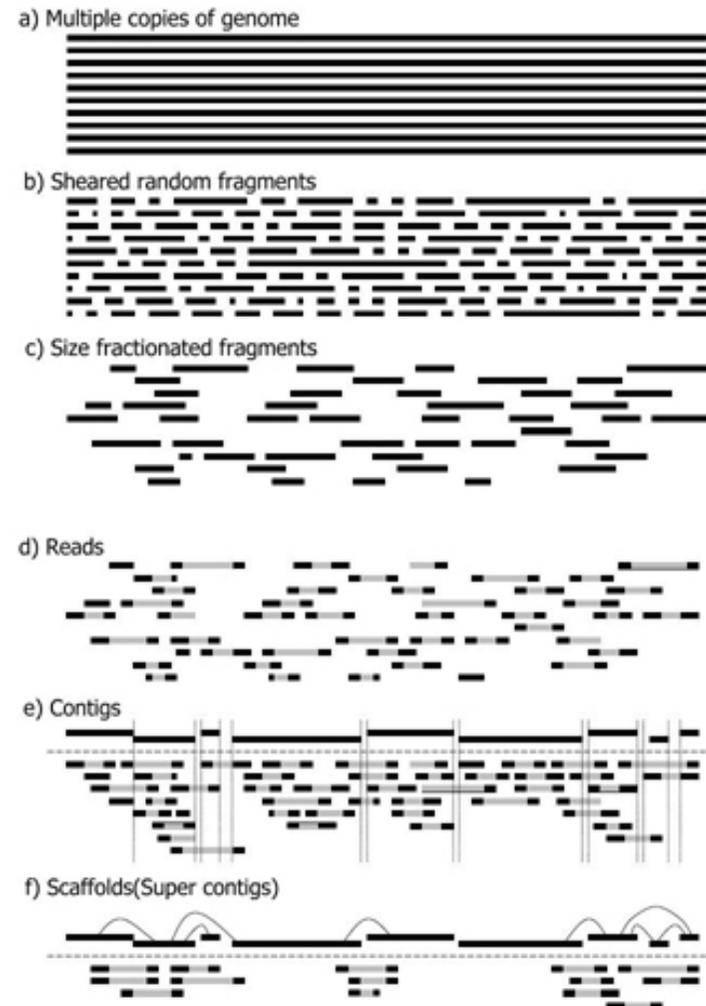
Developments in Sequencing



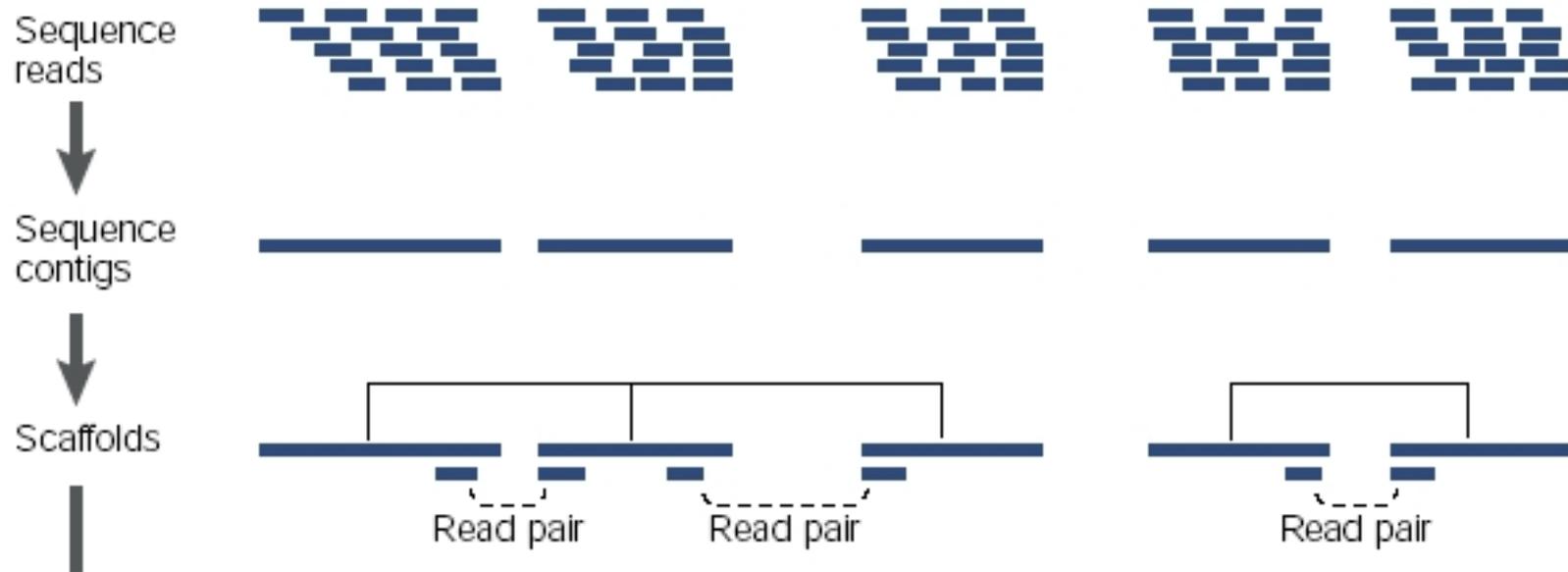
Source: Lex Nederbragt (2012-2016) <https://doi.org/10.6084/m9.figshare.100940>

Whole genome sequence assembly

- Genome sequencing results in millions of small pieces of the full genome
- The challenge is to puzzle these together in the right order
- Genome size ranging from 2Mbp (bacteria) to 3Gbp (human) to 150Gbp (plant)
- Read size from 30 bp to 1000 bp
- Sequencing errors
- Natural variation (allels)
- Repeats and similar regions



All the pieces must be puzzled together



Example: Reads of length 10

nøf, _tidde

snør, _det_

ddeli_bom.

, _den_snør

t_smør, _ti

Det_snør._

Example: Identify overlaps

snør, _det_ nøf, _tidde

ddeli_bom.

, _den_ snør

t_ smør, _ti

Det_ snør. _

Example: Layout

```
Det_snør._  
  snør,_det_  
    ,_den_snør  
      t_smør,_ti  
        nøf,_tidde  
          ddeli_bom.
```

Example: Find consensus sequence

Det_snør.
snør, det
, den snør
t smør, ti
nøf, tidde
ddeli_bom.

Det_snør, _det_snør, _tiddeli_bom.

Repeat of length 9

Overview of the assembly process

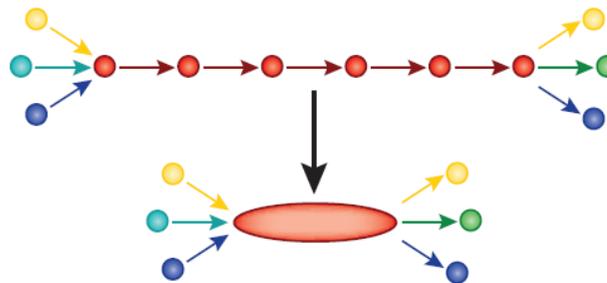
1. Fragment DNA and sequence



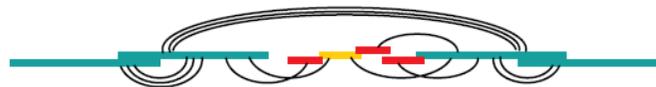
2. Find overlaps between reads

...AGCCTAGACCTACA**GGATGCGCGACACGT**
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

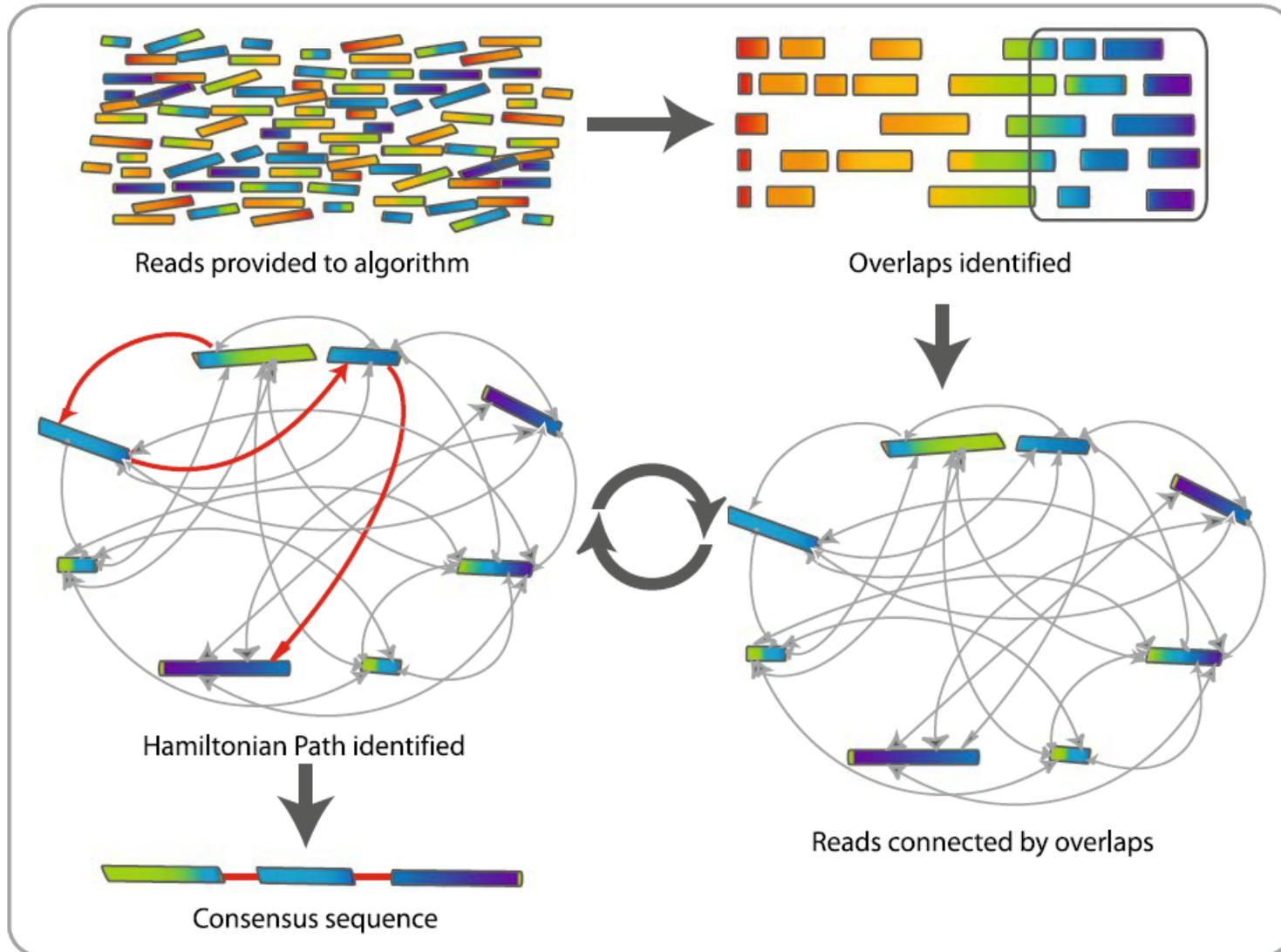


4. Assemble contigs into scaffolds



Genome assembly stitches together a genome from short sequenced pieces of DNA.

Overlap-Layout-Consensus assemblers

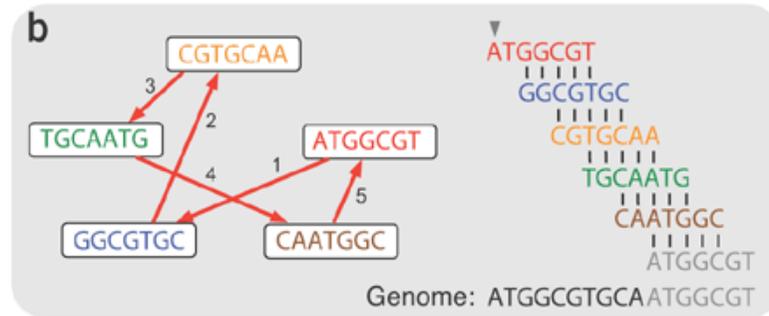
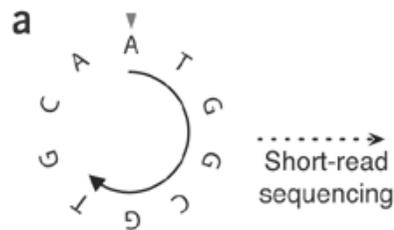


de Bruijn graph assemblers

Strategy:

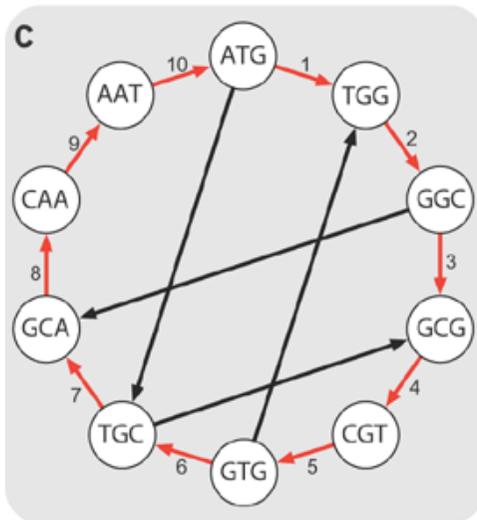
- Shred the reads into k-mers (e.g. $k=31$)
- Connect k-mers that overlap with other k-mers with $k-1$ common nucleotides
- Build a de Bruijn graph where the edges represent the k-mers and the nodes represent the overlap of $k-1$ nucleotides between the edges
- Find an Eulerian path or cycle through the graph. It shall visit all edges once. Nodes may be visited more than once.

Two genome assembly strategies

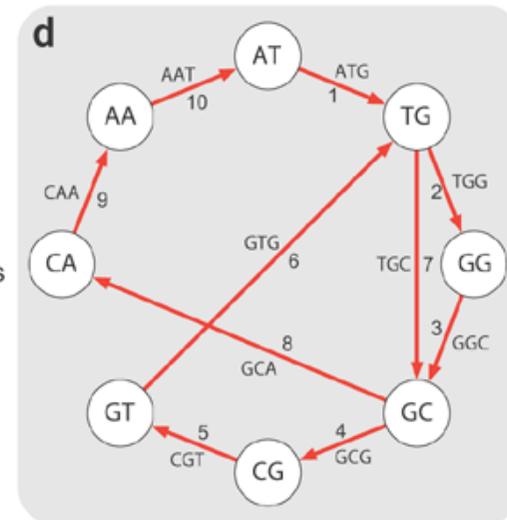
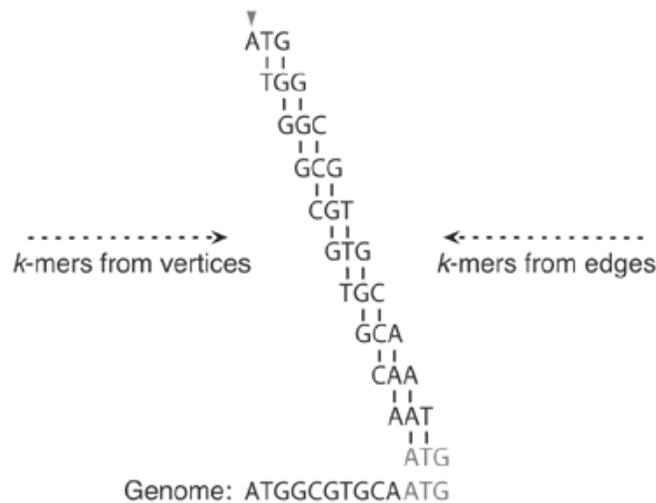


Vertices are k -mers
Edges are pairwise alignments

Vertices are $(k-1)$ -mers
Edges are k -mers



Hamiltonian cycle
Visit each vertex once
(harder to solve)



Eulerian cycle
Visit each edge once
(easier to solve)

Genome browsers

Human chr14:77636798-77743621

https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr14%3A77636798-77743621&hgid=451845769_AgWmCZ1ACUAQXrAwDITR1AVh5Xe

Genomes Genome Browser Tools Mirrors Downloads My Data View Help About Us

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr14:77,636,798-77,743,621 106,824 bp. enter position, gene symbol or search terms go [hg38 replaces hg19 as default human assembly](#)

chr14 (q24.3) 14p13 14p12 14p11.2 14q11.2 14q12 14q21.1 14q21.2 14q21.3 14q22.1 22q3 22q3.1 22q3.2 22q3.3 14q24.2 14q24.3 14q24.31 14q24.32 14q24.33 14q24.34 14q24.35 14q24.36 14q24.37 14q24.38 14q24.39 14q24.4 14q24.41 14q24.42 14q24.43 14q24.44 14q24.45 14q24.46 14q24.47 14q24.48 14q24.49 14q24.5 14q24.51 14q24.52 14q24.53 14q24.54 14q24.55 14q24.56 14q24.57 14q24.58 14q24.59 14q24.6 14q24.61 14q24.62 14q24.63 14q24.64 14q24.65 14q24.66 14q24.67 14q24.68 14q24.69 14q24.7 14q24.71 14q24.72 14q24.73 14q24.74 14q24.75 14q24.76 14q24.77 14q24.78 14q24.79 14q24.8 14q24.81 14q24.82 14q24.83 14q24.84 14q24.85 14q24.86 14q24.87 14q24.88 14q24.89 14q24.9 14q24.91 14q24.92 14q24.93 14q24.94 14q24.95 14q24.96 14q24.97 14q24.98 14q24.99 14q25

Scale chr14: 77,550,000 77,650,000 50 kb 77,670,000 77,680,000 77,690,000 77,700,000 77,710,000 hg38 77,720,000 77,730,000 77,740,000

GENCODE v22 Comprehensive Transcript Set (only Basic displayed by default)

RefSeq Genes

Human mRNAs Human mRNAs from GenBank

Spliced ESTs Human ESTs That Have Been Spliced

Layered H3K27Ac H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)

Cons 100 Verts 100 vertebrates Basewise Conservation by PhyloP

Multiz Alignments of 100 Vertebrates

Common SNPs (142) Simple Nucleotide Polymorphisms (dbSNP 142) Found in $\geq 1\%$ of Samples

Repeating Elements by RepeatMasker

SINE LINE LTR DNR Simple Low Complexity Satellite RNA Other Unknown

move start Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. move end

< 2.0 >

Source: genome.ucsc.edu

Problematic issues

- Sequencing errors
 - Introduces false sequences into the assembly
 - May be alleviated by higher coverage / larger sequencing depth, or by error detection and correction
- Repeats
 - Our genomes are filled with many almost identical repeated sequences
 - Repeats longer than the read length makes it impossible to determine the exact location of the read
 - May cause compression or misassemblies
 - May be alleviated by longer reads or paired-end/mate pair reads
- Heterozygosity
 - Diploid organisms (e.g Humans) actually have two “genomes”, not one. Chromosome pairs 1-22 for all and XX for women (XY for men). One set of chromosomes from our mother and one from our father.
 - The two are mostly identical, but there are some differences

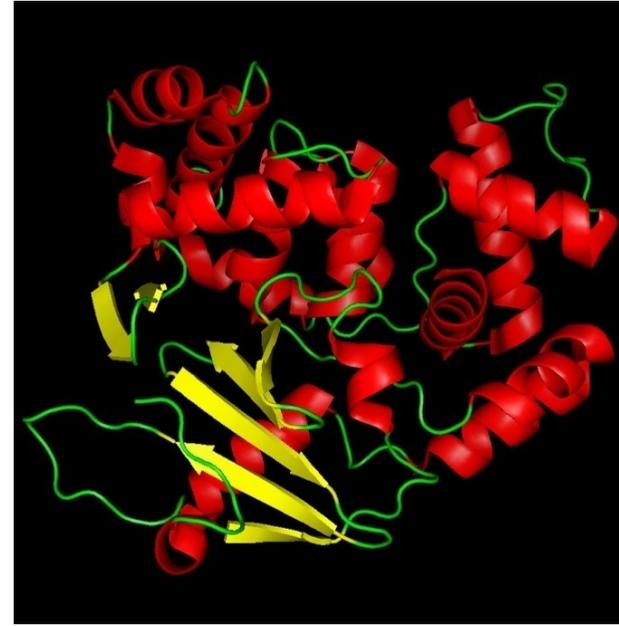
PAIRWISE AND MULTIPLE SEQUENCE ALIGNMENT

Pairwise sequence alignment



E. coli AlkA

Hollis *et al.* (2000) *EMBO J.* 19, 758-766 (PDB ID 1DIZ)



Human OGG1

Source: Bruner *et al.* (2000) *Nature* 403, 859-866 (PDB ID 1EBM)

```

E.c. AlkA 127 SVAMAAKLTARVAQLYGERLDDFPE--YICFPTPQRLAAADPQA-LKALGMPLKRAEALI 183
      ++|    +  |+ | +| ||    +  |  ||+ | ||  + +| |+ ||+  ||  +
H.s. OGG1 151 NIARITGMVERLCQAFGPRLIQLDDVTYHGFP SLQALAGPEVEAHLRKLGLGY-RARYVS 209
E.c. AlkA 184 HLANAALE-----GTLPMTIPGDVEQAMKTLQTFPGIGRWTANYFAL 225
      | | ||      |      | +| | |  ||+|  |+  |
H.s. OGG1 210 ASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICL 256
    
```

Common alignment scoring system

Substitution score matrix

- Score for aligning any two residues to each other
- Identical residues have large positive scores
- Similar residues have small positive scores
- Very different residues have large negative scores

BLOSUM62 amino acid substitution score matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Gap penalties

- Penalty for opening a gap in a sequence (Q)
- Penalty for extending a gap (R)
- Typical gap function: $G = Q + R * L$, where L is length of gap
- Example: Q=11, R=1

```

E.c. Alka 127 SVMAAKLTARVAQLYGERLDDDFPE--YICFPTPQRLAAADPQA-LKALGMPLKRAEALI 183
      ++|    +  |+ | +| ||    +  |  ||+ | ||  + +| |+ ||+  ||  +
H.s. OGG1 151 NIARITGMVERLCQAFGPRLIQLDDVTYHGFP SLQALAGPEVEAHLRKLGLGY-RARYVS 209

E.c. Alka 184 HLANAALE-----GTLPMTIPGDVEQAMKTLQTFPGIGRWTANYFAL 225
      | | ||      |      |+| | |  ||+|  |+  |
H.s. OGG1 210 ASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICL 256
  
```

Amino acid substitution score matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62

How to find the best alignment(s)?

- There are too many possible alignments of two sequences to enable examination of every possible alignment individually
- There is a dynamic programming (DP) type of algorithm to identify the alignment(s) with the highest score

- Global alignments: Needleman and Wunsch (1970)
- Local alignments: Smith and Waterman (1981)

- Two steps:
 - First, identify the highest possible score using DP
 - Then, identify the alignment(s) with the highest score (using temporary results from the initial step)

- Dynamic programming:
 - General method for solving recursive problems by storing temporary results from smaller problems along the way
 - Used to solve many problems in bioinformatics

Needleman-Wunsch alg.: Initialisation

- Consider two strings $S[1..n]$ and $T[1..m]$.
- Define $V(i, j)$ as the score of the optimal alignment between $S[1..i]$ and $T[1..j]$
- Basis:
 - $V(0, 0) = 0$ Empty sequences
 - $V(0, j) = V(0, j-1) + \delta(-, T[j])$ Insert gap j times
 - $V(i, 0) = V(i-1, 0) + \delta(S[i], -)$ Delete gap i times

The alignment matrix, V: Initialisation

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

Match: +2

Mismatch: -1

Gap: -1

Needleman-Wunsch alg.: Recurrence

Recurrence: For $i > 0, j > 0$

$$V(i, j) = \max \begin{cases} V(i-1, j-1) + \delta(S[i], T[j]) & \text{Match/mismatch} \\ V(i-1, j) + \delta(S[i], -) & \text{Delete} \\ V(i, j-1) + \delta(-, T[j]) & \text{Insert} \end{cases}$$

In the alignment, the last pair must be either be a match/mismatch, a delete, or an insert.

xxx...xx	xxx...xx	xxx...x-
yyy...yy	yyy...y-	yyy...yy
match/mismatch	delete	insert

The alignment matrix, V: Filling in

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

Match: +2

Mismatch: -1

Gap: -1

The alignment matrix, V: Complete

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

Final alignment:

A-CAATCC

AGCA-TGC

A-CAATCC

AGC-ATGC

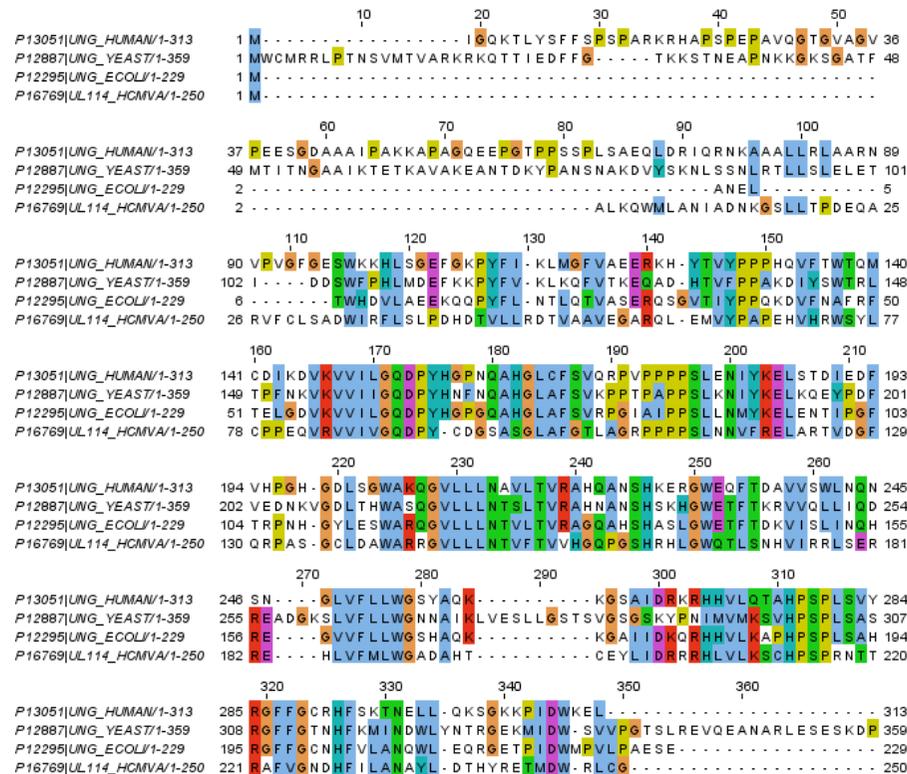
Score: 7

Algorithmic complexity

- Assume that we are aligning two sequences of length m and n , and that the gap penalty is constant
- Memory: $O(nm)$
A fixed number of tables (one or two) with $n*m$ cells: $\text{constant} * nm$
A fixed number of additional variables: constant
Little memory needed if we are only interested in the best score
- Time: $O(nm)$
Calculate $B(i,j)$ and $P(i,j)$ for $n*m$ cells in the table: $\text{constant} * nm$
Perform traceback: $\text{constant} * (n+m)$

Multiple sequence alignment

- Align three or more sequences
- Show corresponding amino acids in the different proteins
- Place gaps at correct positions
- Impossible to solve optimally by brute force for more than a few short sequences



An underwater photograph showing a deep blue ocean. Sunlight filters through the water from the top right, creating a bright, shimmering path of light that illuminates the surrounding water. The water surface is visible at the top, with ripples and reflections of light. The overall scene is serene and peaceful.

Thanks!