3000788 Intro to Comp Molec Biol

Lecture 5: Genome assembly and annotation

Fall 2025

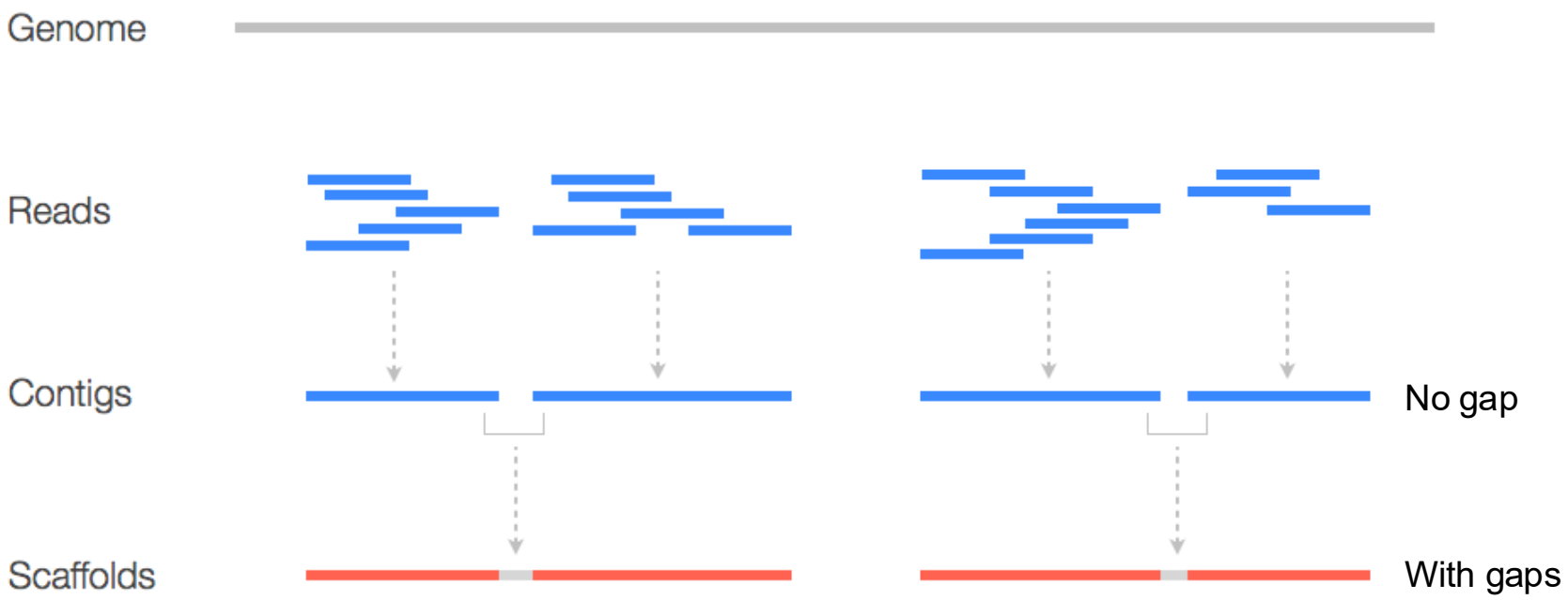
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

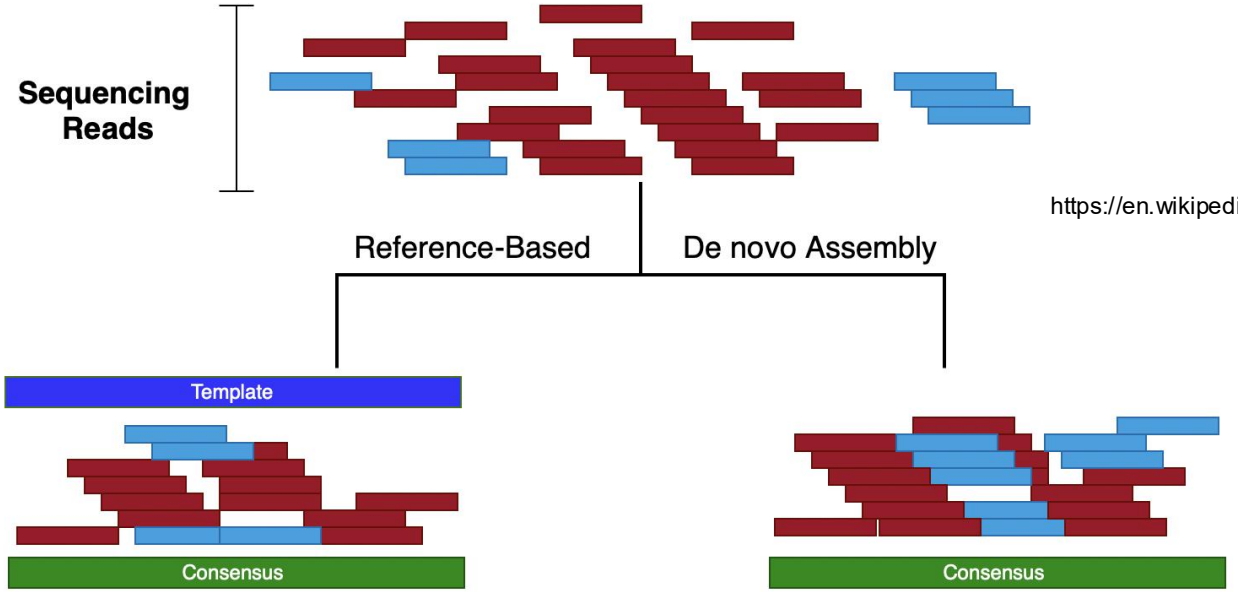
- Sequence assembly
- Structure of genome / genomic elements
- Annotation of genomic features

Sequence assembly



- Reconstruct the original genome from sequencing reads

Alignment versus assembly



https://en.wikipedia.org/wiki/Sequence_assembly

- Alignment trusts the reference genome structure, assembly does not

Assembly of short reads

Assembling short reads is hard

```
GACCTACA
 ACCTACAA
  CCTACAAG
   CTACAAGT
    TACAAGTT
     ACAAGTTA
      CAAGTTAG
    TACAAGTC
     ACAAGTCC
      CAAGTCCG
```

GACCTACAAGTTAG

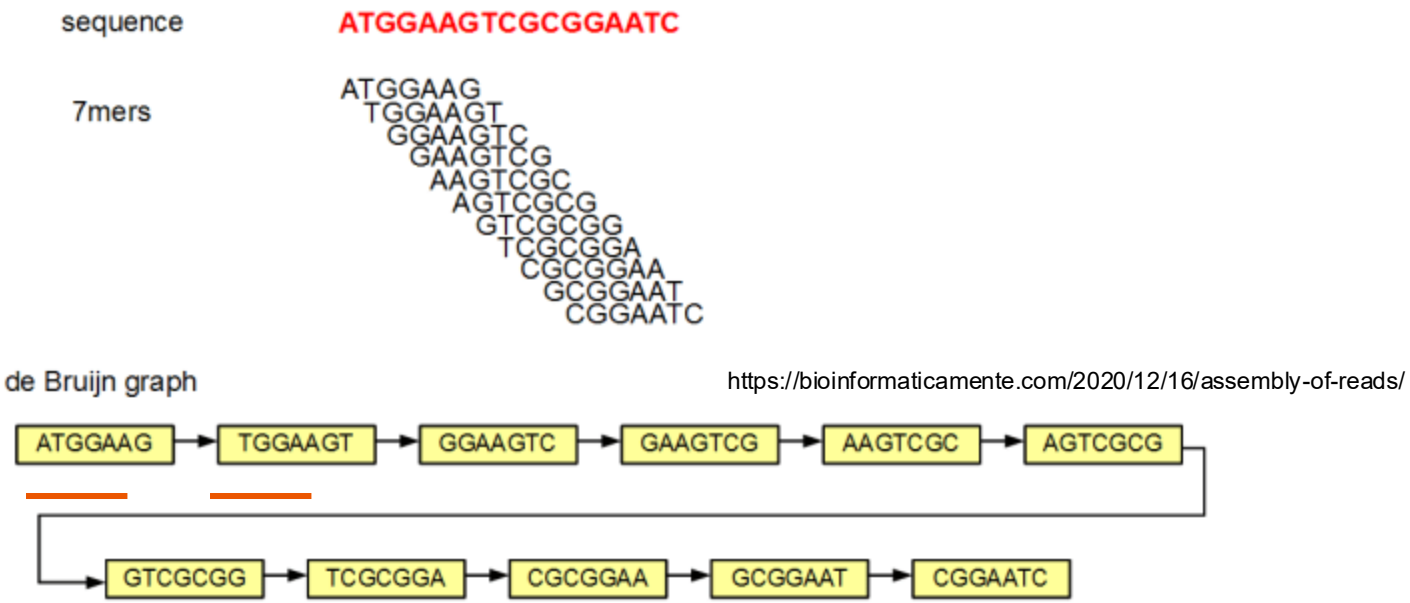
Or

GACCTACAAGTCCG

- Repetitive region? Polyploid genome?

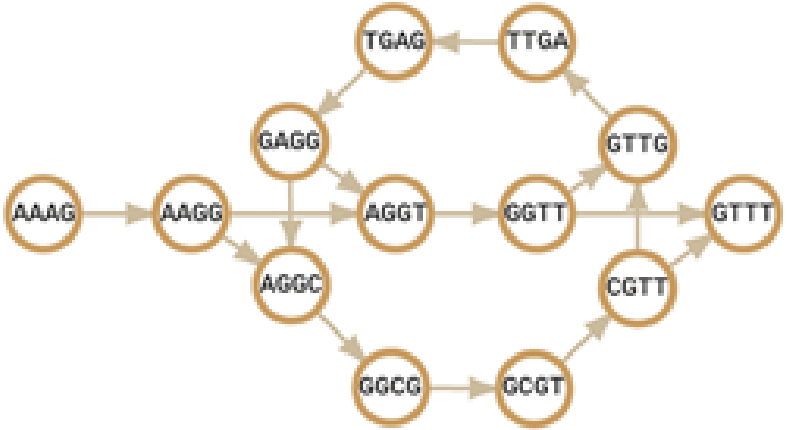
https://dodona.be/en/exercises/649786026/

De Bruijn graph



- Connect read whose **suffix** is identical to the next read's **prefix**
- An assembly is a walk along a **path** in this graph

De Bruijn graph with branches and cycles

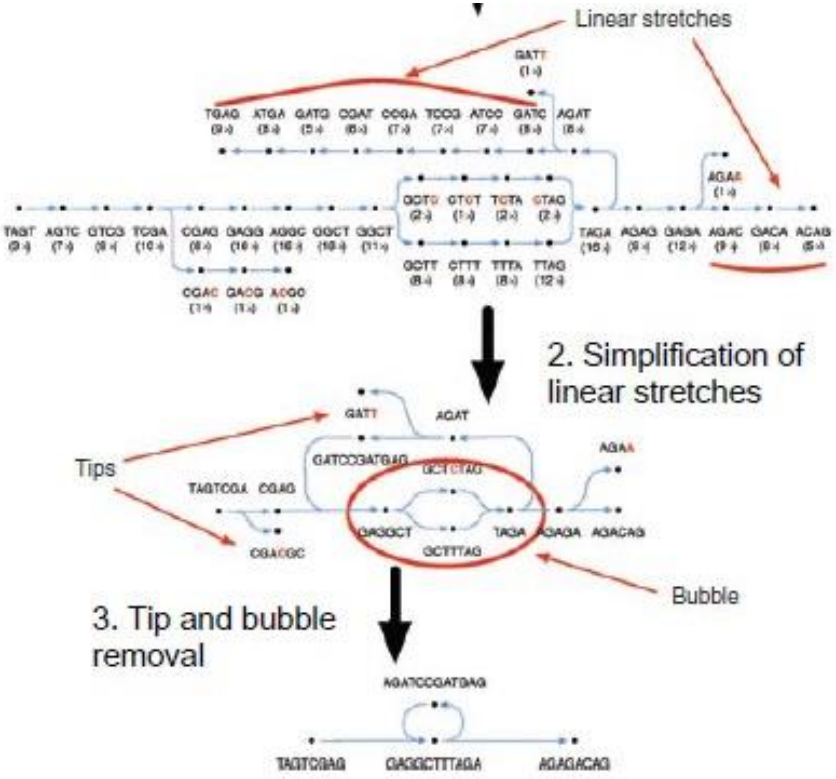


- Which path is best?
  - The longest one that visits as many nodes as possible
  - Cannot visit the same node again (cannot reuse read)
  - Related to the concept of **Hamiltonian Path**

Pruning less probable paths

- **Tips trimming**: remove the less probable branch (shorter sequence, lower read coverage)

- **Bubble removal**: use paired-end information to constrain the length of DNA between the forward read and the reverse read



Assembly of long reads

What changes with long read data?

- **Pros**
  - Less number of reads, smaller de Bruijn graph
  - Less ambiguity of overlapping sequences

- **Cons**
  - More mismatches
  - Multiple possibilities of overlapping sequences

Which one should be connected?

```
AACACATACTCGACTACGACTACGACTAGCACT
                    ACTACGACTAGCACTAGACATCACGCATCA
                        ACGACTAGCACTAGATATAGCTACGACTACGACTACTA
```

Overlap-Layout-Consensus algorithm

- **Overlap**: Connect reads with overlapping sequences (same as before)
- **Layout**: Summarize solvable paths into contigs
- **Consensus**: Call consensus sequence
- Why does OLC fit long read data?
  - Allow flexible overlapping sequences
  - Assume that the reads are not accurate

Finding overlap with dynamic programming

		A	T	G	C	T
	0	0	0	0	0	0
A	0	1	0	0	0	0
G	0	0	0	1	0	0
C	0	0	0	0	2	0
T	0	0	0	0	0	3

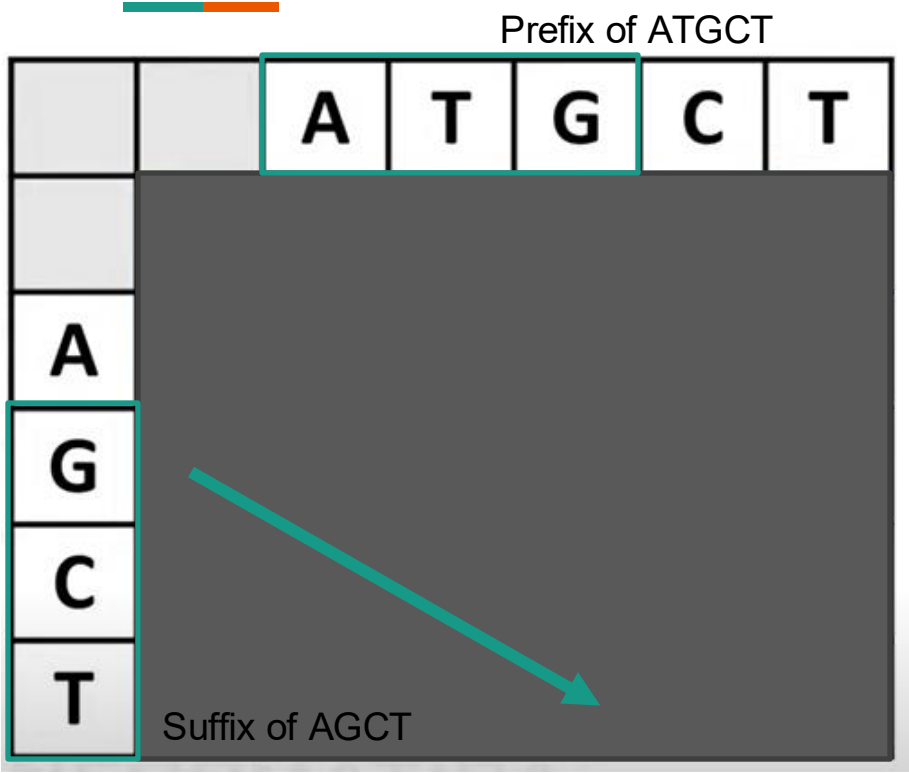
https://www.youtube.com/watch?v=lu9ScxSejSE

Match : 1
Mismatch : -1
GAP : -2

```
Seq1:  ATGCT
          |||
Seq2:   AGCT
```

- For local alignment, we modify the algorithm to **reset negative scores to zeros**
- **How can we modify here**?

Where the path should start and end?



- Overlap = good match between the **suffix of the first sequence (AGCT)** and the **prefix of the second sequence (ATGCT)**

- The alignment path must **start from the first column and end on the last row**

How should the score be initialized?

		A	T	G	C	T
	0	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
A	0					
G	0					
C	0					
T	0					

- **The first column is all zeros** because the alignment can start from anywhere in the first column
- **The first row is minus infinity** (minus large negative) because the alignment cannot start with gap on the second sequence

Putting everything together

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

Match : 1
Mismatch : -1
GAP : -2

```
Seq1:  ATGCT
       | |||
Seq2:  A-GCT
```

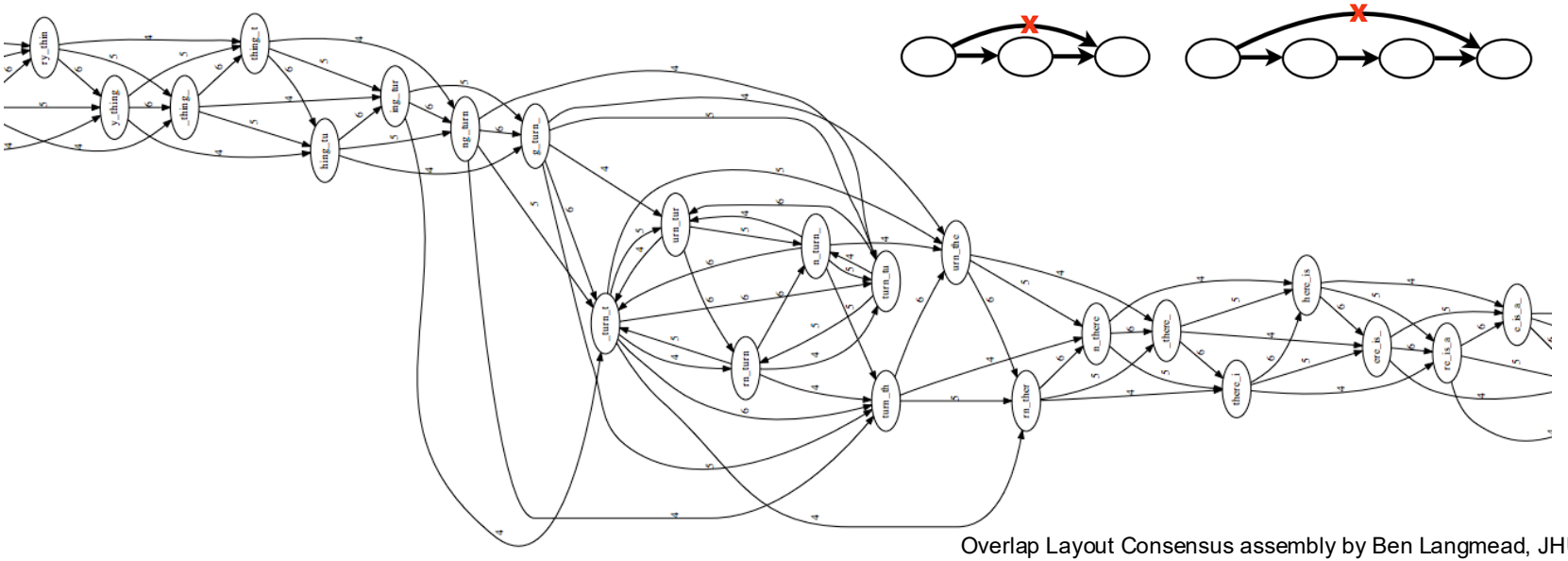
Let's try one more example

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

Match : 1
Mismatch : -1
GAP : -2

```
Seq1:   ATGCT
         |||
Seq2:  ATTG
```

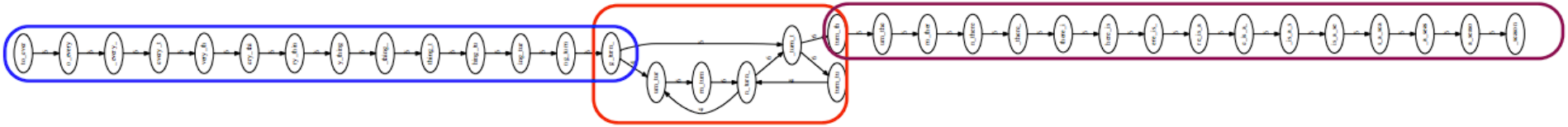
Laying out the contigs



Overlap Layout Consensus assembly by Ben Langmead, JHU

- Long reads will generate a combination of short and long overlaps
- Remove redundant edges to simplify the graph

Calling consensus sequence



```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

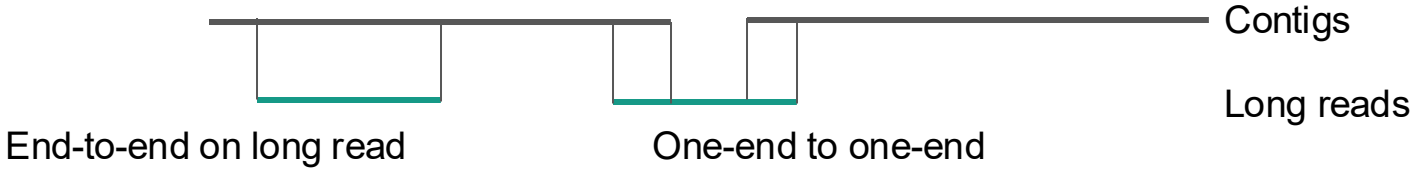
Overlap Layout Consensus assembly by Ben Langmead, JHU

Hybrid assembly

Combining what we already know

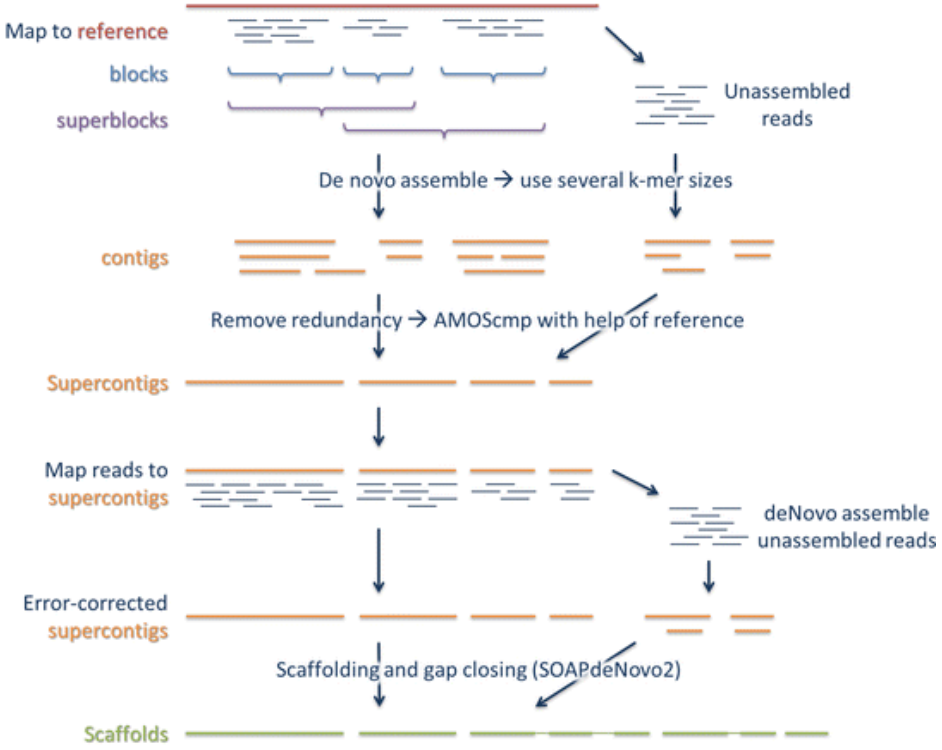
- Generate assembly from short reads

- Align long reads to the assembly
  - **Semi-global alignment**
  - Long read and assembly must reflect a consistent genome structure
  - Long read may span multiple short read contigs



Referenced-guided assembly

Guiding assembly by positions on reference genome



- Aligned reads can be grouped for local assembly
- Unaligned reads can indicate structural variations
- Useful when studying microorganisms, which evolve rapidly

Lischer, H.E.L. and Shimizu, K.K. BMC Bioinformatics 18:474 (2017)

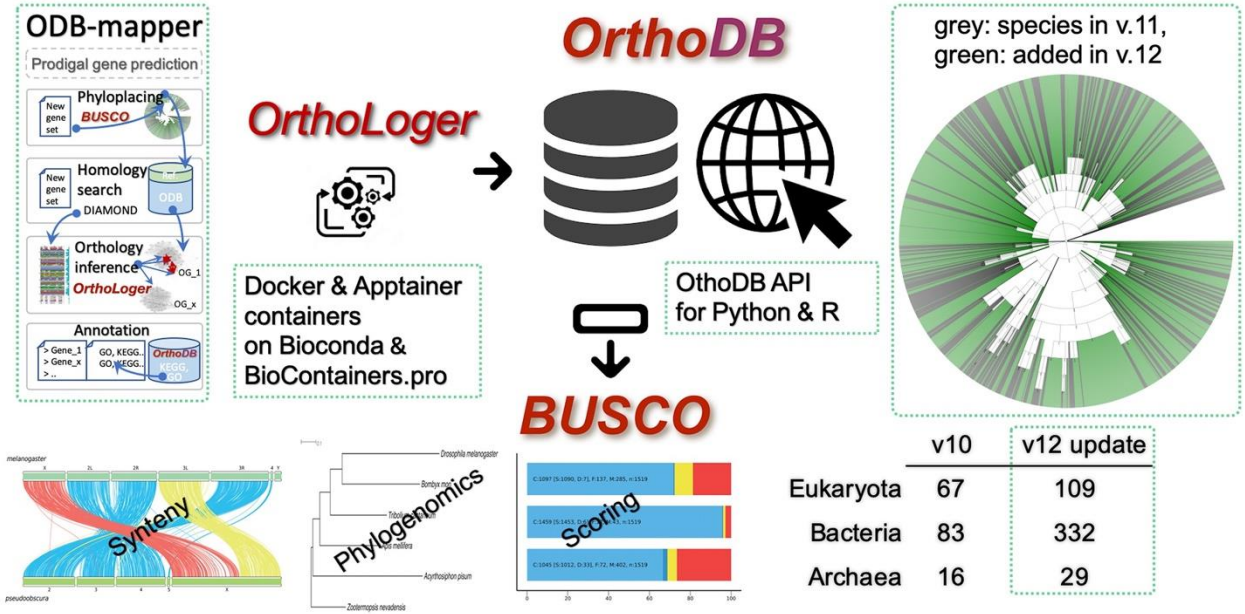
How to check the quality of an assembly?

- Contig count

- **N50**: the length of the shortest contig that form 50% of total contig length
  - Contig lengths: 2, 3, 5, 7, 9
  - 50% of total lengths = 0.5 x (2+3+5+7+9) = 13
  - Contigs that form at least 50% of total length: 7 and 9
  - N50 is 7

- **U50**: N50 for a target genome, in the case where the assembly contains both host/pathogen, plasmid, or background DNA

BUSCO: checking the presence of core genes

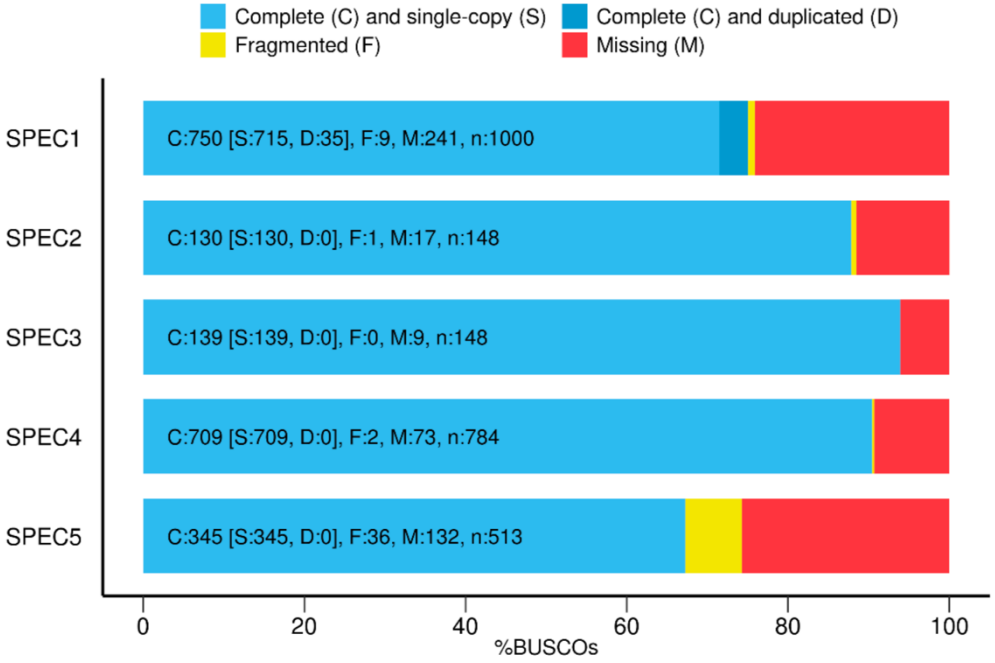


	v10	v12 update
Eukaryota	67	109
Bacteria	83	332
Archaea	16	29

Simao, F.A. et al. Bioinformatics 31:3210-3212 (2015)

- A good assembly should contain all core genes of the related taxonomy

BUSCO assessment result

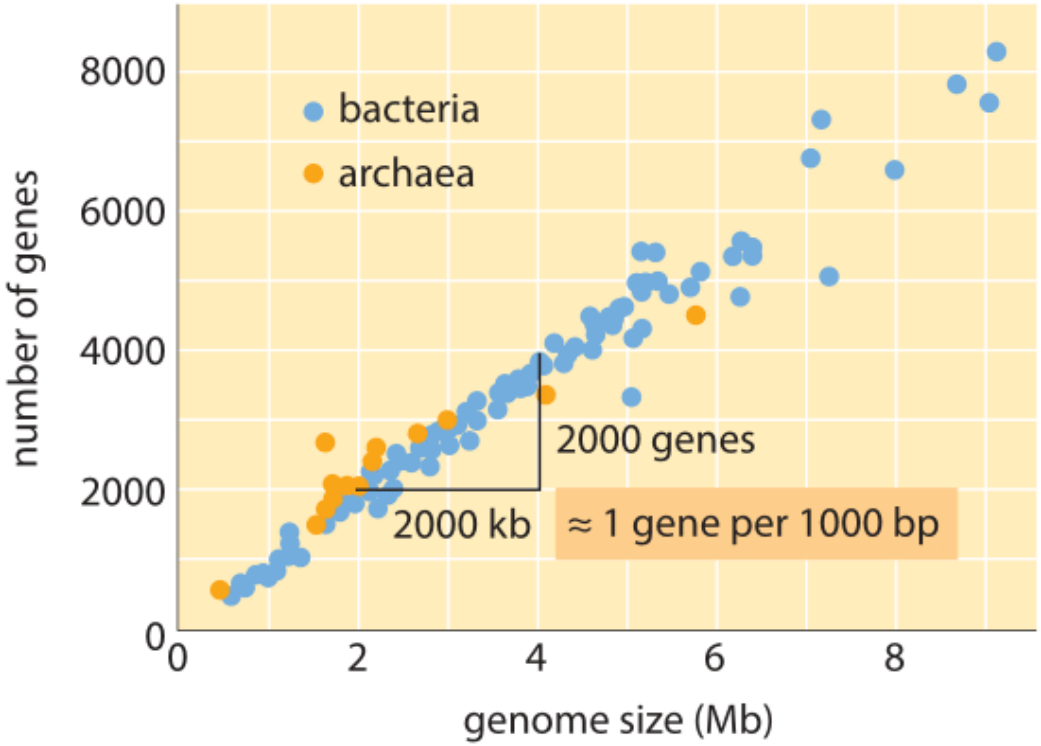


https://busco.ezlab.org/busco_userguide.html

- **C**: Found the complete length of core genes
  - **S/D**: Whether the genes were found as single-copy or duplicated
  - Duplicated can indicate error
- **F**: Found incomplete genes (partial assembly)
- **M**: Did not find the genes

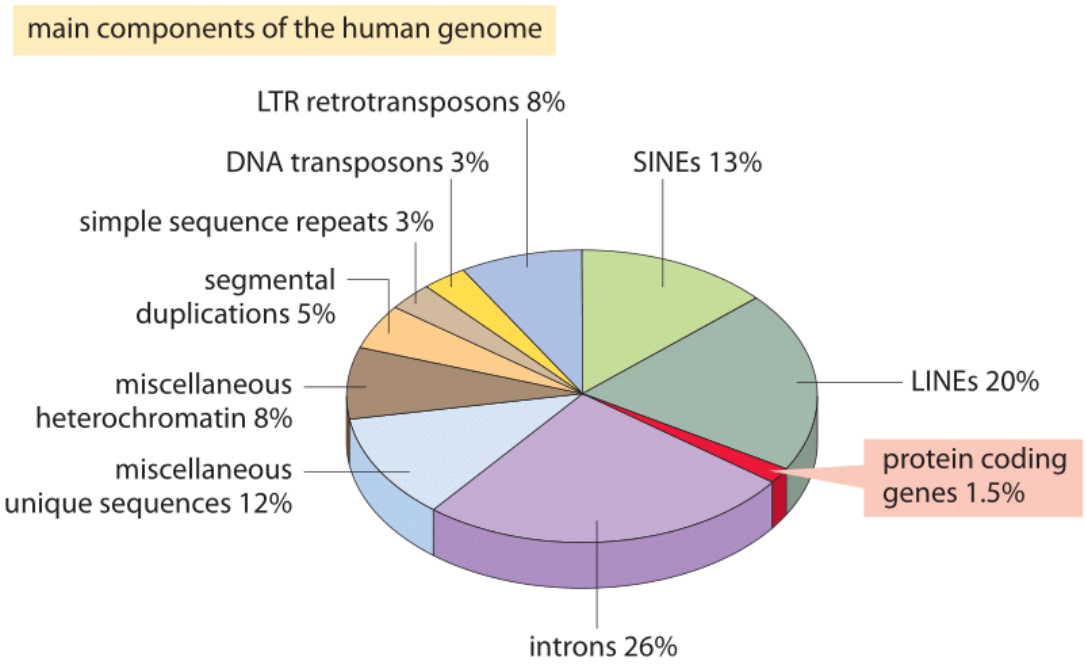
Genomic elements

In microbes, most genomic DNA are genes



Cell biology by the numbers, Milo R. and Phillips, R. https://book.bionumbers.org/

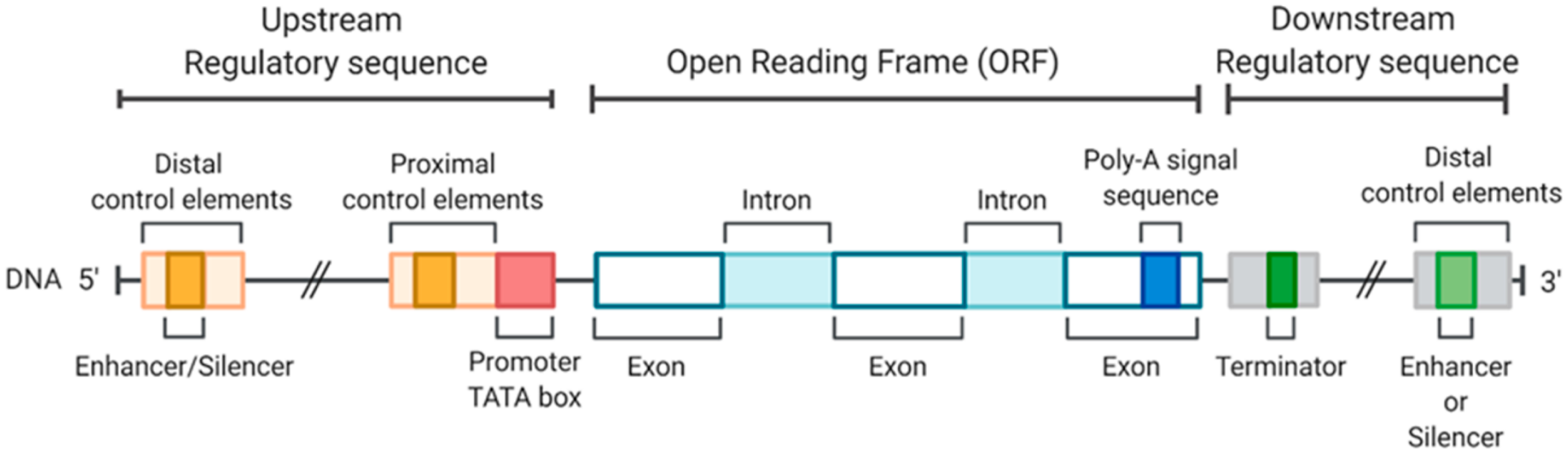
Eukaryotes have complex genomes



	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)
viruses	HIV 1	9	10
	Influenza A virus	10-11	14
	Bacteriophage λ	66	49
	Epstein Barr virus	80	170
prokaryotes	*Buchnera sp.*	610	640
	T. maritima	1,900	1,900
	S. aureus	2,700	2,900
	V. cholerae	3,900	4,000
	B. subtilis	4,400	4,200
	E. coli	4,300	4,600
eukaryotes	*S. cerevisiae*	6,600	12,000
	C. elegans	20,000	100,000
	A. thaliana	27,000	140,000
	D. melanogaster	14,000	140,000
	F. rubripes	19,000	400,000
	Z. mays	33,000	2,300,000
	M. musculus	20,000	2,800,000
	H. sapiens	21,000	3,200,000
	T. aestivum (hexaploid)	95,000	16,800,000

Cell biology by the numbers, Milo R. and Phillips, R. https://book.bionumbers.org/

An example of gene structure

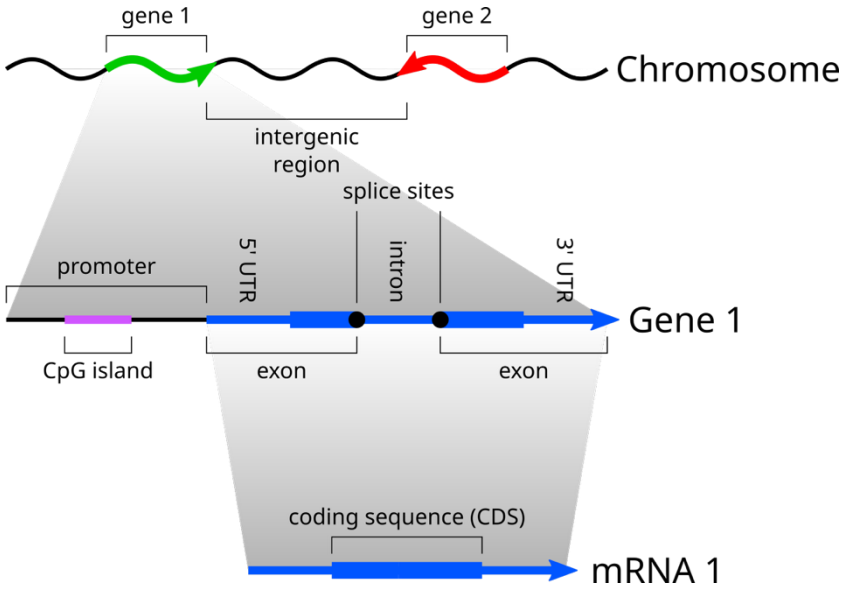


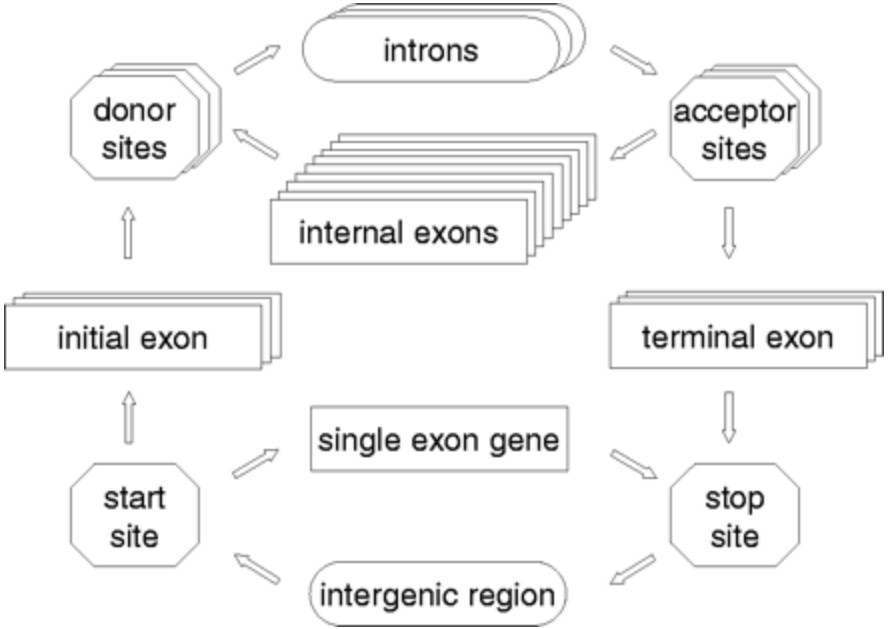
Villao-Uzho, L. et al. Genes 14:1226 (2023)

Genome annotation

- Protein coding genes
  - Exon/intron/splice sites
- Regulatory elements: Promoter/enhancer/silencer
- Repeats
- Non-coding RNA
- **Best if guided by data from other assays, such as RNA-seq and ChIP-seq**

Protein-coding gene annotation



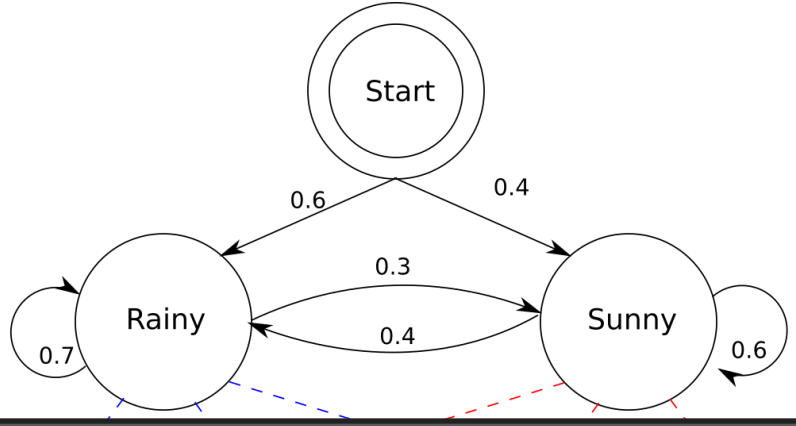


https://cs.rice.edu/~ogilvie/comp571/hidden-markov-models/

Lomsadze, A. et al. NAR 33:6494-6506 (2005)

- Formulate the general structure of protein-coding genes

Markov model



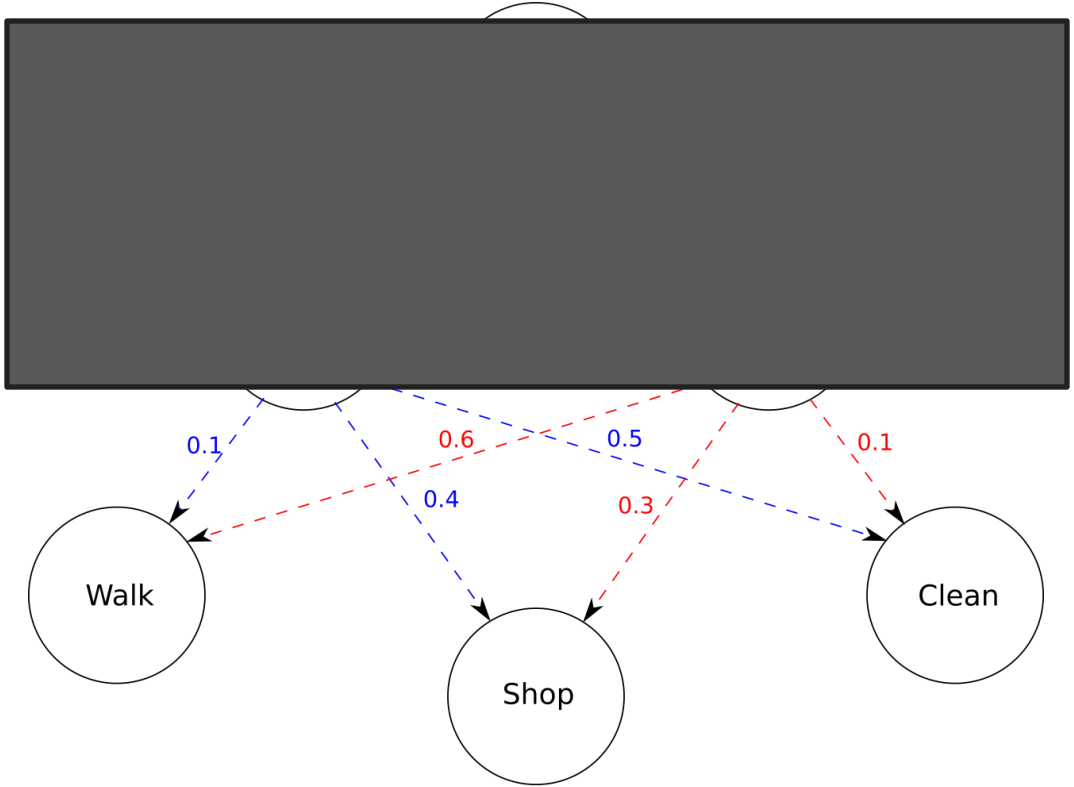
- State changes with specified probability

- P(Sunny | Rainy) = 0.3
- P(Rainy | Rainy) = 0.7

- **Data**: S,R,R,S,S,R

https://en.wikipedia.org/wiki/Hidden_Markov_model

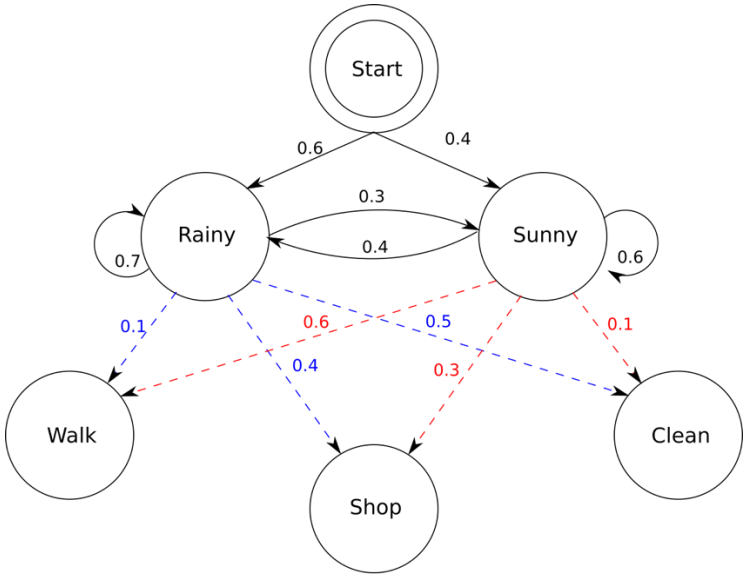
Hidden Markov model



- **Observation** depends on the **hidden** states
- **Data**: W,W,S,C,S,C
- P(Walk | Rainy) = 0.1
- P(Walk | Sunny) = 0.6
- Hidden state changes with specified probability

https://en.wikipedia.org/wiki/Hidden_Markov_model

Decoding a Hidden Markov model

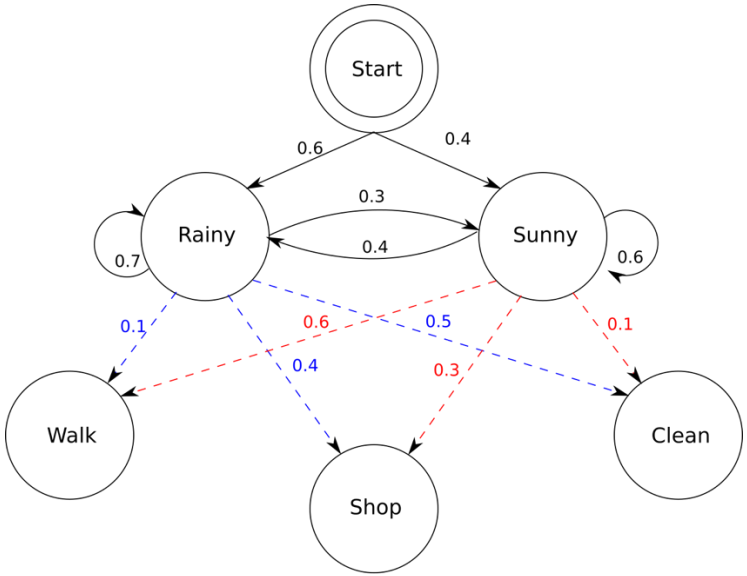


Walk	Walk	Shop	Clean	Shop	Clean
HS1	HS2	HS3	HS4	HS5	HS6

- Given the observation, what is the most likely sequence of hidden states (HS's)?
- Let's consider only the first two observations: Walk-Walk
- We have four possible HS's combinations: RR, RS, SR, and SS

https://en.wikipedia.org/wiki/Hidden_Markov_model

Decoding a Hidden Markov model



Walk	Walk	Shop	Clean	Shop	Clean
HS1	HS2	HS3	HS4	HS5	HS6

Conditional probability:

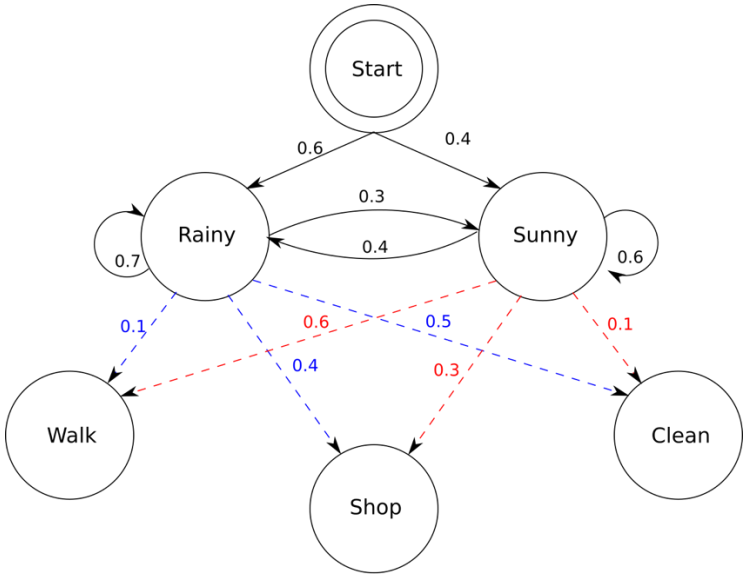
$$P(A \mid B) = P(A, B) / P(B)$$

- $P(RR \mid WW) = P(RR, WW) / P(WW)$
- $P(RS \mid WW) = P(RS, WW) / P(WW)$
- $P(SR \mid WW) = P(SR, WW) / P(WW)$
- $P(SS \mid WW) = P(SS, WW) / P(WW)$

- It suffice to compare joint probabilities

https://en.wikipedia.org/wiki/Hidden_Markov_model

Decoding a Hidden Markov model



Walk	Walk	Shop	Clean	Shop	Clean
HS1	HS2	HS3	HS4	HS5	HS6

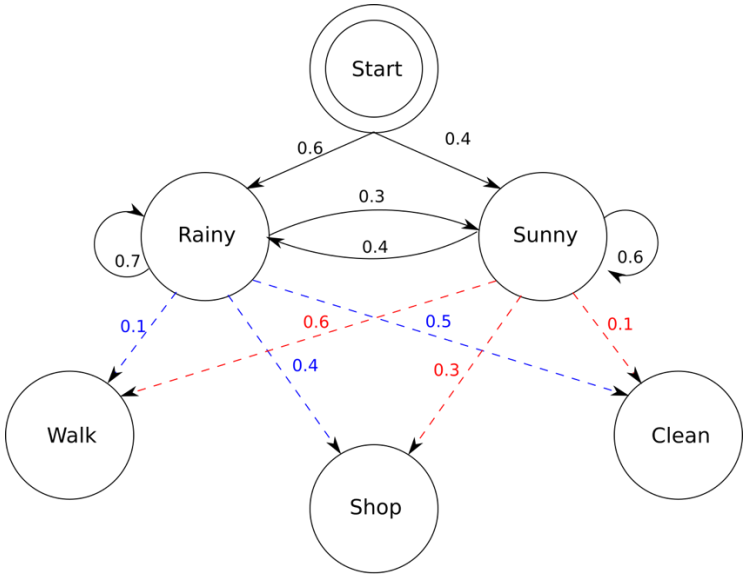
Breakdown the process sequentially:

1. HS1
2. Obs1 given HS1
3. HS2 given HS1
4. Obs2 given HS2

P(RS,WW) = P(Rainy) x P(W | Rainy) x
P(Sunny | Rainy) x
P(W | Sunny)
= 0.6 x 0.1 x 0.3 x 0.6

https://en.wikipedia.org/wiki/Hidden_Markov_model

Decoding a Hidden Markov model



Walk	Walk	Shop	Clean	Shop	Clean
HS1	HS2	HS3	HS4	HS5	HS6

P(RR,WW) = 0.6 x 0.1 x 0.7 x 0.1
P(RS,WW) = 0.6 x 0.1 x 0.3 x 0.6
P(SR,WW) = 0.4 x 0.6 x 0.4 x 0.1
P(SS,WW) = 0.4 x 0.6 x 0.6 x 0.6

Intuition: P(W | Sunny) >> P(W | Rainy)

"Even though it's less likely to observe a Sunny day, it's even less likely to see someone talk a walk on Rainy days"

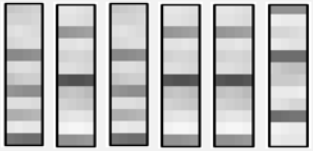
https://en.wikipedia.org/wiki/Hidden_Markov_model

Viterbi algorithm

Input

Observation sequence

$O = (o_1, o_2, o_3, o_4, o_5, o_6)$



β_1 β_3 β_1 β_3 β_3 β_2

Viterbi algorithm

D	$o_1 = \beta_1$	$o_2 = \beta_3$	$o_3 = \beta_1$	$o_4 = \beta_3$	$o_5 = \beta_3$	$o_6 = \beta_2$
α_1	0.4200	0.1008	0.0564	0.0135	0.0033	0
α_2	0.0200	0	0.0010	0	0	0.0006
α_3	0	0.0336	0	0.0045	0.0022	0.0003

E	$o_1 = \beta_1$	$o_2 = \beta_3$	$o_3 = \beta_1$	$o_4 = \beta_3$	$o_5 = \beta_3$
α_1	1	1	1	1	1
α_2	1	1	1	1	3
α_3	1	3	1	3	3

$i_6 = 2$

Output

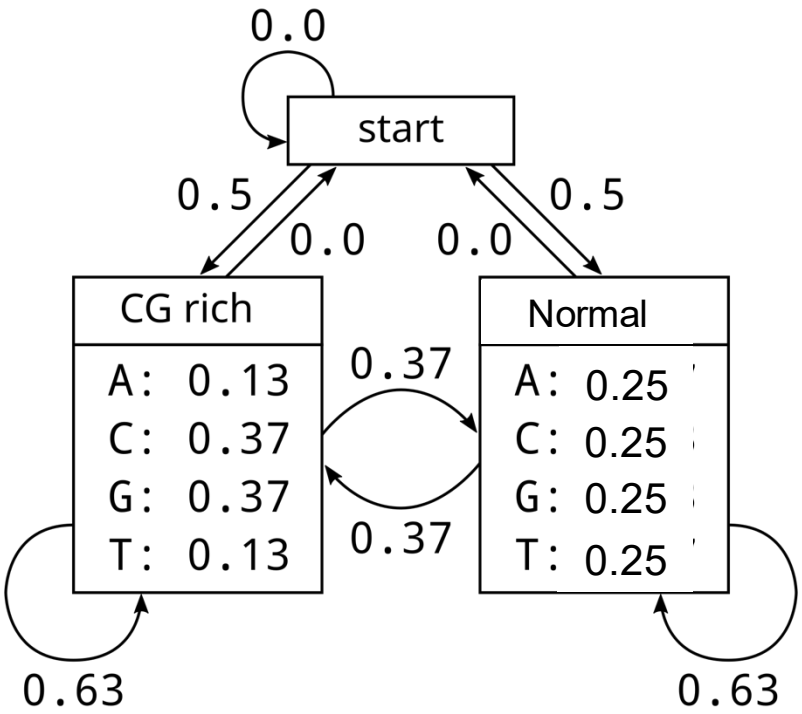
Optimal state sequence

$S^* = (\alpha_1, \alpha_1, \alpha_1, \alpha_3, \alpha_3, \alpha_2)$

Figure 5.28b from [Müller, FMP, Springer 2015]

- Dynamic programming approach
- An optimal solution for *n* observations is based on solutions for *n-1* obs.

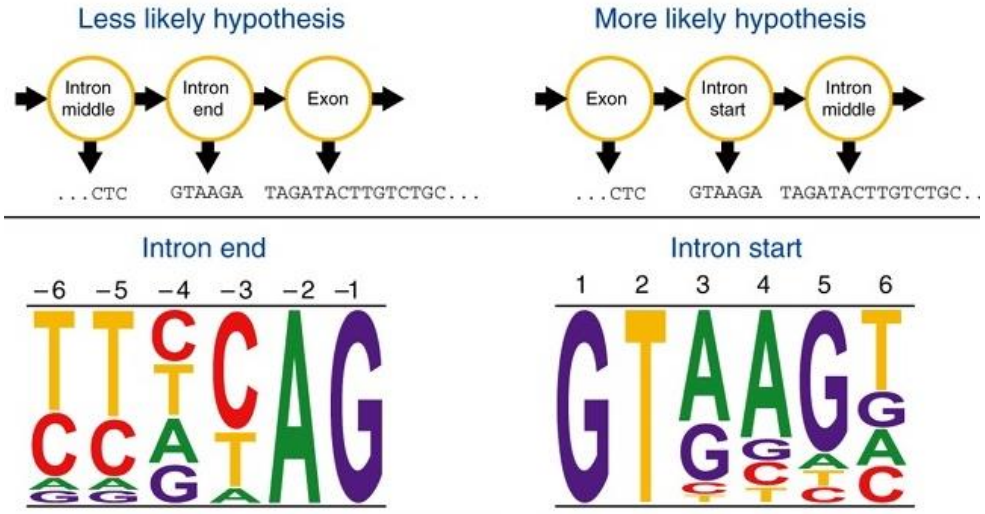
HMM for CpG island annotation



https://cs.rice.edu/~ogilvie/comp571/hidden-markov-models/

- Different genomic elements emit different nucleotide profiles
- CpG islands would consist of more C/G than A/T
- Nucleotide frequencies can be adjusted based on the organism

HMM for protein-coding gene prediction



Brent, M.R. Nature Biotechnology 25:883-885 (2007)

- Consider longer nucleotide segments, not just single bases
- What are the nucleotide frequencies/patterns of each state?
- Must be trained using known genes from related organisms

Any question?

- See you next time