CS 466
Introduction to Bioinformatics
Lecture 18

Mohammed El-Kebir

Nov 7, 2018

Course Announcements

Project Proposal due Nov 14th

Outline

- Hidden Markov Models: Viterbi algorithm

Reading:

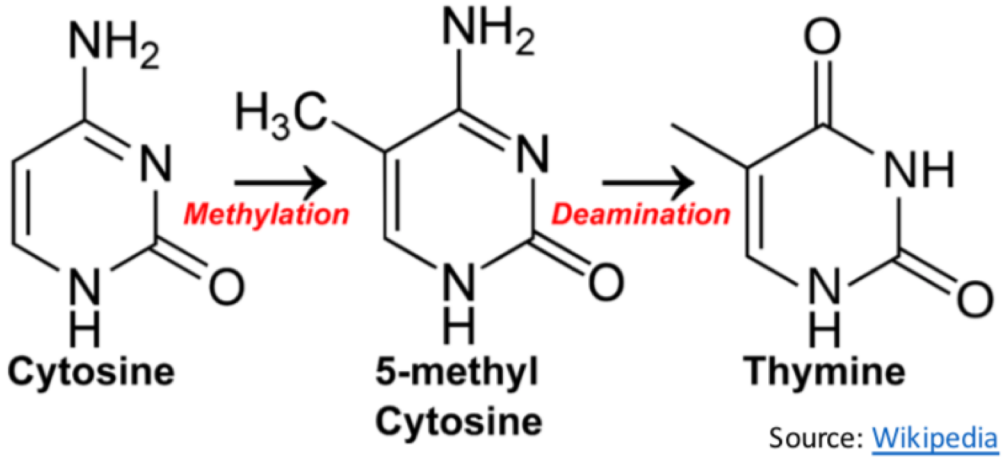
- Jones and Pevzner: Chapters 11.1-11.3
- Lecture notes

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?

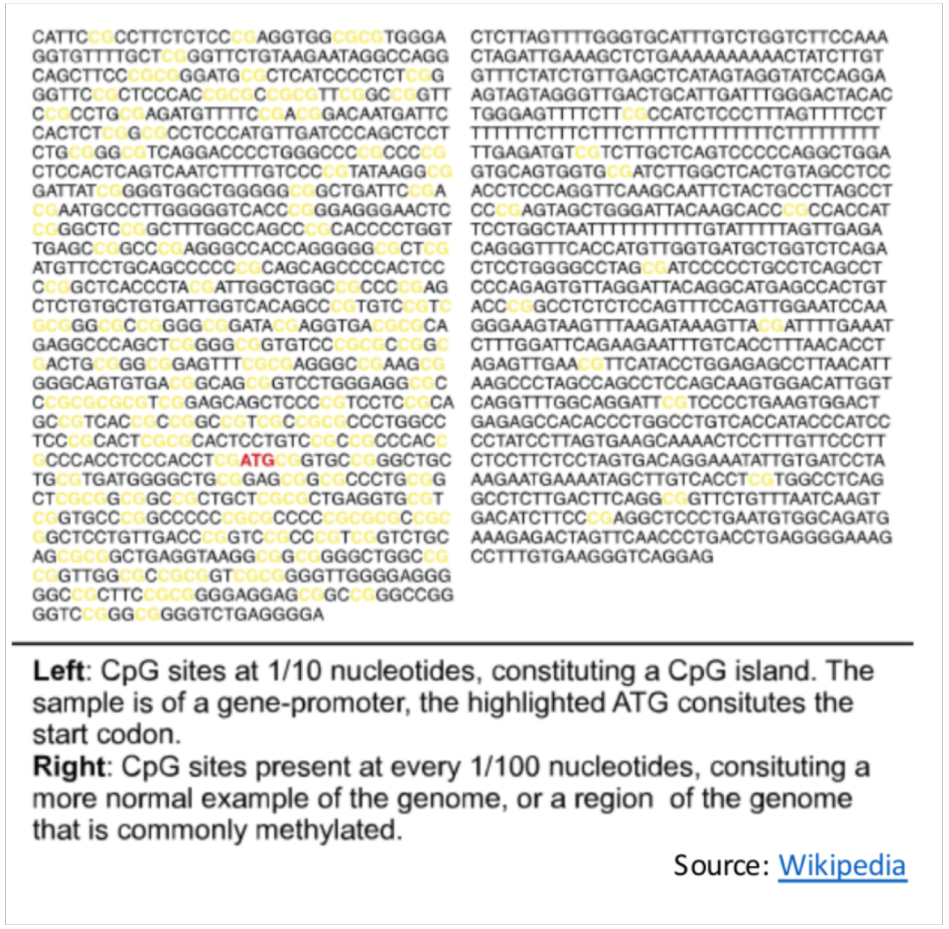


$CG \rightarrow C^*G \rightarrow TG$

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

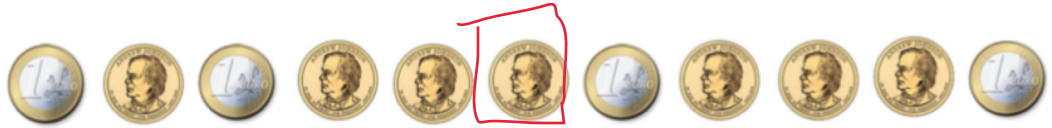
$\mathbf{x} =$ CGAT TG CGA AAAAAAT AACGA TTATATCG

$\pi =$ 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: $\pi : \{1, \ldots, n\} \to \{F, B\}$



$\mathbf{x} =$

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

states

- Set of states Q
  - Markov property:

The next state depends only on the current state

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

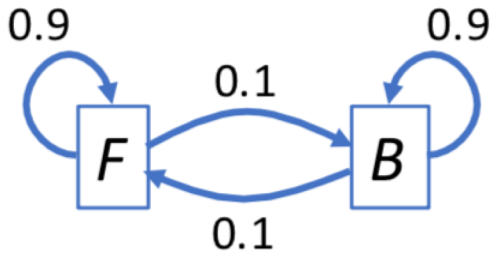
- Transition probabilities $A = [a_{ij}]$ on pairs of states
  - Rows sum to 1

Andrey Markov (source: Wikipedia)

Fair Bet Casino

$Q = \{F, B\}$

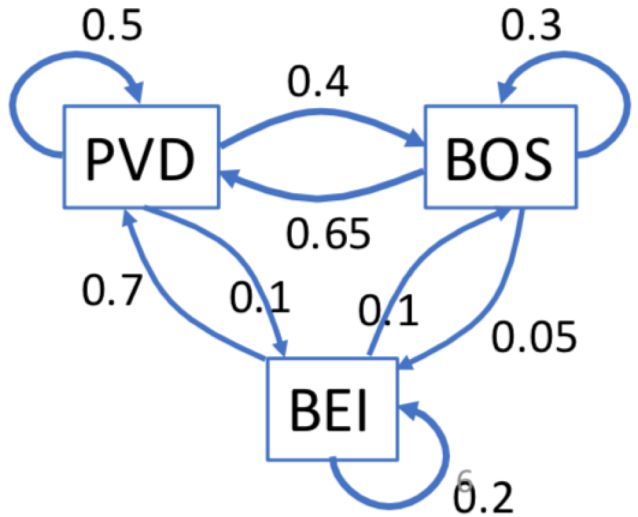
$$A = \begin{matrix} F \\ B \end{matrix} \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$$



Where is the professor?

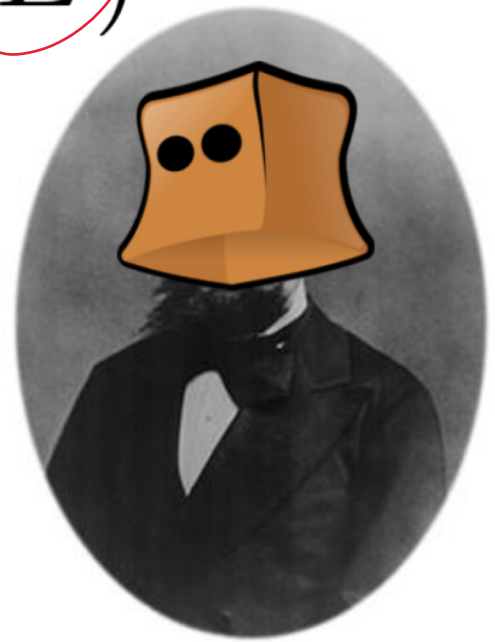
$Q = \{\text{Providence, Boston, Beijing}\}$

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



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]

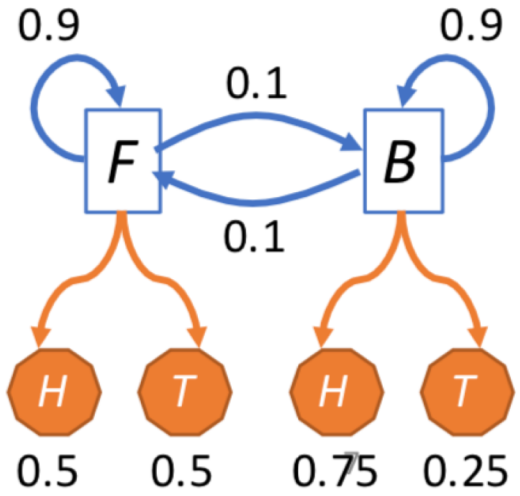
Fair Bet Casino

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

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

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

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



Three Questions

given $M = (Q, A, \Sigma, E)$ $|\vec{\pi}| = |\vec{x}|$

Observations: $x_1, \ldots, x_n \in \Sigma$

Joint probability

Question 1: $\pi^* = \arg\max_{\vec{\pi}} Pr(\vec{x}, \vec{\pi})$ $\vec{x} \in \Sigma^n$

What is the most probable path **π*** that generated observations **x**?

$\vec{\pi}^* \in Q^n$ Find: $\pi_1, \ldots, \pi_n \in Q$

Marginal prob.

Question 2: $Pr(\vec{x}) = \sum_{\vec{\pi}} Pr(\vec{x}, \vec{\pi})$

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

Posterior probability

Question 3: $Pr(\pi_i = s \mid \vec{x})$

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 $\mathbf{x}$?

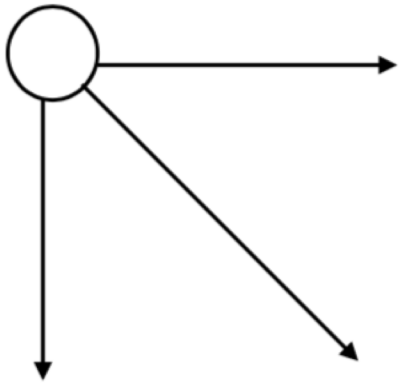
Question 2:
What is probability of observations $\mathbf{x}$ generated by any path $\boldsymbol{\pi}$?

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

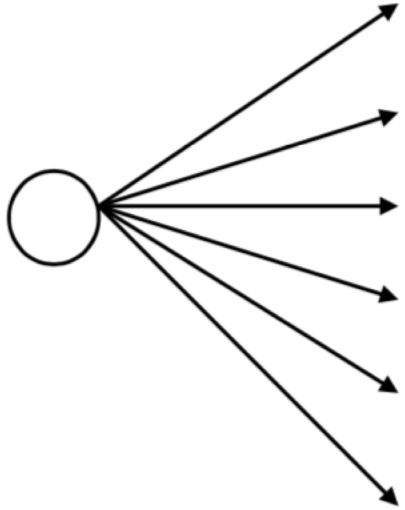
Joint Probability

Recurrence

Alignment vs. Decoding Problem



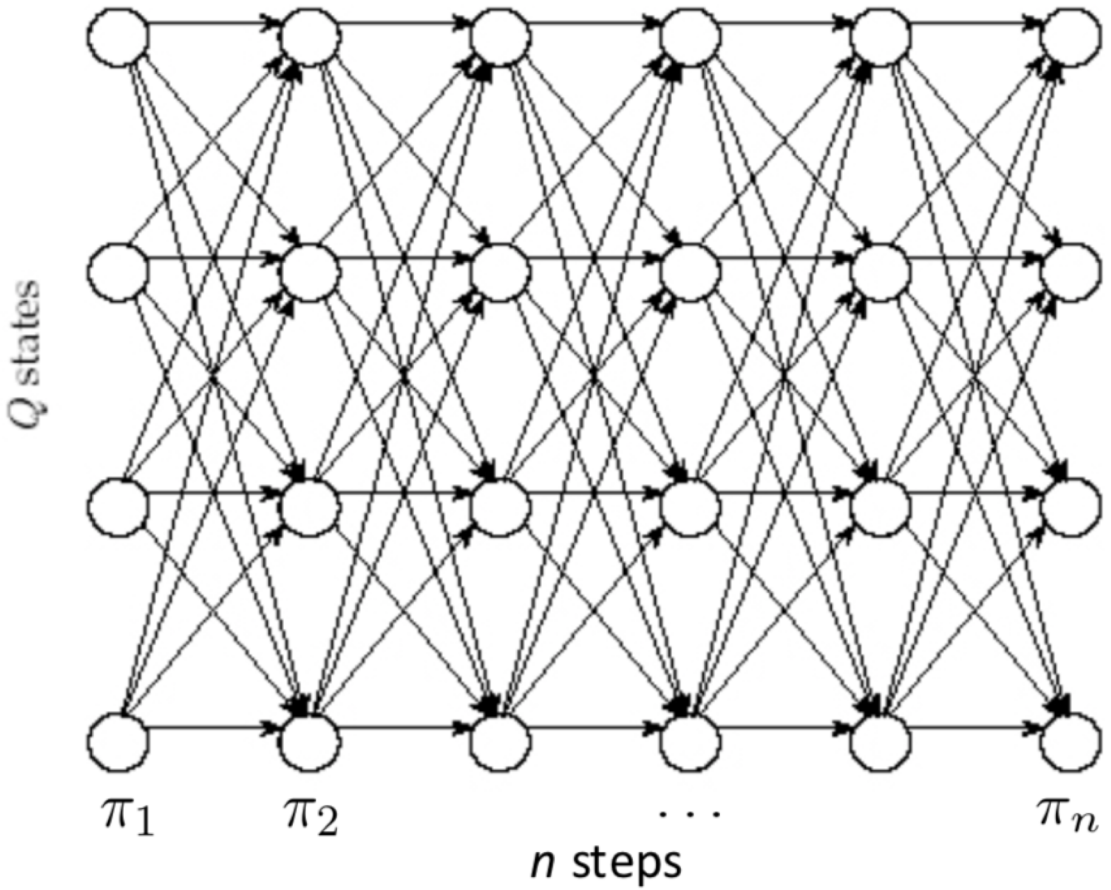
Valid directions in the *alignment problem.*



Valid directions in the *decoding problem.*

Viterbi Algorithm

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



Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

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

$\bar{x}_i = x_1, \ldots, x_i$
$\bar{\pi}_i = \pi_1, \ldots, \pi_i$

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

$v[s,1] = \Pr(x_1, \pi_1 = s) = a_{0,s}\, e_{s,x_1} \qquad \forall s \in Q$

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

Viterbi Algorithm – Numerical Issues

Value of products can become extremely small, leading to underflow

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

Use logarithms!

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

Fair Bet Casino: Example

$\mathbf{x} =$ 

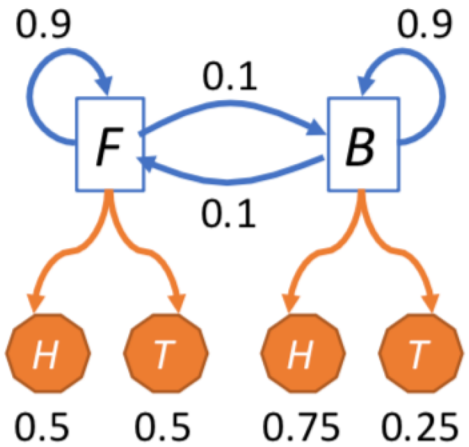
	0	1 (T)	2 (H)	3 (T)	4 (H)	5 (H)	6 (H)	7 (T)	8 (H)	9 (H)	10 (H)	11 (T)
F												
B												

$Q = \{F, B\}$

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

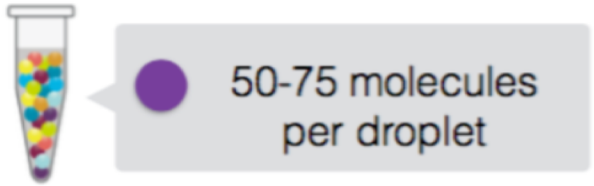
$\Sigma = \{H, T\}$

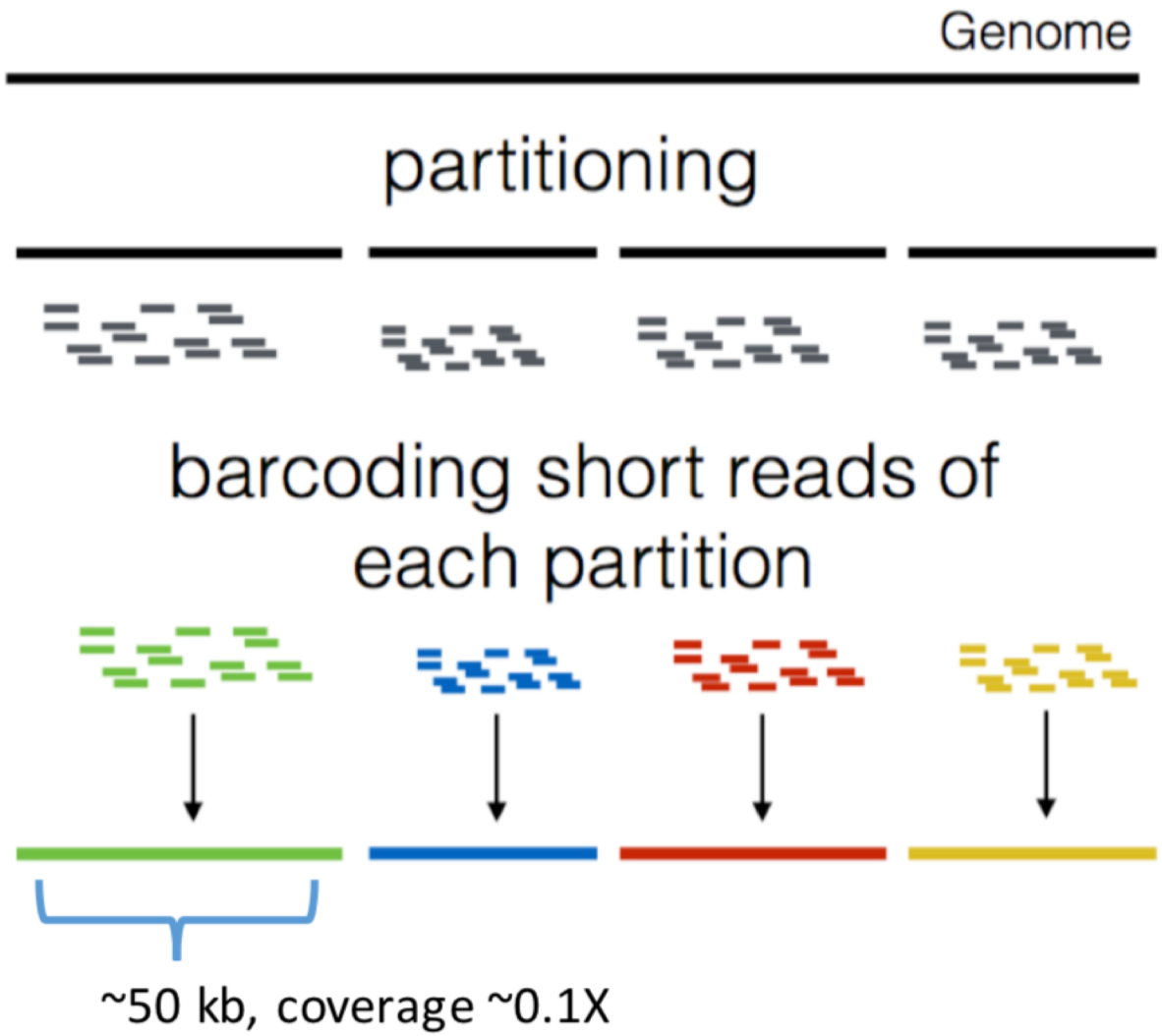
$$E = \begin{array}{c} \begin{matrix} H & T \end{matrix} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix} \end{array} \begin{matrix} F \\ B \end{matrix}$$



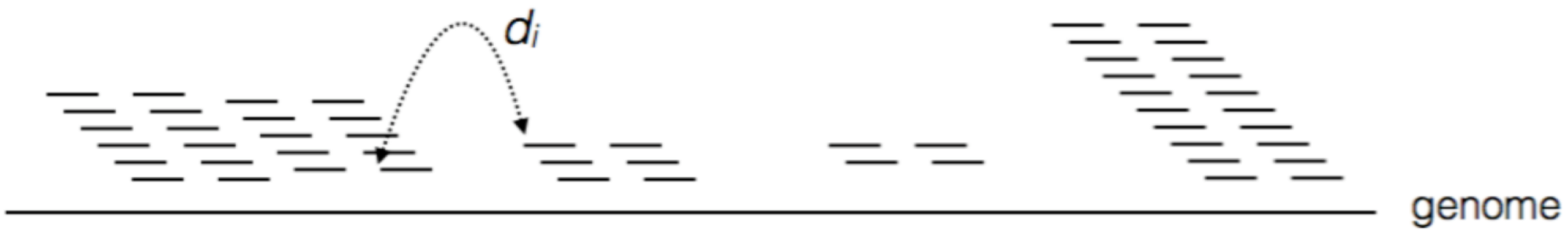
$$v[s,i] = \begin{cases} a_{0,s}, & \text{if } i = 0, \\ e_{s,x_i} \max_{t \in Q}\{v[t, i-1] \cdot a_{s,t}\}, & \text{if } i > 0. \end{cases}$$

10X Genomics: Synthetic Long Reads





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



Sort linked-reads and calculate distances between them

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

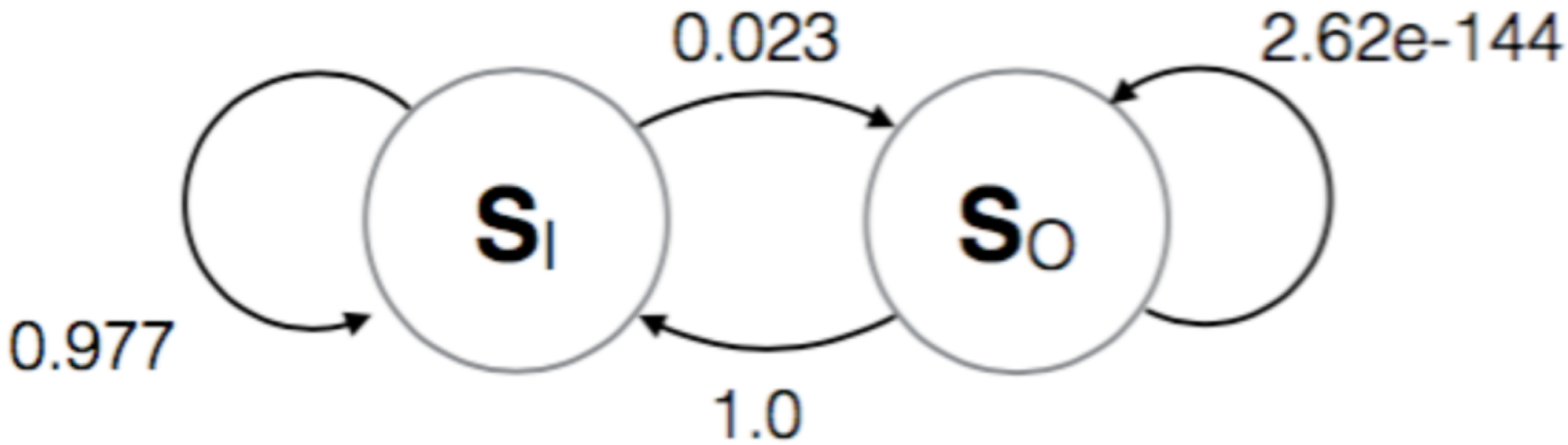
Define distances d_i as intra- or inter- long molecules

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

$Q_j = [\ I,\ \ I,\ \ I,\ \ \ldots,\ \ I,\ \ \ O,\ \ \ I, \ldots,\ \ \ I,\ \ \ O,\ \ \ I, \ldots.\ ]$

I: intra long molecule, O: inter long molecules

HMM



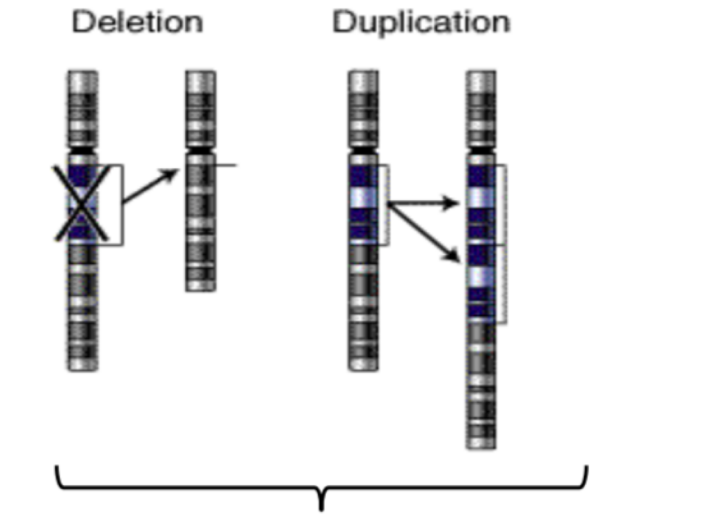
Σ_i =[d_1, d_2, d_3, … d_{500}, d_{501}, d_{502}, … d_{1001}, d_{1002}, d_{1003}, …]

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

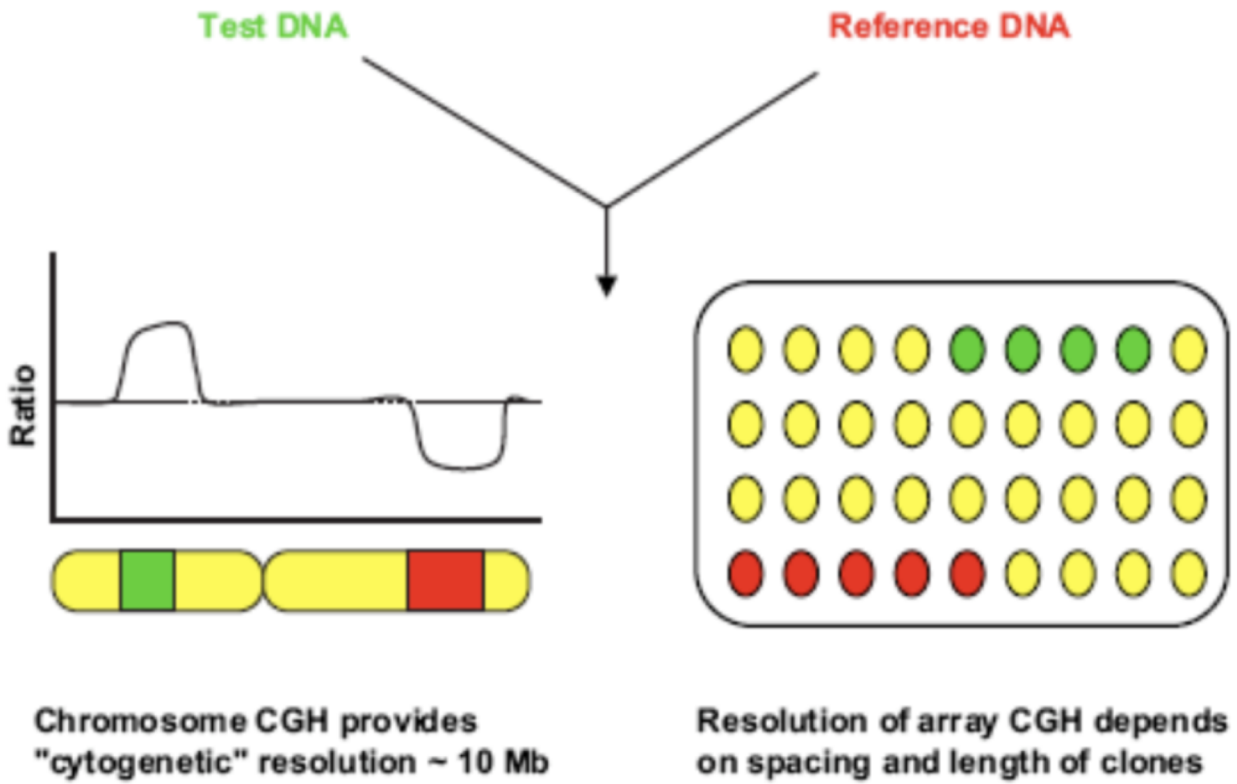
Copy Number Variation



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

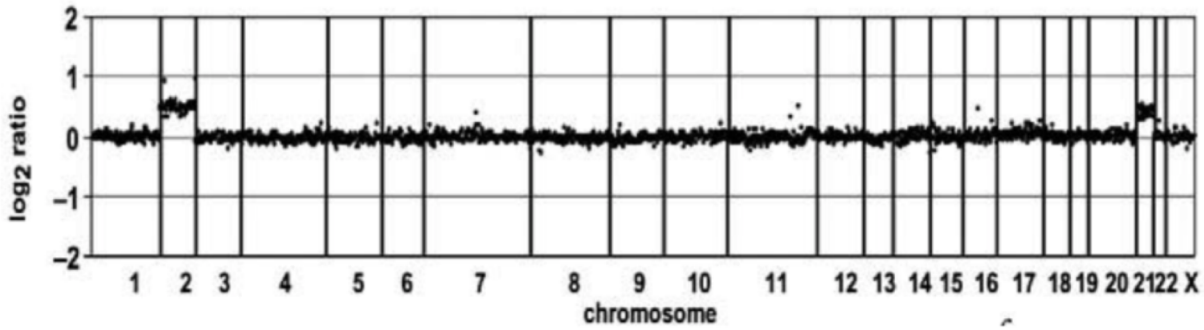
Measuring Copy Number Variants

Comparative Genomic Hybridization (CGH)

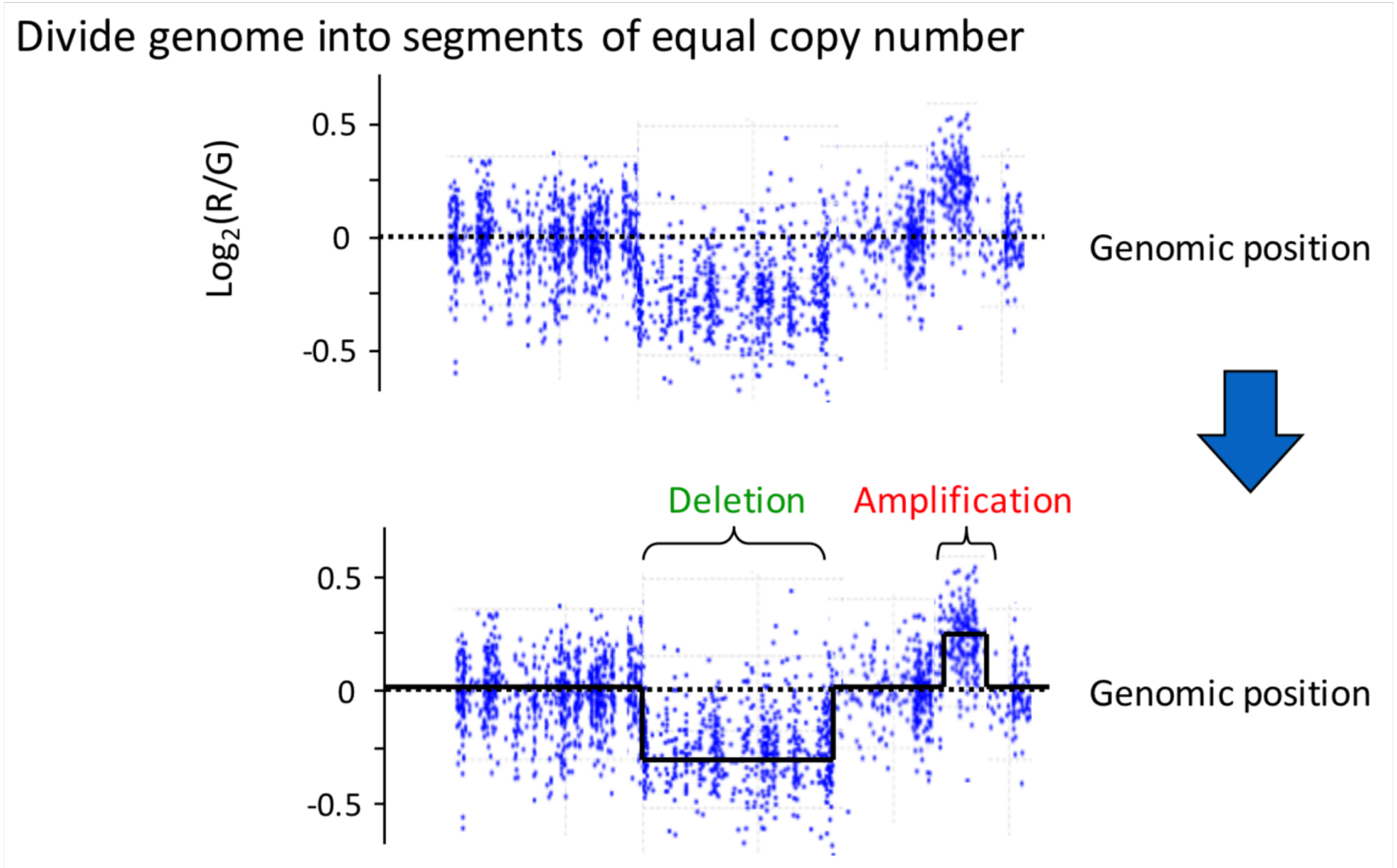


Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb

Resolution of array CGH depends on spacing and length of clones

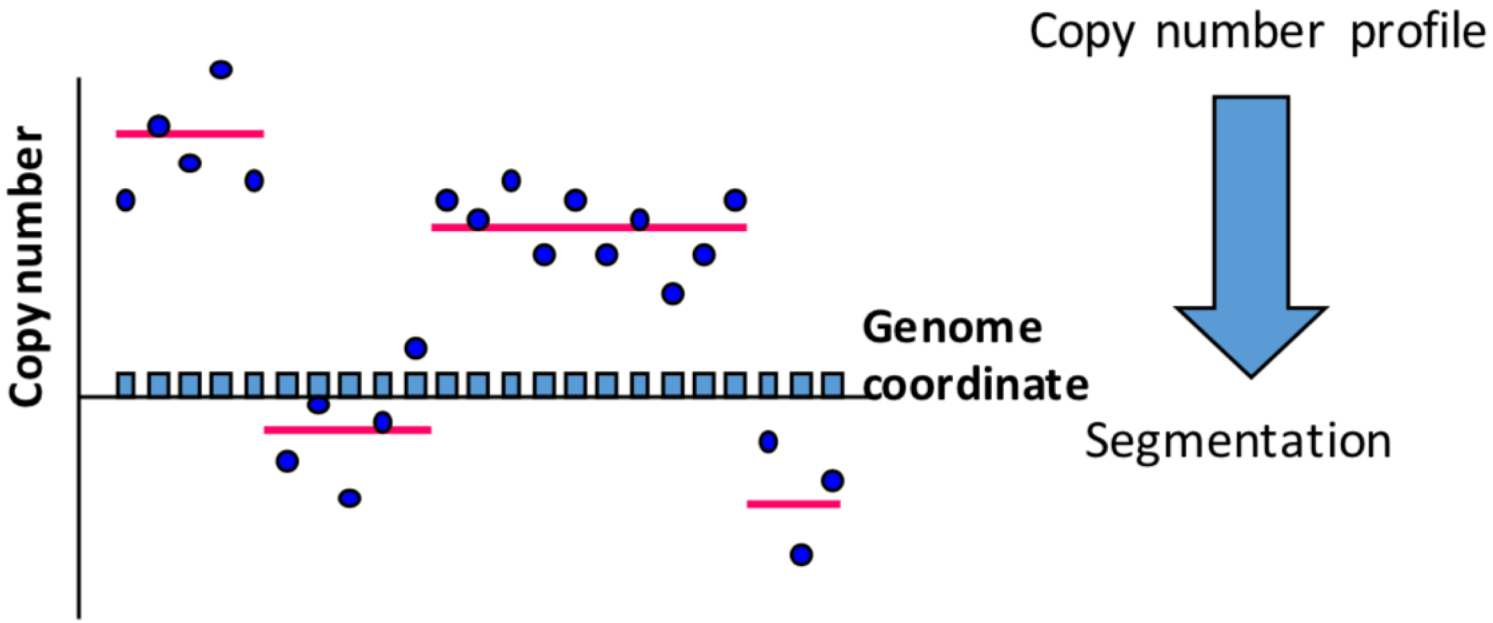


Segmentation and Copy Number Calling



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

Project Proposal due Nov 14th