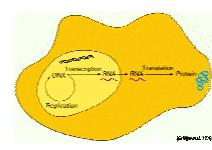


# Machine Learning

10-701/15-781, Spring 2008

## Machine Learning in Computational Biology

Eric Xing

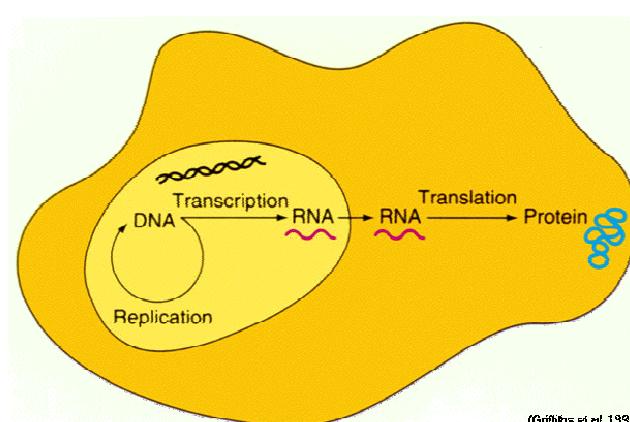


Lecture 20, April 2, 2008

Reading:

1

## The Central Dogma

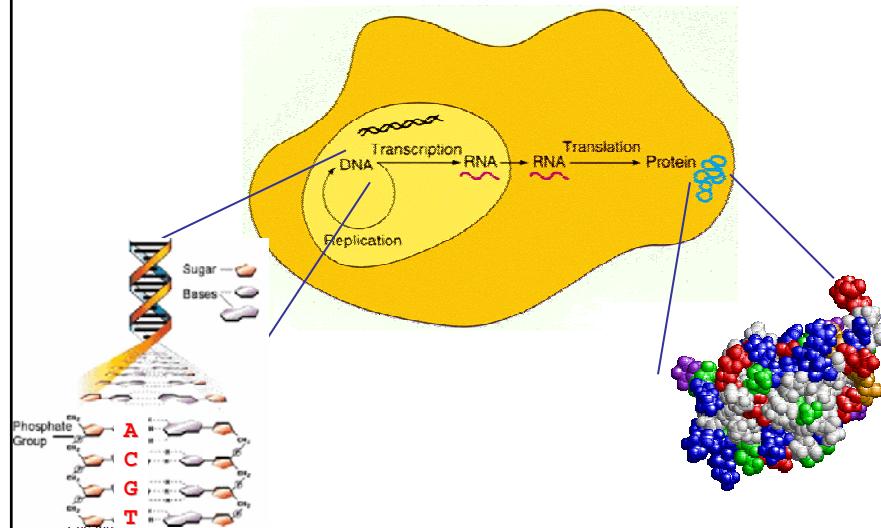


(Griffiths et al. 1998)

Eric Xing

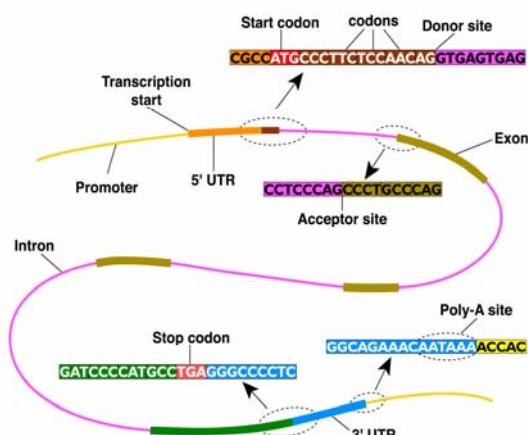
2

## Genome and Proteome



3

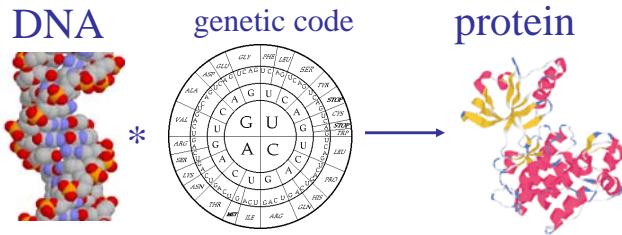
## Gene Structure in DNA



- The inference problem: predicting locations of the genes on DNA

## Proteins are coded by DNA

- There are between 30,000 to 40,000 genes in the human genome



- The human gene inventory corresponds to ~1.5% of the genome (coding regions)

Eric Xing

5

## Protein Structure Hierarchy

Primary Structure	Secondary Structures	Tertiary Structures	Quaternary Structures
 ... LACAAECS...	 Alpha-helix  Anti-parallel beta-sheet Parallel beta-sheet	 Beta-helix	 Triple beta-spiral

- **The inference problem:** predicting the structures from sequences

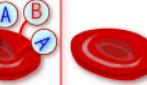
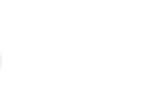


Eric Xing

6

## Genetic Polymorphisms

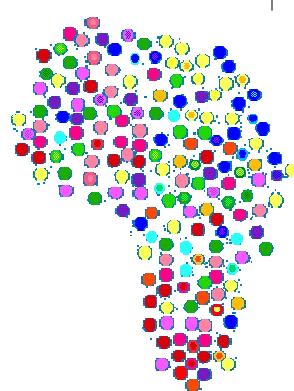
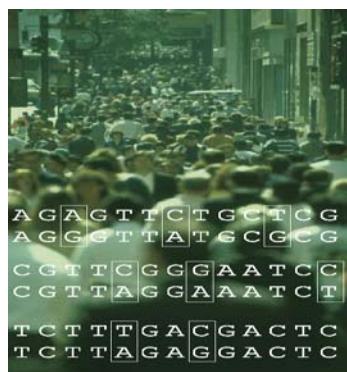
### The ABO Blood System

Blood Type (genotype)	Type A (AA, AO)	Type B (BB, BO)	Type AB (AB)	Type O (OO)
Red Blood Cell Surface Proteins (phenotype)				
Plasma Antibodies (phenotype)	b agglutinin only	a agglutinin only	NONE	a and b agglutinin

Eric Xing

7

## Genetic Demography



- Are there genetic prototypes among them ?
- What are they ?
- How many ? (how many ancestors do we have ?)

Eric Xing

8

# Single Nucleotide Polymorphism (SNP)



...cagttaccgtgcataatctgactatgtggacgttatcc...  
 ...cagttaccgtgcacgataatctgactatgtggacgttatcc...  
 ...cagttaccgtgcacgataatctgactatgtggacgttatcc...  
 ...cagttaccgtgcataatctgactatgtggacgttatcc...  
 ...cagttaccgtgcataatctgactatgtggacgttatcc...



- The inference problem: "haplotypes" and population diversity

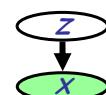
Eric Xing

9

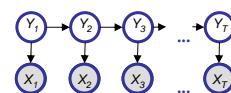
# Computation Biology and ML



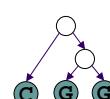
- Mixture and infinite mixture
  - clustering of genetic polymorphisms



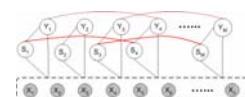
- Hidden Markov Models
  - gene finding



- Trees
  - sequence evolution



- Conditional Random Fields
  - protein structure prediction

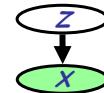


Eric Xing

10

# Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction

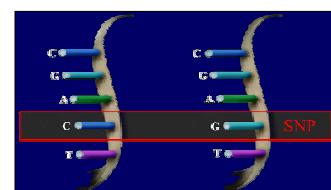


Eric Xing

11

# Biological Terms

- **Genetic polymorphism:** a difference in DNA sequence among individuals, groups, or populations
- **Single Nucleotide Polymorphism (SNP):** DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
  - Each variant is called an "allele"
  - Almost always bi-allelic
  - Account for most of the genetic diversity among different (normal) individuals, e.g. drug response, disease susceptibility

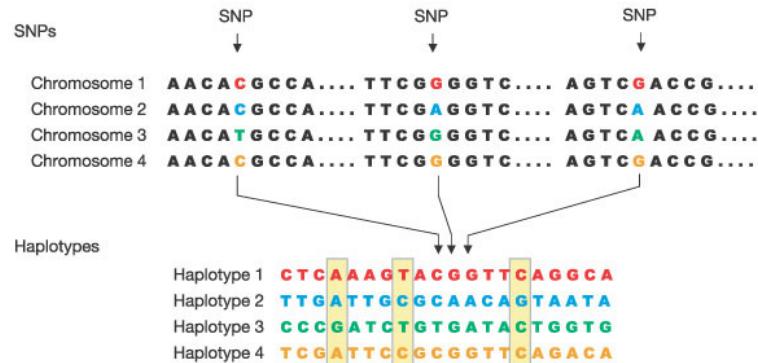


Eric Xing

12

## From SNPs to Haplotypes

- Alleles of adjacent SNPs on a chromosome form **haplotypes**

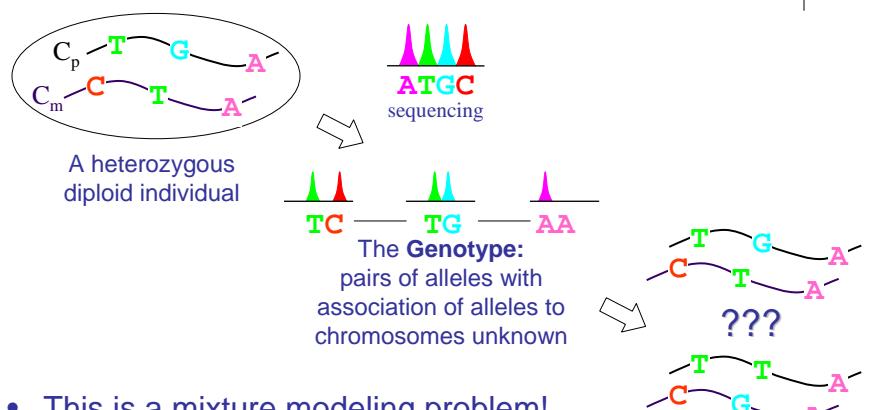


- Useful in the study of **disease association or genetic evolution**

Eric Xing

13

## Phase ambiguity of SNPs "haplotypes"



- This is a mixture modeling problem!

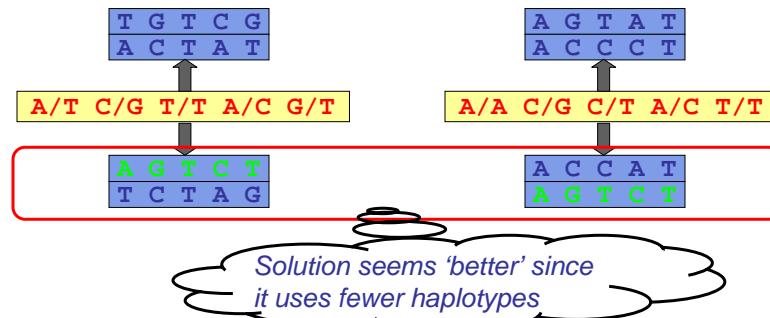
Eric Xing

14

# Haplotype Inference

## Why is it approachable?

- Many of the haplotypes appear many times
- Data for many individuals allows inference



Eric Xing

15

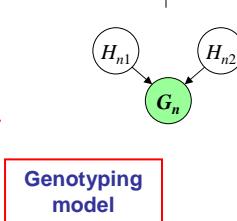
# Finite mixture model

- The probability of a genotype  $g$ :

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1, h_2) p(g | h_1, h_2)$$

Population haplotype pool

Haplotype model



- Standard settings:

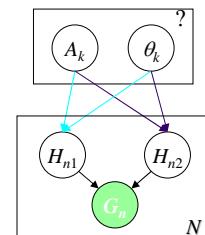
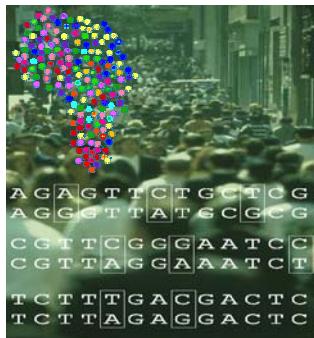
- $p(h1, h2) = p(h1)p(h2)$  Hardy-Weinberg equilibrium
- $|H| = K$  fixed-sized population haplotype pool

- Problem:  $K?$   $H?$

Eric Xing

16

## Ancestral Inference



Essentially a clustering problem, but ...

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities

Eric Xing

17

## Being Bayesian about ...

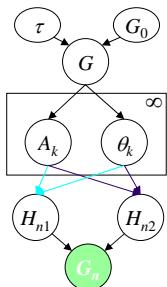
- Population haplotype identities
- Population haplotype frequencies
- Number of population haplotypes
- Associations between population haplotype and individual haplotype/genotype

Eric Xing

18

## A Hierarchical Bayesian Infinite Allele model

*Bayesian Haplotype Inference via the Dirichlet Process (Xing et al. ICML2004)*



- Assume an individual haplotype  $h$  is stochastically derived from a population haplotype  $a_k$  with nucleotide-substitution frequency  $\theta_k$ :

$$h \sim p(h | \{a, \theta\}_k).$$

- Not knowing the correspondences between individual and population haplotypes, each individual haplotype is a mixture of population haplotypes.

- The number and identity of the population haplotypes are unknown

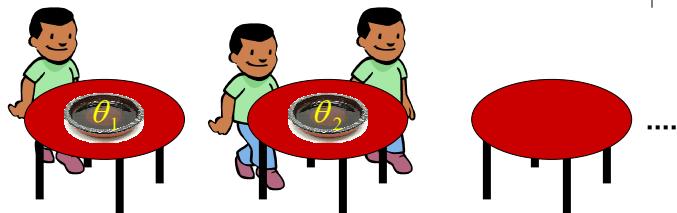
– use a **Dirichlet Process** to construct a prior distribution  $G$  on  $\mathcal{H} \times \mathcal{R}^I$ .

- Inference: Markov Chain Monte Carlo

Eric Xing

19

## Chinese Restaurant Process



$$P(c_i = k | \mathbf{c}_{-i}) = \begin{array}{ccc} \frac{1}{1+\alpha} & \frac{0}{1+\alpha} & \frac{0}{1+\alpha} \\ \frac{1}{2+\alpha} & \frac{1}{2+\alpha} & \frac{\alpha}{2+\alpha} \\ \frac{1}{3+\alpha} & \frac{2}{3+\alpha} & \frac{\alpha}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} & \frac{m_2}{i+\alpha-1} & \dots \end{array}$$

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of integers

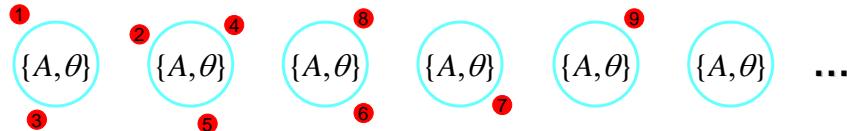
Eric Xing

20

## The DP Mixture of Ancestral Haplotypes



- The customers around a table form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample  $\{a, \theta\}$  at each table from a base measure  $G_0$  to obtain the population haplotype and nucleotide substitution frequency for that component

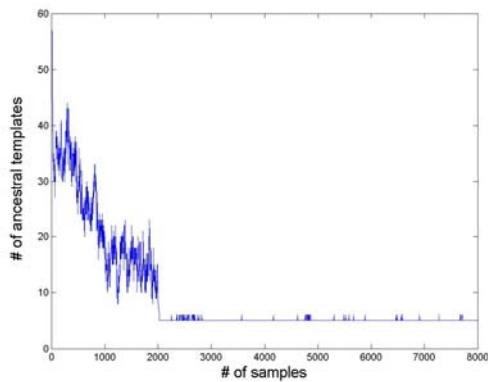


- With  $p(h/\{A, \theta\})$  and  $p(g/h_1, h_2)$ , the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)

Eric Xing

21

## Convergence of Ancestral Inference

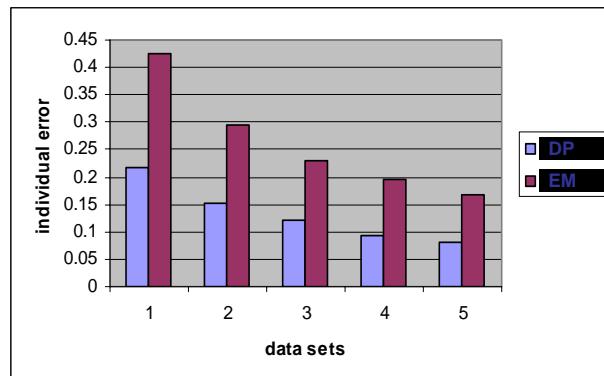


Eric Xing

22

## Results on simulated data

- DP vs. Finite Mixture via EM

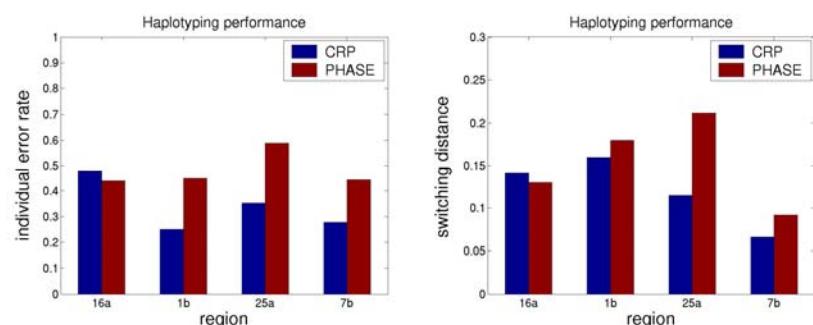


Eric Xing

23

## Results

### The Gabriel data

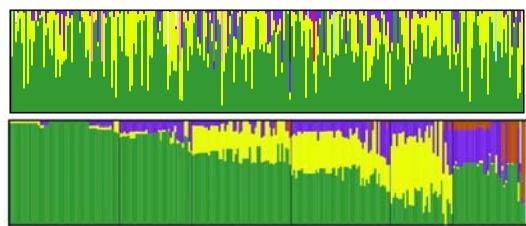


Eric Xing

24

## Population structure

- DATA: 256 European individuals with 103 loci



Eric Xing

25

## Computation Biology and ML

- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction



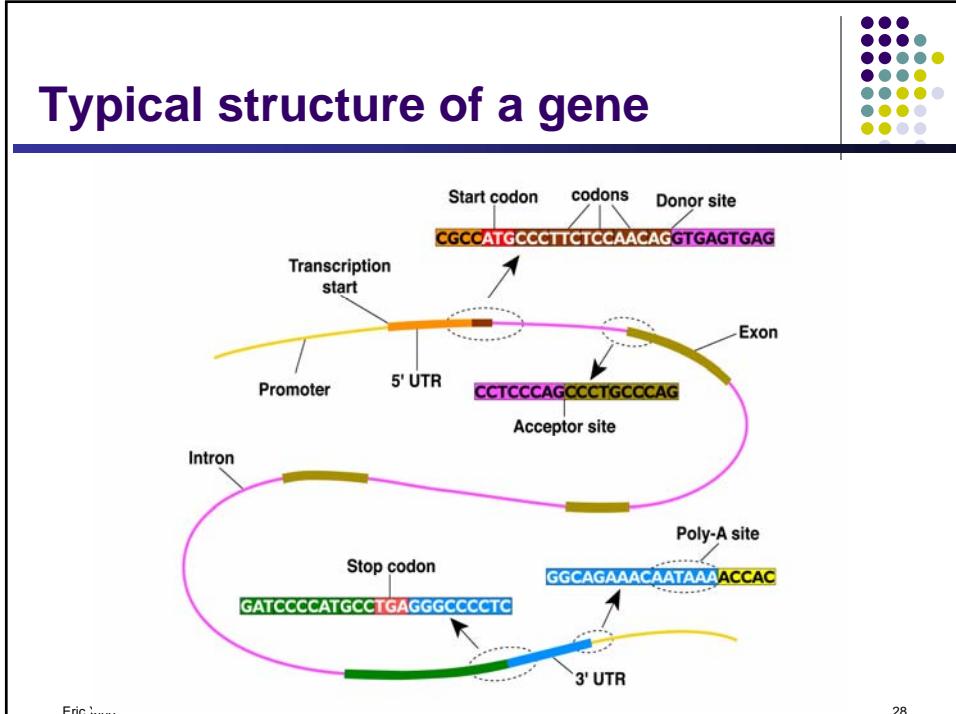
Eric Xing

26

The challenge

Eric Xing

## Typical structure of a gene



# Gene Finding

- Given un-annotated sequences,
  - delineate:
    - transcription initiation site,
    - exon-intron boundaries,
    - transcription termination site,
    - a variety of other motifs: promoters, polyA sites, branching sites, etc.
  - The hidden Markov model (HMM)

Eric Xing

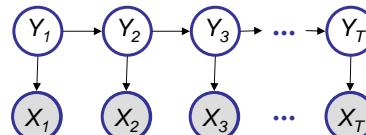
29

# Hidden Markov Models

## The underlying source: genomic entities, dice,

### The sequence:

## Ploy NT, sequence of rolls

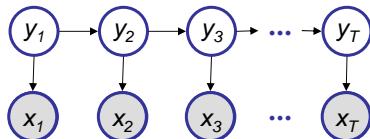


Eric Xing

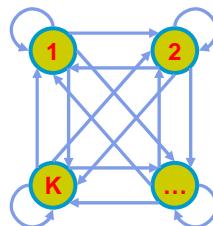
30

## Definition (of HMM)

- **Observation space**
    - Alphabetic set:  $C = \{c_1, c_2, \dots, c_K\}$
    - Euclidean space:  $\mathbb{R}^d$
  - **Index set of hidden states**  
 $I = \{1, 2, \dots, M\}$
  - **Transition probabilities** between any two states  
 $p(y_t^j = 1 | y_{t-1}^i = 1) = a_{i,j}$ ,  
or  $p(y_t^i | y_{t-1}^j = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \dots, a_{i,M}), \forall i \in I.$
  - **Start probabilities**  
 $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M)$ .
  - **Emission probabilities** associated with each state  
 $p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \dots, b_{i,K}), \forall i \in I.$   
or in general:  
 $p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in I.$



## Graphical model

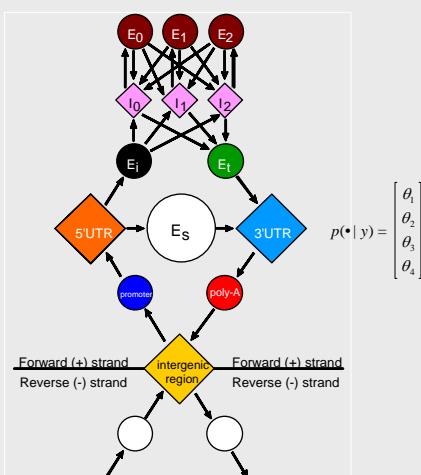


## State automata

Eric Xing

31

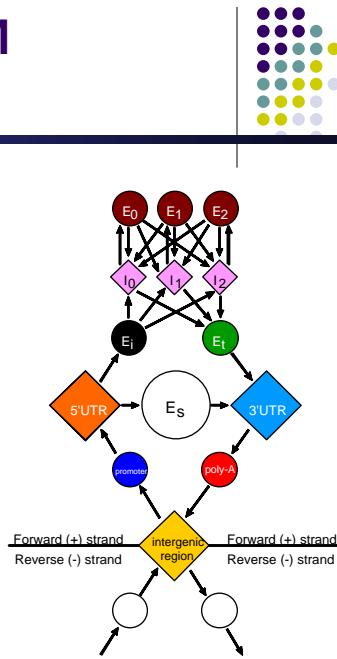
## GENSCAN (Burge & Karlin)



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

## 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.



Eric Xing

33

## The HMM Algorithms

### Questions:

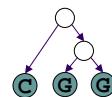
- **Evaluation:** What is the probability of the observed sequence? **Forward**
- **Decoding:** What is the probability that the state of the 3rd position is B<sub>k</sub>, given the observed sequence? **Forward-Backward**
- **Decoding:** What is the most likely parsing? **Viterbi**
- **Learning:** Under what parameterization are the observed sequences most probable? **Baum-Welch (EM)**

Eric Xing

34

# Computation Biology and ML

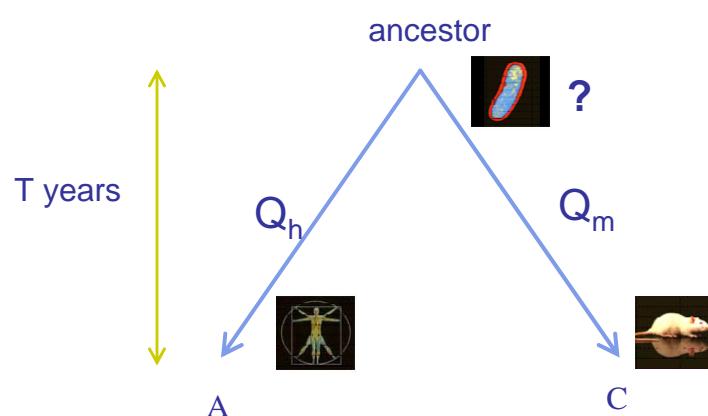
- Mixture and infinite mixture
  - clustering of genetic polymorphisms
- HMMs
  - gene finding
- Trees
  - sequence evolution
- CRMs
  - protein structure prediction



Eric Xing

35

## A pair of homologous bases



Typically, the ancestor is unknown.

Eric Xing

36

# Homology identification via multiple alignment

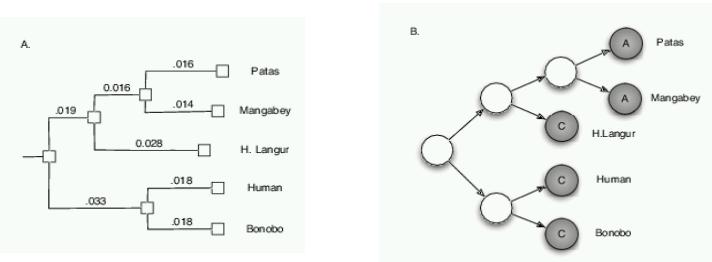


	10	20	30	40	50	60
consensus	1 SAAQKAIVKASUHKVG-----NRELGAEILARLTK-----AYPDTKXPF	46				
IASH	1 ANKIRECLMESLERAVAL-----EARODGIDLYKHE-----NYPPLRKYKS	R-E	49			
ITTH A	3 TAAQIKAQIADPHFLNLUKG-----CLOQAADSIFPKYLT-----AYPDLAFFFH	F-S	48			
g1_1065933	162 DKESCEVVAADSURLVSSseasTSACFELFVEQRYVS-----KIPMLNPLFG	L-S	212			
g1_3877400	71 QWEEKEELIIRFTMSEDEID-----NLVELGSAIYCYIAT-----HNPNCKQOLFP	F-S	115			
g1_3877381	15 TDIEETVTDVVRRA-----KTDNVGKKILQTLIE-----KRPKFAEYFG	Isgf	58			
g1_3874505	230 TCAQIHLIQLRLRQVLT-----KPTVIGASIVHRLFtrivwtk0MKQVE	DPFKFCN	283			
g1_4098133	39 EDPKHLRVLQNAFKL-----DDPELFRFYAHMIA-----LDA3VRLTPP	P---	79			
g1_1707914	18 SPADVK-----KHTYESMCAVW-----DKAQNGIDFYKEFTT-----HHHDOLRKFFKGKA-E	65				
g1_2494780	3 TKEDEFDSLHEIDPKLDe-----eHRMELLGQATELLTA-----AHPEYIKRFSR	L-Q	50			
	70	80	90	100	110	120
consensus	47 LSTAAAALKSSPKFKAHGKVVLGALDEAVKHL-----DDGNGKAALAKLGAP	HAKRG	--H	99		
IASH	50 EYTAEDVWNDFPFAKGOKILLACHYVLCATY-----DDRETINAYTRELLDR	HARDH	--H	103		
ITTH A	49 SVPLYGLRSNPAYKAQTLTVINYLDKVDAL-----GGAGAGLMKAKVPS	HDAMG	--	98		
g1_1065933	213 SDPVFDLJDNHHPVERAHLITSLHLSKQVQ-----ELAQAVAPTVFKYGER	HYRPD	Aip	269		
g1_3877400	116 KXQGEDEKESKEERSALKIVYVTLAQWKMVHMPERTSFLYMVGOKHVK	FADRG	--	170		
g1_3877381	59 SJLIRALNOSKEFRLCAHRJQFLQTAQG3GJL-----CCPISSEVDMAHRIGQI	HFYRG	-N	114		
g1_3874505	284 -----RDNFTRKAELIDQVEMI-----DHLDWVTGELMRIGRV	HAKVL	--	327		
g1_4098133	80 -----DMGAQAAATGQA_HUVV_GELVAQ-----AEPVPAFLAQLCD	HRKYO	--	122		
g1_1707914	66 NFGADDVVK3SKRFKGTALLLAHVHHLANVY-----DNQVHFGFVRELMNR	HEKRG	Dok	121		
g1_2494780	51 EATPANTYMAQDGAKYYAKTILINDLVELLKAS-----DTEATLNTALARTATK	HUKRN	--	103		
	130	140	150	160	170	
consensus	100 VDPAHFLKLGEALL-----VVALEHIL-----DFTPEVKAQAUAKALDVVADAKLSGYR	CYR	147			
IASH	104 MPPEVDTDFMKLFE-----EYLGKKT-----TLDPEPTKQAUHEIAGRFAKEIMKRG	R	150			
ITTH A	99 ITPKHFGOLLKLVC-----GVFOEEF-----ADPTTVAAGCAGA GVVAAMK	--	141			
g1_1065933	270 MTEENWRVFCAOIV-----CTVFPFJ-----EATPKCAE SUIELMEVYLGOKLJLGED	319				
g1_3877400	171 FKHETYDIFODANE-----FALEHRIsmtbLDMDNQERDATVWRTIALYVH	221				
g1_3877381	115 FGADMWLVFKVTV-----DOVITGIC-----DSSKEKED	tsmnotandkwdtdas11	162			
g1_3874505	328 RGELTGKLUNTVTAet1DCTLLEWdg-----tCSSETVRKAWALIVAFVLEKIKAGHH	380				
g1_4098133	123 VLPQTQYDTLFRALY-----TTLRDXL-----HPSRGAWTDAVDEAAQSLNLI	167				
g1_1707914	122 LWKIFDDVVVPFL-----ESEKAKL-----GDAKAIAKELNKNNFSEAQHOL	E	166			
g1_2494780	104 VSGAEQFQTGEPIFI-----KYSFHSV-----TT PANQA FMKEL LTKIFTGVAGQL	148				

Eric Xing

37

# Phylogeny



- The shaded nodes represent the observed nucleotides at a given site for a set of organisms
- The unshaded nodes represent putative ancestral nucleotides
- Transitions between nodes capture the dynamic of evolution

Eric Xing

38



## Phylogeny methods

- Basic principles:

- Degree of sequence difference is proportional to length of independent sequence evolution
- Only use positions where alignment is pretty certain – avoid areas with (too many) gaps

- Major methods:

- Parsimony phylogeny methods
- Likelihood methods

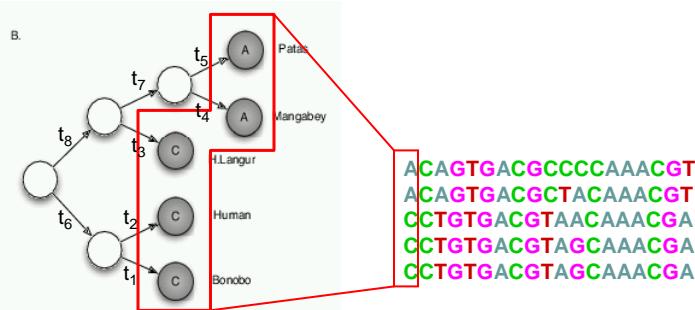
Eric Xing

39



## Likelihood methods

- A tree, with branch lengths, and the data at a single site.



- Since the sites evolve independently on the same tree,

$$L = P(D | T) = \prod_{i=1}^m P(D^{(i)} | T)$$

Eric Xing

40

## Likelihood at one site on a tree

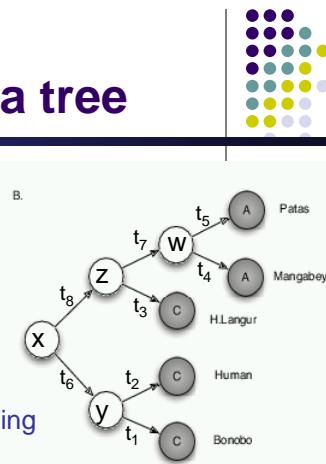
- We can compute this by summing over all assignments of states  $x, y, z$  and  $w$  to the interior nodes:

$$P(D^{(i)} | T) = \sum_x \sum_y \sum_z \sum_w P(A, A, C, C, C, x, y, z, w | T)$$

- Due to the Markov property of the tree, we can factorize the complete likelihood according to the tree topology:

$$\begin{aligned} P(A, A, C, C, C, x, y, z, w | T) = \\ P(x) & P(y | x, t_6) & P(C | y, t_1) P(C | y, t_2) \\ P(z | x, t_8) & P(C | z, t_3) \\ & P(w | z, t_7) P(A | w, t_4) P(A | w, t_5) \end{aligned}$$

- Summing this up, there are 256 terms in this case!



41

## Getting a recursive algorithm

- when we move the summation signs as far right as possible:

$$P(D^{(i)} | T) = \sum_x \sum_y \sum_z \sum_w P(A, A, C, C, C, x, y, z, w | T) =$$

$$\begin{aligned} & \sum_x P(x) \\ & \left( \sum_y P(y | x, t_6) P(C | y, t_1) P(C | y, t_2) \right) \\ & \left( \sum_z P(z | x, t_8) P(C | z, t_3) \right. \\ & \quad \left. \left( \sum_w P(w | z, t_7) P(A | w, t_4) P(A | w, t_5) \right) \right) \end{aligned}$$

Eric Xing

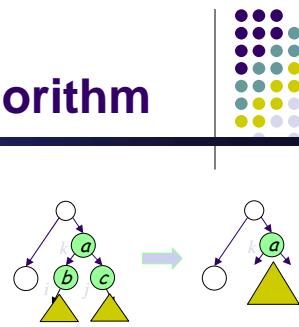
42

## Felsenstein's Pruning Algorithm

- To calculate  $P(x_1, x_2, \dots, x_N | T, t)$

Initialization:

Set  $k = 2N - 1$



Recursion: Compute  $P(L_k | a)$  for all  $a \in \Sigma$

If  $k$  is a leaf node:

Set  $P(L_k | a) = 1(a = x_k)$

If  $k$  is not a leaf node:

1. Compute  $P(L_i | b), P(L_j | b)$  for all  $b$ , for daughter nodes  $i, j$

2. Set  $P(L_k | a) = \sum_{b, c} P(b | a, t_i) P(L_i | b) P(c | a, t_j) P(L_j | c)$

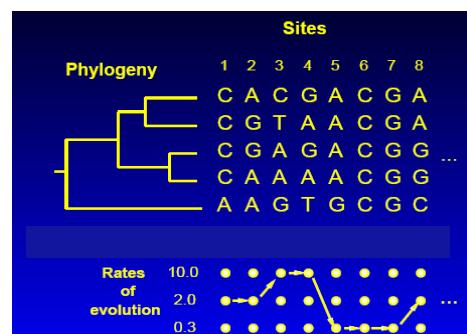
Termination:

Likelihood at this column =  $P(x_1, x_2, \dots, x_N | T, t) = \sum_a P(L_{2N-1} | a) P(a)$

Eric Xing

43

## Modeling rate variation among sites

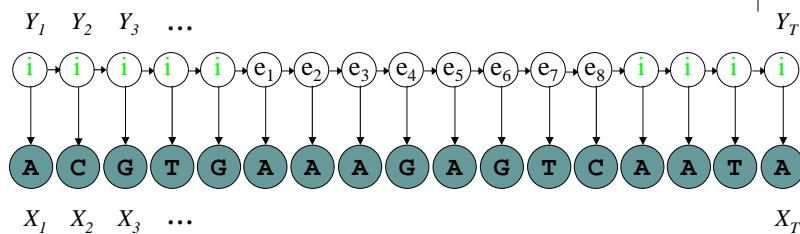


- There are a finite number of rates (denote rate  $i$  as  $r_i$ ).
- There are probabilities  $p_i$  of a site having rate  $i$ .
- A process not visible to us ("hidden") assigns rates to sites.
- The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting appropriately by their probabilities of occurrence.

Eric Xing

44

## Recall the HMM



- The shaded nodes represent the observed nucleotides at particular sites of an organism's genome
- For discrete  $Y_t$ , widely used in computational biology to represent segments of sequences
  - gene finders and motif finders
  - profile models of protein domains
  - models of secondary structure

Eric Xing

45

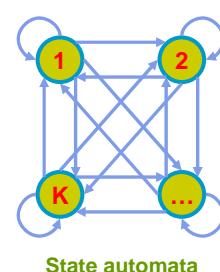
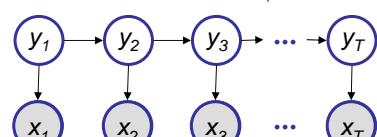
## Definition (of HMM)

- Observation space
 

Alphabetic set:  $C = \{c_1, c_2, \dots, c_K\}$   
Euclidean space:  $\mathbb{R}^d$
- Index set of hidden states  
 $I = \{1, 2, \dots, M\}$
- Transition probabilities between any two states

$$p(y_t^j = 1 | y_{t-1}^i = 1) = a_{i,j}, \text{ or } p(y_t | y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \dots, a_{i,M}), \forall i \in I.$$

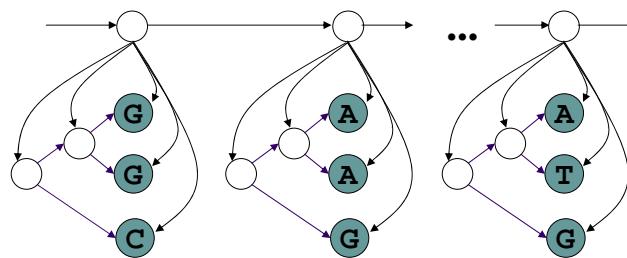
- Start probabilities  
 $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M)$ .
- Emission probabilities associated with each state  
 $p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \dots, b_{i,K}), \forall i \in I.$   
or in general:  
 $p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in I.$



Eric Xing

46

## Hidden Markov Phylogeny

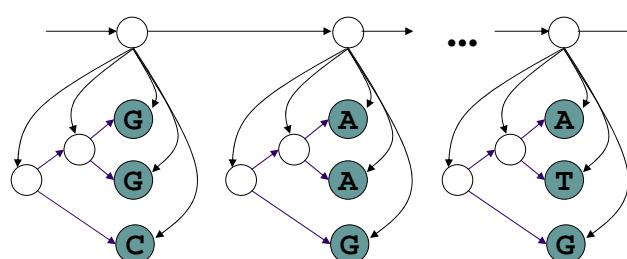


- Replacing the standard emission model with a tree
  - A process not visible to us (.hidden") assigns rates to sites. It is a Markov process working along the sequence.
  - For example it might have transition probability Prob ( $j|i$ ) of changing to rate  $j$  in the next site, given that it is at rate  $i$  in this site.
- These are the most widely used models allowing rate variation to be correlated along the sequence.

Eric Xing

47

## Hidden Markov Phylogeny



- this yields a gene finder that exploits evolutionary constraints

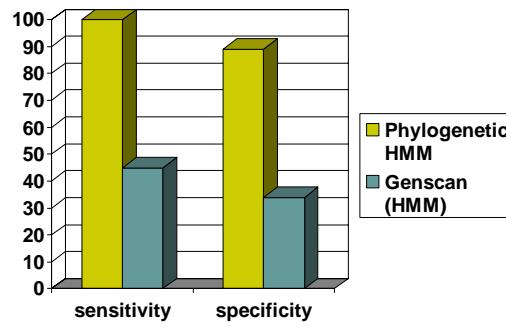
Eric Xing

48

## A Comparison of comparative genomic gene-finding and isolated gene-finding



- Based on sequence data from 12-15 primate species, McAuliffe et al (2003) obtained sensitivity of 100%, with a specificity of 89%.
  - Genscan (state-of-the-art gene finder) yield a sensitivity of 45%, with a specificity of 34%.



Eric Xing

49