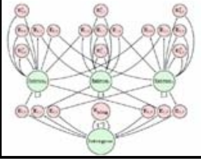
Computational Genomics

10-810/02-710, Spring 2009

Gene Finding and HMM (cont'd)

Eric Xing

Lecture 7, February 4, 2009



Reading: Durbin Chap 3, class assignment

The HMM Algorithms

Questions:

- **Decoding**: What is the most likely DNA parsing? Viterbi
- **Evaluation**: What is the probability of the observed sequence? Forward
- **Decoding**: What is the probability that the state of the 3rd position is Bk or gene, given the observed sequence? Forward-Backward
- **Learning**: Under what parameterization are the observed sequences most probable? Baum-Welch (EM)

Learning HMM: two scenarios

- **Supervised learning**: estimation when the “right answer” is known
  - **Examples:**

    GIVEN: a genomic region $x = x_1 \ldots x_{1,000,000}$ where we have good (experimental) annotations of the CpG islands

    GIVEN: the casino player allows us to observe him one evening, as he changes dice and produces 10,000 rolls

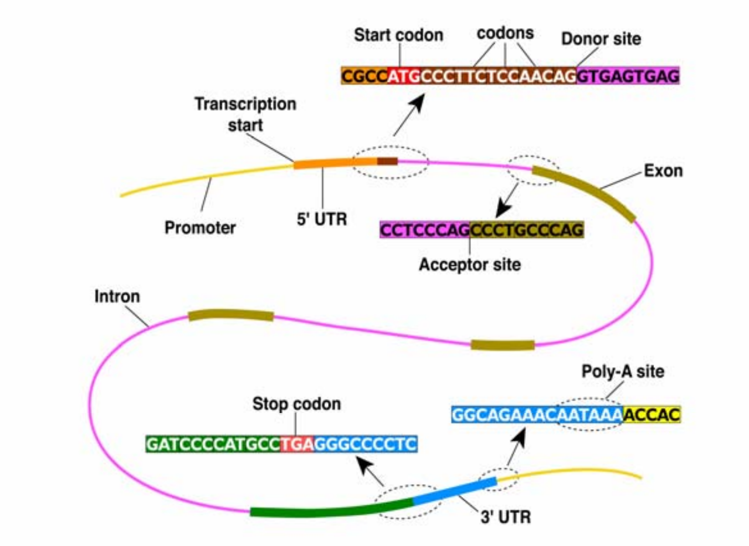
- **Unsupervised learning**: estimation when the “right answer” is unknown
  - **Examples:**

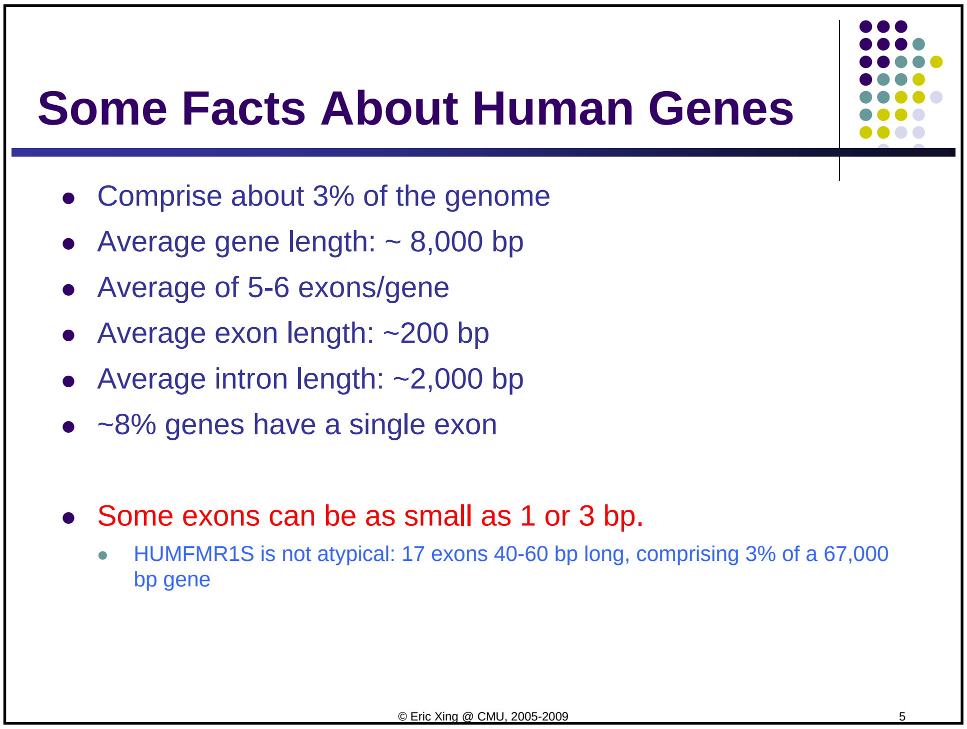
    GIVEN: the porcupine genome; we don’t know how frequent are the CpG islands there, neither do we know their composition

    GIVEN: 10,000 rolls of the casino player, but we don’t see when he changes dice

- **QUESTION:** Update the parameters θ of the model to maximize $P(x|\theta)$ --- Maximal likelihood (ML) estimation

Typical structure of a gene

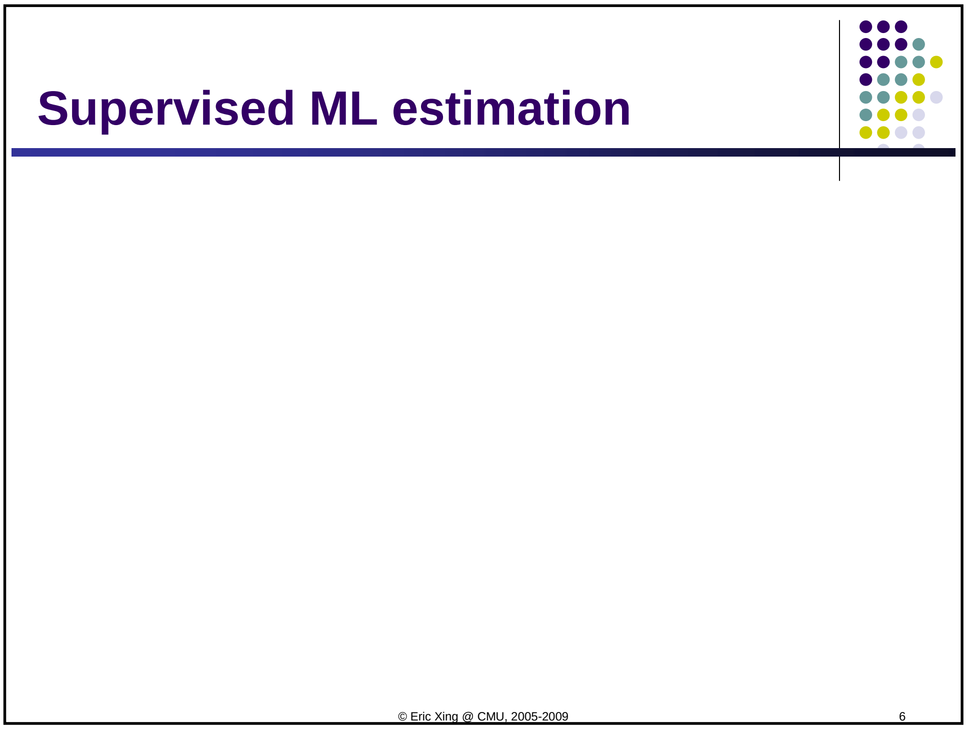




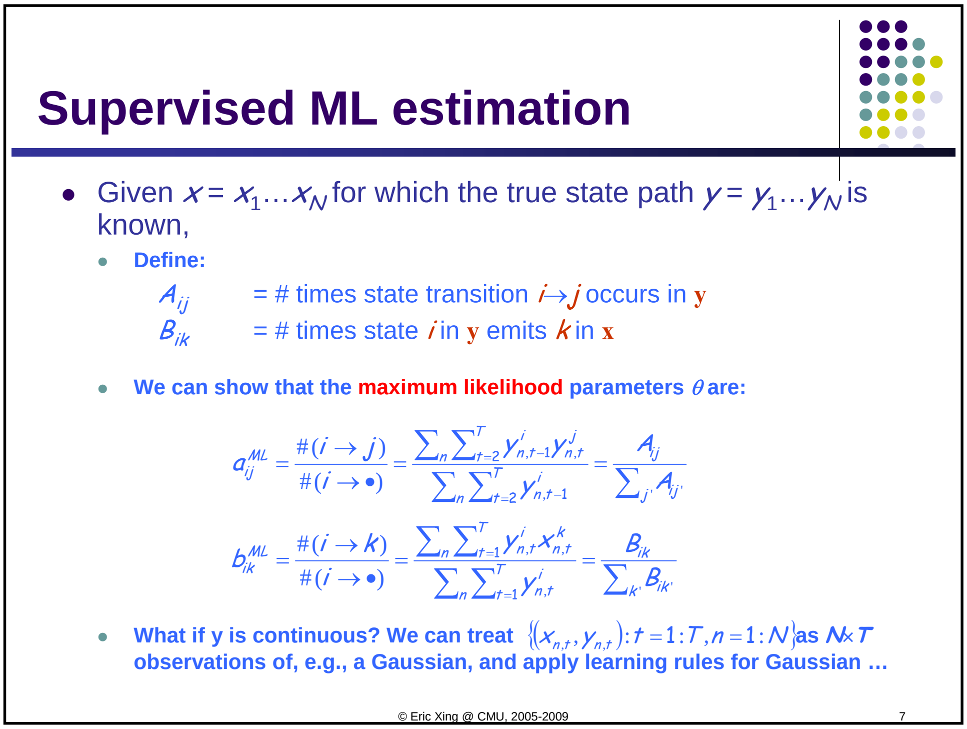
Some Facts About Human Genes

- Comprise about 3% of the genome
- Average gene length: ~ 8,000 bp
- Average of 5-6 exons/gene
- Average exon length: ~200 bp
- Average intron length: ~2,000 bp
- ~8% genes have a single exon

- Some exons can be as small as 1 or 3 bp.
  - HUMFMR1S is not atypical: 17 exons 40-60 bp long, comprising 3% of a 67,000 bp gene



Supervised ML estimation



Supervised ML estimation

- Given $x = x_1 \dots x_N$ for which the true state path $y = y_1 \dots y_N$ is known,
  - **Define:**

    A_{ij} = # times state transition $i \rightarrow j$ occurs in $\mathbf{y}$

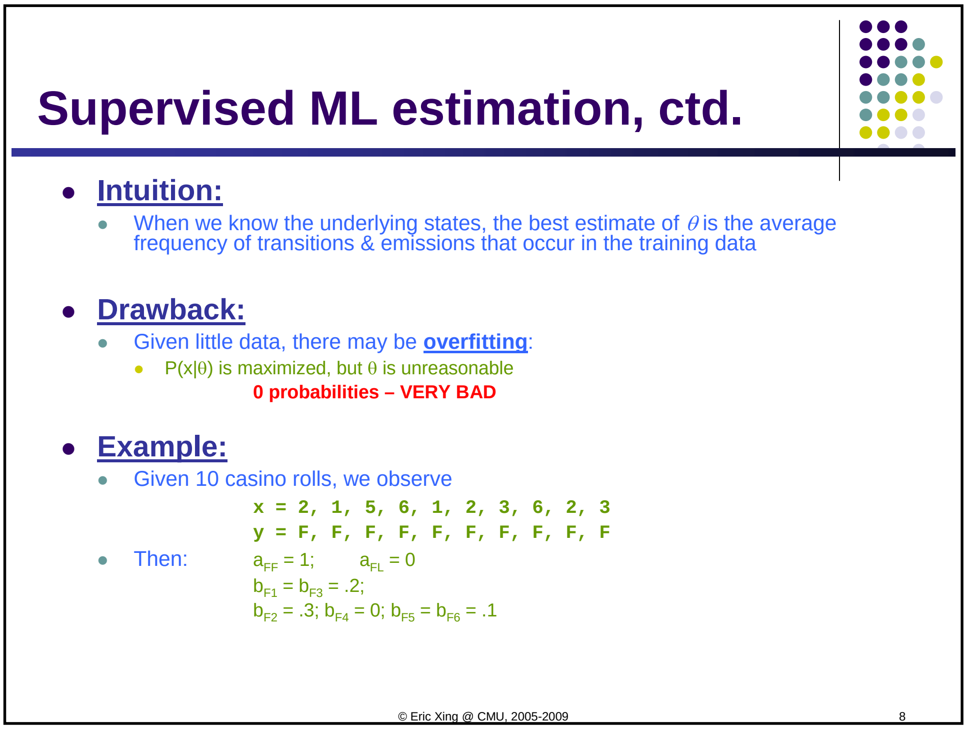
    B_{ik} = # times state i in $\mathbf{y}$ emits k in $\mathbf{x}$

  - **We can show that the maximum likelihood parameters θ are:**

$$a_{ij}^{ML} = \frac{\#(i \rightarrow j)}{\#(i \rightarrow \bullet)} = \frac{\sum_n \sum_{t=2}^{T} y_{n,t-1}^{i} y_{n,t}^{j}}{\sum_n \sum_{t=2}^{T} y_{n,t-1}^{i}} = \frac{A_{ij}}{\sum_{j'} A_{ij'}}$$

$$b_{ik}^{ML} = \frac{\#(i \rightarrow k)}{\#(i \rightarrow \bullet)} = \frac{\sum_n \sum_{t=1}^{T} y_{n,t}^{i} x_{n,t}^{k}}{\sum_n \sum_{t=1}^{T} y_{n,t}^{i}} = \frac{B_{ik}}{\sum_{k'} B_{ik'}}$$

  - **What if y is continuous? We can treat $\{(x_{n,t}, y_{n,t}) : t = 1:T, n = 1:N\}$ as $N \times T$ observations of, e.g., a Gaussian, and apply learning rules for Gaussian …**



Supervised ML estimation, ctd.

- **Intuition:**
  - When we know the underlying states, the best estimate of θ is the average frequency of transitions & emissions that occur in the training data
- **Drawback:**
  - Given little data, there may be **overfitting**:
    - P(x|θ) is maximized, but θ is unreasonable

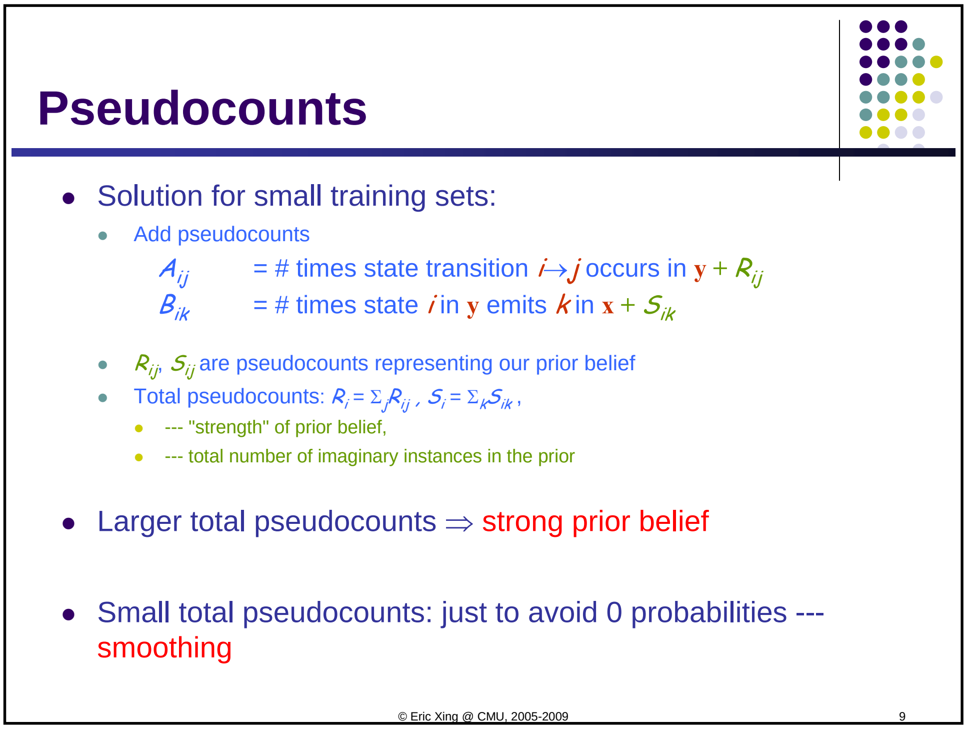
      0 probabilities – VERY BAD
- **Example:**
  - Given 10 casino rolls, we observe

```
x = 2, 1, 5, 6, 1, 2, 3, 6, 2, 3
y = F, F, F, F, F, F, F, F, F, F
```

  - Then: $a_{FF} = 1$; $a_{FL} = 0$

    $b_{F1} = b_{F3} = .2$;

    $b_{F2} = .3$; $b_{F4} = 0$; $b_{F5} = b_{F6} = .1$



Pseudocounts

- Solution for small training sets:
  - Add pseudocounts

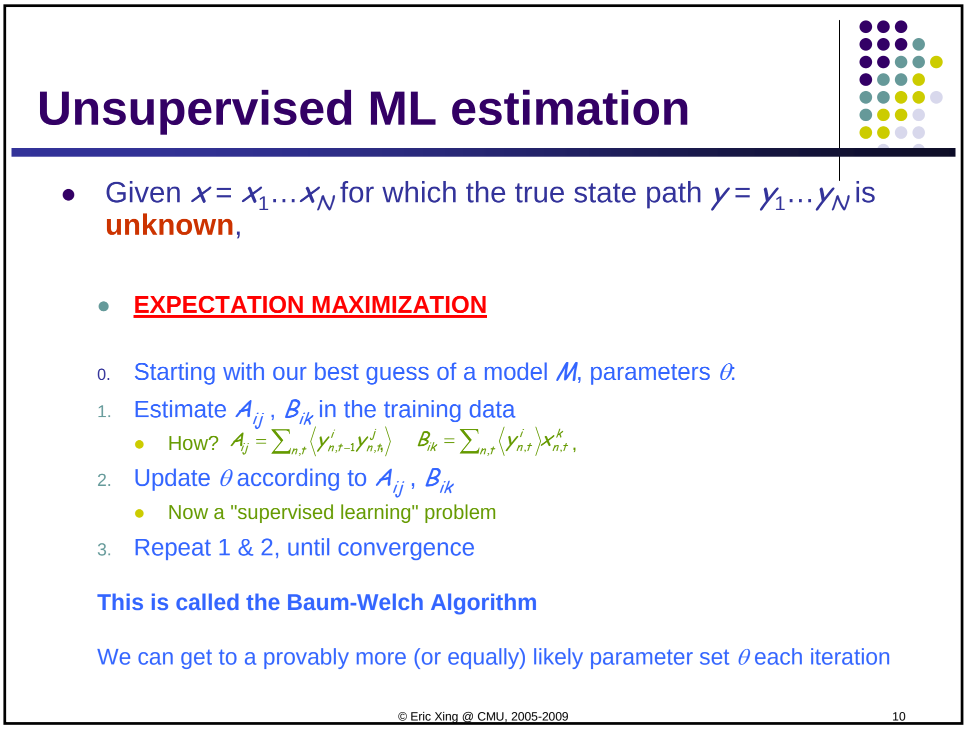
    A_{ij} = # times state transition $i \rightarrow j$ occurs in $\mathbf{y}$ + R_{ij}

    B_{ik} = # times state i in $\mathbf{y}$ emits k in $\mathbf{x}$ + S_{ik}

  - R_{ij}, S_{ij} are pseudocounts representing our prior belief
  - Total pseudocounts: $R_i = \Sigma_j R_{ij}$, $S_i = \Sigma_k S_{ik}$,
    - --- "strength" of prior belief,
    - --- total number of imaginary instances in the prior

- Larger total pseudocounts ⇒ strong prior belief

- Small total pseudocounts: just to avoid 0 probabilities --- smoothing



Unsupervised ML estimation

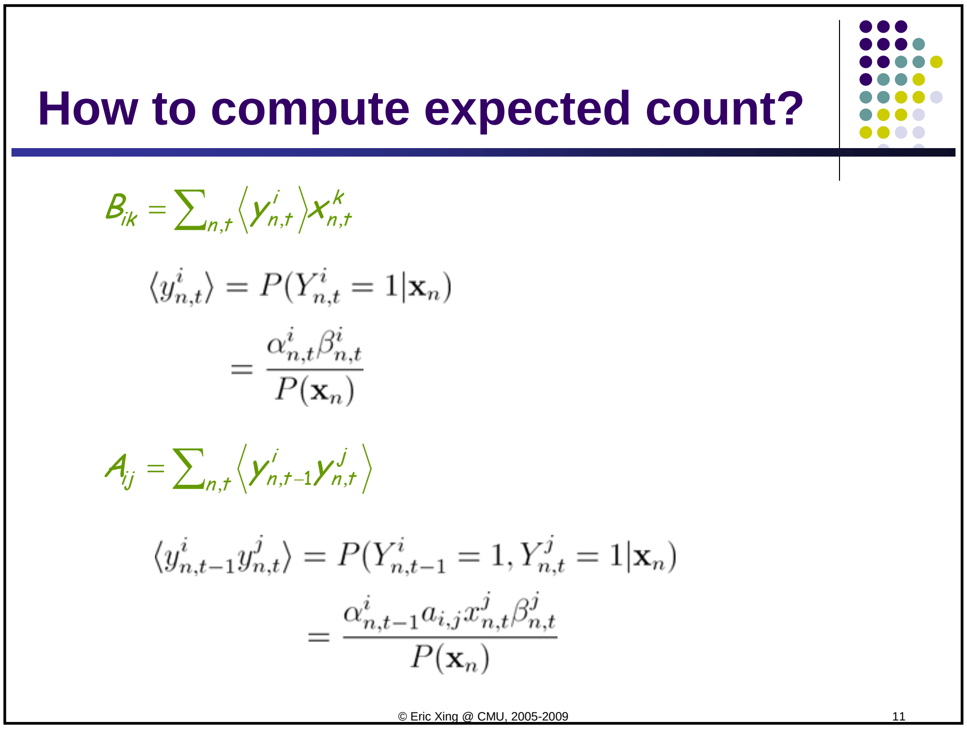
- Given $x = x_1 \dots x_N$ for which the true state path $y = y_1 \dots y_N$ is **unknown**,

  - **EXPECTATION MAXIMIZATION**

0. Starting with our best guess of a model M, parameters θ:
1. Estimate A_{ij} , B_{ik} in the training data
   - How? $A_{ij} = \sum_{n,t} \langle y^i_{n,t-1} y^j_{n,t} \rangle$ $B_{ik} = \sum_{n,t} \langle y^i_{n,t} \rangle x^k_{n,t}$,
2. Update θ according to A_{ij} , B_{ik}
   - Now a "supervised learning" problem
3. Repeat 1 & 2, until convergence

This is called the Baum-Welch Algorithm

We can get to a provably more (or equally) likely parameter set θ each iteration



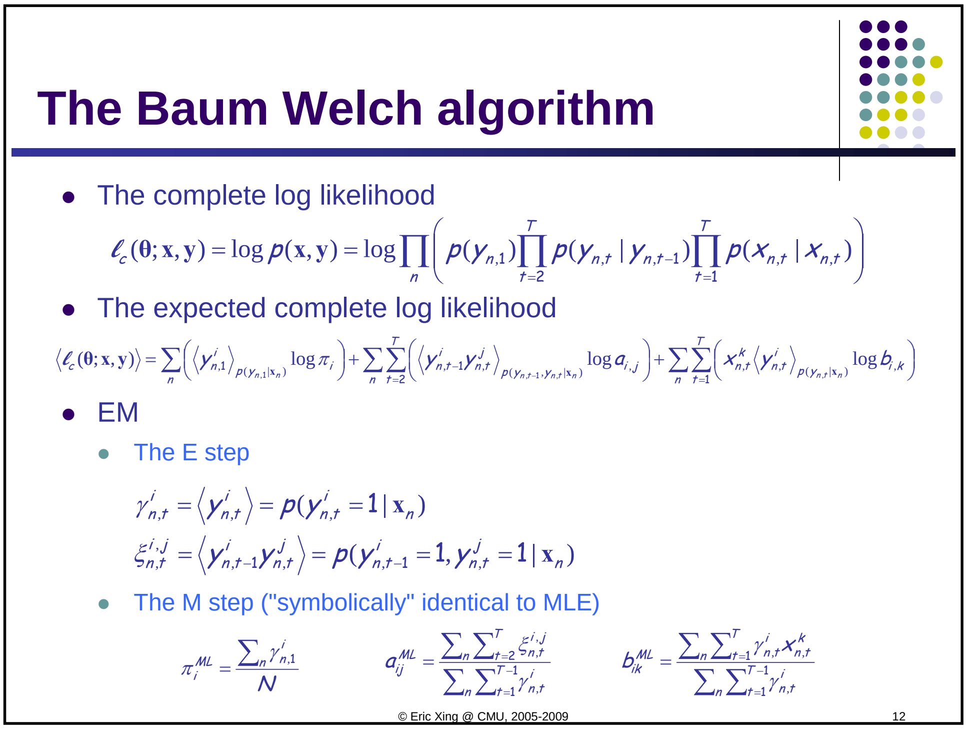
How to compute expected count?

$$B_{ik} = \sum_{n,t} \left\langle y^i_{n,t} \right\rangle x^k_{n,t}$$

$$\begin{aligned}\langle y^i_{n,t} \rangle &= P(Y^i_{n,t} = 1 | \mathbf{x}_n) \\ &= \frac{\alpha^i_{n,t}\beta^i_{n,t}}{P(\mathbf{x}_n)}\end{aligned}$$

$$A_{ij} = \sum_{n,t} \left\langle y^i_{n,t-1} y^j_{n,t} \right\rangle$$

$$\begin{aligned}\langle y^i_{n,t-1} y^j_{n,t} \rangle &= P(Y^i_{n,t-1} = 1, Y^j_{n,t} = 1 | \mathbf{x}_n) \\ &= \frac{\alpha^i_{n,t-1} a_{i,j} x^j_{n,t} \beta^j_{n,t}}{P(\mathbf{x}_n)}\end{aligned}$$



The Baum Welch algorithm

- The complete log likelihood

$$\ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}, \mathbf{y}) = \log \prod_n \left(p(y_{n,1}) \prod_{t=2}^{T} p(y_{n,t} | y_{n,t-1}) \prod_{t=1}^{T} p(x_{n,t} | x_{n,t}) \right)$$

- The expected complete log likelihood

$$\left\langle \ell_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}) \right\rangle = \sum_n \left(\left\langle y^i_{n,1} \right\rangle_{p(y_{n,1}|\mathbf{x}_n)} \log \pi_i \right) + \sum_n \sum_{t=2}^{T} \left(\left\langle y^i_{n,t-1} y^j_{n,t} \right\rangle_{p(y_{n,t-1}, y_{n,t}|\mathbf{x}_n)} \log a_{i,j} \right) + \sum_n \sum_{t=1}^{T} \left(x^k_{n,t} \left\langle y^i_{n,t} \right\rangle_{p(y_{n,t}|\mathbf{x}_n)} \log b_{i,k} \right)$$

- EM
  - The E step

$$\gamma^i_{n,t} = \left\langle y^i_{n,t} \right\rangle = p(y^i_{n,t} = 1 | \mathbf{x}_n)$$

$$\xi^{i,j}_{n,t} = \left\langle y^i_{n,t-1} y^j_{n,t} \right\rangle = p(y^i_{n,t-1} = 1, y^j_{n,t} = 1 | \mathbf{x}_n)$$

  - The M step ("symbolically" identical to MLE)

$$\pi^{ML}_i = \frac{\sum_n \gamma^i_{n,1}}{N} \qquad a^{ML}_{ij} = \frac{\sum_n \sum_{t=2}^{T} \xi^{i,j}_{n,t}}{\sum_n \sum_{t=1}^{T-1} \gamma^i_{n,t}} \qquad b^{ML}_{ik} = \frac{\sum_n \sum_{t=1}^{T} \gamma^i_{n,t} x^k_{n,t}}{\sum_n \sum_{t=1}^{T-1} \gamma^i_{n,t}}$$

The Baum-Welch algorithm -- comments

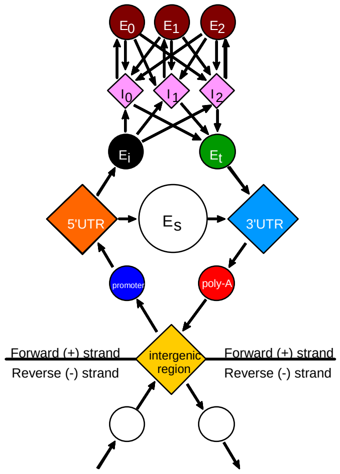
Time Complexity:

# iterations × $O(K^2N)$

- Guaranteed to increase the log likelihood of the model
- Not guaranteed to find globally best parameters
- Converges to local optimum, depending on initial conditions
- Too many parameters / too large model: Overt-fitting

The Idea Behind a GHMM GeneFinder

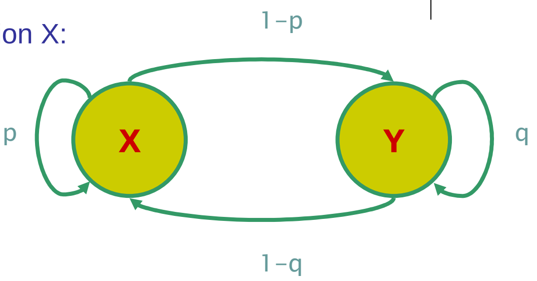
- **States** represent standard gene features: intergenic region, exon, intron, perhaps more (promotor, 5'UTR, 3'UTR, Poly-A,..).
- **Observations** embody state-dependent base composition, dependence, and signal features.
- In a GHMM, **duration** must be included as well.
- Finally, **reading frames** and **both strands** must be dealt with.



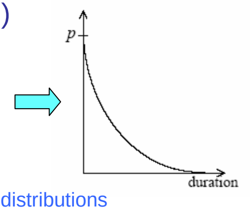
Modeling the Duration of States

- Length distribution of region X:

  $E[l_X] = 1/(1-p)$

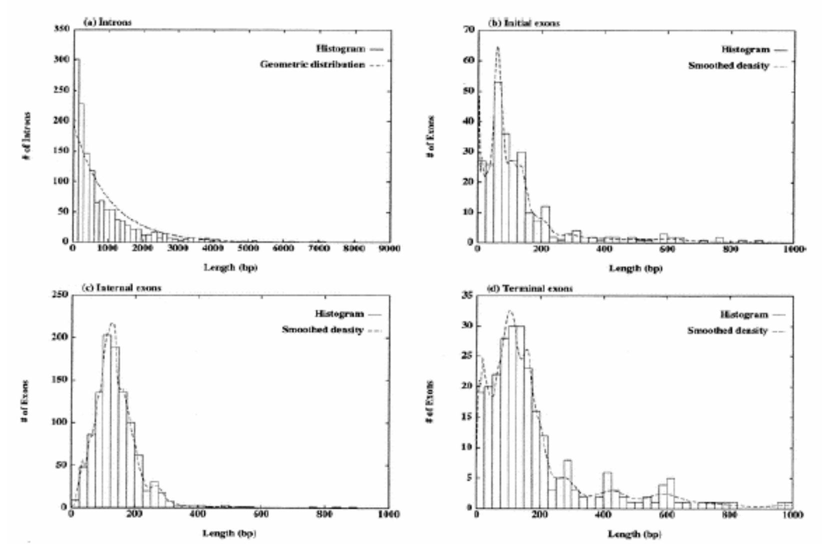


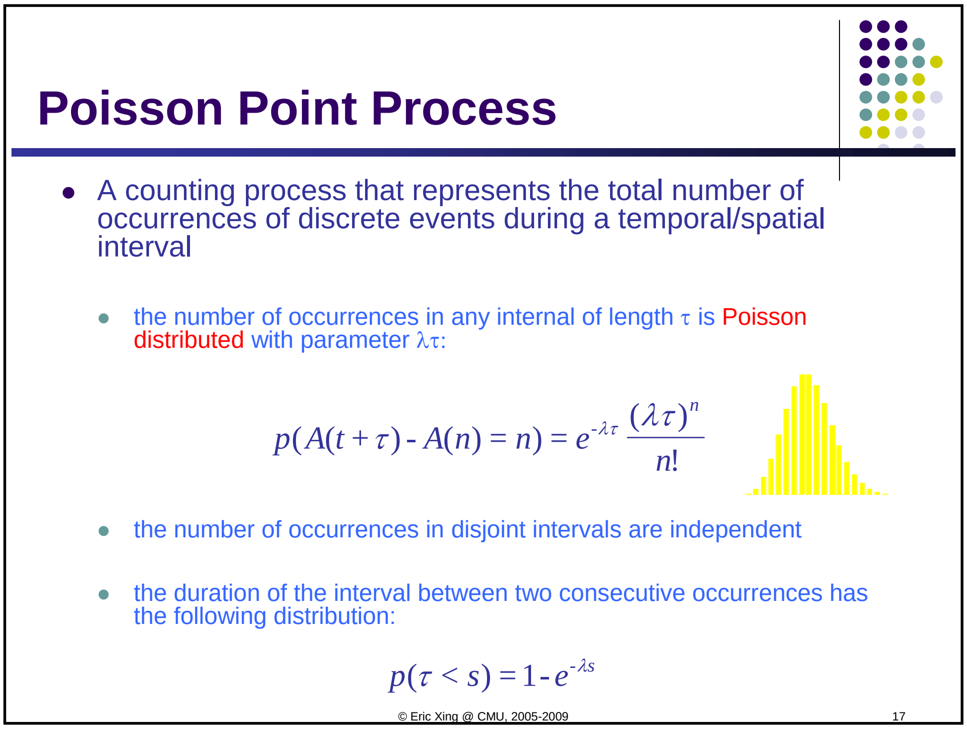
- Geometric distribution, with mean 1/(1-p)
  - (homework: derive this)



  - This is a significant disadvantage of HMMs
  - Several solutions exist for modeling different length distributions

Observed Duration Time





Poisson Point Process

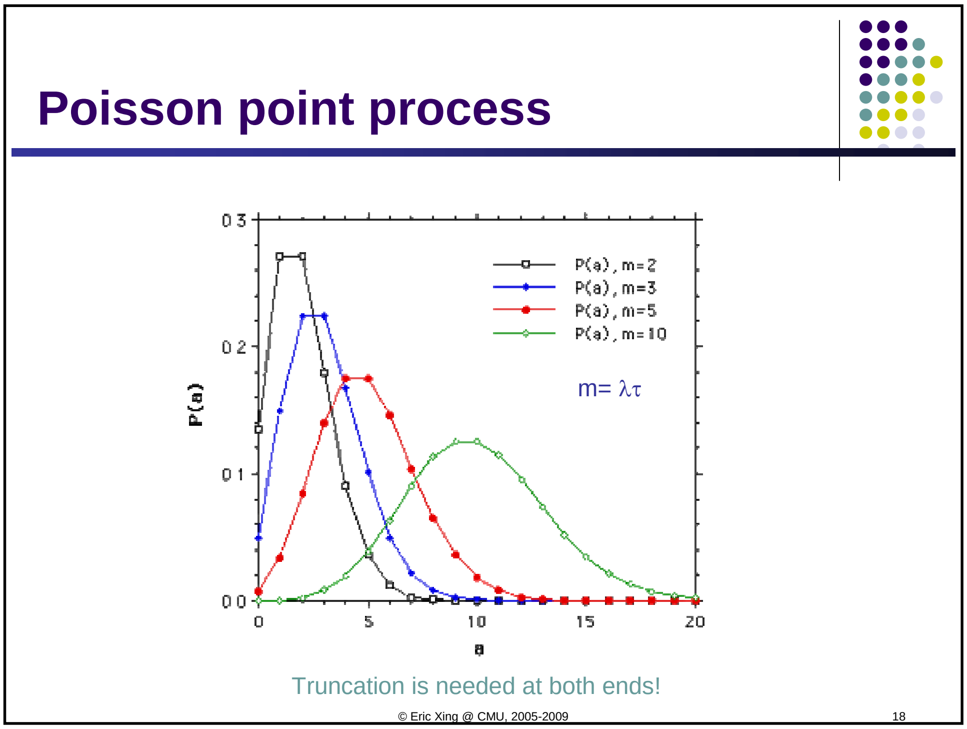
- A counting process that represents the total number of occurrences of discrete events during a temporal/spatial interval
  - the number of occurrences in any internal of length τ is Poisson distributed with parameter $\lambda\tau$:

$$p(A(t+\tau) - A(n) = n) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$$

  - the number of occurrences in disjoint intervals are independent
  - the duration of the interval between two consecutive occurrences has the following distribution:

$$p(\tau < s) = 1 - e^{-\lambda s}$$

© Eric Xing @ CMU, 2005-2009 17



Poisson point process

m= $\lambda\tau$

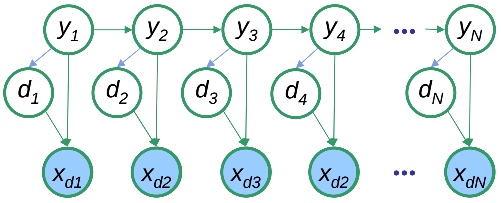
Truncation is needed at both ends!

© Eric Xing @ CMU, 2005-2009 18

Generalized HMM

Upon entering a state:

1. Choose duration d, according to probability distribution
2. Generate d letters according to emission probs
3. Take a transition to next state according to transition probs



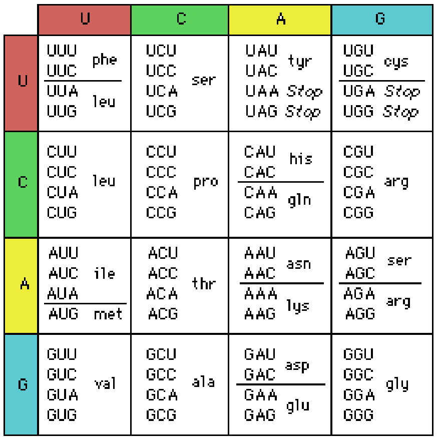
Disadvantage: Increase in complexity:

Time: $O(D^2)$
Space: $O(D)$

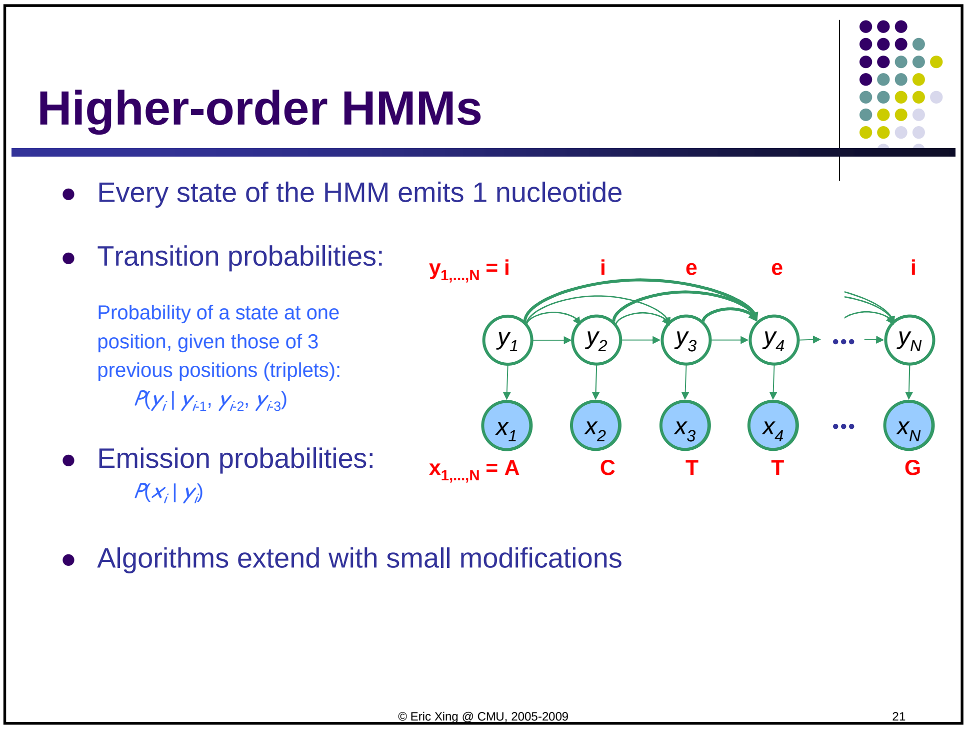
where D = maximum duration of state

Higher-order HMMs

- **The Genetic Code**
  - 3 nucleotides make 1 amino acid
  - Statistical dependencies in triplets
- **Question:**
  - Recognize protein-coding segments with an HMM



	U	C	A	G
U	UUU phe UUC UUA leu UUG	UCU UCC ser UCA UCG	UAU tyr UAC UAA Stop UAG Stop	UGU cys UGC UGA Stop UGG Stop
C	CUU CUC leu CUA CUG	CCU CCC pro CCA CCG	CAU his CAC CAA gln CAG	CGU CGC arg CGA CGG
A	AUU AUC ile AUA AUG met	ACU ACC thr ACA ACG	AAU asn AAC AAA lys AAG	AGU ser AGC AGA arg AGG
G	GUU GUC val GUA GUG	GCU GCC ala GCA GCG	GAU asp GAC GAA glu GAG	GGU GGC gly GGA GGG



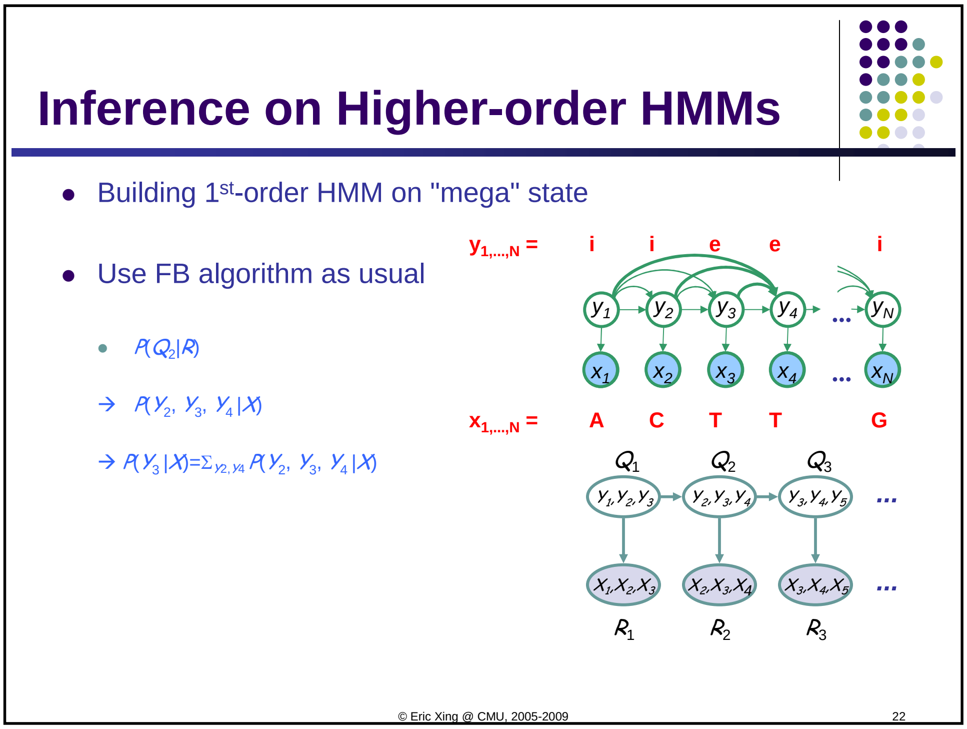
Higher-order HMMs

- Every state of the HMM emits 1 nucleotide
- Transition probabilities:

  Probability of a state at one position, given those of 3 previous positions (triplets):

  $P(y_i \mid y_{i-1}, y_{i-2}, y_{i-3})$
- Emission probabilities:

  $P(x_i \mid y_i)$
- Algorithms extend with small modifications



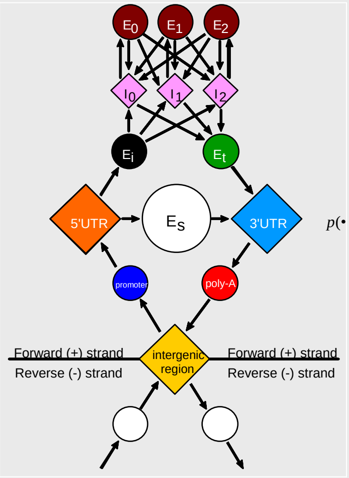
Inference on Higher-order HMMs

- Building 1st-order HMM on "mega" state
- Use FB algorithm as usual
  - $P(Q_2 \mid R)$

  → $P(Y_2, Y_3, Y_4 \mid X)$

  → $P(Y_3 \mid X) = \Sigma_{y2,y4} P(Y_2, Y_3, Y_4 \mid X)$

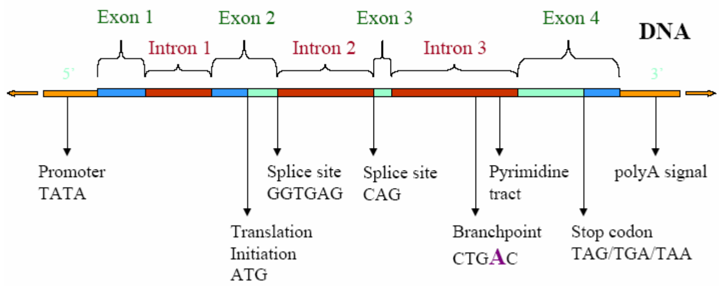
GENSCAN (Burge & Karlin)



$$p(\bullet \mid y) = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$$

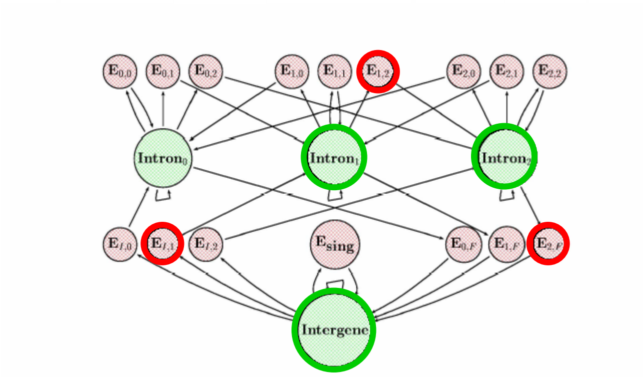
```
GAGAACGTGTGAGAGAGAGGCAAGCCGAAAAATCAGCCGC
CGAAGGATACACTATCGTCGTCCTTGTCCGACGAACCGGT
GGTCATGCAAACAACGCACAGAACAAATTAATTTTCAAAT
TGTTCAATAAATGTCCCACTTGCTTCTGTTGTTCCCCCCT
TTCCGCTAGCGGAATTTTTTATATTCTTT


                     GATACACAGCGCACACAT
ATAAGCTTGCACACTGATGCACACACACCGACACGTTGTC
ACCGAAATGAACGGGACGGCCATATGACTGGCTGGCGCTC
GGTATGTGGGTGCAAGCGAGATACCGCGATCAAGACTCGA
ACGAGACGGGTCAGCGAGTGATACCGATTCTCTCTCTTTT
GCGATTGGGAATAATGCCCGACTTTTTACACTACATGCGT
TGGATCTGGTTATTTAATTATGCCATTTTTCTCAGTATAT
CGGCAATTGGTTGCATTAATTTTGCCGCAAAGTAAGGAAC
ACAAACCGATAGTTAAGATCCAACGTCCCTGCTGCGCCTC
GCGTGCACAATTTGCGCCAATTTCCCCCCTTTTCCAGTTT
TTTTCAACCCAGCACCGCTCGTCTCTTCCTCTTCTTAACG
TTAGCATTCGTACGAGGAACAGTGCTGTCATTGTGGCCGC
TGTGTAGCTAAAAAGCGTAATTATTCATTATCTAGCTATC
TTTTCGGATATTATTGTCATTTGCCTTTAATCTTGTGTAT
TTATATGGATGAAACGTGCTATAATAACAATGCAGAATGA
AGAACTGAAGAGTTTCAAAACCTAAAAATAATTGGAATAT
AAAGTTTGGTTTTACAATTTGATAAAACTCTATTGTAAGT
GGAGCGTAACATAGGGTAGAAAACAGTGCAAATCAAAGTA
CCTAAATGGAATACAAATTTTAGTTGTACAATTGAGTAAA
ATGAGCAAAGCGCCTATTTTGGATAATATTTGCTGTTTAC
AAGGGGAACATATTCATAATTTTCAGGTTTAGGTTACGCA
TATGTAGGCGTAAAGAAATAGCTATATTTGTAGAAGTGCA
TATGCACTTTATAAAAAATTATCCTACATTAACGTATTTT
ATTTGCTTTAAAACCTATCTGAGATATTCCAATAAGGTAA
GTGCAGTAATACAATGTAAATAATTGCAAATAATGTTGTA
ACTAAATACGTAAACAATAATGTAGAAGTCCGGCTGAAAG
CCCCAGCAGCTATAGCCGATATCTATATGATTTAAACTCT
TGTCTGCAACGTTCTAATAAATAAATAAAATGCAAAATAT
AACCTATTGAGACAATACATTTATTTTATTTTTTTATATC
ATCAATCATCTACTGATTTCTTTCGGTGTATCGCCTAATC
CATCTGTGAAATAGAAATGGCGCCACCTAGGTTAAGAAAA
GATAAACAGTTGCCTTTAGTTGCATGACTTCCCGCTGGAT
```

Gene Finding

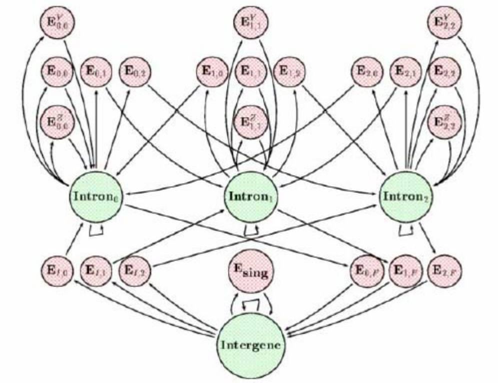


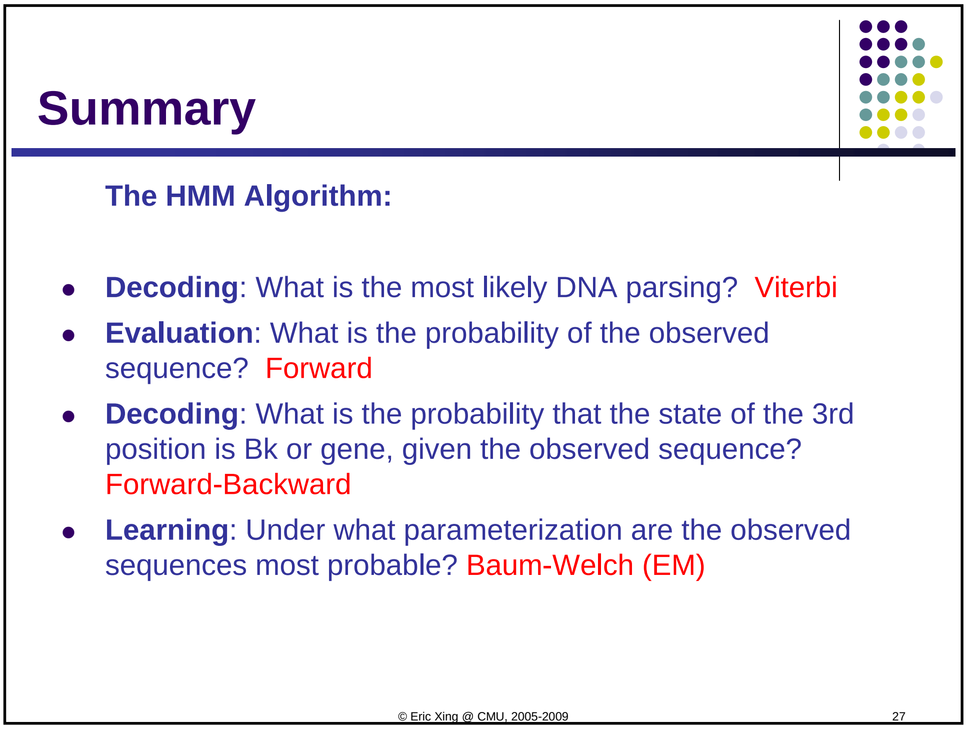
Generalized HMM Gene finder

TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



Allowing for inserted exons

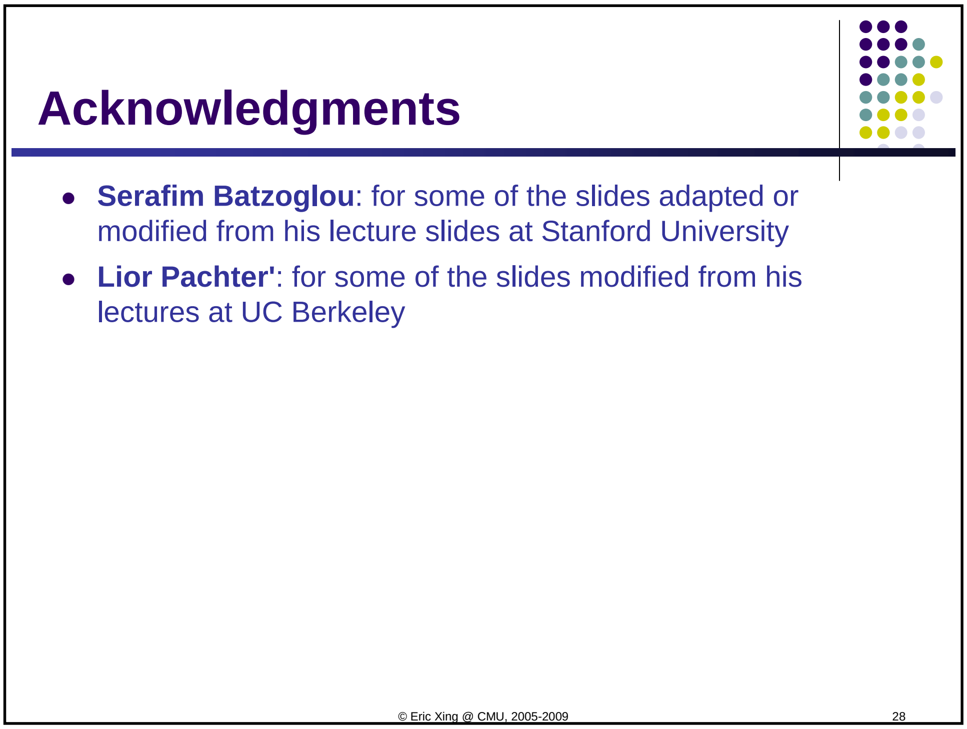




Summary

The HMM Algorithm:

- **Decoding**: What is the most likely DNA parsing? Viterbi
- **Evaluation**: What is the probability of the observed sequence? Forward
- **Decoding**: What is the probability that the state of the 3rd position is Bk or gene, given the observed sequence? Forward-Backward
- **Learning**: Under what parameterization are the observed sequences most probable? Baum-Welch (EM)



Acknowledgments

- **Serafim Batzoglou**: for some of the slides adapted or modified from his lecture slides at Stanford University
- **Lior Pachter'**: for some of the slides modified from his lectures at UC Berkeley