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

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

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



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

- 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.

CATTCCGCCTTCTCTCCCGAGGTGGCGCGTGGGA CTCTTAGTTTTGGGTGCATTTGTCTGGTCT GGTGTTTTGCTCGGGTTCTGTAAGAATAGGCCAGG GGATGCGCTCATCCCCTCTCGG GGTTCCGCTCCCACCGCGCCGCGTTCGGCCGGTT GCCTGCGAGATGTTTTCCGACGGACAATGATTC GCGCCTCCCATGTTGATCCCAGCTCCT GGCGTCAGGACCCCTGGGCCCCGCCCC CTCCACTCAGTCAATCTTTTGTCCCCGTATAAGG GGGTGGCTGGGGGGCGGCTGATTCCGA AATGCCCTTGGGGGGTCACCCGGGAGGGAACTC GGCTCCGGCTTTGGCCAGCCCGCACCCCTGGT TGAGCCGGCCCGAGGGCCACCAGGGGGCGCTCC CTGCAGCCCCCCGCAGCAGCCCCACTCC CACCCTACGATTGGCTGGCCGCCCCGAG CTCTGTGCTGTGATTGGTCACAGCCCGTGTCCGTC GGCCCCCGGGCCGGATACGAGGTGACGC CA GAGGCCCAGCTCGGGGCGGTGTCCC GGC GCGCC ACTGCGGGCGGAGTTTC AGGGCCGAAGC GGGCAGTGTGACGGCAGCGGTCCTGGGAGGCGC TCGGAGCAGCTCCCCGTCCTCCGCA GCCGGCCGTCGCCG CCCTGGCC TCCCGCACT CACTCCTGTCCGCCGCCCACC CCCACCTCCCACCTCGATGCGGTGCCGGGCTGC TGCGTGATGGGGCTGCGGAGCGGCGCCCTGCGG GCGGCCGCTGCTCGCGCTGAGGTGCGT GTGCCCGGCCCCCC SCGCCCC GCC GCTCCTGTTGACCCGGTCGGCCCGTCGGTCTGC GCTGAGGTAAGGCGGCGGGGGCTGGCC GGGTTGGGGAGGG GTTGGCGCC GTCG GGCCGCTTCCGC GGGAGGAGCGGCCGGGCCGG GGTCCGGGCGGGGTCTGAGGGGA

CTAGATTGAAAGCTCTGAAAAAAAAAAACTATCTTGT GTTTCTATCTGTTGAGCTCATAGTAGGTATCCAGG. AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC TGGGAGTTTTCTTCGCCATCTCCCT TTTTTCTTCTTCTTTCTTTCTTT TTGAGATGTCGTCTTGCTCAGTCCCCC GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC ACCTCCCAGGTTCAAGCAATTCTACTGCCTTAGCCT CCCGAGTAGCTGGGATTACAAGCACCCGCCACCA TCCTGGCTAATTTTTTTTTTTTGTATTTTTAGTTGAGA CAGGGTTTCACCATGTTGGTGATGCTGGTCTCAG/ CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT CCCAGAGTGTTAGGATTACAGGCATGAGCCACTG ACCCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA GGGAAGTAAGTTTAAGATAAAGTTACGATTTTGAAAT CTTTGGATTCAGAAGAATTTGTCACCTTTAACACCT AGAGTTGAACGTTCATACCTGGAGAGCCTTAACAT AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT CAGGTTTGGCAGGATTCGTCCCCTGAAGTGGACT GAGAGCCACACCCTGGCCTGTCACCATACCCATCC CCTATCCTTAGTGAAGCAAAACTCCTTTGTTCCCTT CTCCTTCTCCTAGTGACAGGAAATATTGTGATCCTA AAGAATGAAAATAGCTTGTCACCTCGTGGCCTCAG GCCTCTTGACTTCAGGCGGTTCTGTTTAATCAAGT GACATCTTCCCGAGGCTCCCTGAATGTGGCAGATG AAAGAGACTAGTTCAACCCTGACCTGAGGGGAAAG CCTTTGTGAAGGGTCAGGAG

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

- 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.

CCTTCTCTCCCCAGGTGG TGGGA GGTTCTGTAAGAATAGGCCAGG GGATGCGCTCATCCCCTCTCGG CTCCCAC GCCGC GTTCGGCCGGTT AGATGTTTTCCGACGGACAATGATTC CCTCCCATGTTGATCCCAGCTCC1 TGGGGGGTCACC GCTTTGGCCAGCCCGCACCCCTGGT GCCCGAGGGCCACCAGGGGGC FGCAGCCCCCCGCAGCAGCCCCCACTCC CCCTACGATTGGCTGGCCGCCCCGAG GCTGTGATTGGTCACAGCC GTGTCC GTC CCGGGGCGGATACGAGGTGAC CA CAGCTCGGGGGCGGTGTCC GC CC GGAGTTT AGGGCCGAAG GTGACGGCAGCGGTCCTGGGAGGC GC GAGCAGCTCCCCGTCCTCCGCA GGCCGTCC C CACTCCTGTCCGCCCCCCC ATGCGGTGCC GGGCTGC GGAGCGGCGCCCTGCGG TGATGGGGGCTG GCGGCCGCTGCTC CTGAGGTG GTGCCCGGCCCCC CCCC C CTCCTGTTGACCCGGTCGGCCCGTCGGTCTGC GCTGAGGTAAGGCGGCGGGGCTGGC GGGTTGGGGAGGG GTTGGCGC GTC GGGAGGAGCGGCCGGGCCGG GCTTC GGGCGGGGTCTGAGGGGA

CTCTTAGTTTTGGGTGCATTTGTCTGGT CTAGATTGAAAGCTCTGAAAAAAAAAAACTATCTTGT GTTTCTATCTGTTGAGCTCATAGTAGGTA AGTAGTAGGGTTGACTGCATTGATT TTGAGATGTCGTCTTGCTCAGT GTGCAGTGGTGCGATCTTGGCTCACTGTA ACCTCCCAGGTTCAAGCAATTCTACTGC CCCGAGTAGCTGGGATTACAAGCACC TCCTGGCTAATTTTTTTTTTTTTTTTGTATTTTTAGTTGAGA CAGGGTTTCACCATGTTGGTGATGCT CTCCTGGGGCCTAGCGATCCCCCTGCC1 CCCAGAGTGTTAGGATTACAGGCATGAGCCACTG ACCCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA GGGAAGTAAGTTTAAGATAAAGTTA CTTTGGATTCAGAAGAATTTGTCACCTTTAACACCT AGAGTTGAACGTTCATACCTGGAGAGCCTTAACAT AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT CAGGTTTGGCAGGATTCGTCCCCTGAAGTGGACT GAGAGCCACACCCTGGCCTGTCACCATACCCATCC CCTATCCTTAGTGAAGCAAAACTCCTTTGTTCCCT CTCCTTCTCCTAGTGACAGGAAATATTGTGATCCTA AAGAATGAAAATAGCTTGTCACCTCGTGGCCTCAG GCCTCTTGACTTCAGGCGGTTCTGTTTAATCAAGT GACATCTTCCCGAGGCTCCCTGAATGTGGCAGATG AAAGAGACTAGTTCAACCCTGACCTGAGGGG CCTTTGTGAAGGGTCAGGAG

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 $\mathbf{x} = x_1 x_2 \dots x_n$ Output: $\pi : \{1, \ldots, n\} \rightarrow \{\text{yes, no}\}$ AACGA CGAT ΤG CGA AAAAAAT TTATATCG $\mathbf{x} =$ $\pi =$ ves 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:



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 $\mathbf{x} = x_1 x_2 \dots x_n$ where $x_i \in \{A, T, C, G\}$

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

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

Fair Bet Casino

Input: Coin flips $\mathbf{x} = x_1 x_2 \dots x_n$ where $x_i \in \{H, T\}$ Output: $\pi : \{1, \dots, n\} \rightarrow \{F, B\}$ $\mathbf{x} = \bigcup_{F \in F} \bigotimes_{F \in B} \bigotimes_{B \in B} \bigotimes_{B \in F} \bigotimes_{F \in F}$

Question: Given 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)

Fair Bet Casino

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



Where is the professor?

 $Q = \{$ 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

Two decisions:

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





Andrey Markov



Three Questions

$$free Questions$$

 $free Questions$
 $free Questi$

$$M_{x_{1}} \wedge M_{y_{2}} \wedge M_{y_{3}} \rightarrow M_{y_{3}} + M_{y$$

 γ of terior γ ofQuestion 3: $\Pr(\pi; = \varsigma \mid \overline{\chi})$ 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 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 $P_r(\bar{x}, \bar{\pi}) = 1\bar{\lambda} = (\bar{\pi})$ $P_r(\overline{\chi}, \overline{\pi}) = P_r(\underline{\chi}_1, \overline{\pi}_1, \cdots, \overline{\chi}_n, \overline{\pi}_n) \qquad P_r(A,B) = P_r(A,B) P_r(B)$ $= \Pr(\chi_{n}, \pi_{n} | \underline{\chi_{n-1}}, \pi_{n-1}, \dots, \chi_{1}, \pi_{1}) \cdots \Pr(\chi_{2}\pi_{2} | \underline{\chi_{1}}, \pi_{1}) \mathbb{P}(\chi_{1}, \pi_{1})$ $= P_r(\chi_n, \pi_n | \pi_{n-1}) \cdots P_r(\chi_2 \pi_2 | \pi_1) P_r(\chi_1 | \pi_2) P_r(\pi_1)$ $= N e_{\pi_n, \chi_n} \alpha_{\pi_n-1, \pi_n}$ $TT e_{\pi_i, \chi_i} a_{\pi_{i-1}, \pi_i}$ In practice TT is hidden to us. Goal: Find To maximizing tr(X,T)



Joint Probability

 $\overline{\chi} = \overline{\chi}_{\mu}$ $\overline{\chi}_i = \chi_1, \dots, \chi_c$ $\overline{\pi_i} = \pi_1, \dots, \pi_i$ $\overline{\pi} = \overline{\pi}$ $V[s_i]$ denote the prob. $Pr(\overline{x}_i, \pi_{i-1}^*, \pi_i = s)$ state path π_i^* of the first i Ubservations Zi with knd state T:=5 $Pr(\overline{n},\overline{\pi}^*) = \max Pr(\overline{n},\overline{n},\overline{n}) = \max Pr(\overline{n},\overline{n},\overline{n}) = \max Pr(\overline{n},\overline{n}) = \max Pr(\overline{n}) = \max Pr(\overline{n},\overline{n}) = \max Pr(\overline{n}) = \max Pr(\overline{n},\overline{n}) = \max Pr(\overline{n}) = \max Pr(\overline{n}) = \max Pr(\overline{n},\overline{n}) = \max Pr(\overline{n}) = \max Pr$

v[F, 1]ao, F e Joint Probability $= P_r(\overline{x_1}, \overline{\overline{x_0}}, \overline{x_1} = s)$ Base case v [s, 1 90,5 PS, 21 11>1 $V[s,i] = P_r(\overline{x_i}, \pi_{i-1}^*, \pi_i = s)$ $= P_{i}(\chi_{1}, \dots, \chi_{i-1}, \chi_{i}, \pi_{1}^{*}, \dots, \pi_{i-1}^{*}, \pi_{i-1}^{*}, \pi_{i-1}^{*}, \pi_{i-1}^{*}, \pi_{i-1}^{*})$ $= Pr(\chi_{i}, \pi_{i} = s \mid \chi_{1, -i}, \chi_{i-1}, \pi_{1, -i}, \pi_{i-1}) Pr(B)$ $= \Pr[\mathcal{N}(\pi_{i}=\mathcal{N}) \pi_{i-1}) \Pr(\mathbf{B})$

Recurrence $v(x_i)$ (i>1) $P_r(x_i, \pi_i = s) \frac{\pi_i^*}{\pi_i^* - 2} P_r(x_1, \dots, x_{i-1}, \pi_1, \dots, \pi_{i-1})$ $v\Sigmat, c-1J$ = max $Pr(x_i | \pi_i = s) Pr(\pi_i = s | \pi_{i-1} = f) V [t, i-1]$ max $(e_{s,ni}) \cdot a_{t,s} \cdot v[t_{i,i-1}]$ $t \in Q$ = $e_{s,ni} \cdot max a_{t,s} v[t_{i,i-1}]$





Valid directions in the alignment problem.

Valid directions in the *decoding problem*.



- 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

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

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

$$\mathbf{X} = \bigcup_{i=1}^{n} \bigcup_{i=1}^{$$

ς.

$$\begin{array}{c} Q_{0} \not F = Q_{0} & g = \frac{1}{2} \\ Q = \{F, B\} \\ A = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.9 \end{pmatrix}_{B}^{F} \\ \Sigma = \{H, T\} \\ E = \begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix}_{B}^{F} \\ O.5 & 0.5 & 0.5 & 0.75 & 0.25 \end{array} \quad \begin{array}{c} V[F, L] = \\ V[F, L] = \\ V[F, L] = \\ W_{1} & V[F, L] = \\ W_{2} & V[F, L] = \\ W_{1} & V[F, L] = \\ W_{2} & V[F, L] = \\ W_$$

 $P_r(\overline{x}) = \sum_{\overline{n}} P_r(\overline{x}, \overline{n})$ $= \sum_{\{T_{2}, \dots, T_{n}\}} \Pr(\chi_{2}, \pi_{2}, \dots, \chi_{n}, \pi_{n}) = \sum_{\{T_{2}, \dots, T_{n}\}} (T_{2}, \dots, T_{n})$ k-1 , N; T-S þ D(y 1 $6)^{\prime}$ Goal : 26



 $\overline{\chi_i} = \chi_1, \dots, \chi_i$ $\overline{\pi_i} = \tau_{i_1}, \dots, \tau_i$

 $P_{r}(\overline{\pi}_{i}, \pi_{i} = s) = P_{r}(\pi_{1}, \dots, \chi_{i}, \pi_{i} = s) = \underbrace{FIs_{i}}_{Ts_{i-1}}$ $= \underbrace{P_{r}(\overline{\pi}_{i}, \pi_{i} = s)}_{\overline{\pi}_{i-1}} \underbrace{P_{r}(\overline{\pi}_{i}, \pi_{i} = s)}_{T_{i}}$ $\frac{\Pr(\bar{n})}{\Pr(\bar{n})} = \Pr(\chi_{1}, \dots, \chi_{n}) = \sum_{s \in Q} \Pr(\chi_{2}, \dots, \chi_{n}, \eta_{s} = s)$ 2 Z_ F[Sin].

Base [=1 $F[s,1] = Pr(x_1, \pi_1 = s) = a_{0,s} \cdot e_{s,x_1}$ Step i>1 $F[s,i] = Pr(\bar{x}_i, \pi_i = s) = Pr(x_1, ..., x_i, \pi_i = s)$ $= \sum_{\{\pi_{2}, \dots, \pi_{i-1}\}} P_{r}(\pi_{2}, \dots, \pi_{i-1}, \pi_{i}, \pi_{2}, \dots, \pi_{i-1}, \pi_{i-1}) \prod_{i=1}^{\pi_{i-1}} P_{r}(\pi_{2}, \dots, \pi_{i-1}, \pi_{i-1})$ $= \sum_{n=1}^{\infty} R_{n}(\pi_{1}, \dots, \pi_{r-1}, \pi_{2}, \dots, \pi_{i-1}, \pi_{i-1}, \pi_{i-1}) \cdot R_{s}, \pi_{i}$

 $\sum_{\pi_{i-1}} P_r(\pi_{1}, \dots, \pi_{i-1}, \pi_{2}, \dots, \pi_{i-1}, \pi_{i}=s) a_{s,\pi_{i}}$

$$\sum_{t \in Q} \sum_{\{\pi_{1}, \dots, \pi_{i-2}\}} \Pr(\pi_{1}, \dots, \pi_{i-1}, \pi_{1}, \dots, \pi_{i-2}, \pi_{i-1} = f) | a_{t,s} e_{s, \pi_{i}}$$

$$\Pr(\pi_{1}, \dots, \pi_{i-2}) = \Pr(\pi_{1}, \dots, \pi_{i-1}, \pi_{i-1} = f) = \Pr[f, i-1]$$

$$= e_{s, n_i} Z_{f[t_i; -1]} a_{t_is}$$

$$= e_{s, n_i} Z_{f[t_i; -1]} a_{t_is}$$

EF C = 1,

$$P_{r}(\tau_{1i} = 5 \mid \overline{\chi}) = \frac{P_{r}(\overline{\chi}, \tau_{1i} = 5)}{P_{r}(\overline{\chi})} \qquad P_{r}(A \mid B) = \frac{P_{r}(A \mid B)}{P_{r}(B)}$$

$$We have hole to compute $P_{r}(\overline{\chi}) = \frac{P_{r}(A \mid B)P_{r}(B)}{P_{r}(B)}$

$$(forward alg.) \qquad P_{r}(\overline{\chi}, \tau_{i} = 5) = \frac{P_{r}(\tau_{1}, \dots, \tau_{i}, \overline{\chi}_{i+1}, \dots, \tau_{n}, \tau_{i} = 5)}{P_{r}(\tau_{1}, \tau_{i}, \tau_{i} = 5)} = \frac{P_{r}(\tau_{1}, \dots, \tau_{i}, \overline{\chi}_{i+1}, \dots, \tau_{n}, \tau_{i} = 5)}{P_{r}(\chi_{1+1}, \dots, \chi_{n} \mid \tau_{i} = 5)}$$

$$= \frac{P_{r}(\tau_{1}, \dots, \tau_{i}, \tau_{i}, \tau_{i} = 5)}{P_{r}(\chi_{1+1}, \dots, \chi_{n} \mid \tau_{i} = 5)}$$$$

 $b[s_i] = P(z_{i+1}, \dots, z_n \mid \pi_i = s)$ i = n: $b[s, n] = Pr(x_{n+1}, ..., x_n | \pi_n = s)$ Q = sample space = Pr(D Tin = 5) $bI_{s,n}J = 1$ $F_r(\underline{\Omega}, \pi_n = s)$ $P_r(\pi_n = s)$ $P_r(\pi_n = s)$ $P_r(\pi_n = s)$ $P_r(\pi_n = s)$

i>1.

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



$$\begin{split} \mathbf{L}[s_{i}i] &= \begin{cases} 1, & \text{if } c = n, \\ z & a_{s_{i}t}e_{t,x_{t+1}}b[t_{i}t_{t}], & \text{if } 1 \leq i < n. \end{cases} \\ & \epsilon \epsilon q \end{split}$$
 $P_{r}(\pi_{i}=s|\pi) = \frac{P_{r}(\pi,\pi_{i}=s)}{P_{r}(\pi)} = \frac{F_{r}(\pi,\pi_{i}=s)}{P_{r}(\pi)} = \frac{F_{r}(\pi,\pi_{i}=s)}{F_{r}(\pi)} = \frac{F_{r}(\pi,\pi_{i}=s)}{F_{r}(\pi)}$ $(1) \text{ for wordto complete f } O(|P|^{2}n| f_{r}) = \frac{F_{r}(\pi,\pi_{i}=s)}{F_{r}(\pi,\pi_{i}=s)} = \frac{F_{$ D'hun bachward to compute b O(16 n)

Posterior de co day T_1 , T_2 , ... $\frac{\Lambda}{\Pi_{i}} = \arg \max \Pr(\Pi_{i} = S \mid \overline{x})$ $S \in Q$ Viterbi de codory: $\overline{a}^* = argmax P_r(\overline{\mathcal{I}}, \overline{T})$ $(\overline{a}_2, \dots, \overline{a}_n)$ -T1, ..., T*h



10X Genomics: Synthetic Long Reads ~ 2016-17

Genome indexing by partitioning and molecular barcoding





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



Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb Resolution of array OGH depends on spacing and length of clones





Input: $X_i = \log_2 T_i / R_i$, clone i = 1, ..., NOutput: Assignment $s(i) \in \{S_1, ..., S_K\}$ 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