CS 466
Introduction to Bioinformatics
Lecture 21

Mohammed El-Kebir

Nov 6, 2020

Outline

- Hidden Markov Models: Viterbi algorithm

Reading:

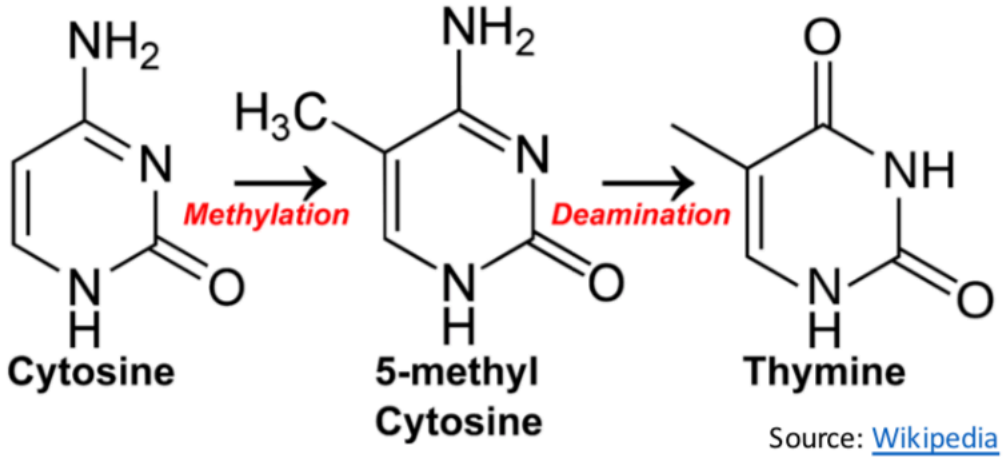
- Jones and Pevzner: Chapters 11.1-11.3
- Lecture notes

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

$CG \rightarrow C^*G \rightarrow TG$

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.

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG consitutes the start codon.
Right: CpG sites present at every 1/100 nucleotides, consituting a more normal example of the genome, or a region of the genome that is commonly methylated.

Source: Wikipedia

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.

Left: CpG sites at 1/10 nucleotides, constituting a CpG island. The sample is of a gene-promoter, the highlighted ATG consitutes the start codon.
Right: CpG sites present at every 1/100 nucleotides, consituting a more normal example of the genome, or a region of the genome that is commonly methylated.

Source: Wikipedia

Input: DNA sequence $\mathbf{x} = x_1 x_2 \ldots x_n$

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

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

Question: How do we identify CpG islands?

A Related Problem: Fair Bet Casino

- Game is to flip coins, two outcomes:

Head or **T**ail

- Two coins: **F**air and **B**iased

$$\Pr(H \mid F) = \Pr(T \mid F) = 1/2$$
$$\Pr(H \mid B) = 3/4, \Pr(T \mid B) = 1/4$$

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

CpG Islands and Fair Bet Casino

CG Islands

Input: DNA sequence $\mathbf{x} = x_1 x_2 \ldots x_n$ where $x_i \in \{A, T, C, G\}$

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

$\mathbf{x} =$ CGAT TG CGA AAAAAAT AACGA TTATATCG

$\pi =$ yes no yes no yes no

Fair Bet Casino

Input: Coin flips $\mathbf{x} = x_1 x_2 \ldots x_n$ where $x_i \in \{H, T\}$

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

$\mathbf{x} =$

$\pi =$ 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: $\boldsymbol{\pi}$ or $\boldsymbol{\pi}'$?

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

- Set of states Q
  - Markov property:

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

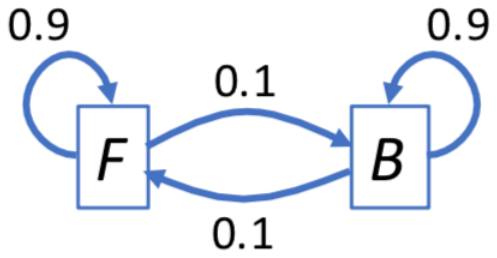
- Transition probabilities $A = [a_{ij}]$ on pairs of states
  - Rows sum to 1

Andrey Markov (source: Wikipedia)

Fair Bet Casino

$Q = \{F, B\}$

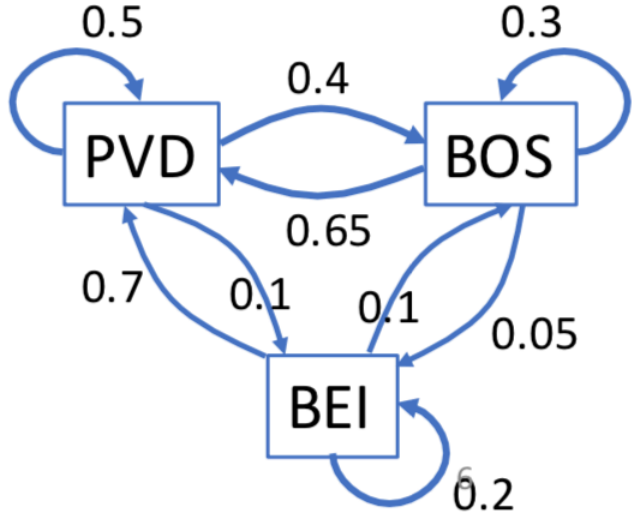
$$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$



Where is the professor?

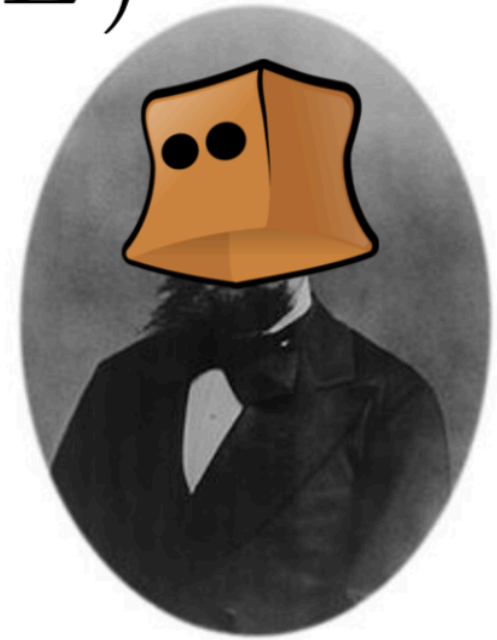
$Q = \{\text{Providence}, \text{Boston}, \text{Beijing}\}$

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



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]

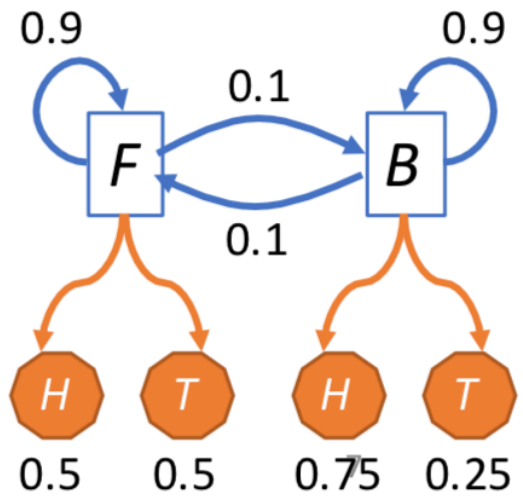
Fair Bet Casino

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

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

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

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



Three Questions

Question 1:
What is the most probable path $\boldsymbol{\pi}^*$ that generated observations $\mathbf{x}$?

Question 2:
What is probability of observations $\mathbf{x}$ generated by any path $\boldsymbol{\pi}$?

Question 3:
What is the probability of observation x_i generated by state s?

Three Questions

Question 1:
What is the most probable path $\boldsymbol{\pi}^*$ that generated observations $\mathbf{x}$?

Question 2:
What is probability of observations $\mathbf{x}$ generated by any path $\boldsymbol{\pi}$?

Question 3:
What is the probability of observation x_i generated by state s?

Joint Probability

$Pr(\bar{x}, \bar{\pi})$ $n = |\bar{x}| = |\bar{\pi}|$

Q, A, Σ, ϵ

$Q \times Q$ $\quad Q \times S$

$Pr(A, B) = Pr(A|B) Pr(B)$

$$Pr(\bar{x}, \bar{\pi}) = Pr(x_1, \pi_1, \cdots, x_n, \pi_n)$$

$$= Pr(x_n, \pi_n \mid x_{n-1}, \pi_{n-1}, \ldots, x_1, \pi_1) \cdots Pr(x_2 \pi_2 \mid x_1 \pi_1) Pr(x_1, \pi_1)$$

$$= Pr(x_n, \pi_n \mid \pi_{n-1}) \cdots Pr(x_2 \pi_2 \mid \pi_1) Pr(x_1 \mid \pi_1) Pr(\pi_1)$$

$$= \prod_{i=1}^{n} e_{\pi_i, x_i} \, a_{\pi_{i-1}, \pi_i}$$

In practice $\bar{\pi}$ is hidden to us.

Goal: find $\bar{\pi}$ maximizing $Pr(\bar{x}, \bar{\pi})$

Joint Probability

We are interested in identifying $\bar{\pi}^* \in Q^n$ with max prob $Pr(\bar{x}, \bar{\pi}^*)$.

$$\prod_{i=1}^{n} e_{\pi_i, x_i} \, a_{\pi_{i-1}, \pi_i}$$

$e_{\pi_1, x_1} \; a_{\pi_0, \pi_1}$

$\bar{\pi} = \pi_1, \ldots, \pi_n$

H	H		T
x_1	x_2	- - - -	x_n
π_1^*	π_2^*	- - - -	π_n^*
F	F		F
B	B		B

optimal substructure

Joint Probability

$\bar{x}_i = x_1, \cdots, x_i$ $\quad\quad$ $\bar{x} = \bar{x}_n$

$\bar{\pi}_i = \pi_1, \cdots, \pi_i$ $\quad\quad$ $\bar{\pi} = \bar{\pi}_n$

$v[s, i]$ denote the prob. $Pr(\bar{x}_i, \bar{\pi}^*_{i-1}, \pi_i = s)$

prob. of the most probable state path $\bar{\pi}^*_i$ of the first i observations $\bar{x}_i$ with final state $\pi_i = s$

$$Pr(\bar{x}, \bar{\pi}^*) = \max_{s \in Q} Pr(\bar{x}_n, \bar{\pi}_{n-1}, \pi_n = s) = \max_{s \in Q} v[s, n]$$

Joint Probability

$x_1 = T$ $\quad V[F,1]$ $\quad a_{0,F}$ $e_{F,T}$ $\quad 0.5 \cdot 0.5$

Base case: $i = 1$

$$V[s,1] = Pr(x_1, \pi_0^*, \pi_1 = s)$$

$$= a_{0,s} \, e_{s,x_1}$$

$i > 1$

$$V[s,i] = Pr(x_i, \pi_{i-1}^*, \pi_i = s)$$

$Pr(A,B) = Pr(A|B)\,Pr(B)$

$$= Pr(\underbrace{x_1, \ldots, x_{i-1}}_{B}, \underbrace{x_i}_{A}, \underbrace{\pi_1^*, \ldots, \pi_{i-1}^*}_{B}, \underbrace{\pi_i = s}_{A})$$

$$= Pr(x_i, \pi_i = s \mid \underbrace{x_1, \ldots, x_{i-1}, \pi_1^*, \ldots, \pi_{i-1}^*}_{B}) \, Pr(B)$$

$$= Pr(x_i, \pi_i = s \mid \pi_{i-1}^*) \; Pr(B)$$

Recurrence

$v[s,i]$ $(i > 1)$

$$= \Pr(x_i, \pi_i = s \mid \pi_{i-2}^*) \Pr(x_1, \dots, x_{i-1}, \pi_1^*, \dots \pi_{i-1}^*)$$

$$= \max_{t \in Q} \Pr(x_i, \pi_i = s \mid \pi_{i-1} = t) \Pr(\bar{x}_{i-1}, \bar{\pi}_{i-2}^*, \pi_{i-1} = t)$$

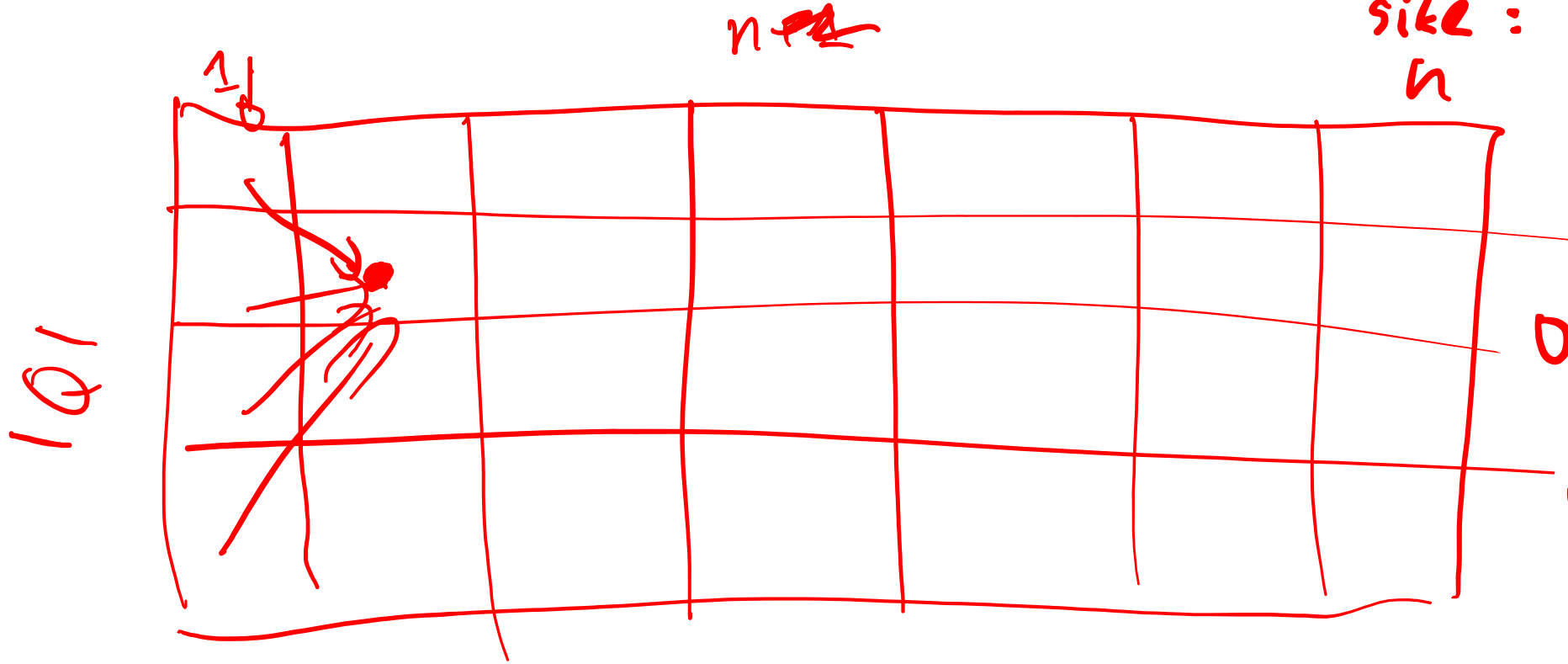
cond. independence

$$= \max_{t \in Q} \Pr(x_i \mid \pi_i = s) \Pr(\pi_i = s \mid \pi_{i-1} = t) \; v[t, i-1]$$

$$= \max_{t \in Q} e_{s,x_i} \cdot a_{t,s} \cdot v[t, i-1]$$

$$= e_{s,x_i} \cdot \max_{t \in Q} a_{t,s} \; v[t, i-1]$$

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

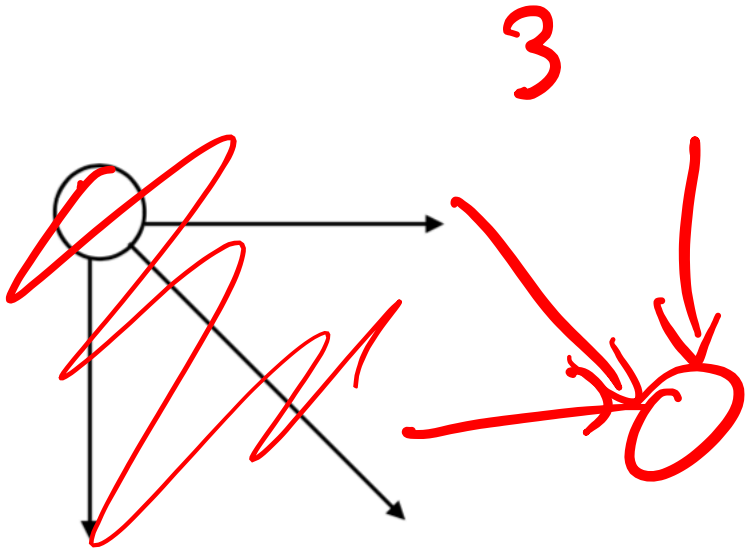


size : $|Q|\, n$

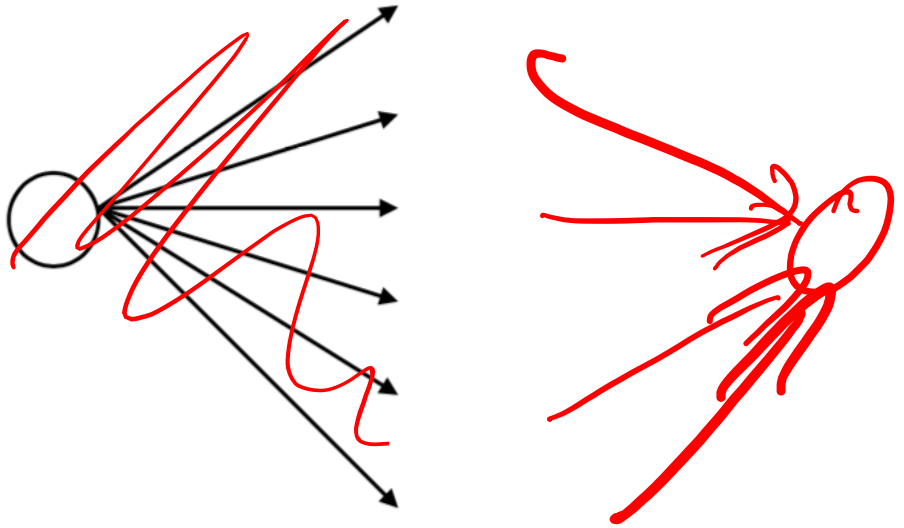
$O(|Q|^2\, n)$

Viterbi alg.

Alignment vs. Decoding Problem



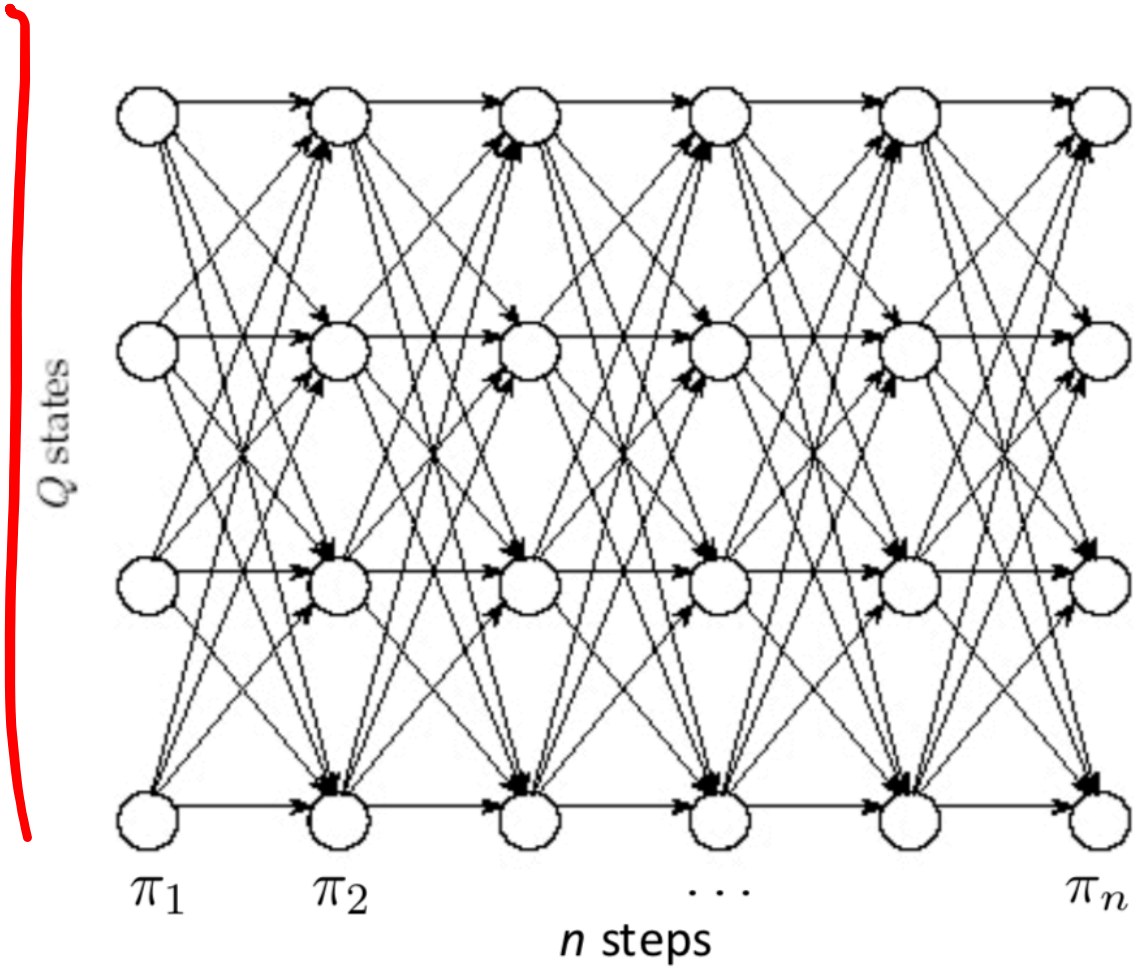
Valid directions in the *alignment problem.*



Valid directions in the *decoding problem.*

Viterbi Algorithm

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



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

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

$[0, 1]$

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

$\mathbf{x} =$ 

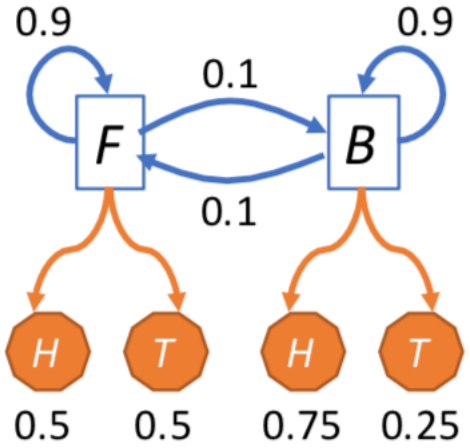
	0	1 (T)	2 (H)	3 (T)	4 (H)	5 (H)	6 (H)	7 (T)	8 (H)	9 (H)	10 (H)	11 (T)
F												
B												

$Q = \{F, B\}$

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

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

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



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

$$\Pr(\bar{x}) = \sum_{\bar{\pi}} \Pr(\bar{x}, \bar{\pi})$$

$$= \sum_{(\pi_1, \cdots, \pi_n)} \Pr(x_1, \pi_1, \cdots, x_n, \pi_n) = \sum_{(\pi_2, \cdots, \pi_n)} \prod_{i=1}^{n} e_{\pi_i, x_i} a_{\pi_{i-1}, \pi_i}$$

$|(Q, A, \Sigma, E)$

$\cdot \quad \cdot \quad \cdot \quad \cdots \quad n$

$O(n\,|Q|^n)$

Goal: $O(n\,|Q|^2)$

$$P_\theta(\bar{x}) = \sum_{\bar{\pi}} P_\theta(\bar{x}, \bar{\pi}) \geq P_\theta(\bar{x}, \bar{\pi}^*)$$

$$\bar{\pi}^* = \arg\max_{\bar{\pi}} P_\theta(\bar{x}, \bar{\pi})$$

$\bar{x}_i = x_1, \dots, x_i$

$\bar{\pi}_i = \pi_1, \dots, \pi_i$

$$Pr(\bar{x}_i, \pi_i = s) = Pr(x_1, \dots, x_i, \pi_i = s) = \boxed{F[s, i]}$$

$$= \sum_{\bar{\pi}_{i-1}} Pr(\bar{x}_i \mid \bar{\pi}_{i-1}, \pi_i = s)$$

$$Pr(\bar{\pi}) = Pr(x_1, \dots, x_n) = \sum_{s \in Q} Pr(x_1, \dots, x_n, \pi_n = s)$$

$$= \sum_{s \in Q} F[s, n].$$

Base $i=1$

$$F[s,1] = Pr(x_1, \pi_1 = s) = a_{0,s} \cdot e_{s,x_1}$$

Step $i > 1$

$$F[s,i] = Pr(\bar{x}_i, \pi_i = s) = Pr(x_1, \dots, x_i, \pi_i = s)$$

$$= \sum_{(\pi_1, \dots, \pi_{i-1})} Pr(x_1, \dots, x_{i-1}, x_i, \pi_1, \dots, \pi_{i-1}, \pi_i = s)$$

$$= \sum_{\bar{\pi}_{i-1}} Pr(x_1, \dots, x_{i-1}, \pi_1, \dots, \pi_{i-1}, \pi_i = s) \cdot e_{s,x_i}$$

$$\sum_{\pi_{i-1}} \Pr(x_1, \dots, x_{i-1}, \pi_1, \dots, \pi_{i-1}, \pi_i = s)\, a_{s,x_i}$$

$$= \sum_{t \in Q} \left[\frac{\sum_{(\pi_1, \dots, \pi_{i-2})} \Pr(x_1, \dots, x_{i-1}, \pi_1, \dots, \pi_{i-2}, \pi_{i-1} = t)}{\Pr(x_1, \dots, x_{i-1}, \pi_{i-1} = t)} = F[t, i-1] \right] a_{t,s}\, e_{s,x_i}$$

$$= e_{s,x_i} \sum_{t \in Q} F[t, i-1]\, a_{t,s}$$

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

$$Pr(\pi_i = s \mid \bar{x}) = \frac{Pr(\bar{x}, \pi_i = s)}{Pr(\bar{x})}$$

$$Pr(A \mid B) = \frac{Pr(A, B)}{Pr(B)}$$

$$Pr(A, B) = Pr(A \mid B) Pr(B)$$

we know how to compute $Pr(\bar{x}) = \sum_{s \in Q} F[s, n]$

(forward alg.)

$$Pr(\bar{x}, \pi_i = s) = Pr(\overbrace{x_1, \dots, x_i}^{B}, \overbrace{x_{i+1}, \dots, x_n}^{A}, \overbrace{\pi_i = s}^{B})$$

$$= Pr(x_1, \dots, x_i, \pi_i = s) \; Pr(x_{i+1}, \dots, x_n \mid x_1, \dots, x_i, \pi_i = s)$$

$$= F[s, i] \cdot \underbrace{Pr(x_{i+1}, \dots, x_n \mid \pi_i = s)}_{b[s, i]}$$

$$b[s,i] = Pr(x_{i+1}, \dots, x_n \mid \pi_i = s)$$

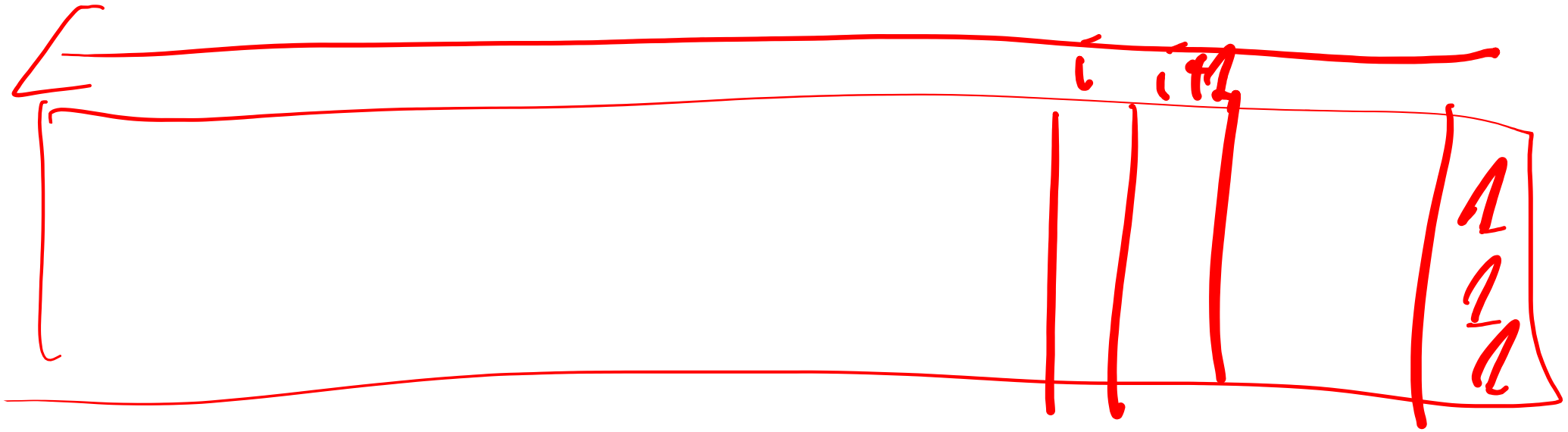
$i = n$: $b[s,n] = Pr(x_{n+1}, \dots, x_n \mid \pi_n = s)$

Ω = sample space = $Pr(\Omega \mid \pi_n = s)$

$$\boxed{b[s,n] = 1 \quad \forall s \in Q} = \frac{Pr(\Omega, \pi_n = s)}{Pr(\pi_n = s)} = \frac{Pr(\pi_n = s)}{Pr(\pi_n = s)} = 1$$

Step: $i > 1$.

$$b[s, i] = \sum_{t \in Q} a_{s,t} \cdot e_{t, x_{i+1}} \cdot b[t, i+1]$$



$$b[s,i] = \begin{cases} 1, & \text{if } i = n, \\ \sum_{t \in Q} a_{s,t}\, e_{t,x_{i+1}}\, b[t,i+1], & \text{if } 1 \le i < n. \end{cases}$$

$$Pr(\pi_i = s \mid \bar{x}) = \frac{Pr(\bar{x}, \pi_i = s)}{Pr(\bar{x})} = \frac{f[s,i] \cdot b[s,i]}{\sum_{s \in Q} f[s,n]}$$

(1) Run forward to compute f $O(|Q|^2 n)$

(2) Run backward to compute b $O(|Q|^2 n)$

Posterior decoding

$\bar{\pi}^* = \bar{\pi}^{\wedge}$

$\hat{\pi}_1 \rightarrow \hat{\pi}_2 \quad \ldots \quad \hat{\pi}_n$

$$\hat{\pi}_i = \arg\max_{s \in Q} \Pr(\pi_i = s \mid \bar{x})$$

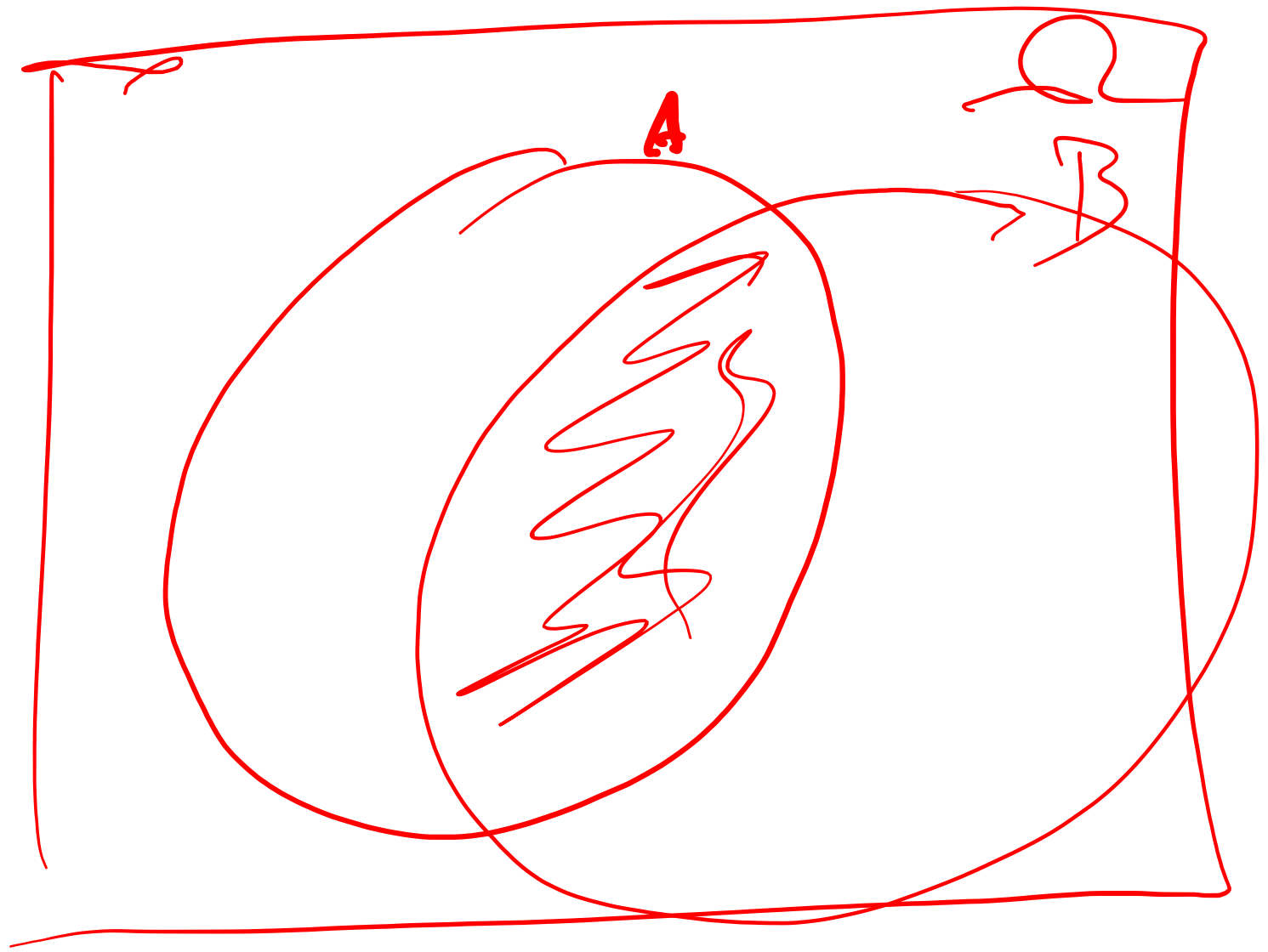
Viterbi decoding: $\bar{\pi}^* = \arg\max_{(\pi_1, \ldots, \pi_n)} \Pr(\bar{x}, \bar{\pi})$

$\pi_1^*, \ldots, \pi_n^*$

$Pr(A, B)$

$Pr(A)$

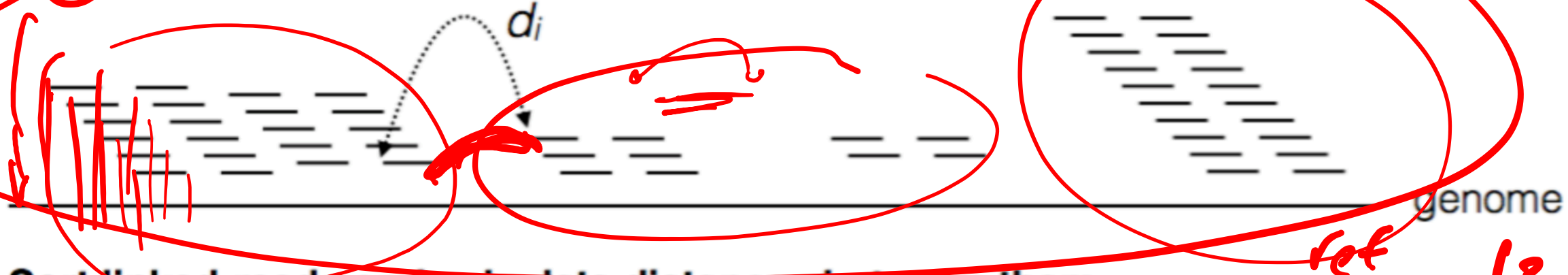
$Pr[\Omega, A \cap B]$



10X Genomics: Synthetic Long Reads

Genome indexing by partitioning and molecular barcoding

R_i={r_i | ∀ i, r_i contains barcode j}: Paired-reads possessing barcode j



Sort linked-reads and calculate distances between them

D_j =[d_i | ∀ i, d_i : distance between r_i and r_{i+1}, $r_i < r_{i+1}$]

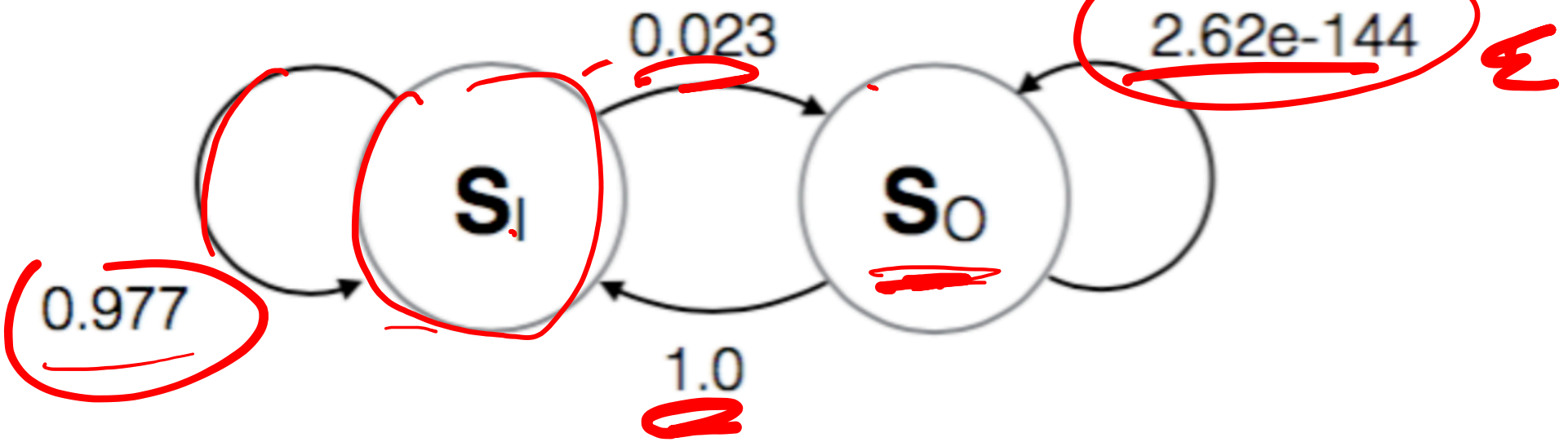
Define distances d_i as intra- or inter- long molecules

Σ_i =[d_1, d_2, d_3, … d_{500}, d_{501}, d_{502}, … d_{1001}, d_{1002}, d_{1003}, …]

Q_j =[I, I, I, …, I, O, I, …, I, O, I, ….]

I: intra long molecule, O: inter long molecules

HMM



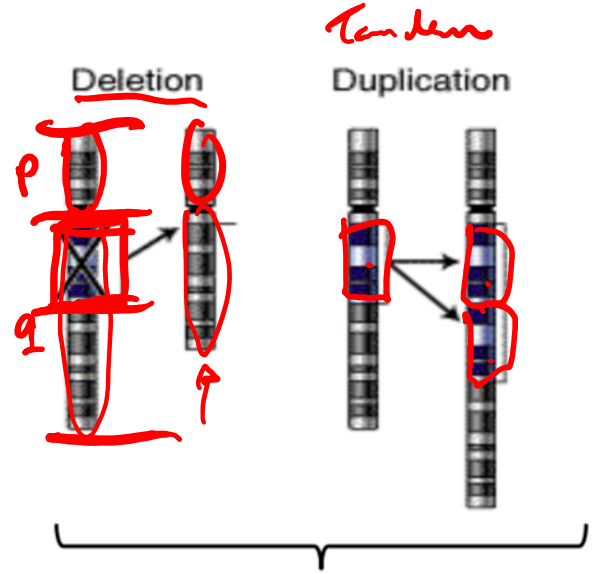
Σ_i =[d_1, d_2, d_3, … d_{500}, d_{501}, d_{502}, … d_{1001}, d_{1002}, d_{1003}, …]

Q_j =[I, I, I, …, I, O, I, …, I, O, I, ….]

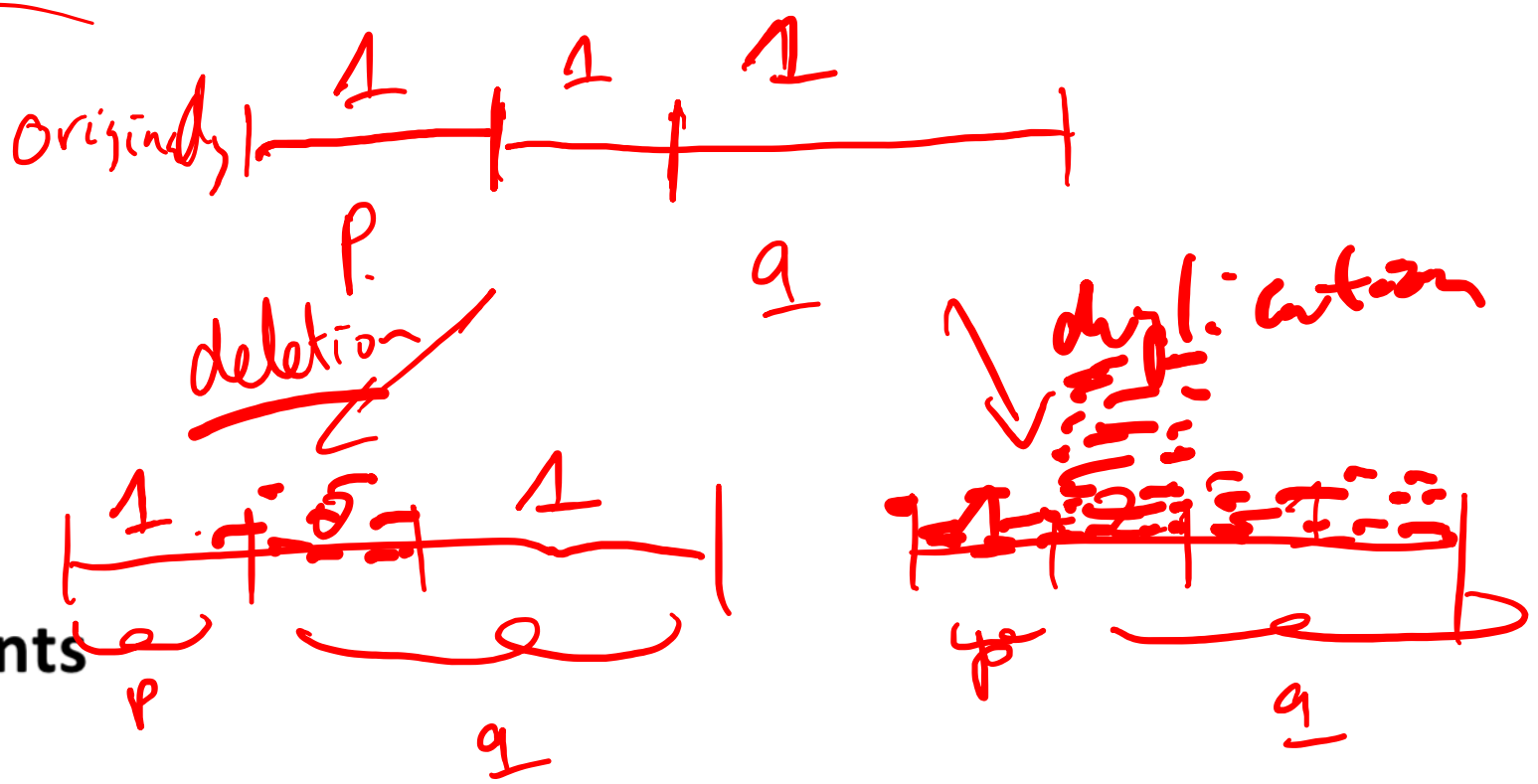
ℓ_{j1} ℓ_{j2} ℓ_{j3}

I: intra long molecule, O: inter long molecules

Copy Number Variation



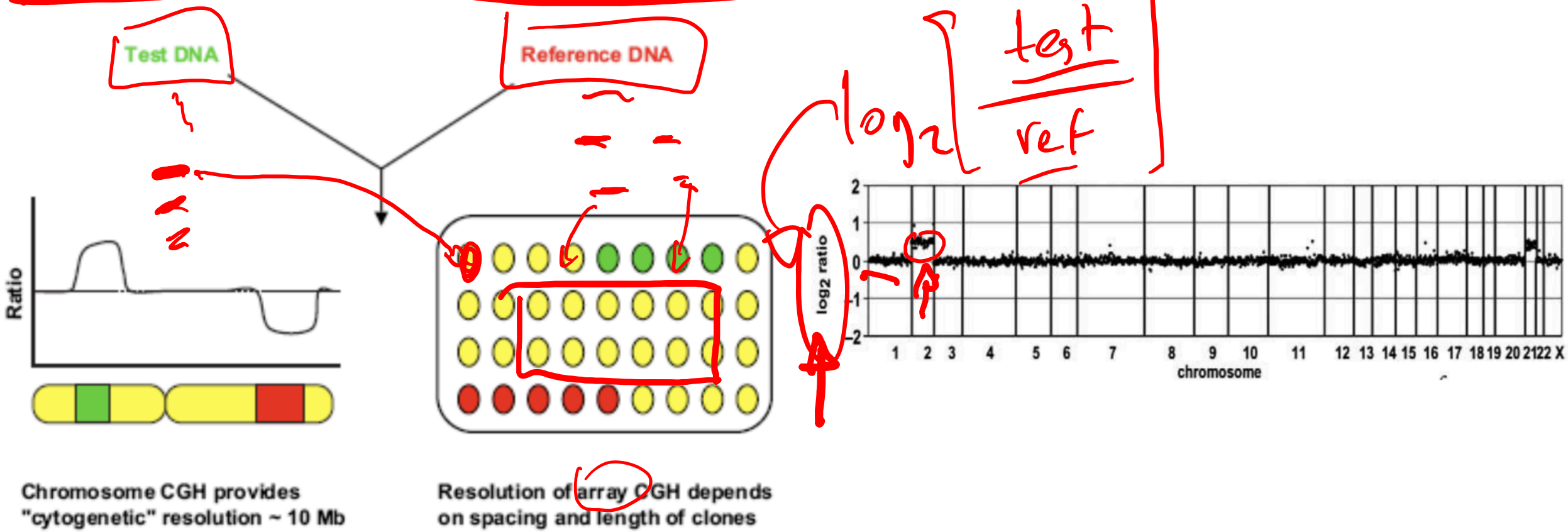
Copy Number Variants



- 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

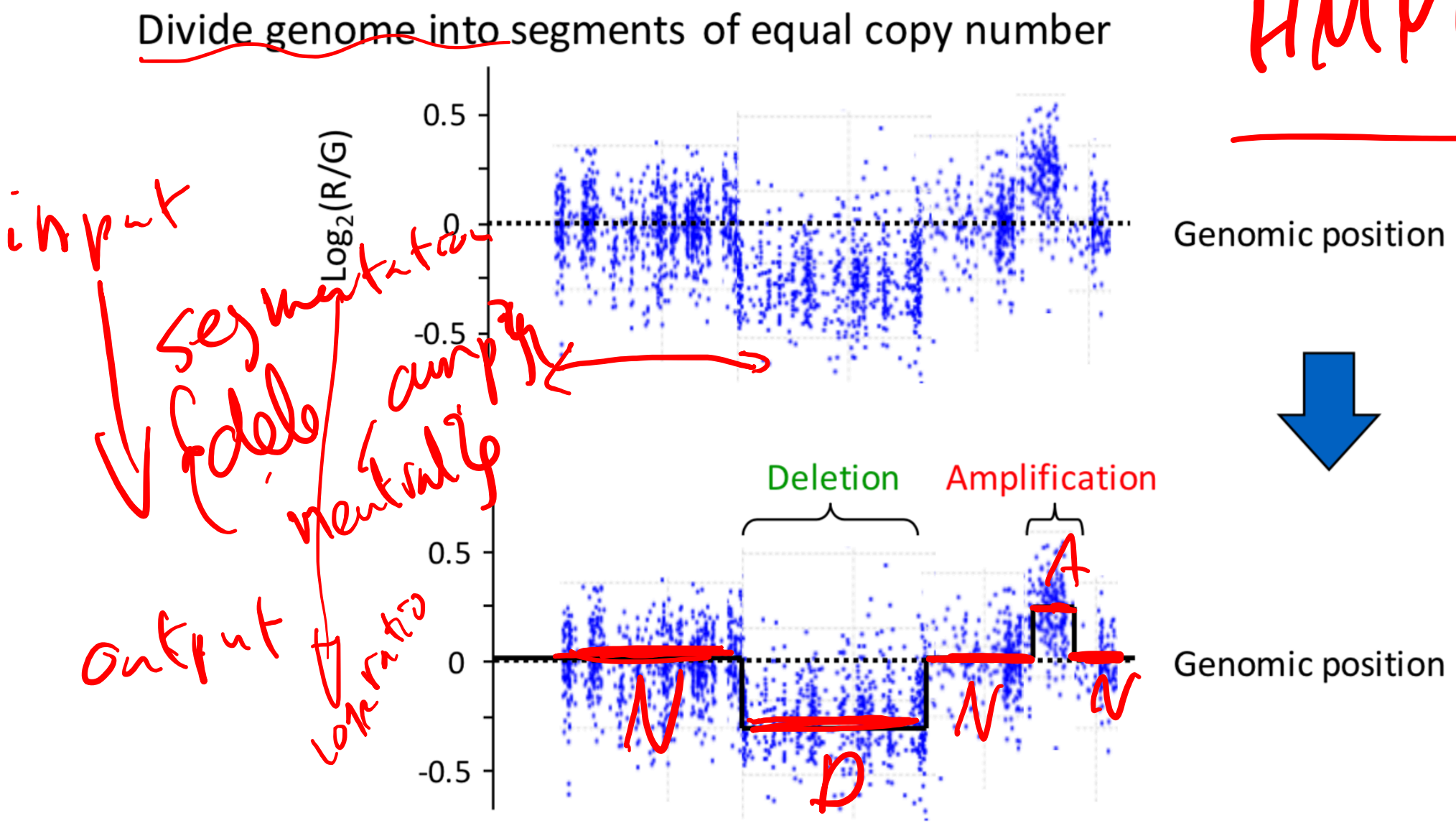
Comparative Genomic Hybridization (CGH)



Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb

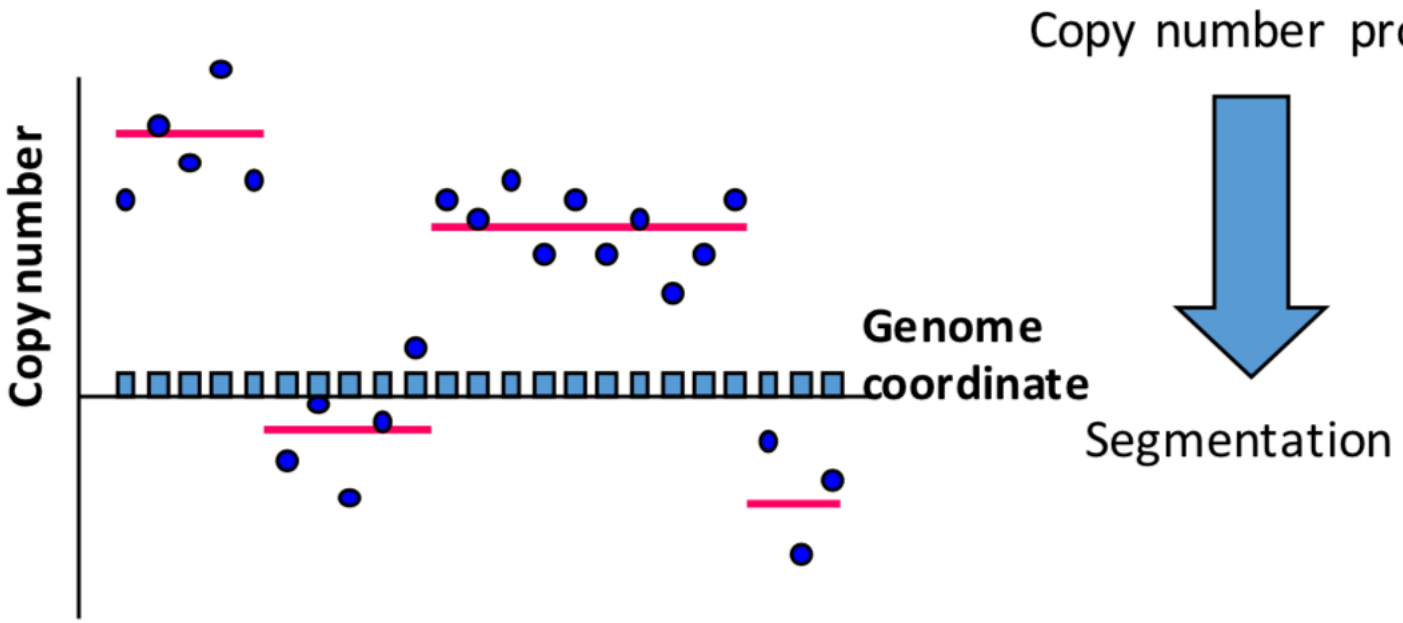
Resolution of array CGH depends on spacing and length of clones

Segmentation and Copy Number Calling



Segmentation and Copy Number Calling

Divide genome into segments of equal copy number



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

Output: Assignment $s(i) \in \{S_1, \ldots, S_K\}$ 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