

Machine Learning

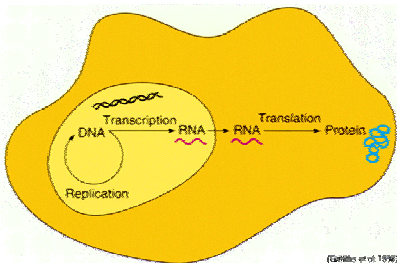
10-701/15-781, Fall 2006

Machine Learning in Computational Biology

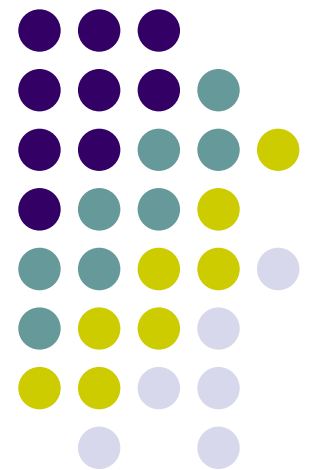
Eric Xing

Lecture 21, November 28, 2006

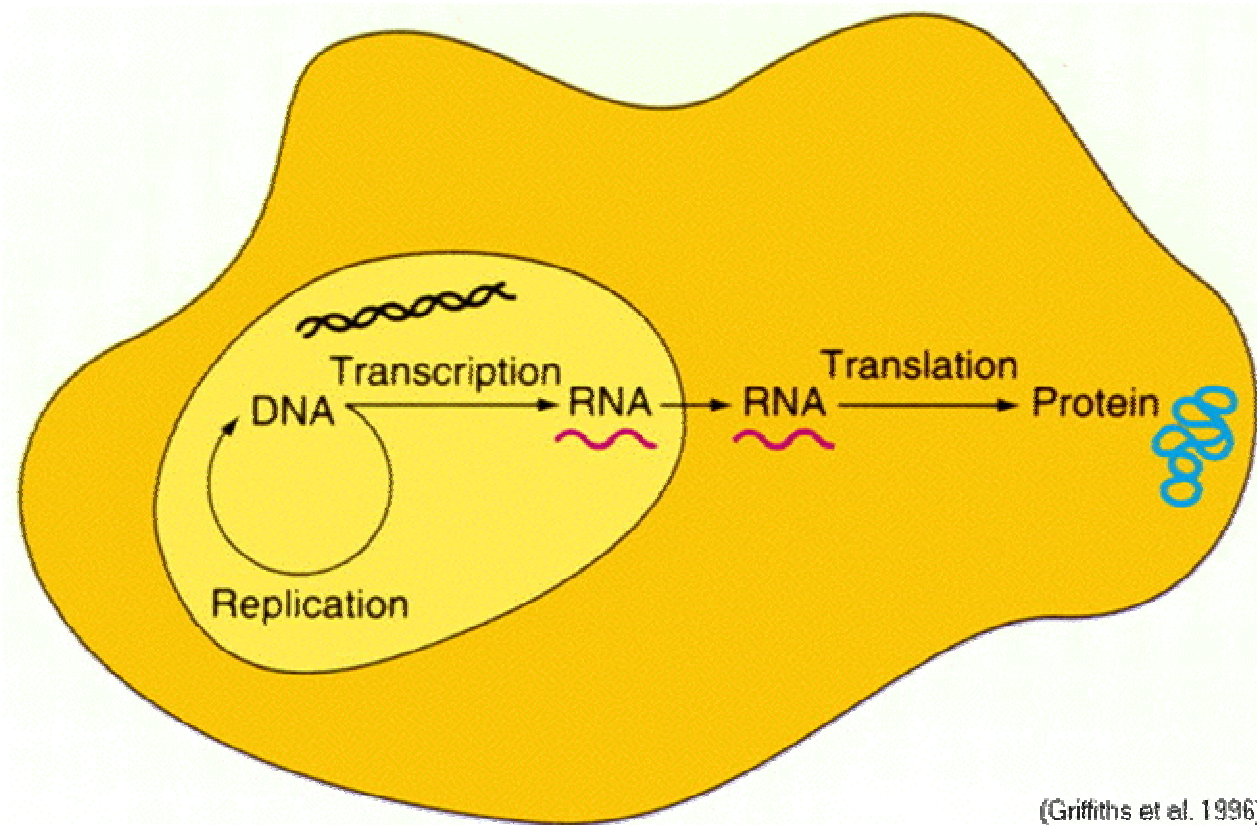
Reading:



Eric Xing

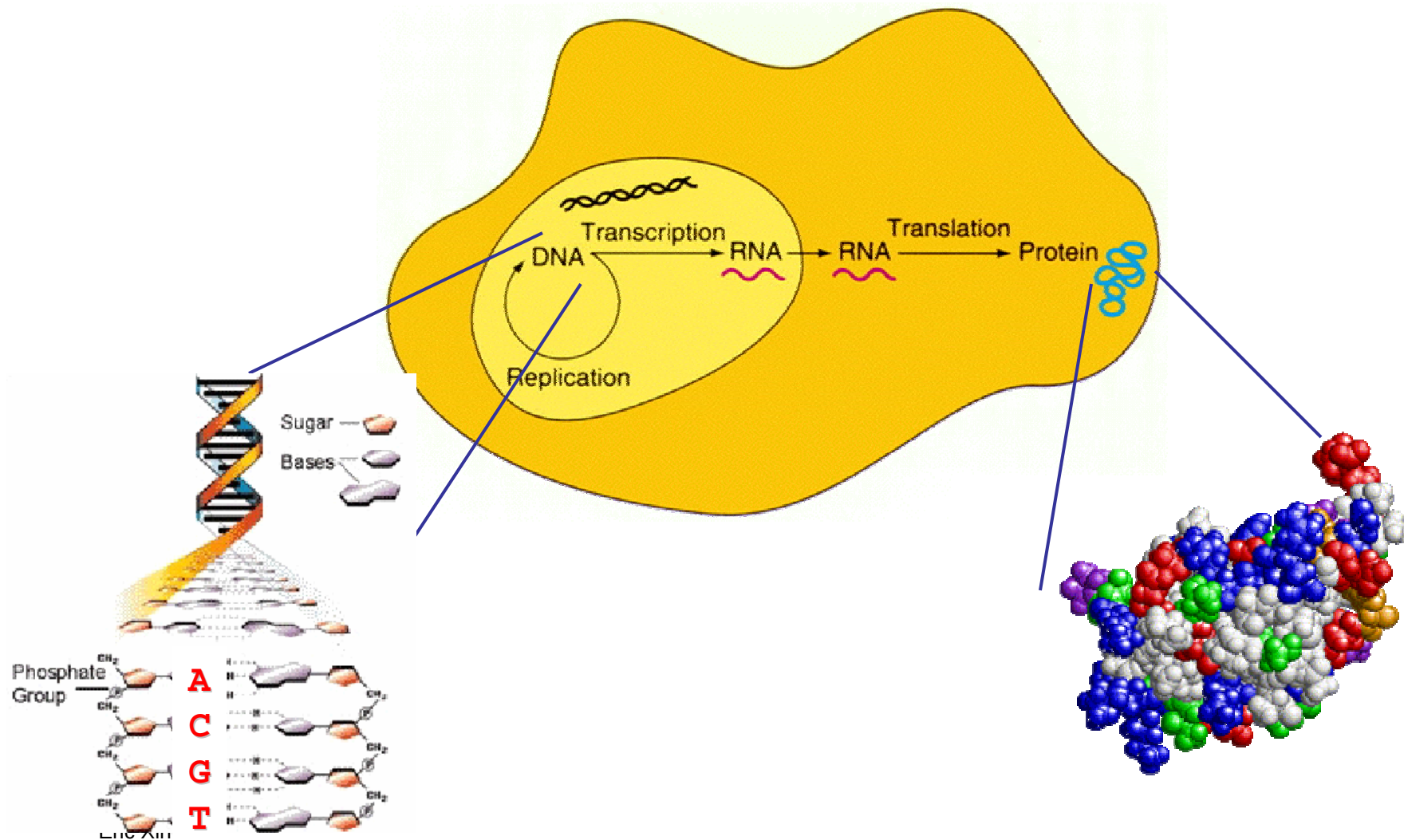


The Central Dogma

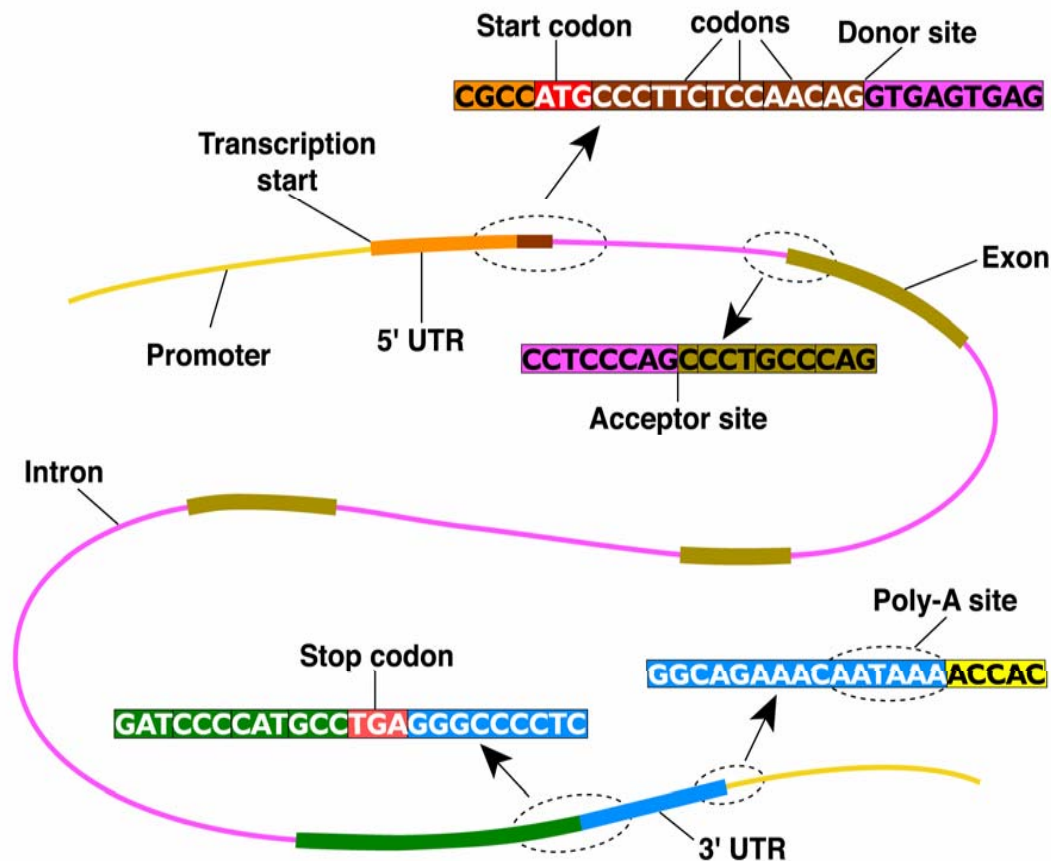


(Griffiths et al. 1996)

Genome and Proteome



Gene Structure in DNA

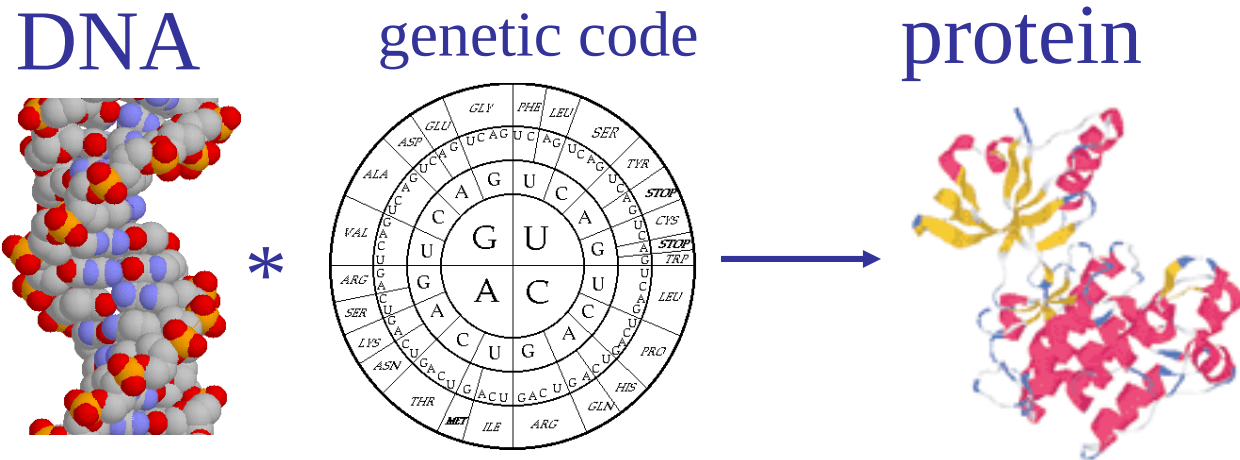


- The inference problem: predicting locations of the genes on DNA

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)

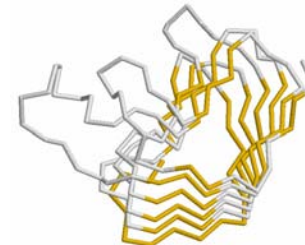
Protein Structure Hierarchy



Primary Structure	Secondary Structures	Tertiary Structures	Quaternary Structures
amino group carboxy group R H₂N-C-C-OH H α carbon ... LACA(A)EECS...	Alpha-helix Anti-parallel beta-sheet Parallel beta-sheet	Beta-helix	Triple beta-spiral

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

APAFSVSPASGACGPECA



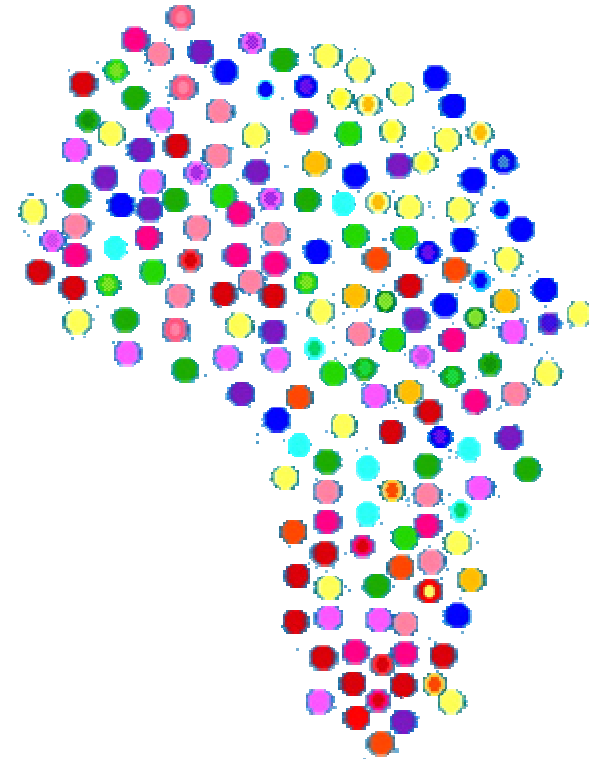
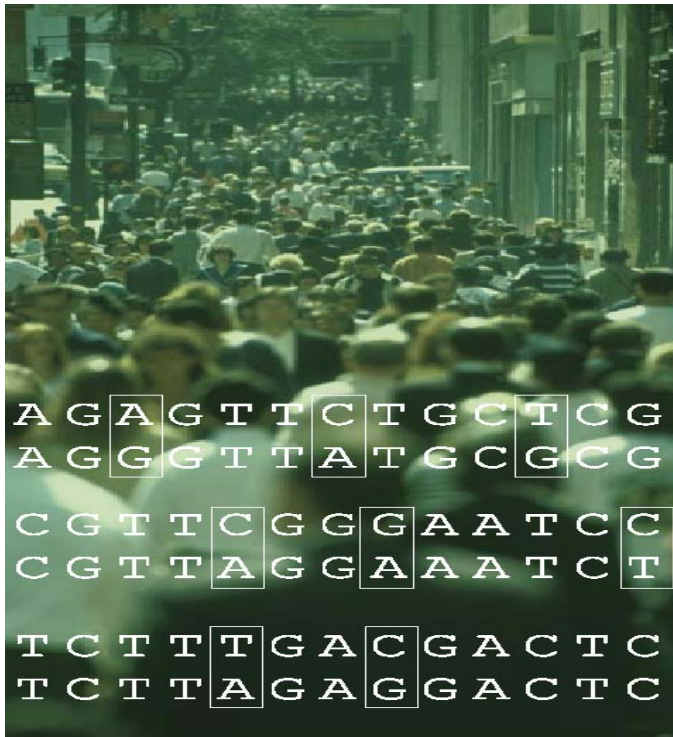
Genetic Polymorphisms



The ABO Blood System

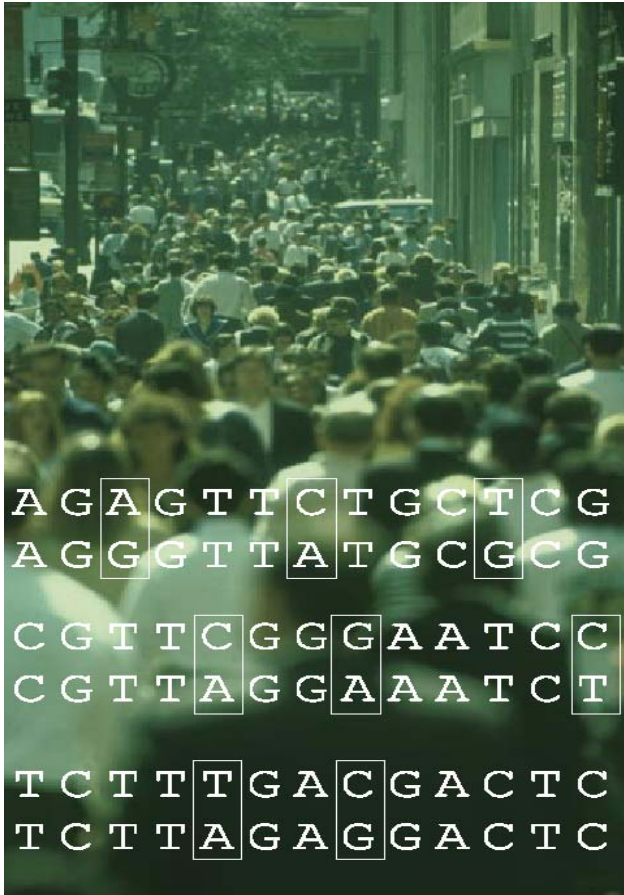
Blood Type (genotype)	Type A (AA, AO)	Type B (BB, BO)	Type AB (AB)	Type O (OO)
Red Blood Cell Surface Proteins (phenotype)	 A agglutinogens only	 B agglutinogens only	 A and B agglutinogens	 No agglutinogens
Plasma Antibodies (phenotype)	 b agglutinin only	 a agglutinin only	NONE. No agglutinin	 a and b agglutinin

Genetic Demography



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

Single Nucleotide Polymorphism (SNP)



...cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc...
...cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc...
...cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc...
...cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc...
...cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc...
...cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc...
...cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc...
...cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc...
...cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc...
...cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc...
...cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc...

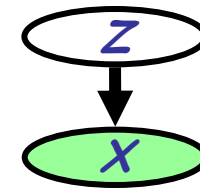


- The inference problem: "haplotypes" and population diversity

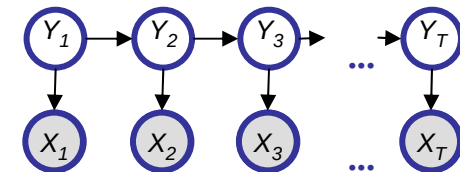
Computation Biology and ML



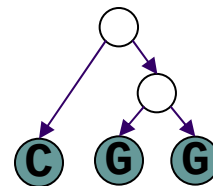
- Mixture and infinite mixture
 - clustering of genetic polymorphisms



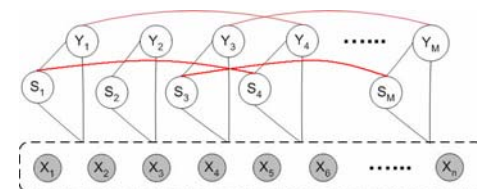
- Hidden Markov Models
 - gene finding



- Trees
 - sequence evolution



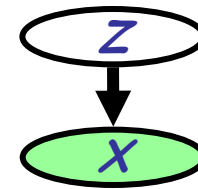
- Conditional Random Fields
 - protein structure prediction



Computation Biology and ML



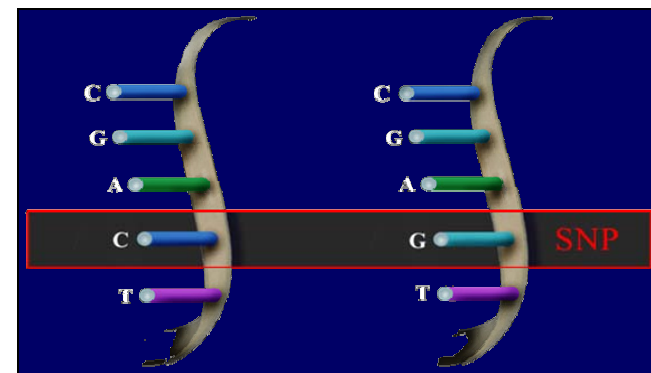
- Mixture and infinite mixture
 - clustering of genetic polymorphisms
- HMMs
 - gene finding
- Trees
 - sequence evolution
- CRMs
 - protein structure prediction

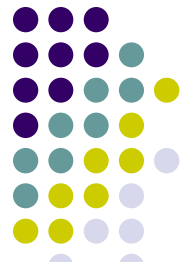


Biological Terms



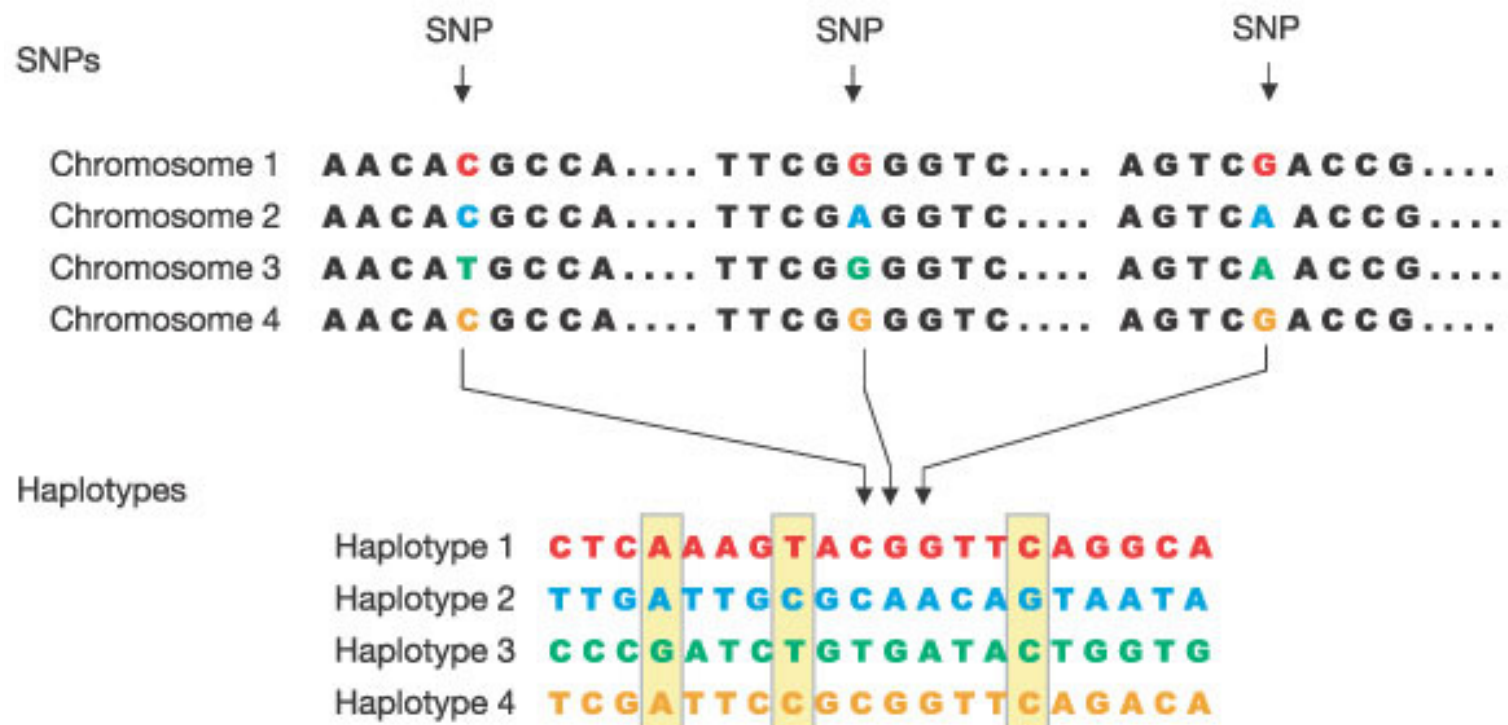
- **Genetic polymorphism**: a difference in DNA sequence among individuals, groups, or populations
- **Single Nucleotide Polymorphism (SNP)**: DNA sequence variation occurring when a single nucleotide - A, T, C, or G - differs between members of the species
 - Each variant is called an “allele”
 - Almost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





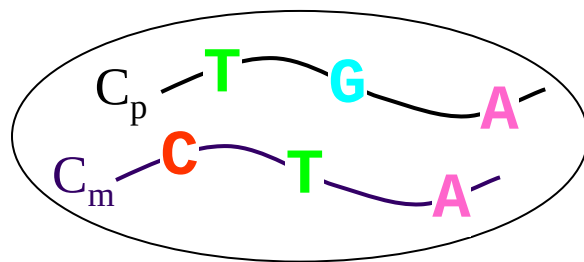
From SNPs to Haplotypes

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

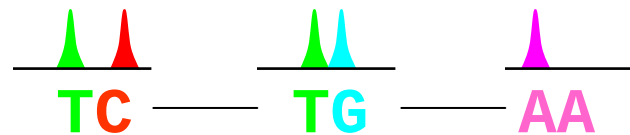
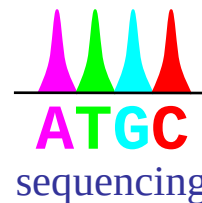


- Useful in the study of **disease association** or **genetic evolution**

Phase ambiguity of SNPs "haplotypes"

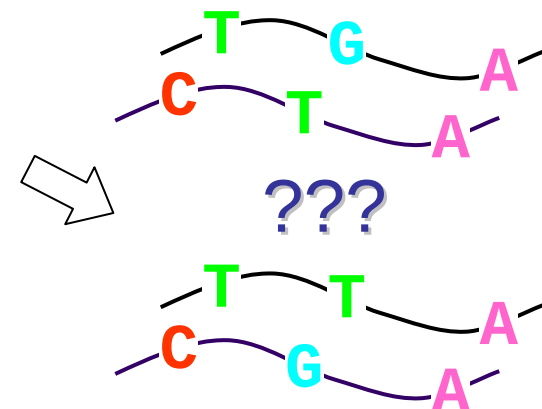


A heterozygous
diploid individual



The **Genotype**:
pairs of alleles with
association of alleles to
chromosomes unknown

- This is a mixture modeling problem!



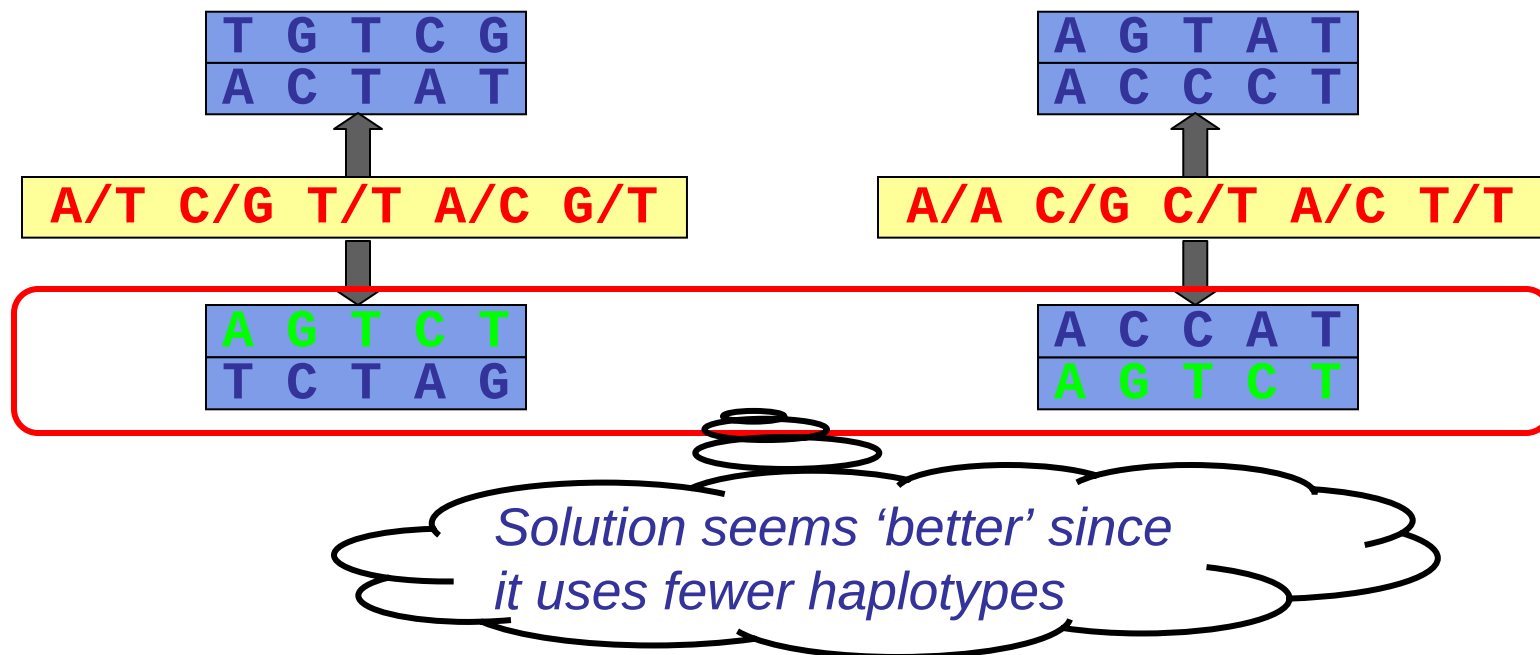
Haplotype $h \equiv (h_1, h_2)$
possible associations of
alleles to chromosome

Haplotype Inference

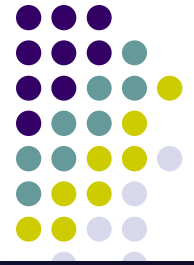


Why is it approachable?

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

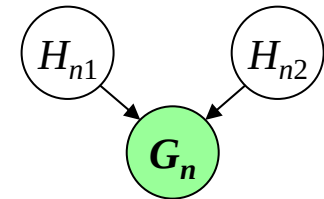


Finite mixture model



- The probability of a genotype g :

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



- Standard settings:

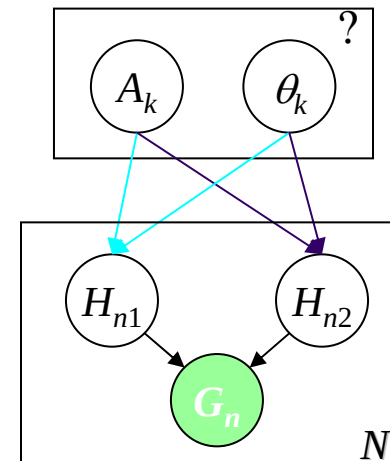
- $p(h_1, h_2) = p(h_1)p(h_2)$
- $|\mathcal{H}| = K$

Hardy-Weinberg equilibrium

fixed-sized population haplotype pool

- Problem: K ? $\mathcal{H}$?

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

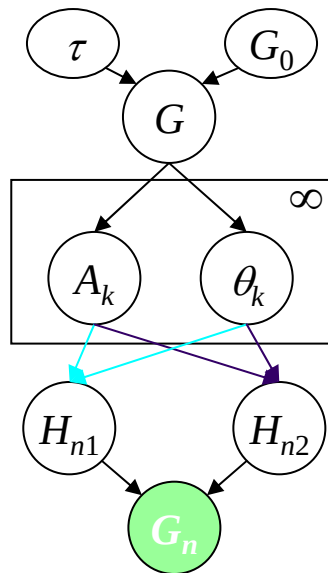
Being Bayesian about ...



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

A Hierarchical Bayesian Infinite Allele model

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



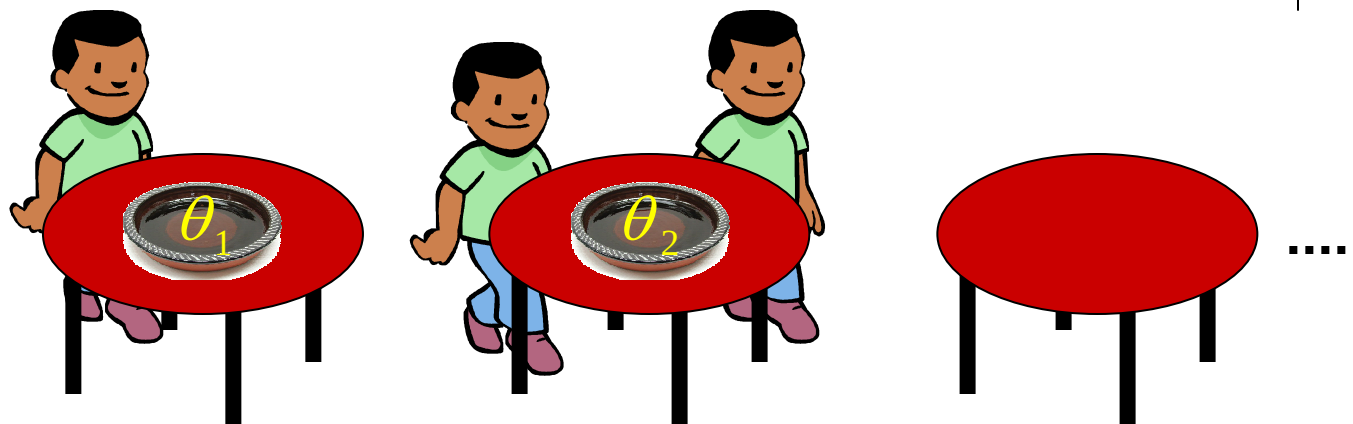
- Assume an individual haplotype h is stochastically derived from a population haplotype a_k with nucleotide-substitution frequency θ_k :

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

- 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 $\mathcal{H} \times \mathcal{R}^J$.
- Inference: Markov Chain Monte Carlo



Chinese Restaurant Process



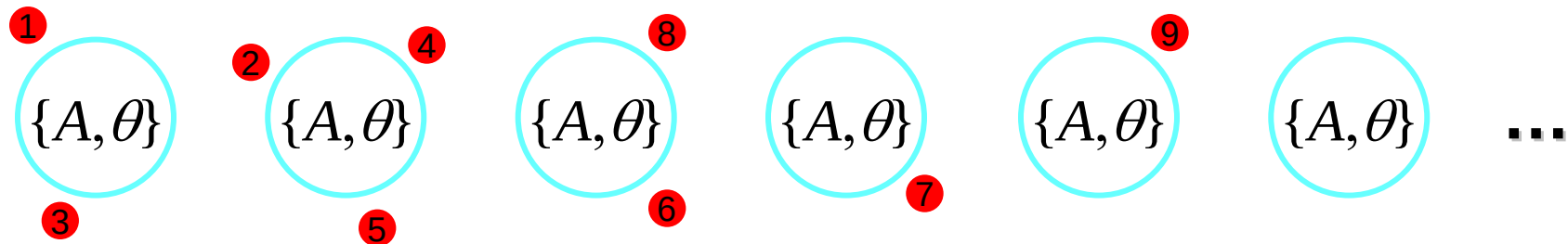
$$P(c_i = k | \mathbf{c}_{-i}) = \begin{array}{ccc} \frac{1}{1+\alpha} & \frac{0}{1+\alpha} & \frac{0}{1+\alpha} \\ \frac{1}{2+\alpha} & \frac{\alpha}{2+\alpha} & \frac{0}{2+\alpha} \\ \frac{1}{3+\alpha} & \frac{1}{3+\alpha} & \frac{\alpha}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} & \frac{m_2}{i+\alpha-1}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

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

The DP Mixture of Ancestral Haplotypes

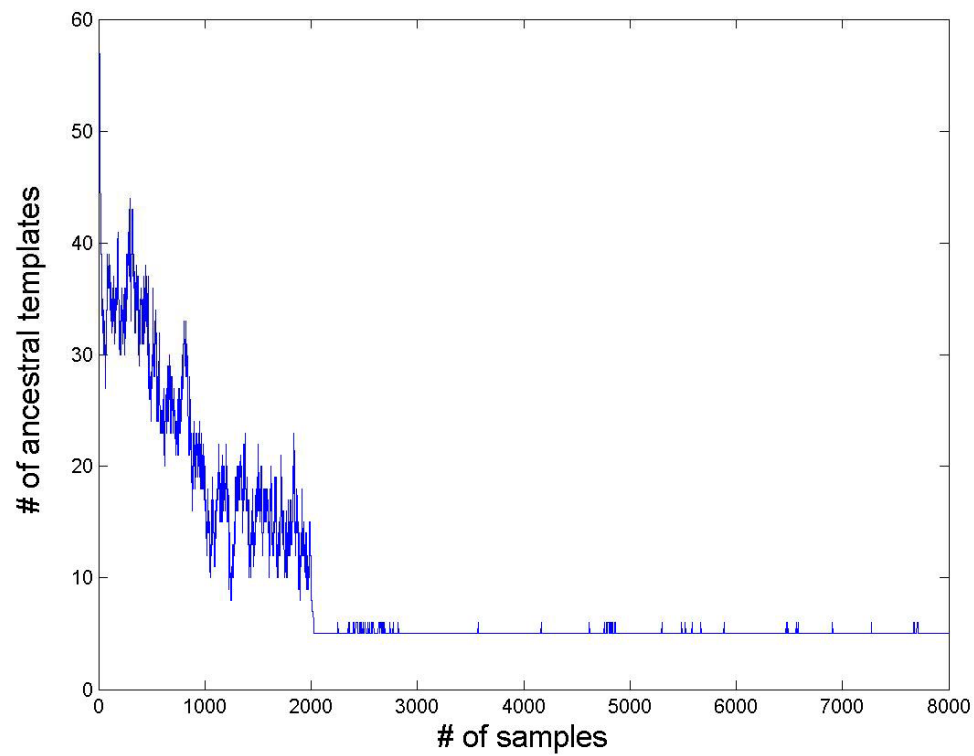


- The customers around a table form a cluster
 - associate a mixture component (*i.e.*, a population haplotype) with a table
 - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component



- 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)

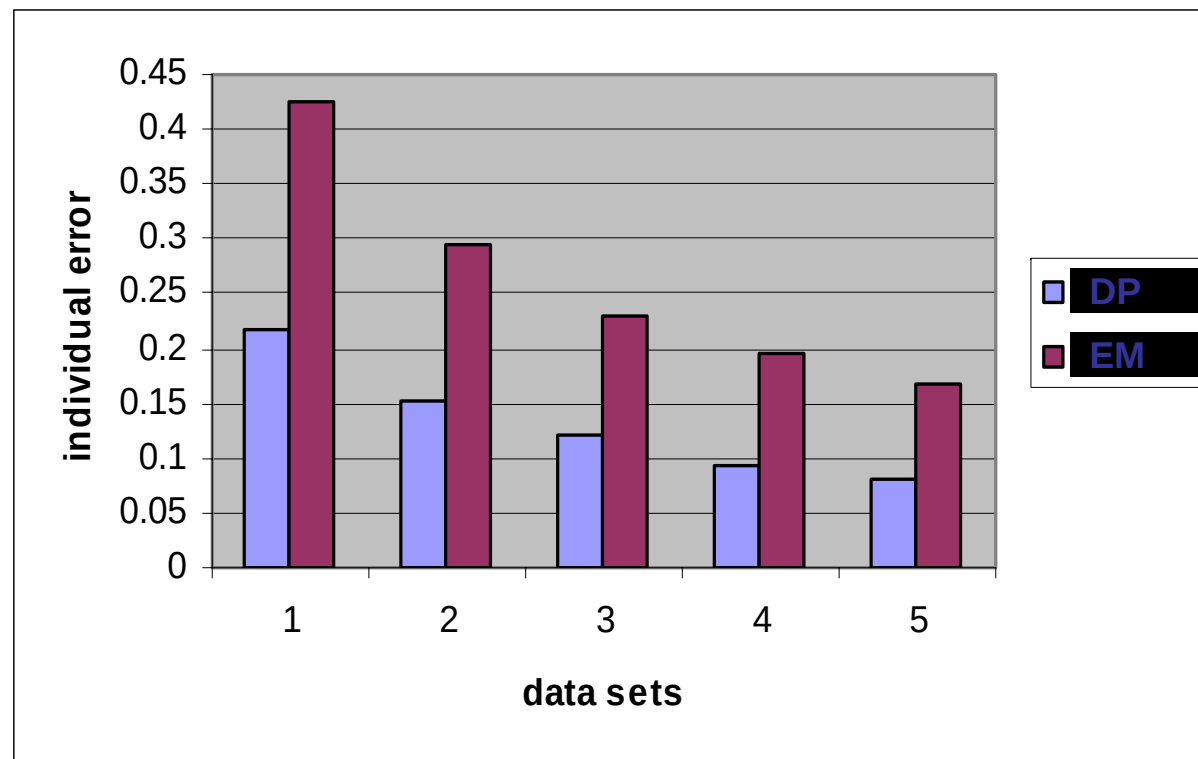
Convergence of Ancestral Inference





Results on simulated data

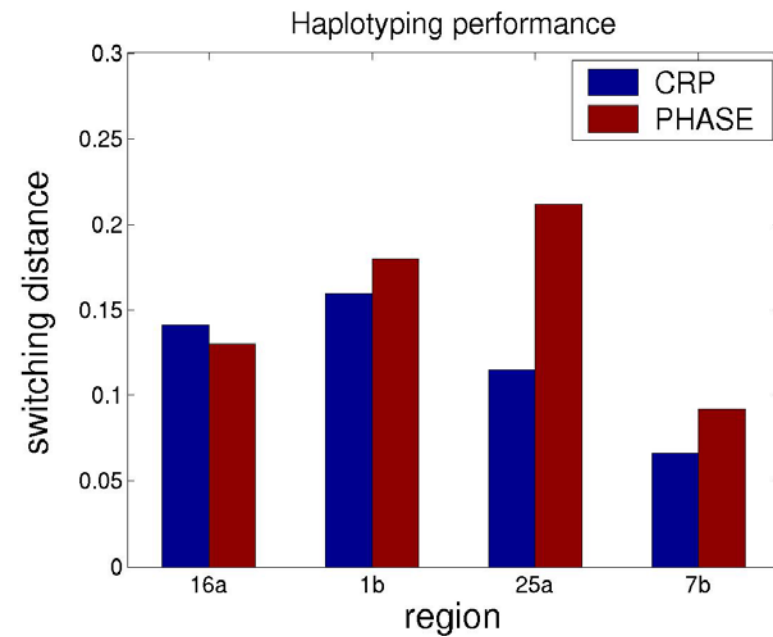
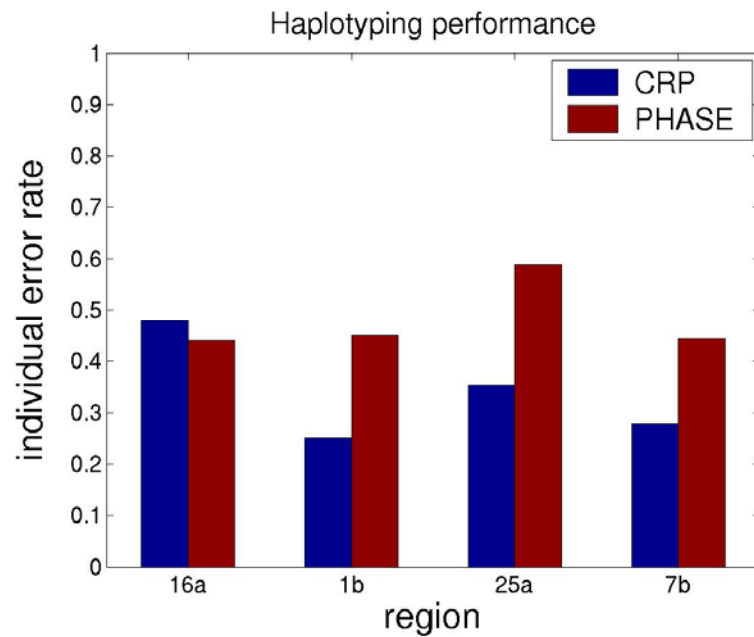
- DP vs. Finite Mixture via EM



Results



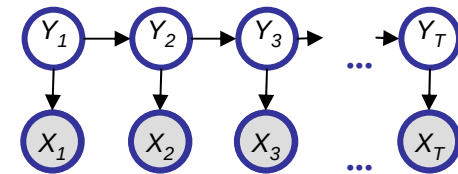
The Gabriel data



Computation Biology and ML



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

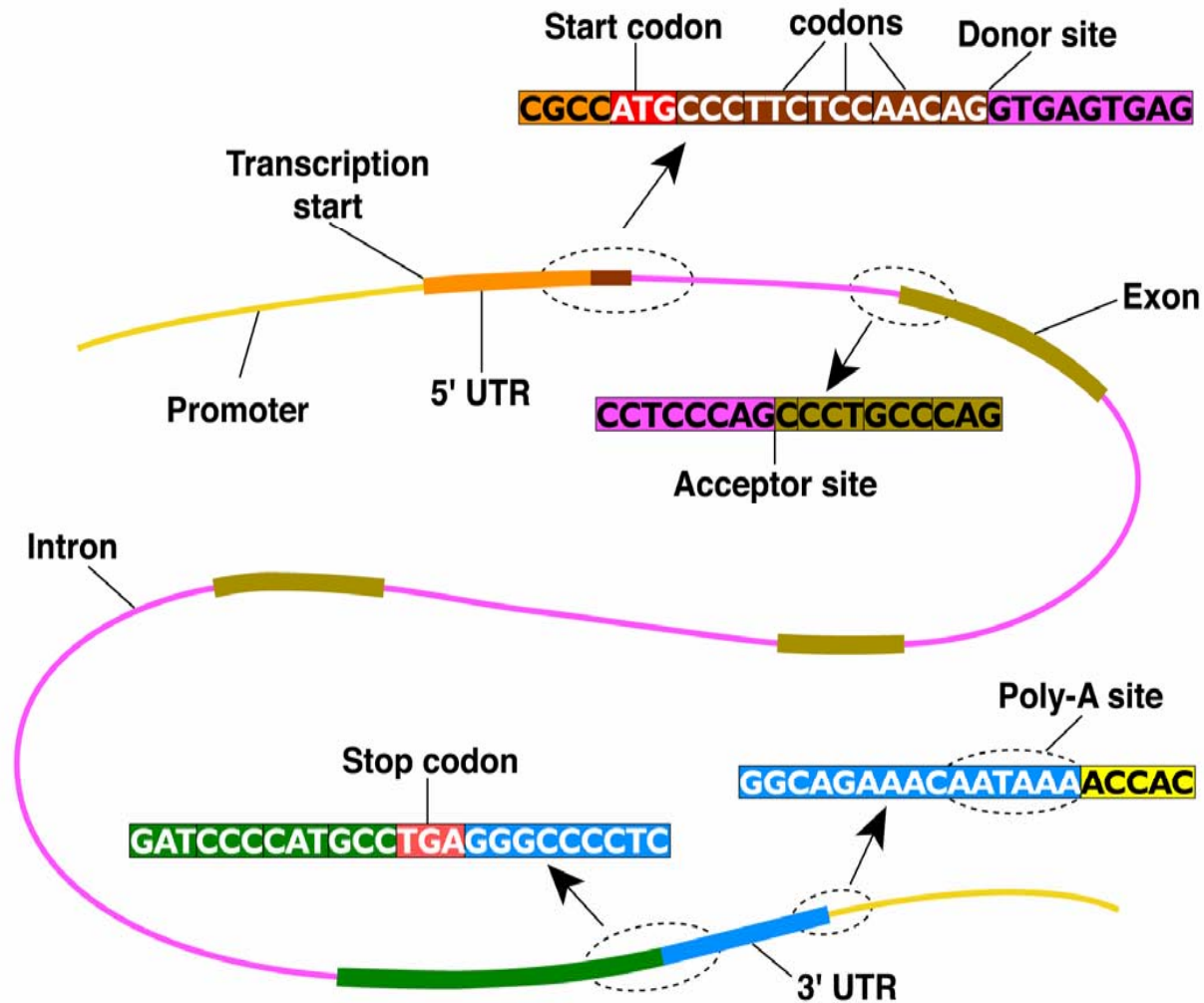


The challenge

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)

```

GAGAACGTGTGAGAGAGAGGCAAGCCGAAAAATCAGCCGC
CGAAGGATACACTATCGTCGTCCTTGTCGACGAACCGGT
GGTCATGCAAAACAACGCACAGAACAAATTAATTTTCAAAT
TGTTCAATAAATGTCCCACTTGCTTCTGT GTTCCCCCT
TTCCGCTAGCGGAATTTTATATTCTTT
[REDACTED]
[REDACTED] ATACACAGCGCACACAT
ATAAGCTTGCACACTGATGCACACACCGACACGTTGTC
ACCGAAATGAACGGGACGGCCATATGACTGGCTGGCGCTC
GGTATG GGGTGCAAGCGAGATACCGGATCAAGACTCGA
ACGAGACGGGTGAGCGAGTGATACCGATTCTCTCTTTT
GCGATTGGGAATAATGCCGACTTTTACACTACATGCGT
TGGATCTGGTTATTTAATTATGCCATTTTCTCAGTATAT
CGGCAATTGGTTGCATTAATTTGCCGCAAAGTAAGGAAC
ACAAACCGATAGTTAAGATCCAACG CCCTGCTGCCCTC
GCGTGCACAATTTGCGCAATTTCCCCCTTTTCCAGTTT
TTTTCAACCCAGCACCGCTCGTCTCTTCTCTTCTTAACG
TTAGCATTCGTACGAGGAACAGTGCTGTCATTGTGGCCGC
TGTGTAGCTAAAAAGCGTAATTATTCATTATCTAGCTATC
TTTTCGGATATTATTGTCATTTGCCCTTAATCTTGTGTAT
TTATATGGATGAAACGTGCTATAATAACAATGCAGAATGA
AGAACTGAAGAGTTTCAAAACCTAAAAATAATTGGAATAT
AAAGTTTGGTTTTACAATTTGATAAACTCTATTGTAAGT
GGAGCGTAACATAGGGTAGAAAACAGTGCAATCAAAGTA
CCTAAATGGAATACAAATTTTAGTTGTACAATTGAGTAAA
ATGAGCAAAGCGCCTATTTTGGATAAATTTGCTGTTTAC
AAGGGGAACATATTCATAATTTTCAGGTTTAGGTTACGCA
TATGTAGGCGTAAAGAAATAGCTATATTTGTAGAAGTGCA
TATGCACTTTATAAAAAATTATCCTACATTAAACGTATTTT
ATTTGCTTTAAACCTATCTGAGATATTTCAATAAGGTAA
GTGCAGTAATACAATGTAATAATTGCAATAATGTTGTA
ACTAAATACGTAAACAATAATGTAG AGTCCGGCTGAAAG
CCCCAGCAGCTATAGCCGATATCTATATGATTTAACTCT
TGTCTGCAACGTTCTAATAAAATAAAATGCAAAATAT
AACCTATTAGACAATACATTTATTTTATTTTATATC
ATCAATCATCTACTGATTTCTTTCGGTGTATCGCCTAATC
CATCTGTGAAATAG AAATGGCGCCACCTAGGTT AGAAAA
GATAAACAGTTGCCTTTAGTTGCATGACTTCCCGCTGGAT
    
```



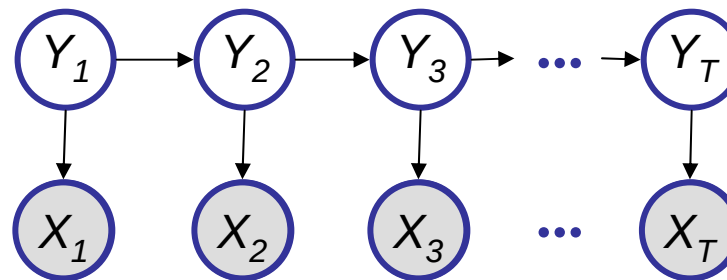
Hidden Markov Models

The underlying source:

genomic entities,
dice,

The sequence:

Ploy NT,
sequence of rolls,



Definition (of HMM)



- **Observation space**

Alphabetic set:

$$\mathbb{C} = \{c_1, c_2, \dots, c_K\}$$

Euclidean space:

$$\mathbb{R}^d$$

- **Index set of hidden states**

$$\mathbb{I} = \{1, 2, \dots, M\}$$

- **Transition probabilities** between any two states

$$p(y_t^j = 1 | y_{t-1}^i = 1) = a_{i,j},$$

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

- **Start probabilities**

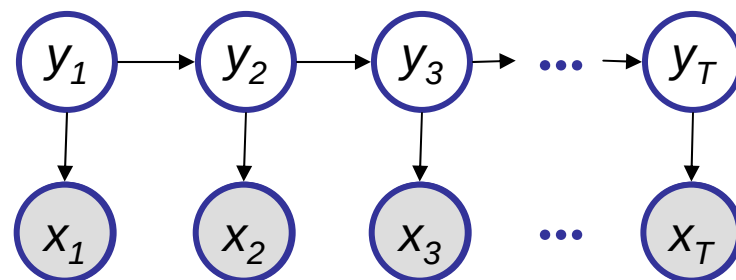
$$p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M).$$

- **Emission probabilities** associated with each state

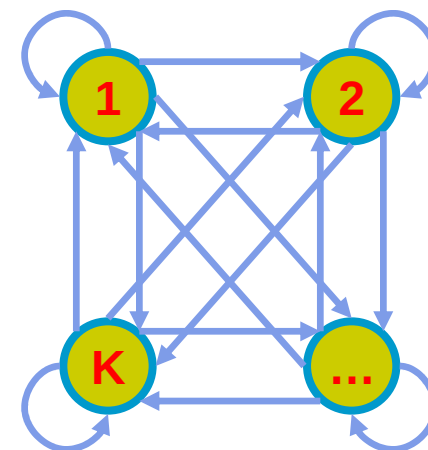
$$p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \dots, b_{i,K}), \forall i \in \mathbb{I}.$$

or in general:

$$p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in \mathbb{I}.$$

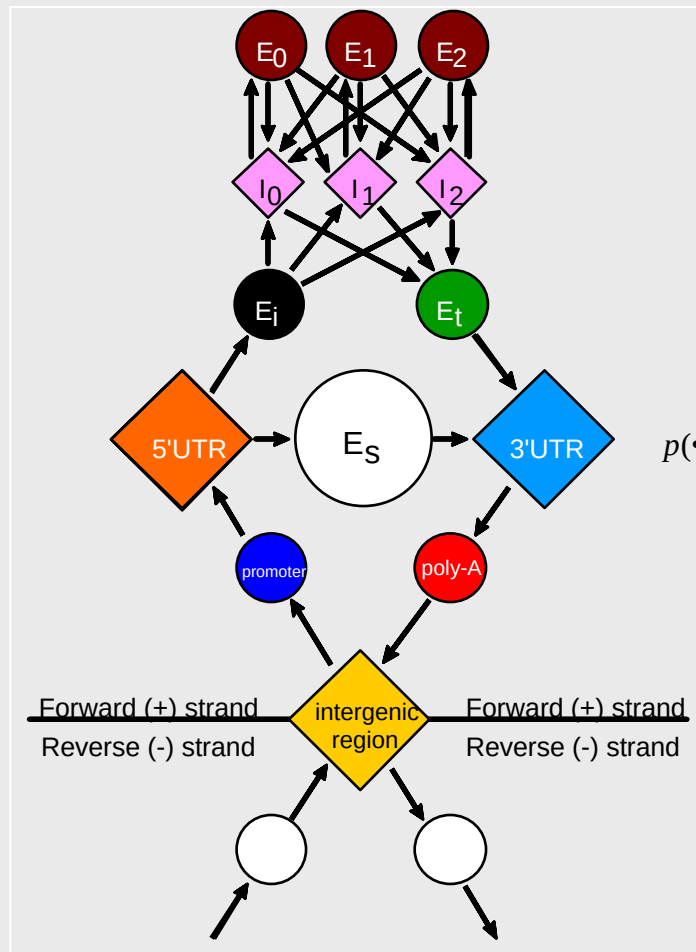


Graphical model



State automata

GENSCAN (Burge & Karlin)



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

```

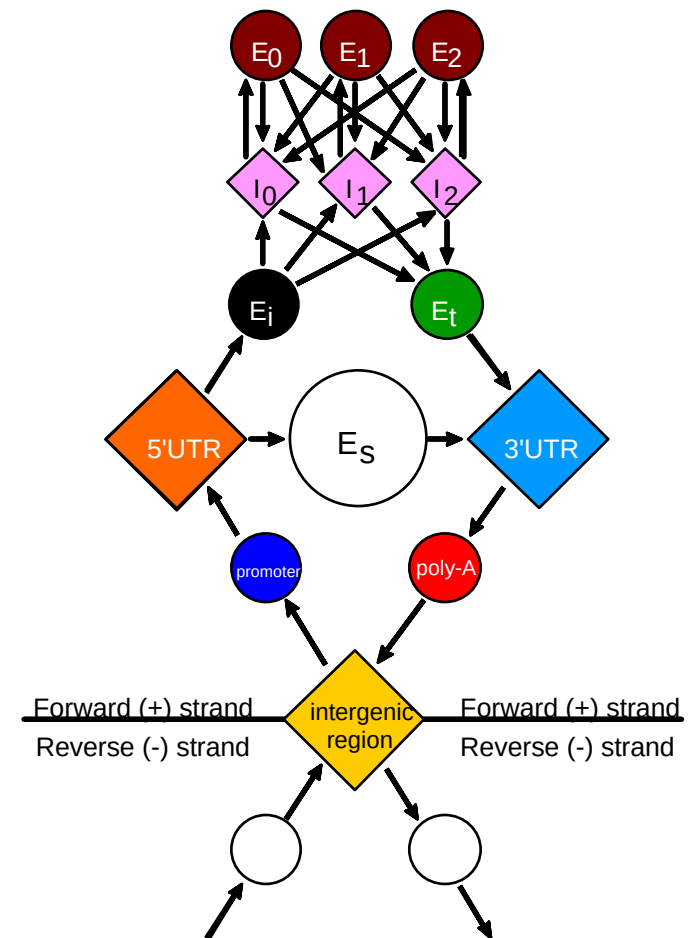
GAGAACGTGTGAGAGAGAGGCAAGCCGAAAAATCAGCCGC
CGAAGGATACACTATCGTCGTCCTTGTCGACGAACCGGT
GGTCATGCAAACAACGCACAGAACAATAATTTTCAAAT
TGTTCATAAATGTCCCACTTGCTTCTGT GTTCCCCCCT
TTCCGCTAGCGGAATTTTTATATTCTTT

                                BATACACAGCGCACACAT
ATAAGCTTGCACACTGATGCACACACACCGACACGTTGTC
ACCGAAATGAACGGGACGGCCATATGACTGGCTGGCGCTC
GGTATG GGGTGCAAGCGAGATACCGCATCAAGACTCGA
ACGAGACGGGTCAGCGAGTGATACCGATTCTCTCTCTTT
GCGATTGGGAATAATGCCCGACTTTTACACTACATGCGT
TGGATCTGGTTATTTAATTATGCCATTTTCTCAGTATAT
CGGCAATTGGTTGCATTAATTTGCGCGAAAGTAAGGAAC
ACAAACCGATAGTTAAGATCCAACG TCCCTGCTGCGCCTC
GCGTGCACAATTTGCGCAATTTCCCCCTTTTCCAGTTT
TTTTCAACCCAGCACCGCTCGTCTCTTCTCTCTTAACG
TTAGCATTCGTACGAGGAACAGTGCTGTCTTGTGGCCGC
TGTGTAGCTAAAAAGCGTAATTATTCATTATCTAGCTATC
TTTTCCGATATTATTGTCATTGCGCTTAATCTTGTGTAT
TTATATGGATGAAACGTGCTATAATAACAATGCAGAATGA
AGAACTGAAGAGTTTTCAAAACCTAAAAATAATTGGAATAT
AAAGTTTGGTTTTACAATTTGATAAACTCTATTGTAAGT
GGAGCGTAACATAGGGTAGAAAACAGTGCAAATCAAAGTA
CCTAAATGGAATACAAATTTAGTTGTACAATTGAGTAAA
ATGAGCAAAGCGCTATTTTGGATAATATTGCTGTTTTAC
AAGGGGAACATATTCATAATTTTCAGGTTTAGGTTACGCA
TATGTAGGCGTAAAGAAATAGCTATATTTGTAGAAGTGCA
TATGCACTTTATAAAAAATTATCCTACATTAACTATTTT
ATTTGCTTTAAACCTATCTGAGATATTCCAATAAGGTAA
GTGCAGTAATACAATGTAATAATTGCAAATAATGTTGTA
ACTAAATACGTAAACAATAATGTAGA AGTCCGGCTGAAAG
CCCCAGCAGCTATAGCCGATATCTATATGATTTAACTCT
TGTCTGCAACGTTCTAATAAATAAATAAATGCAAAATAT
AACCTATTGAGACAATACATTATTTATTTTATATATC
ATCAATCATCTACTGATTTCTTTCCGGTGATCGCCTAATC
CATCTGTGAAATAG AAATGGCGCCACCTAGGTTA AGAAAA
GATAAACAGTTGCCTTTAGTTGCATGACTTCCCCTGGAT
    
```

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.





The HMM Algorithms

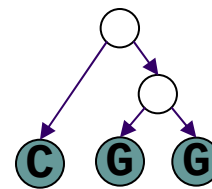
Questions:

- **Evaluation:** What is the probability of the observed sequence? **Forward**
- **Decoding:** What is the probability that the state of the 3rd position is B_k , 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)**

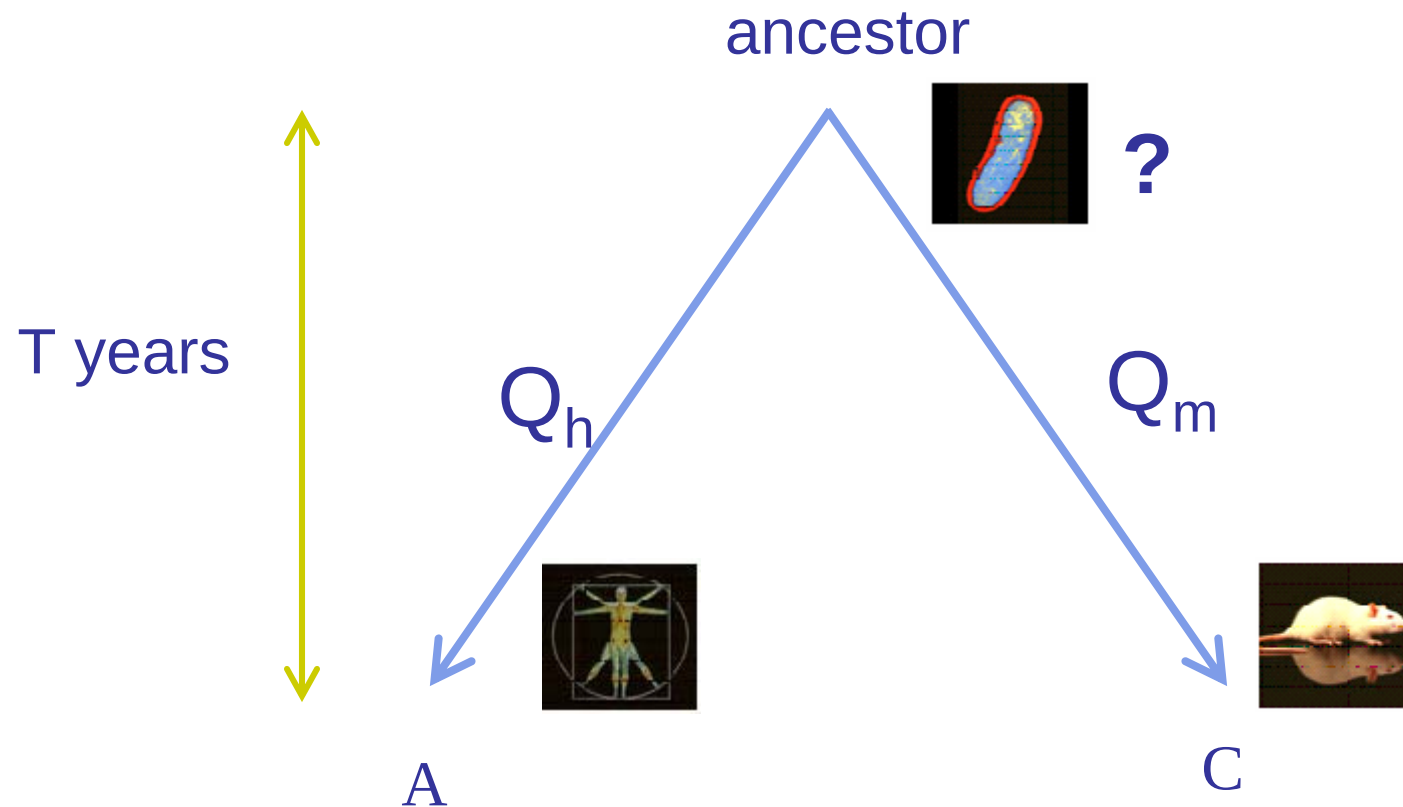
Computation Biology and ML



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



A pair of homologous bases



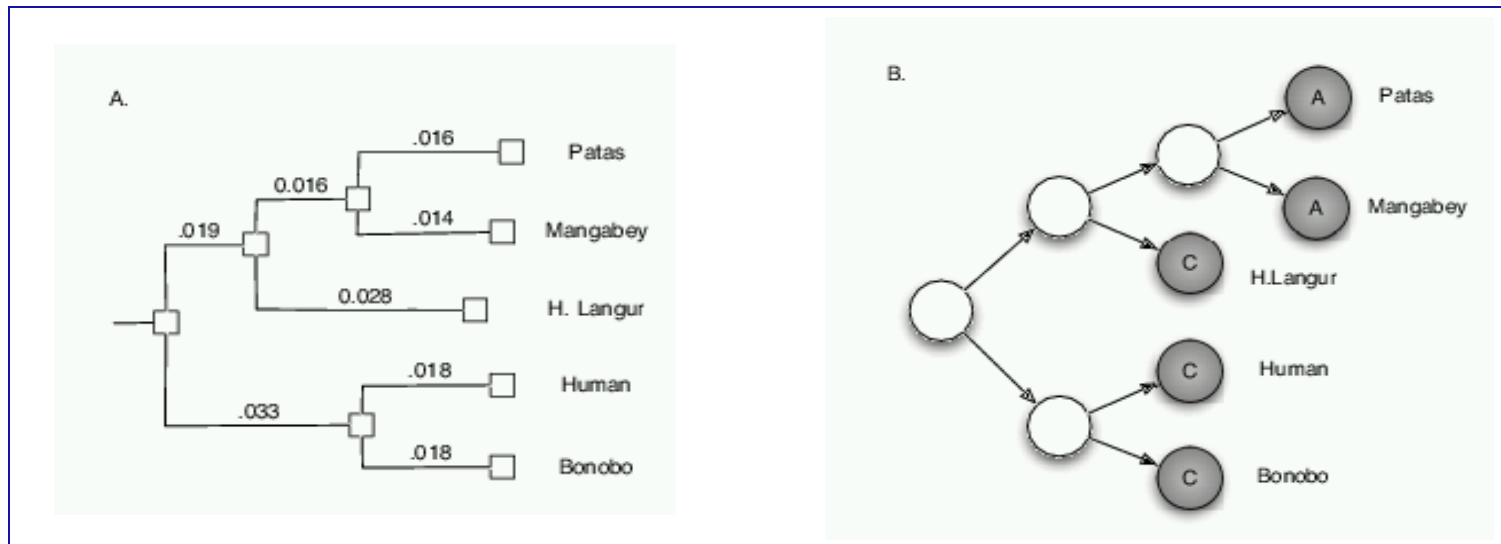
Typically, the ancestor is unknown.

Homology identification via multiple alignment



		10	20	30	40	50	60	
		*	*	*	*	*	*	
consensus	1	SAAQKALVKASWGKVKG	----	NREELGAEILARL	FK----	AYPDTKAYFPKF	g-D	46
LASH	1	ANKTRELCMKSLFAHAKVdt	--sn	EARODGIDLYKHM	FE----	NYPLRKYFKSR	--E	49
LIITH A	3	TAAQIKATQDHUFLNIK	g-----	CLQAAADSIFPKY	LT----	AYPGDLAFFHKF	--S	48
gi_1065933	162	DKESCEVVADSWRLVESrassaa	TSACFGLFVFORV	FS----	KIPMLRPLFG	-Ls-E	212	
gi_3877400	71	nvvEKELLRRTWSDEFD	----	NLYELGSAIYCYI	FD----	HNPNCQLFP	-Fi-S	115
gi_3877381	15	TDEEVTAIRDVWRRRA	----	KTDNVGKKILQTL	IE----	KRPKFAEYFG	-JgsE	58
gi_3874505	230	TCAQIHLVRALNRQVYt	----k	GPTVIGASLVHRL	CFknyvmvke	QMKOVE	-LPPKFg-N	283
gi_4098133	39	EDRDALRVLQNAFKL	----	DDPELVRRFYAHW	FA----	LDASVRDLFP	-P----	79
gi_1707914	18	SPADVK--KHTVESMKAvnv-gr	DKAONGIDFYKFF	FT----	HHKDLRKFFKGA	--E	65	
gi_2494780	3	TKDEFDSLHLELDPKIDt	--e	HRMELGLGAYTEL	FA----	AHPEYIKKFSRL	--Q	50
		70	80	90	100	110	120	
		*	*	*	*	*	*	
consensus	47	LSTAAALKSSPKFKAHGGKVL	GALDEAVKHL	----	DDGNLKAALAKL	GAR-HAKRG	--H	99
LASH	50	EYTAEDVQNDPFFAKQGQKIL	LACHVLCATY	----	DDRTFNAYTRELLDR	-HARDH	--H	103
LIITH A	49	SVPLYGLRNPAYKAQTLTV	INYLDKVV	DAL-----	GGNAGALMKAKVPS	HDAMG	---	98
gi_1065933	213	SDPVFDLPDMHPVRRHARL	FTSILHISVKNV	d-----	ELEAQVAPT	VFKYGER	HYRPDtpH	269
gi_3877400	116	KYQGDENKESKEFRSQALKFV	QTLAQVVKNI	vhmERTESFL	YMWGOKHV	KFADRG	----	170
gi_3877381	59	SIDIRALNOSKEFHLQHR	IONFLDTAVGSL	g-fCP	ISSVFDMAHRIGOI	HFYRG	---N	114
gi_3874505	284	-----RDNFIKAHCKA	VAELIDQV	VENL---	DHLDNVTGELMRIGRV	HAKVL	----	327
gi_4098133	80	-----DMGAQRAAF	GQALHWVY	YGE	LVAAQr-----	AEEPVAFLAQLGRD	HRKYG	---122
gi_1707914	66	NFGADDVQKSKRFEKQGTALL	LAVHVL	ANVY---	DNQAVFHGFVRE	LMNR-HEKRG	dpK	121
gi_2494780	51	EATPANVMAQDGAKYYAKTL	LINDLVE	LLKAS---	TDEATLNTA	IARTATKSHKPRN	---	103
		130	140	150	160	170		
		*	*	*	*	*	*	
consensus	100	VDPANFKLFGEALL--	VVLAEHL	g--DFT	PEVKA	AWDKALD	VVADALKSGYR	147
LASH	104	MPPEVWTDFWKLFE	----	EYLGKKT	----TL	DEPTKQAWHEI	GREFAKEINKHGR	150
LIITH A	99	ITPKHFGQLLKLVG	----	GVFOEEF	s--ADP	TTVAAWGDAAGVL	VAAAMK	141
gi_1065933	270	MTEENVRVFCAQIV--	CTVFDFL	rd-tEATPKCAE	SWIELMRYL	GOKLLDGF		319
gi_3877400	171	FKHEYUDIFQDAME--	FALEHRL	simtDLD	DMOKRDAVTV	WRTLALYTTVHMR		221
gi_3877381	115	FGADNULVFKKVTV	----	DQVITGTT	----DSSKEKED	tnsngtanckvdt	dasli	162
gi_3874505	328	RGELTGKLU	NTVAETiDCTLEWG	dr-rCRSETYRK	AWALI	VAFVIEKIK	AGHH	380
gi_4098133	123	VLPTQYDTLRRALY	----	TTLRDYL	g-----	HPSRGAWTDA	VDEAAGQSLN	LI 167
gi_1707914	122	LWKIFFDDVWVPFL	----	ESKGAKL	s-----	GDAKAAWKE	LNMNFN	SEAQHOLE 166
gi_2494780	104	VSGAEFQTGEPIFI	----	KYFSHVL	----	TTPANQA	FMEKL	LTKIFTGVAGQL- 148

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



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

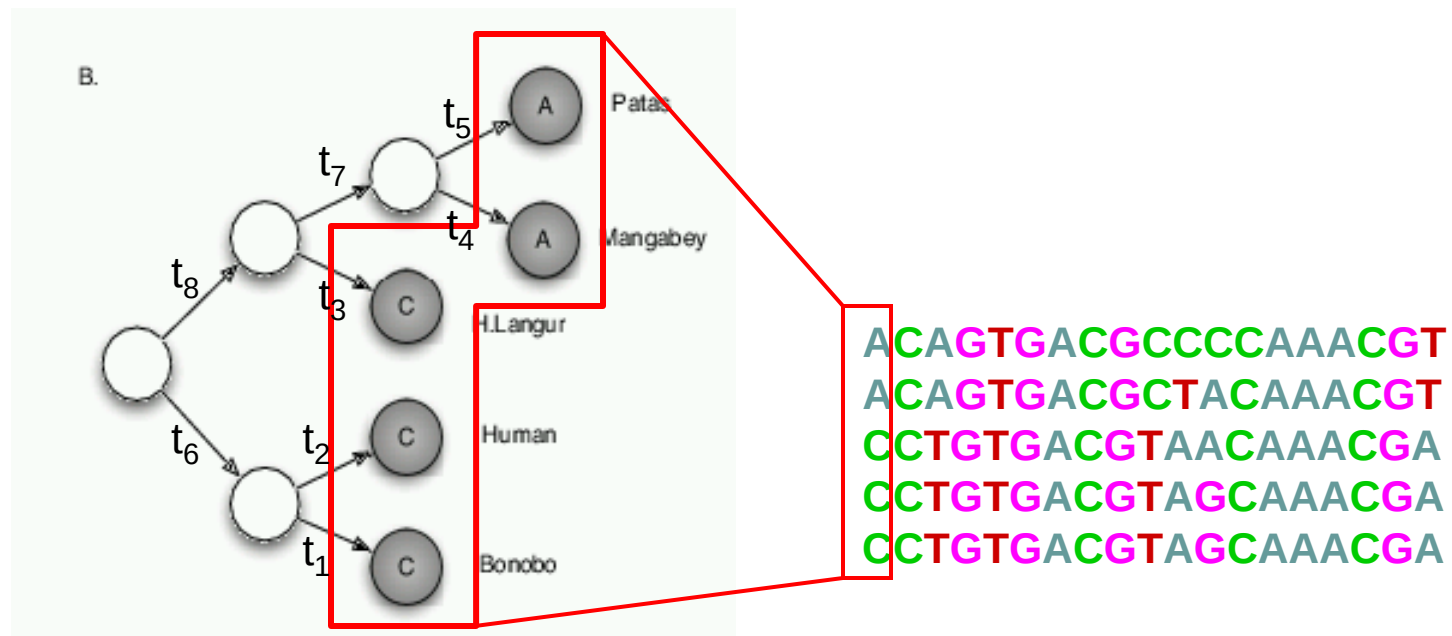
- Major methods:

- Parsimony phylogeny methods
- Likelihood methods

Likelihood methods

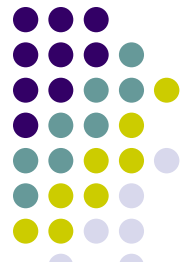


- A tree, with branch lengths, and the data at a single site.



- Since the sites evolve independently on the same tree,

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



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:

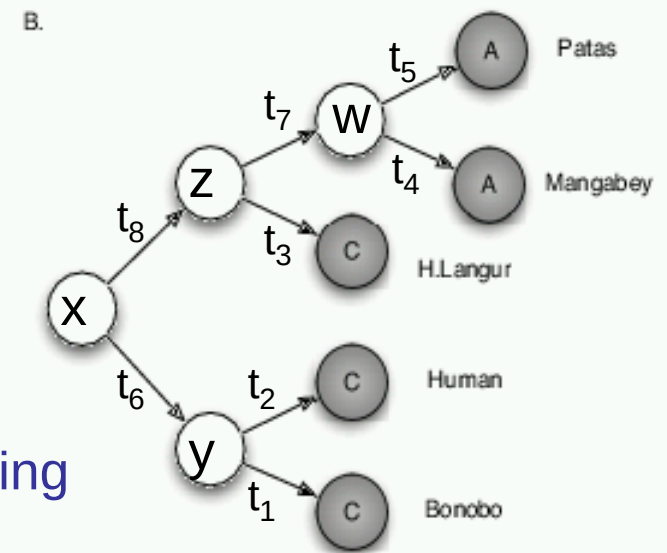
$$P(A, A, C, C, C, x, y, z, w | T) =$$

$$P(x) P(y | x, t_6) P(C | y, t_1) P(C | y, t_2)$$

$$P(z | x, t_8) P(C | y, t_3)$$

$$P(w | z, t_7) P(A | y, t_4) P(A | y, t_5)$$

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



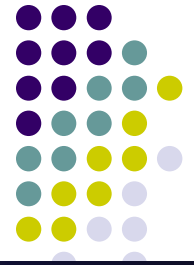


Getting a recursive algorithm

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

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

Felsenstein's Pruning Algorithm



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

Initialization:

Set $k = 2N - 1$

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

If k is a leaf node:

Set $P(L_k \mid a) = 1(a = x_k)$

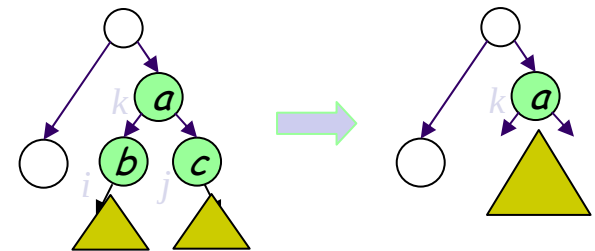
If k is not a leaf node:

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

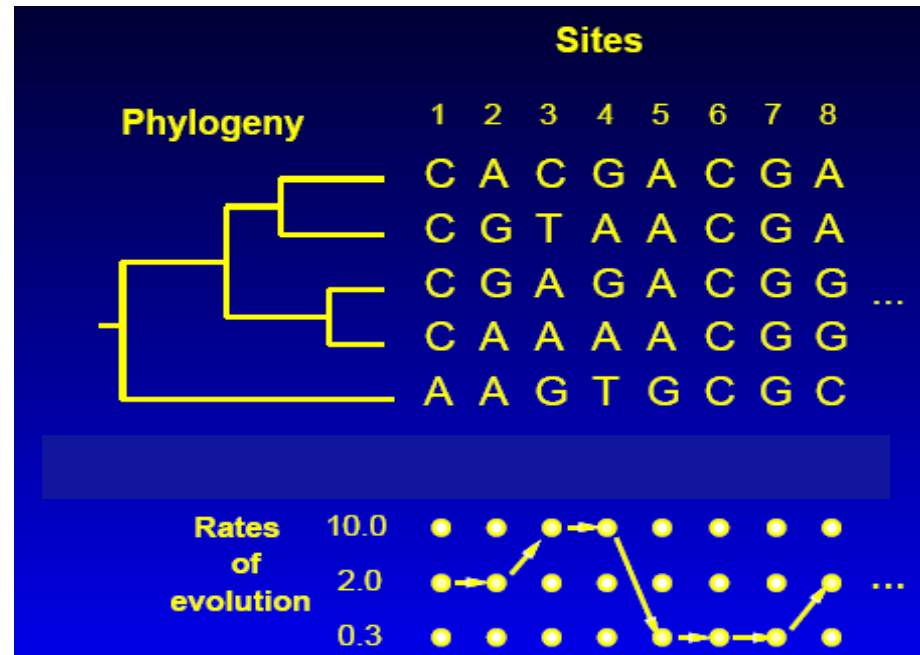
2. Set $P(L_k \mid a) = \sum_{b, c} P(b \mid a, t_i) P(L_i \mid b) P(c \mid a, t_j) P(L_j \mid c)$

Termination:

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

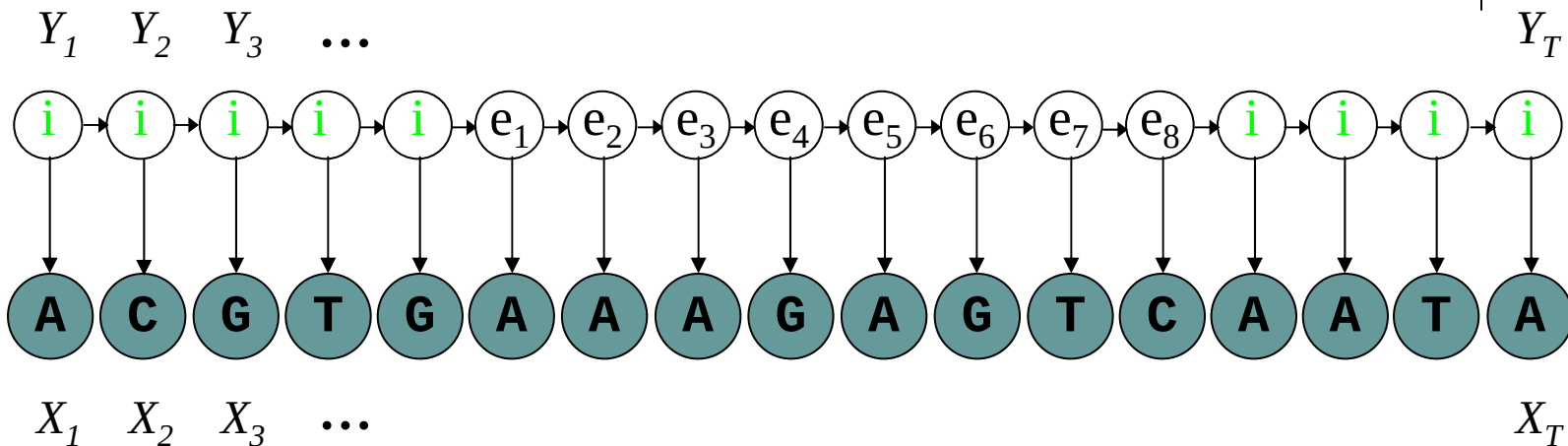


Modeling rate variation among sites



- 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.

Rocall the HMM



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

Definition (of HMM)



- **Observation space**

Alphabetic set:

$$\mathbb{C} = \{c_1, c_2, \dots, c_K\}$$

Euclidean space:

$$\mathbb{R}^d$$

- **Index set of hidden states**

$$\mathbb{I} = \{1, 2, \dots, M\}$$

- **Transition probabilities** between any two states

$$p(y_t^j = 1 | y_{t-1}^i = 1) = a_{i,j},$$

or

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

- **Start probabilities**

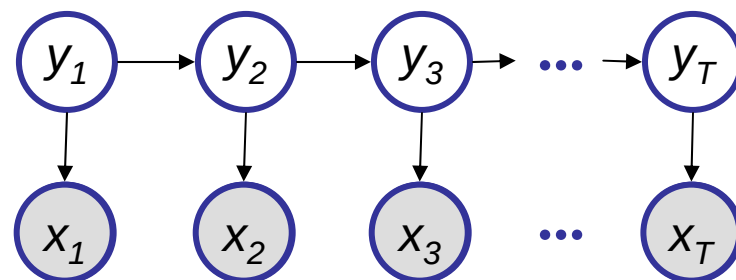
$$p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M).$$

- **Emission probabilities** associated with each state

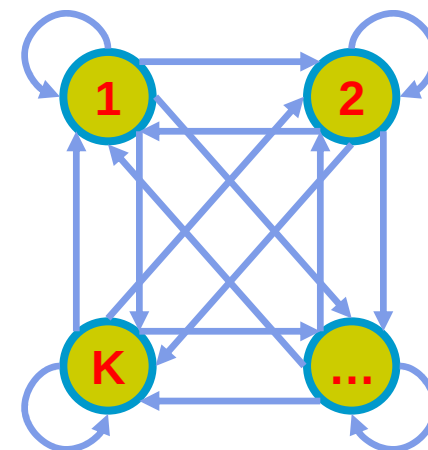
$$p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \dots, b_{i,K}), \forall i \in \mathbb{I}.$$

or in general:

$$p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in \mathbb{I}.$$

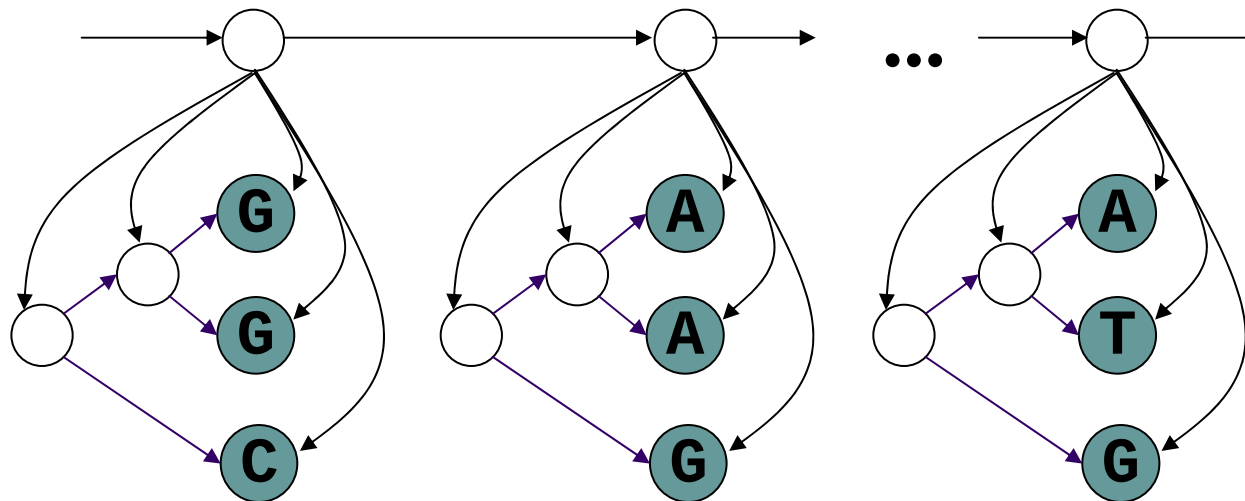


Graphical model

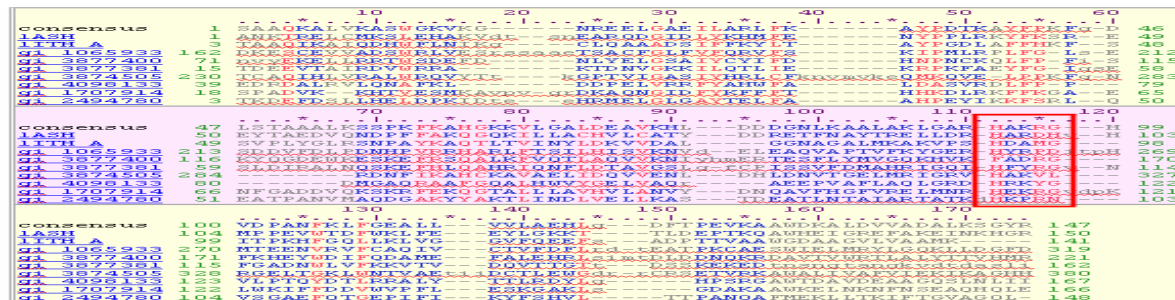


State automata

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 $\text{Prob}(j|i)$ 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

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%.

