

CS 466

Introduction to Bioinformatics

Lecture 21

Mohammed El-Kebir

Nov 6, 2020



Outline

- Hidden Markov Models: Viterbi algorithm

Reading:

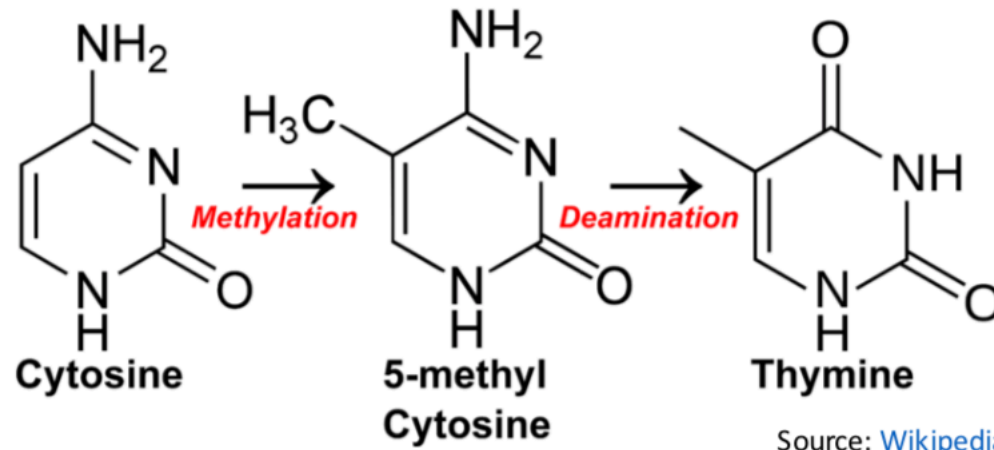
- Jones and Pevzner: Chapters 11.1-11.3
- Lecture notes

CpG Islands

Question: Given four nucleotides $\Sigma = \{A, T, C, G\}$, what is the probability of observing dinucleotide CG ?

CpG Islands

Question: Given four nucleotides $\Sigma = \{A, T, C, G\}$, what is the probability of observing dinucleotide CG ?



Source: [Wikipedia](#)



CG is least observed dinucleotide as C is easily methylated and has tendency to mutate into a T afterwards

CpG Islands

- Methylation is suppressed around promoter regions of genes in a genome. So CG appears at relatively high frequency within these CpG island.
- Finding CpG islands in a genome is an important problem for annotating genes and regulatory regions.

```

CATTCCGGCCTTCTCTCCGGAGGTGGCGCGTGGGA
GGTGTGTTTGTCTGGTTCTGTAAGAAATAGGCCAGG
CAGCTTCCCGCGGATGCGCTCATCCCTCTCGG
GGTTCCCTCCACCGCGCCCGCGTTTGGCGCGGTT
CCGCGTGGAGATGTTTTCCGACGGACAATGATTC
CACTCTCGGCGCCTCCCATGTTGATCCAGCTCCT
CTGCGGGCGTCAGGACCCCTGGGCCCGCGCCCG
CTCCACTCAGTCAATCTTTTGTCCCGTATAAGGCG
GATTATCGGGGTGGCTGGGGGCGGCTGATTCCGA
CGAATGCCCTTGGGGGTACCCCGGAGGGAATC
CGGGCTCGGCTTTGGCCAGCCCGCACCCCTGGT
TGAGCCCGGCCCGAGGGCCACAGGGGGCGCTCG
ATGTTCTCTGCAGCCCCCGCAGCAGCCCCACTCC
CCCGCTCACCCTACGATTGGCTGGCCCGCCCGAG
CTCTGTGCTGTGATTGGTCACAGCCCGTGTCCGT
CGCGGGCGCGGGCGGATACAGGTGAAGCGCA
GAGGCCAGCTCGGGCGGTGTCCCGCGCGCG
GACTGCGGGCGGAGTTTCCCGAGGGCCGAAGCG
GGGCAGTGTGAAGCAGCGGTCTGGGAGGCC
CCGGCGCGCGTCCGAGCAGCTCCCGTCTCGCA
GCCTCACCGCGCGCGTCCCGCCCTGGCC
TCCCGCACTCGCGCACTCCTGTCCCGCCAC
GCCACCTCCACCTCGATGCGGTGCGGGCTGC
TGCGTGATGGGGCTCGGAGCGGCGCCTGCGG
CTCGCGCGCGCGCTGCTCGCGCTGAGGTGCGT
CGGTGCCCGGCCCGCGCCCCCGCGCGCGCG
GGCTCCTGTTGACCCTGTCGGCCGTCTGCTGC
AGCGCGGCTGAGGTAAGGCGCGGGGCTGGCG
CGGTTGGCGCGCGGTCCCGGGGTTGGGGAGGG
GGCCTTCCCGCGGGGAGGAGCGGCCGGCCGG
GGTCCGGCGGGGTCTGAGGGGA
CTCTTAGTTTTGGGTGCATTTGTCTGGTCTTCCAAA
CTAGATTGAAAGCTCTGAAAAAAATATCTTGT
GTTTCTATCTGTTGAGCTCATGATAGGTATCCAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCCCATCTCCCTTAGTTTTCT
TTTTTCTTTCTTTCTTTCTTTTCTTTTCTTTTCT
TTGAGATGTCTCTTGTCTAGTCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCAGGTTCAAGCAATTCTACTGCCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACCCCGCACCAT
TCCTGGCTAATTTTTTTTTGTATTTTAGTTGAGA
CAGGGTTTCCACATGTTGGTGTGCTGGTCTCAGA
CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCAGAGTGTAGGATTACAGGCATGAGCCACTGT
ACCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTTAAGATAAAGTTACGATTTGAAAT
CTTTGGATTGAGAAAGATTTGTACCTTTAACACCT
AGAGTTGAACTTCATACCTGGAGAGCCTTAACATT
AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCTCCCTGAAGTGGACT
GAGAGCCACACCCTGGCCTGTACCATACCCATCC
CCTATCCTTAGTGAAGCAAACTCCTTTGTTCCCTT
CTCCTTCTCCTAGTGACAGGAAATATTGTATCCTA
AAGAATGAAAATAGCTTGTACCTCGTGGCCTCAG
GCCTCTTGACITTCAGGCGTTCTGTTTAATCAAGT
GACATCTTCCCGAGGCTCCTGAATGTGGCAGATG
AAAGAGACTAGTTCAACCTGACCTGAGGGGAAAG
CCTTTGTGAAGGGTCAGGAG

```

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG constitutes the start codon.

Right: CpG sites present at every 1/100 nucleotides, constituting a more normal example of the genome, or a region of the genome that is commonly methylated.


Source: [Wikipedia](https://en.wikipedia.org/wiki/CpG_island)

CpG Islands

- Methylation is suppressed around promoter regions of genes in a genome. So CG appears at relatively high frequency within these CpG island.
- Finding CpG islands in a genome is an important problem for annotating genes and regulatory regions.

Input: DNA sequence $\mathbf{x} = x_1x_2 \dots x_n$

Output: $\pi : \{1, \dots, n\} \rightarrow \{\text{yes}, \text{no}\}$

$\mathbf{x} =$	CGAT	TG	CGA	AAAAAAT	AACGA	TTATATCG
						
$\pi =$	yes	no	yes	no	yes	no

```

GATTCGGCCTTCTCTCCGGAGGTGGCGCGTGGGA
GGTGTCTTGTCTCGGTTCTGTAAGAAATAGGCCAGG
CAGCTTCCCGCGGATGCGCTCATCCCTCTCGG
GGTTCCCTCCACCGCGCGCGGTTGGCGCGGTT
CCGCTGCGAGATGTTTTCCGACGACAATGATTC
CACTCTCGGCGCCTCCCATGTTGATCCAGCTCCT
CTGCGGGCGTCAAGGACCCCTGGGCCCGCGCGG
CTCCACTCAGTCAATCTTTGTCCCGGTATAAGGCG
GATTATCGGGGTGGCTGGGGGCGGCTGATTCCGA
CGAATGCCCTTGGGGGTACCCCGGAGGGAATC
CGGGCTCGGCTTTGGCCAGCCCGCACCCCTGGT
TGAGCCCGGCCCGAGGGCCACAGGGGGCGCTCG
ATGTTCTGTCAGCCCCCGCAGCAGCCCCACTCC
CCCGCTCACCCTAAGATTGGCTGGCCCGCCCGAG
CTCTGTGCTGTGATTGGTCACAGCCGTGTCCGT
CGCGGGCGCGGGCGGATACAGGTGAAGCGCA
GAGGCCAGCTCGGGCGGTGTCCCGCGCGCGG
GACTGCGGGCGGAGTTTCCCGAGGGCGGAAGCG
GGGCAGTGTGACGGCAGCGGTCTGGGAGGCC
CCCGCGCGCGTCCGAGCAGCTCCCGTCTCCGCA
GCCTCACCGCGCGCGTCCCGCGCCCTGGCC
TCCCGCACTCGCGCACTCCTGTCCCGCGCCAC
GCCACCTCCACCTCGATGCGGTGCGGGCTGC
TGCGTGATGGGGCTGCGAGCGGCGCCTGCGG
CTCGCGGCGCGCGCTGCTCGCGCTGAGGTGCGT
CGGTGCCCGGCCCGCGCGCCCGCGCGCGCG
GGCTCCTGTTGACCCTGCGCGCGCTCGGTCTGC
AGCGCGGCTGAGGTAAGGCGCGGGGCTGGCG
CGGTTGCGCGCGGTCCCGGGGTTGGGGAGGG
GGCGCTTCCCGCGGGGAGGCGCGCGGCCCG
GGTCCGGCGGGGTCTGAGGGGA
CTCTAGTTTTGGGTGCATTTGTCTGGTCTTCCAA
CTAGATTGAAGCTCTGAAAAAAATATCTTGT
GTTTCTATCTGTTGAGCTCATAGTAGGTACAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCCCATCTCCCTTAGTTTTCT
TTTTTCTTTCTTTCTTTCTTTTCTTTTCTTTT
TTGAGATGTCTCTTGTCTCAGTCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCAGGTTCAAGCAATTCTACTGCCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACCCCGCACCAT
TCCTGGCTAATTTTTTTTGTATTTTGTGAGTGA
CAGGGTTTCCACATGTTGGTGTGCTGGTCTCAGA
CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCCAGAGTGTAGGATTACAGGCATGAGCCACTGT
ACCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTTAAGATAAAGTTACGATTTGAAAT
CTTTGGATTGAGAAGAAATTTGTACCTTTAACACCT
AGAGTTGAAGTTTCATACCTGGGAGAGCCTTAACATT
AAGCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCTCCCTGAAAGTGGACT
GAGAGCCACACCCTGGCCTGTACCATACCCATCC
CCTATCCTTAGTGAAGCAAACTCCTTTGTTCCCTT
CTCCTTCTCCTAGTGACAGGAAATATTGTATCCTA
AAGAATGAAAATAGCTTGTACCTCGTGGCCTCAG
GCCTCTTGACITTCAGGCGGTTCTGTTTAATCAAGT
GACATCTTCCCGAGGCTCCTGAATGTGGCAGATG
AAAGAGACTAGTTCAACCTGACCTGAGGGGAAAG
CCTTTGTGAAGGGTCAAGAG
    
```

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG constitutes the start codon.

Right: CpG sites present at every 1/100 nucleotides, constituting a more normal example of the genome, or a region of the genome that is commonly methylated.

Source: [Wikipedia](https://en.wikipedia.org/wiki/CpG_island)

Question: How do we identify CpG islands?

A Related Problem: Fair Bet Casino

- Game is to flip coins, two outcomes:



Head or Tail

- Two coins: **Fair** and **Biased**

$$\Pr(H \mid F) = \Pr(T \mid F) = 1/2$$

$$\Pr(H \mid B) = 3/4, \Pr(T \mid B) = 1/4$$

- The crooked dealer changes between Fair and Biased coins with probability 10%









CpG Islands and Fair Bet Casino

CG Islands

Input: DNA sequence $\mathbf{x} = x_1x_2 \dots x_n$ where $x_i \in \{A, T, C, G\}$

Output: $\pi : \{1, \dots, n\} \rightarrow \{\text{yes}, \text{no}\}$

$\mathbf{x} =$	CGAT	TG	CGA	AAAAAAT	AACGA	TTATATCG
						
$\pi =$	yes	no	yes	no	yes	no

Fair Bet Casino

Input: Coin flips $\mathbf{x} = x_1x_2 \dots x_n$ where $x_i \in \{H, T\}$

Output: $\pi : \{1, \dots, n\} \rightarrow \{F, B\}$

$\mathbf{x} =$											
$\pi =$	F	F	F	B	B	B	B	B	F	F	F
$\pi' =$	F	F	F	B	B	B	F	F	F	F	F

Question: Given \mathbf{x} , what is more likely: π or π' ?

Markov Model $\mathcal{M} = (Q, A)$

- Set of states Q

- Markov property:

$$\Pr(Q_i = q_i \mid Q_1 = q_1, \dots, Q_{i-1} = q_{i-1}) = \Pr(Q_i = q_i \mid Q_{i-1} = q_{i-1})$$

- Transition probabilities $A = [a_{ij}]$ on pairs of states

- Rows sum to 1

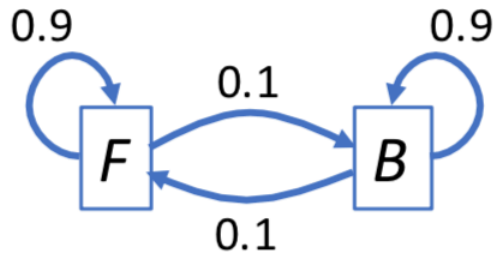


Andrey Markov (source: [Wikipedia](https://en.wikipedia.org/wiki/Andrey_Markov))

Fair Bet Casino

$$Q = \{F, B\}$$

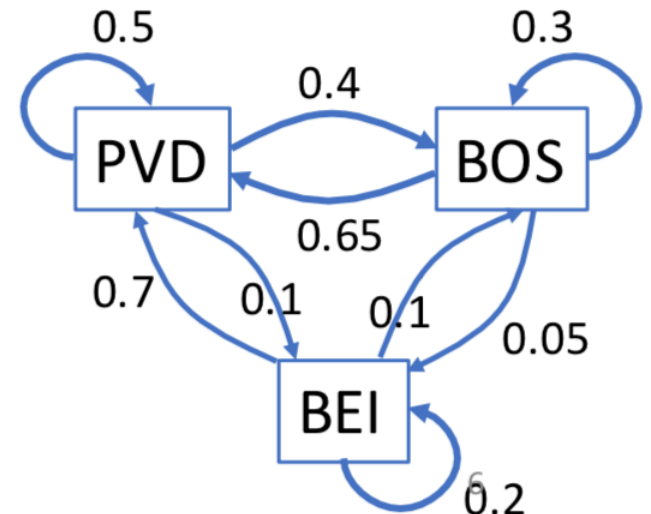
$$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$



Where is the professor?

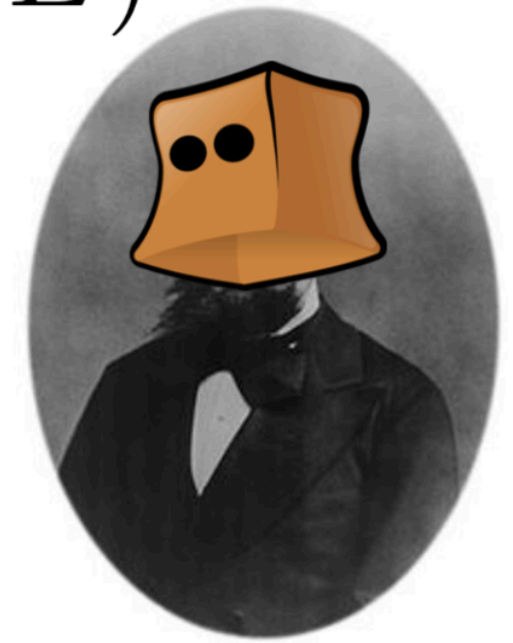
$$Q = \{\text{Providence, Boston, Beijing}\}$$

$$A = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.65 & 0.3 & 0.05 \\ 0.7 & 0.1 & 0.2 \end{pmatrix}$$



Hidden Markov Model $\mathcal{M} = (Q, A, \Sigma, E)$

- Set of *hidden* states Q
 - Markov property
- Transition probabilities $A = [a_{ij}]$ on pairs of states
- Set of *emitted* symbols Σ
- Emission probabilities $E = [e_{ik}]$ on state-symbol pairs



Andrey Markov

Two decisions:

1. What symbol should I emit?
[emission probabilities E]
2. What state should I move to next?
[transition probabilities A]

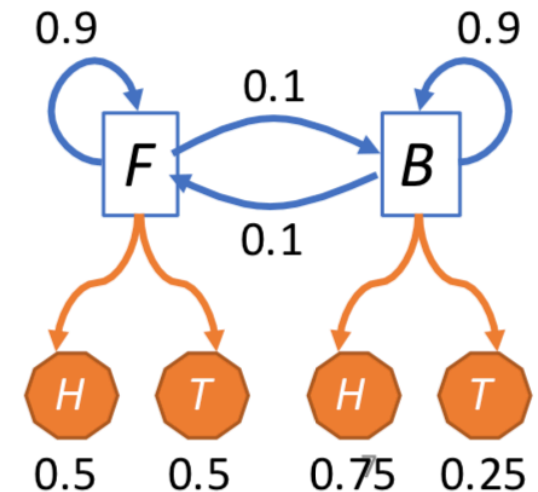
Fair Bet Casino

$$Q = \{F, B\}$$

$$A = \begin{pmatrix} & F & B \\ \begin{matrix} F \\ B \end{matrix} & \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \end{pmatrix}$$

$$\Sigma = \{H, T\}$$

$$E = \begin{pmatrix} & H & T \\ \begin{matrix} F \\ B \end{matrix} & \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix} \end{pmatrix}$$



Three Questions

Question 1:

What is the most probable path π^* that generated observations \mathbf{x} ?

Question 2:

What is probability of observations \mathbf{x} generated by any path π ?

Question 3:

What is the probability of observation x_i generated by state s ?

Three Questions

Question 1:

What is the most probable path π^* that generated observations \mathbf{x} ?

Question 2:

What is probability of observations \mathbf{x} generated by any path π ?

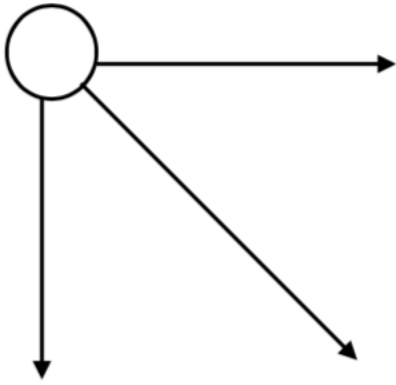
Question 3:

What is the probability of observation x_i generated by state s ?

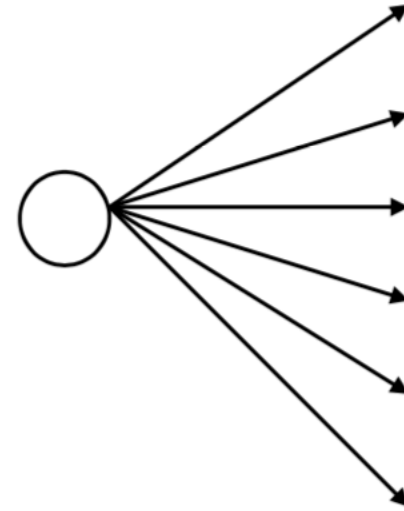
Joint Probability

Recurrence

Alignment vs. Decoding Problem



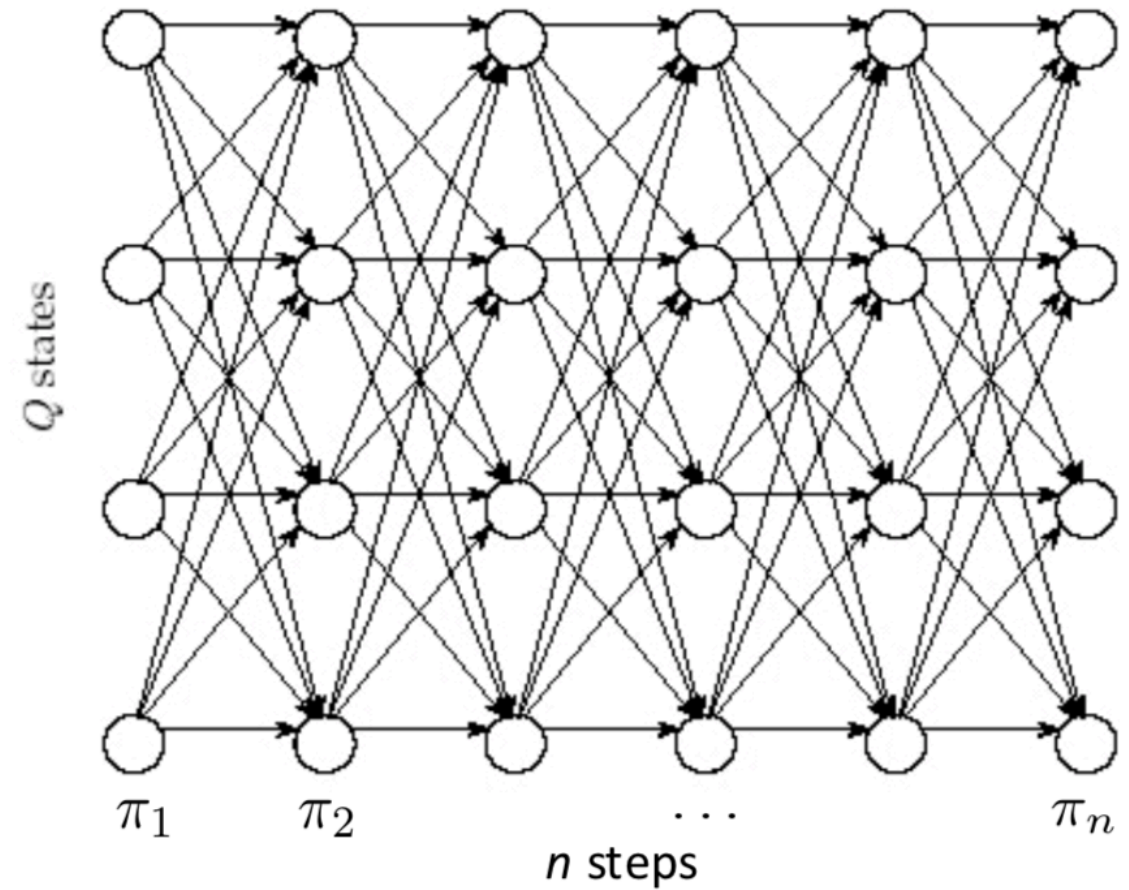
Valid directions in the
alignment problem.



Valid directions in the
decoding problem.

Viterbi Algorithm

- Finds path π^* with maximum $\Pr(\mathbf{x}, \pi^*)$
- Dynamic Programming algorithm
- Runs in $O(\#edges) = O(n|Q|^2)$



Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

$$v[s, i] = \begin{cases} a_{0,s} \cdot e_{s,x_1}, & \text{if } i = 1, \\ e_{s,x_i} \max_{t \in Q} \{v[t, i-1] a_{t,s}\}, & \text{if } i > 1. \end{cases}$$

Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

$$v[s, i] = \begin{cases} a_{0,s} \cdot e_{s,x_1}, & \text{if } i = 1, \\ e_{s,x_i} \max_{t \in Q} \{v[t, i-1] a_{t,s}\}, & \text{if } i > 1. \end{cases}$$

Use logarithms!

$$\log(v[s, i]) = \begin{cases} \log(a_{0,s}) + \log(e_{s,x_1}), & \text{if } i = 1, \\ \log(e_{s,x_i}) + \max_{t \in Q} \{\log(v[t, i-1]) + \log(a_{t,s})\}, & \text{if } i > 1. \end{cases}$$

Fair Bet Casino: Example



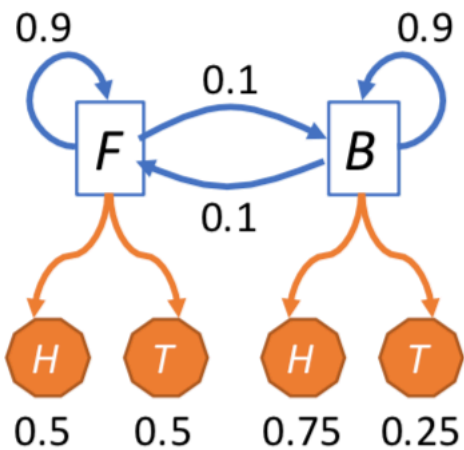
	0	1 (T)	2 (H)	3 (T)	4 (H)	5 (H)	6 (H)	7 (T)	8 (H)	9 (H)	10 (H)	11 (T)
F												
B												

$Q = \{F, B\}$

$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \begin{matrix} F \\ B \end{matrix}$

$\Sigma = \{H, T\}$

$E = \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix} \begin{matrix} H \\ T \end{matrix} \begin{matrix} F \\ B \end{matrix}$



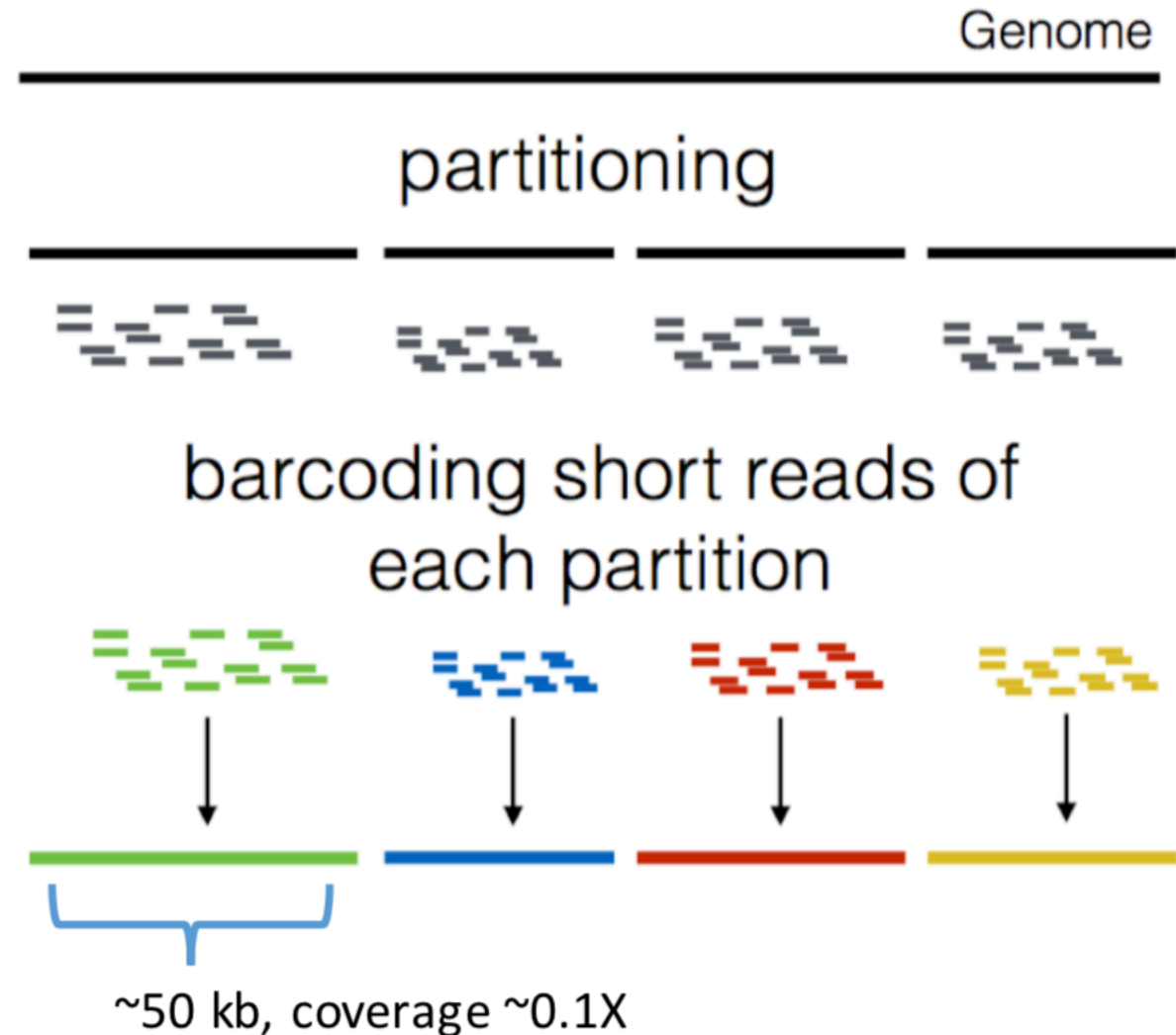
$$v[s, i] = \begin{cases} a_{0,s} \cdot e_{s,x_1}, & \text{if } i = 1, \\ e_{s,x_i} \max_{t \in Q} \{v[t, i - 1] a_{t,s}\}, & \text{if } i > 1. \end{cases}$$

10X Genomics: Synthetic Long Reads

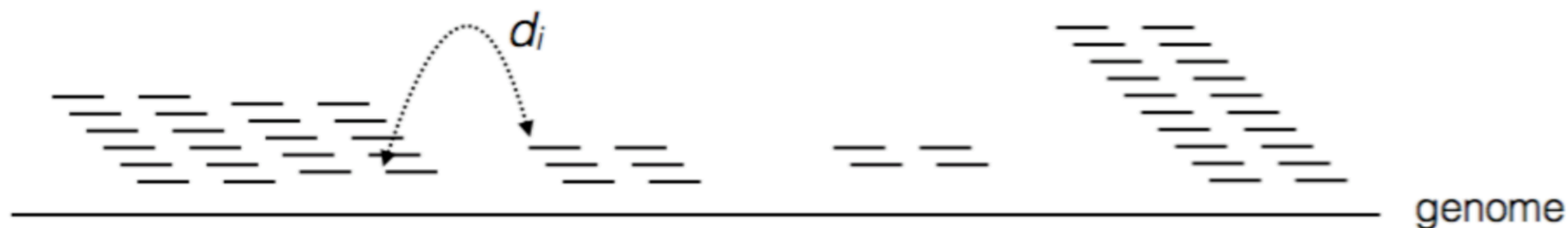
Genome indexing by partitioning and molecular barcoding



50-75 molecules
per droplet



$R_j = \{r_i \mid \forall i, r_i \text{ contains barcode } j\}$: Paired-reads possessing barcode j



Sort linked-reads and calculate distances between them

$$D_j = [d_i \mid \forall i, d_i : \text{distance between } r_i \text{ and } r_{i+1}, r_i < r_{i+1}]$$

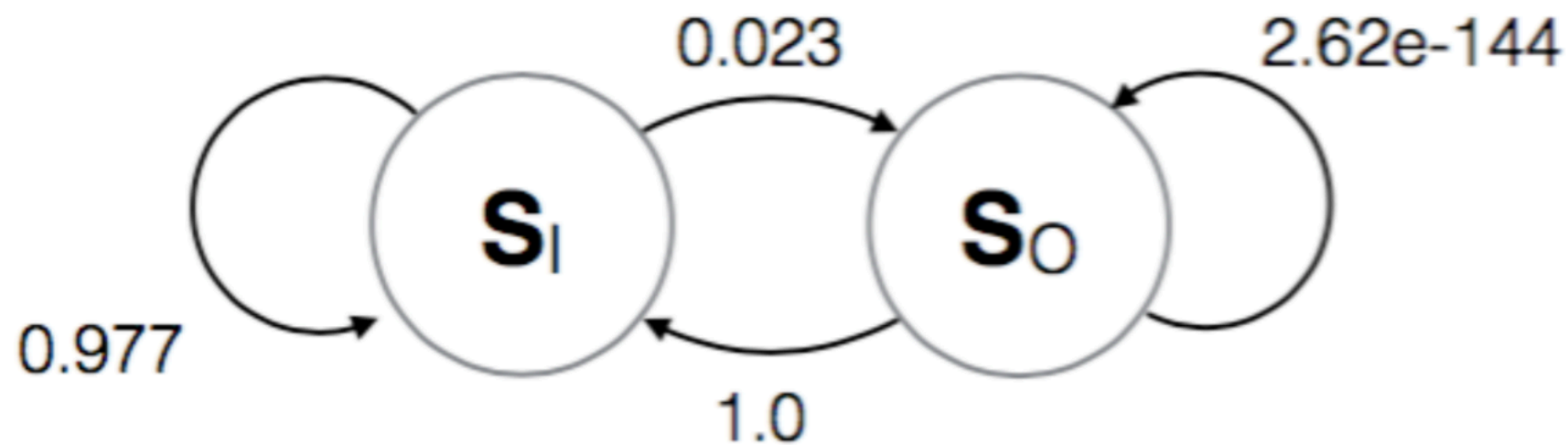
Define distances d_i as intra- or inter- long molecules

$$\Sigma_i = [d_1, d_2, d_3, \dots, d_{500}, d_{501}, d_{502}, \dots, d_{1001}, d_{1002}, d_{1003}, \dots]$$

$$Q_j = [I, I, I, \dots, I, O, I, \dots, I, O, I, \dots]$$

I: intra long molecule, O: inter long molecules

HMM

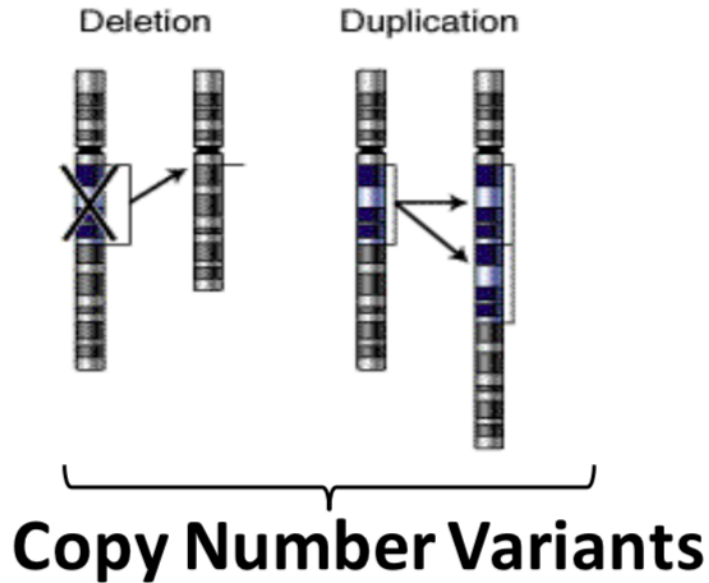


$$\Sigma_i = [d_1, d_2, d_3, \dots, d_{500}, d_{501}, d_{502}, \dots, d_{1001}, d_{1002}, d_{1003}, \dots]$$

$$Q_j = [\underbrace{I, I, I, \dots, I}_{\ell_{j1}}, \underbrace{O, I, \dots, I}_{\ell_{j2}}, \underbrace{O, I, \dots}_{\ell_{j3}}]$$

I: intra long molecule, O: inter long molecules

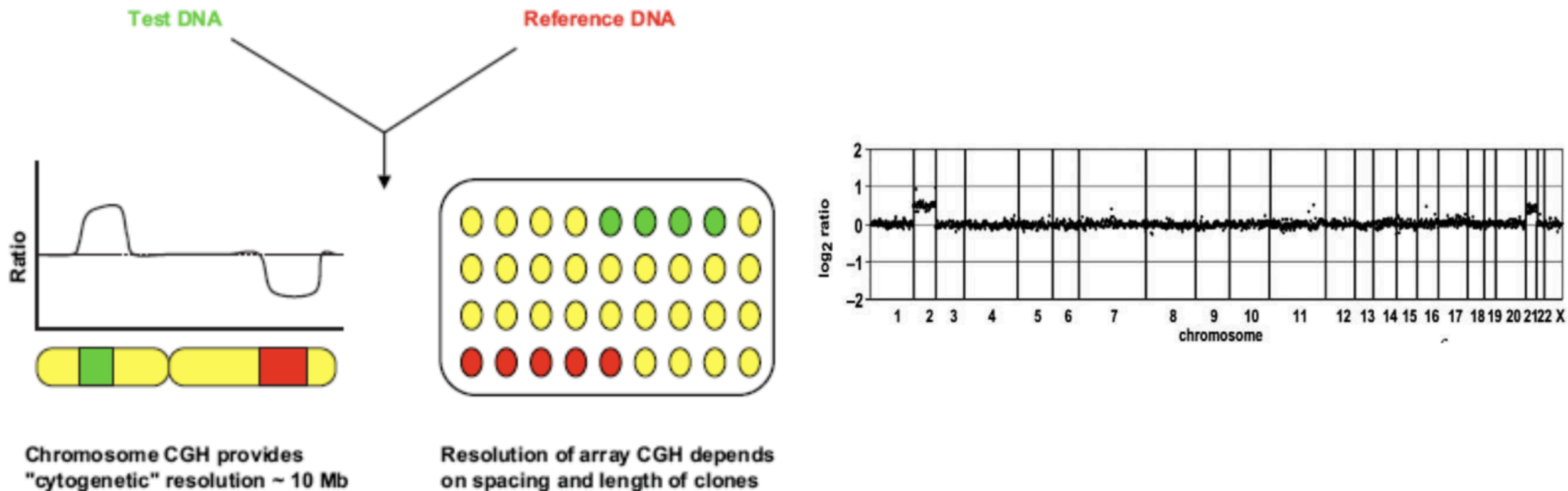
Copy Number Variation



- Different individuals may have different number of copies of segments of genome.
- These variants are associated with various diseases: autism, schizophrenia, cancer

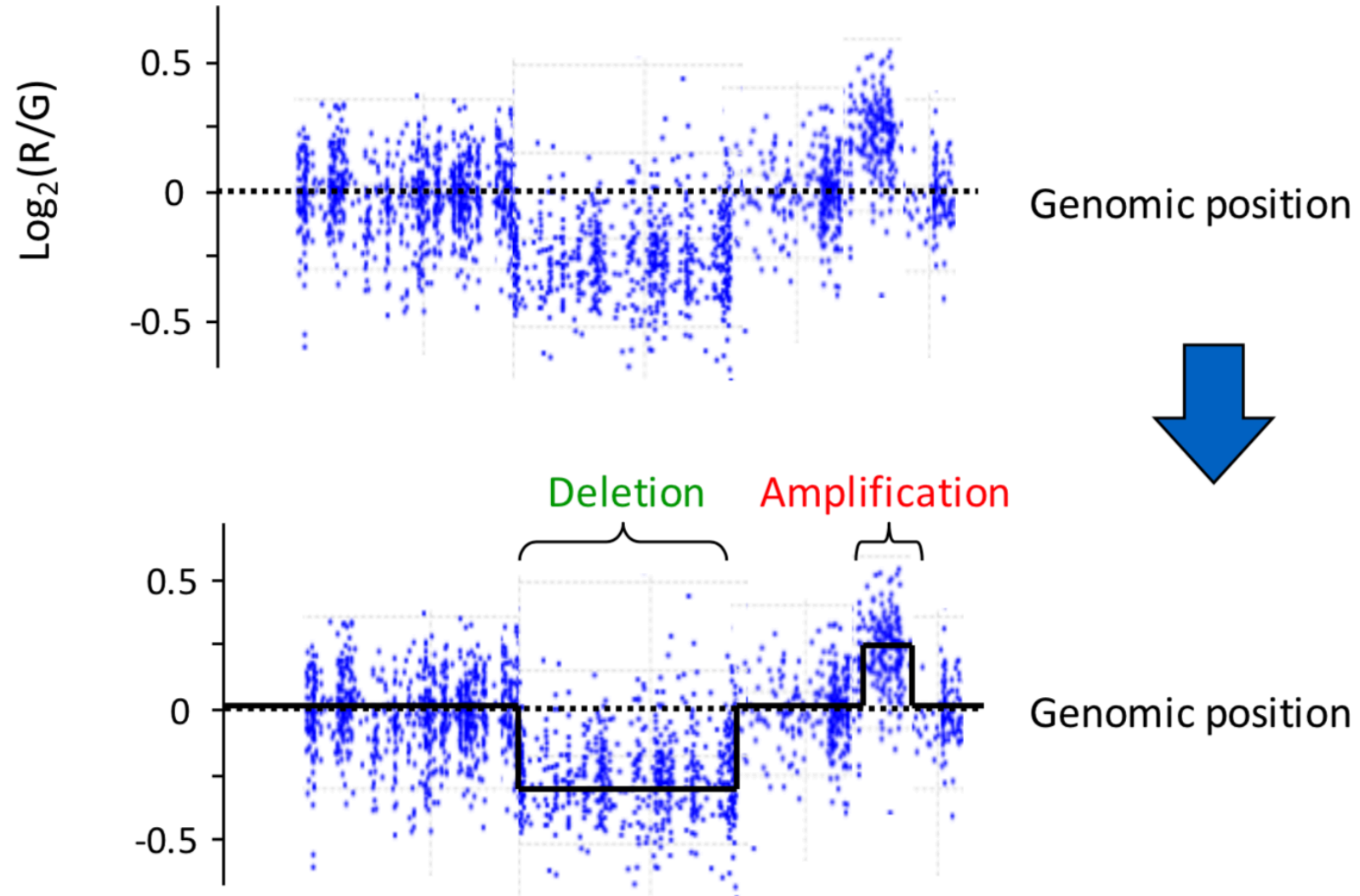
Measuring Copy Number Variants

Comparative Genomic Hybridization (CGH)



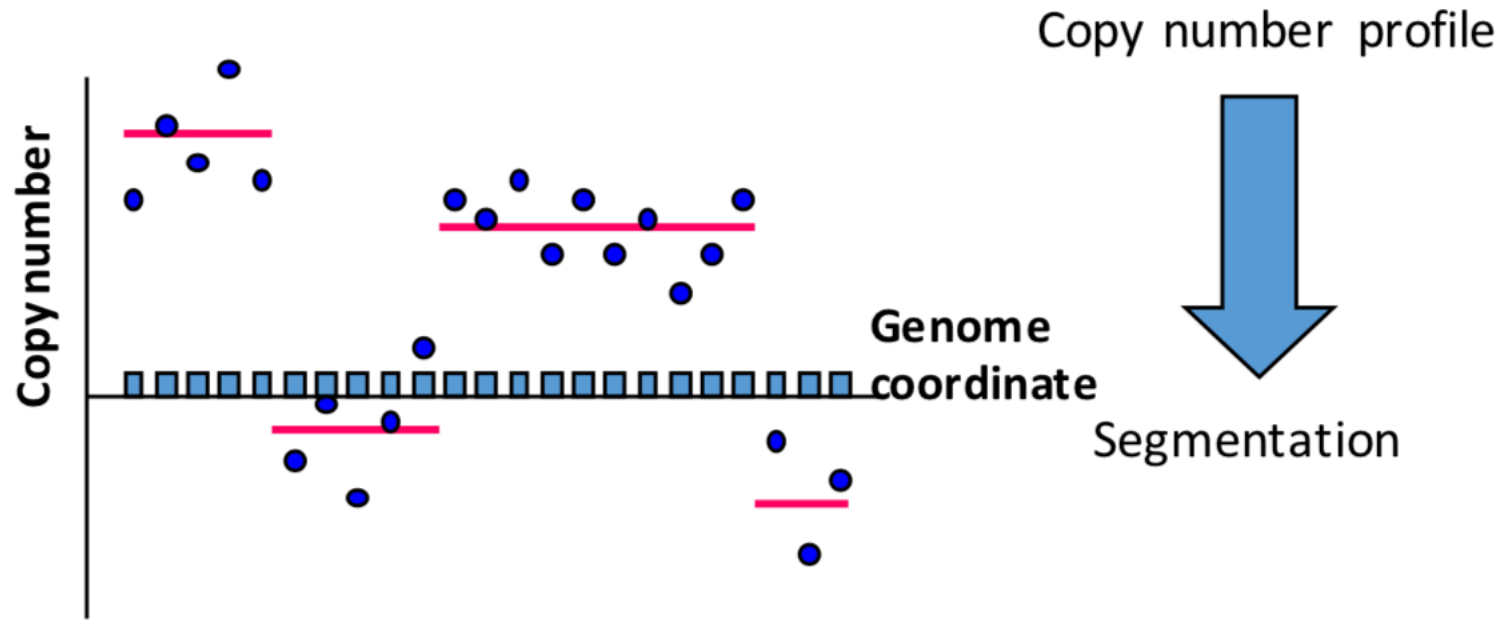
Segmentation and Copy Number Calling

Divide genome into segments of equal copy number



Segmentation and Copy Number Calling

Divide genome into segments of equal copy number



Input: $X_i = \log_2 T_i / R_i$, clone $i = 1, \dots, N$

Output: Assignment $s(i) \in \{S_1, \dots, S_K\}$ where S_i represent *copy number states*

Summary

- Markov property – Current state depends only on previous state
- Hidden Markov Models: states are not given only emitted symbols
- Viterbi algorithm: Find the most likely sequence of states given a set of observations

Reading:

- Jones and Pevzner: Chapters 11.1-11.3
- Lecture notes