기계학습 모델링 기법

Bayesian Inference & Probabilistic Graphical Models

손경아

아주대학교

Bioinformatics & Machine Learning for Life Scientists 2016

Contents

- Introduction to Bayesian Inference
- Naïve Bayes

- Bayesian network
- Probabilistic graphical model

- Hidden Markov Model (HMM)
- Latent Dirichlet Allocation (LDA)

Classification problem

Men

vs.

Women

Classification performance

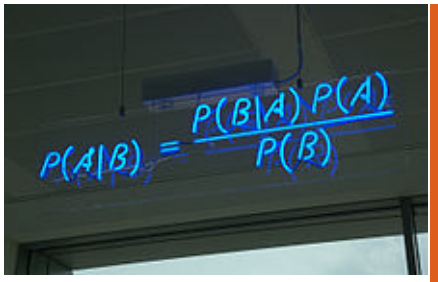
- Example: gender classification
  - Classify a person based on his/her body fat percentage (BFP).

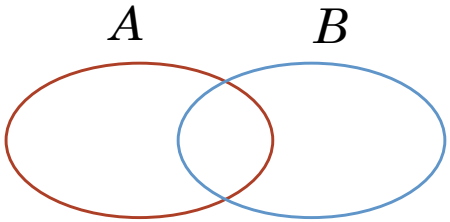
10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.5	23.2

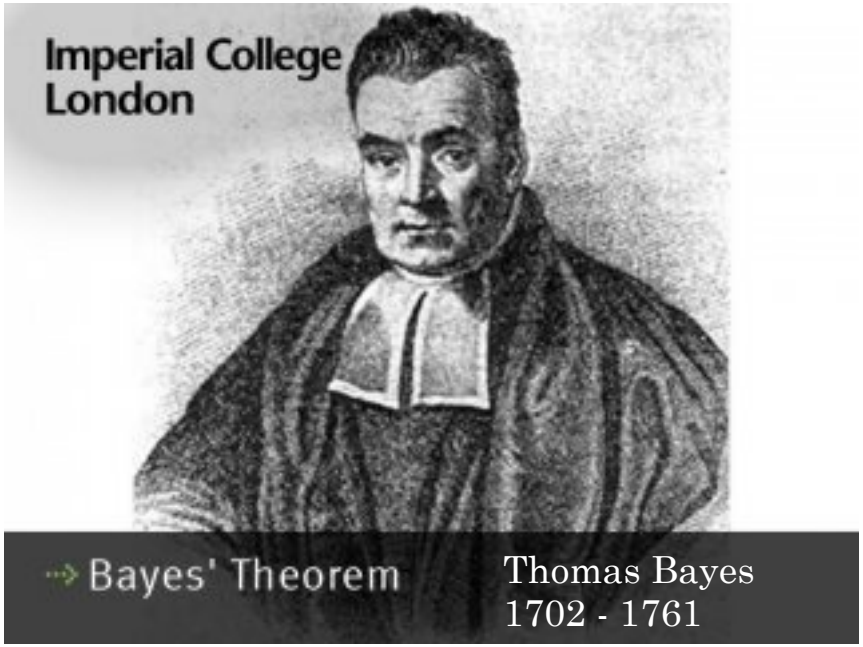
  - Simple classifier: if BFP > 20 then female else male.

10.0	21.7	8.9	19.9	23.4	28.9	15.7	21.6	21.8	23.2
M	F	M	M	F	F	M	F	F	F

Bayes Theorem







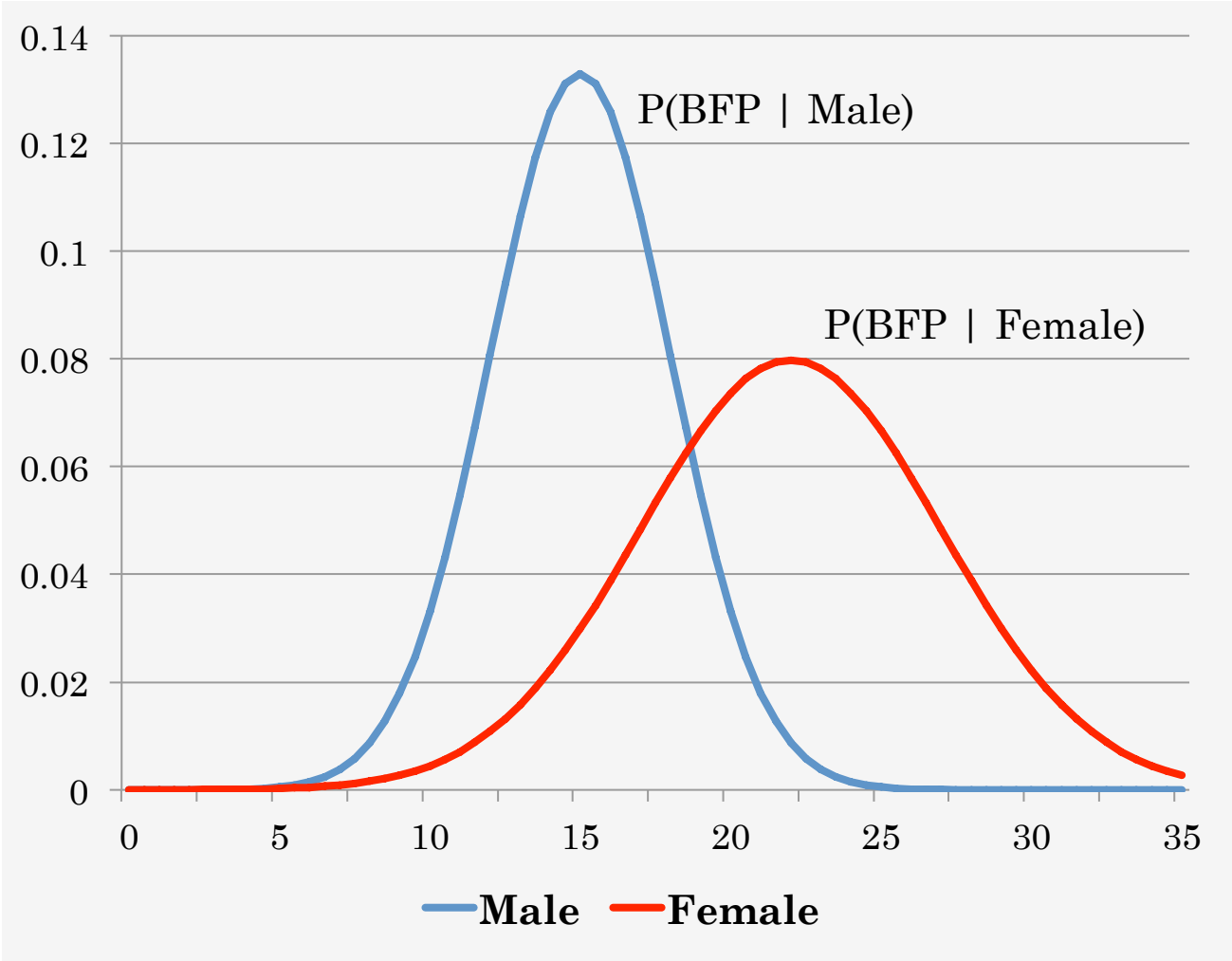
Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A)P(B|A)$$

Bayesian Classification

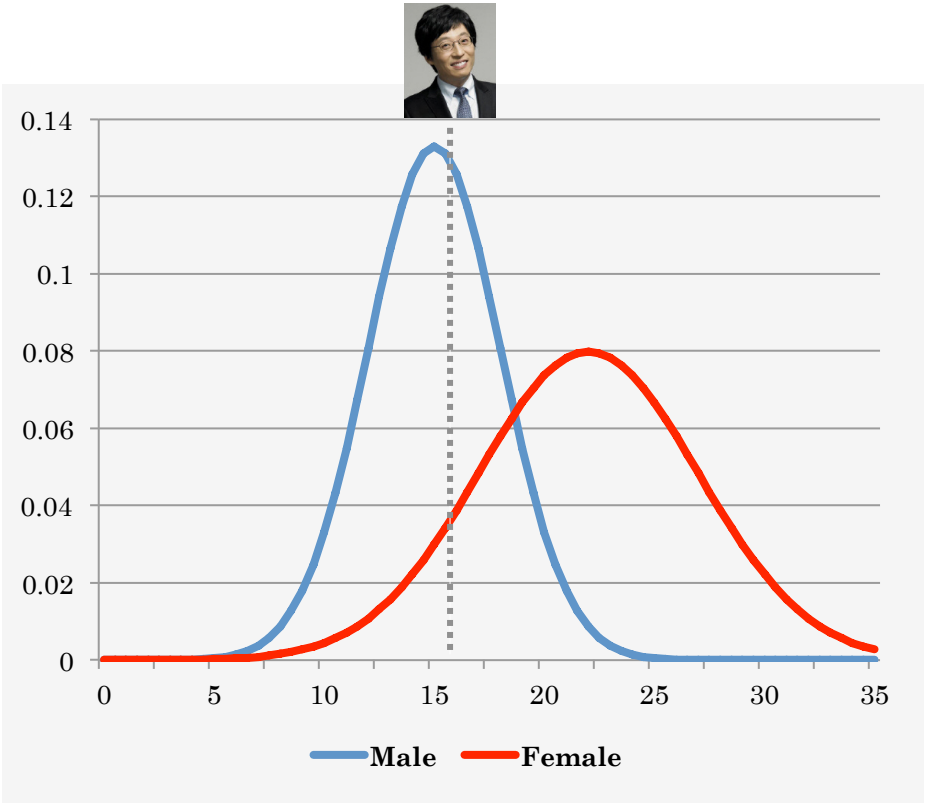
- What if the BFP distributions are …



Class-conditional probability

- In case of 

(15.7)

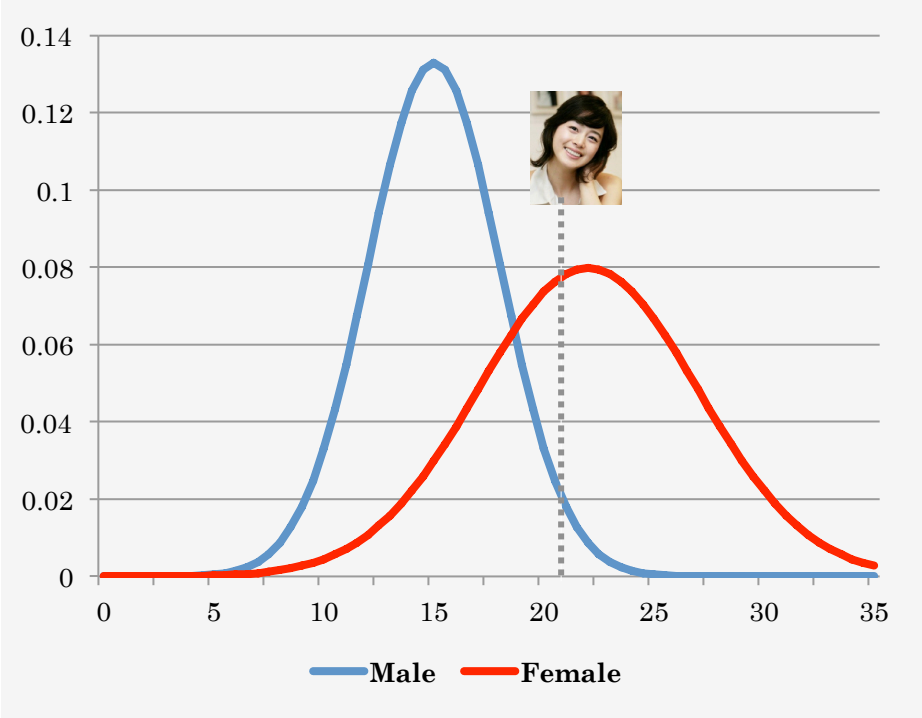


➔ Classify him as male

Class-conditional probability

- In case of 

(21.7)



➔ Classify her as female

Bayesian Classification: procedure

1. Prepare the training data

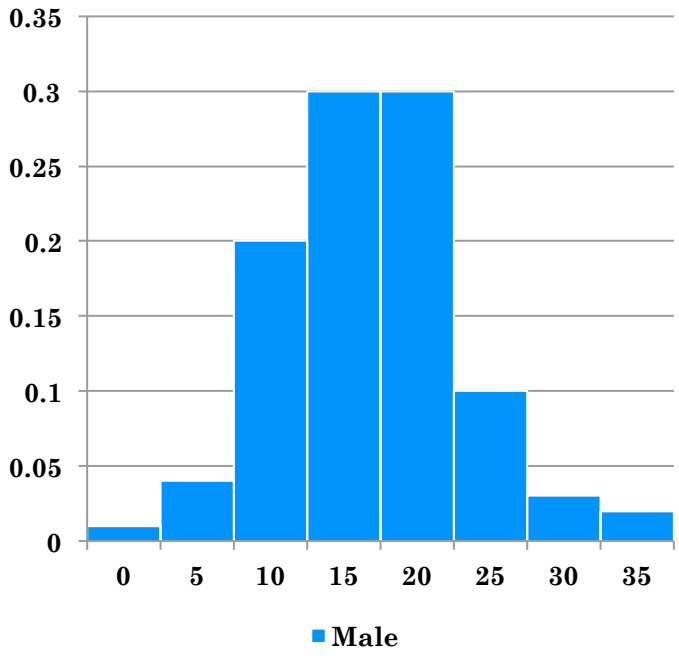
- Define attributes and collect data
  - Total training data: 200 (100 males, 100 females)
  - Attribute (feature): BFP

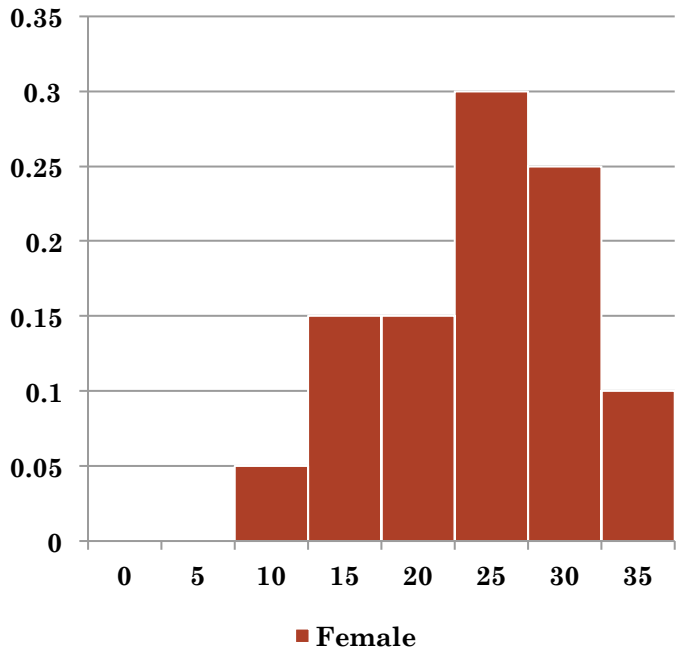
	BFP	Class
1	15	M
2	25	F
3	14	M
4	29	F
…	…	…
N	12	M

Bayesian Classification: procedure

2. Estimate the probability distribution of each variable

- Only one distribution in this one-variable case
- Count the number of training samples in each bin and divide it by the total count



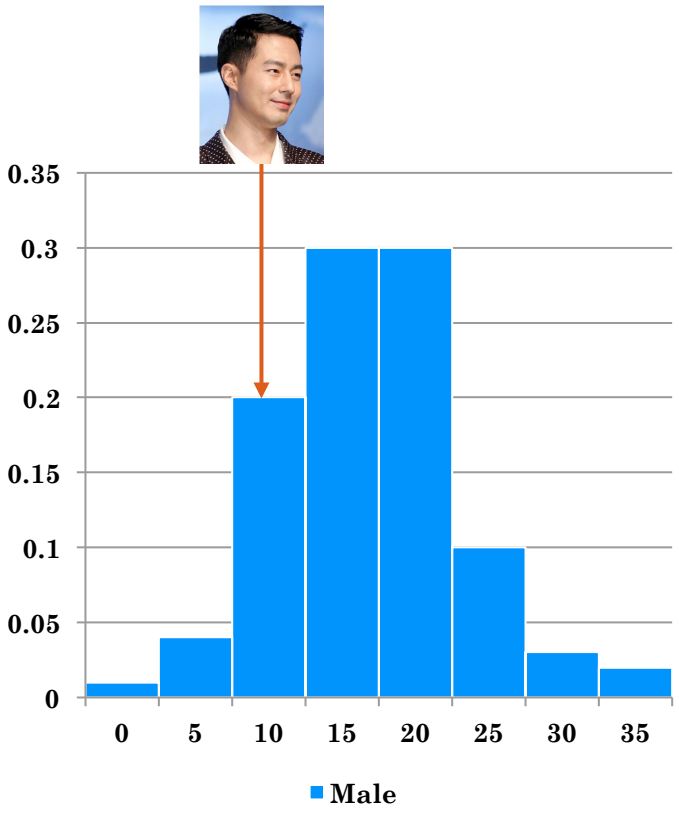


Bayesian Classification: procedure

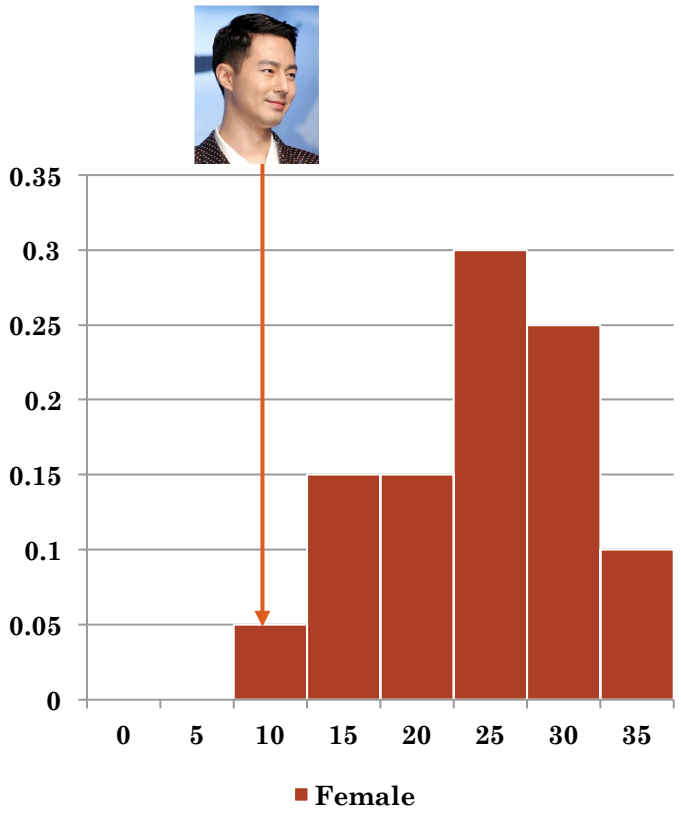
3. For a given input  , compute the conditional probability

(11)

P(BFP = 11 | Male) =



P(BFP = 11 | Female) =



Bayesian Classifier: procedure

3. Compute the posterior probability

P(BFS = 11 | Male) = 0.2

P(BFS = 11 | Female) = 0.05

➔ P(BFS = 11 | Male) > P(BFS = 11 | Male)

➔ Male…?

➔ This is the so-called "likelihood", not "posterior"

- What if there are 400 males and 100 females in the training data?

  Consider the prior probabilities P(Male) & P(Female):

  ✓ P(BFS=11 | Male)*P(Male) = 0.2*0.8 = 0.16

  ✓ P(BFS=11 | Female)*P(Female) = 0.05*0.2 = 0.01

0.16 > 0.01 ➔ Classify as Male

Bayes classifier by posterior probability

$$P(Female|BFP = 11) > P(Male|BFP = 11)??$$

$$P(Male|BFP = 11) = \frac{P(Male \cap BFP = 11)}{P(BFP = 11)} = \frac{P(Male)P(BFP = 11|Male)}{P(BFP = 11)}$$

$$P(Female|BFP = 11) = \frac{P(Female \cap BFP = 11)}{P(BFP = 11)} = \frac{P(Female)P(FP = 11|Female)}{P(BFP = 11)}$$

$$P(Female|BFP = 11) > P(Male|BFP = 11)??$$

Classify by checking the posterior probability

$$P(Male)P(BFP = 11|Male) > P(Female)P(BFP = 11|Female)??$$

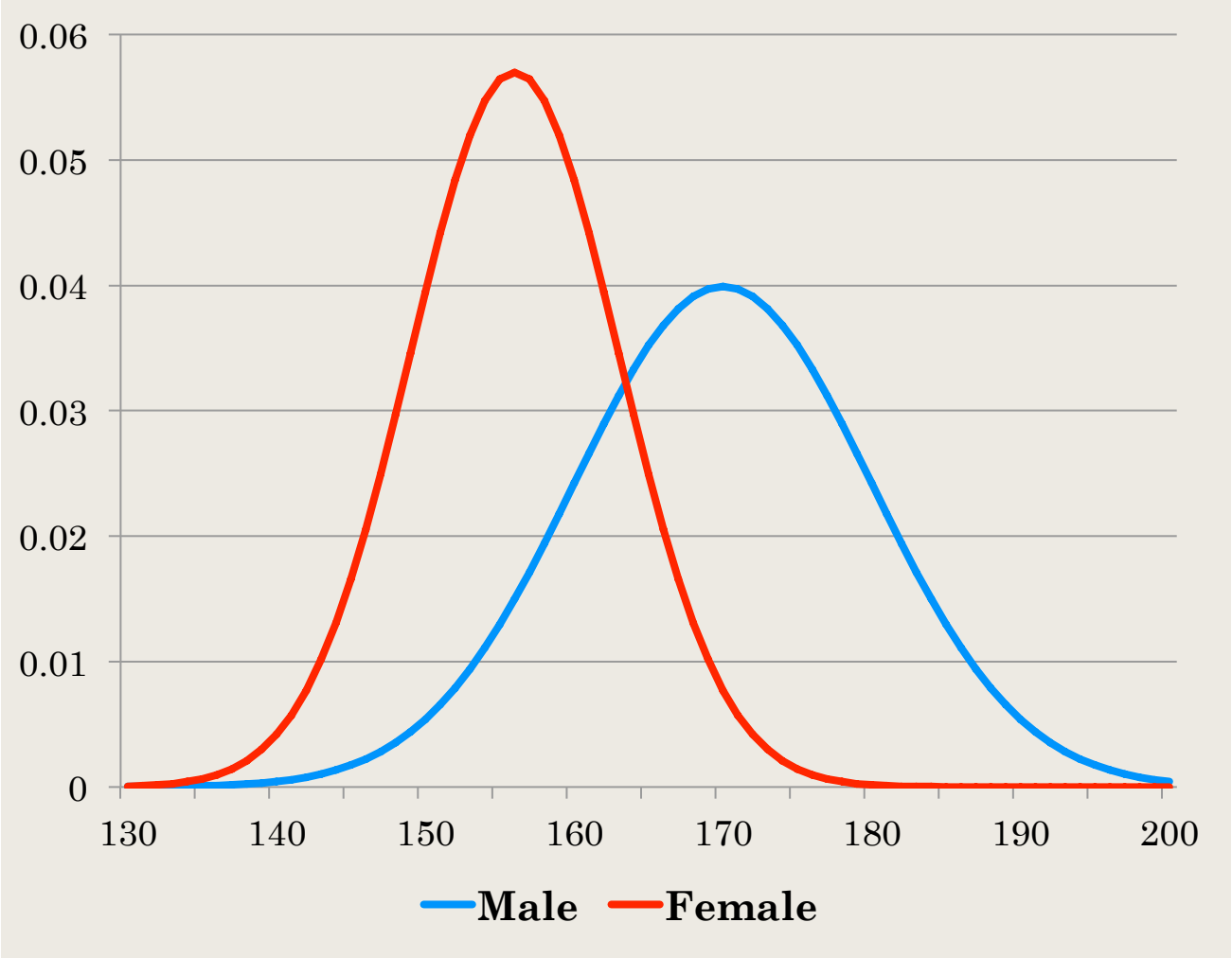
Prior likelihood

Naïve Bayes Classifier

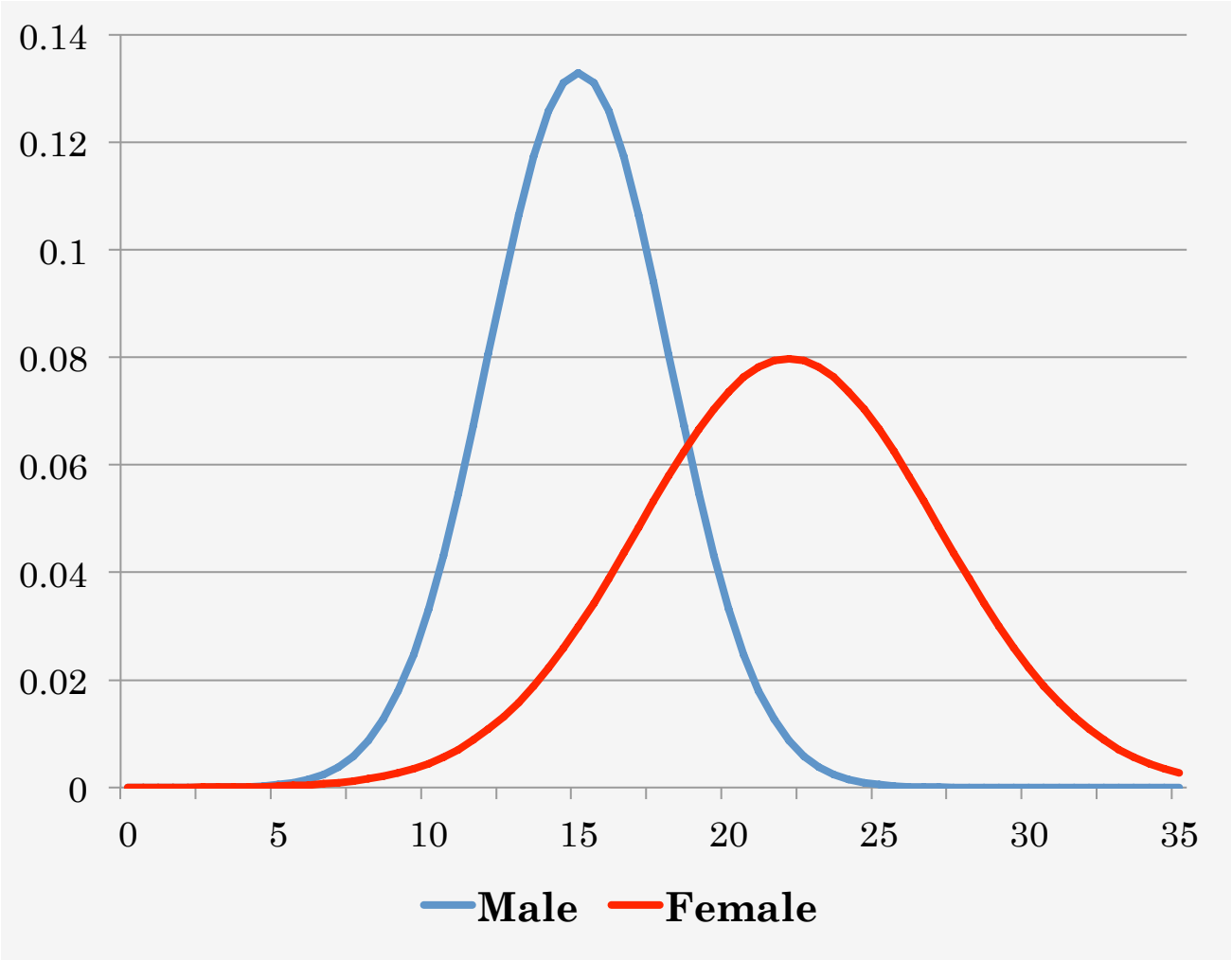
- What if we have multivariate data?
- Example: suppose we measured *Height* of each person as well as *BFP* for gender classification

	Height	Gender	BFS
1	187	M	15
2	165	F	25
3	174	M	14
4	156	F	29
…	…	…	…
N	168	M	12

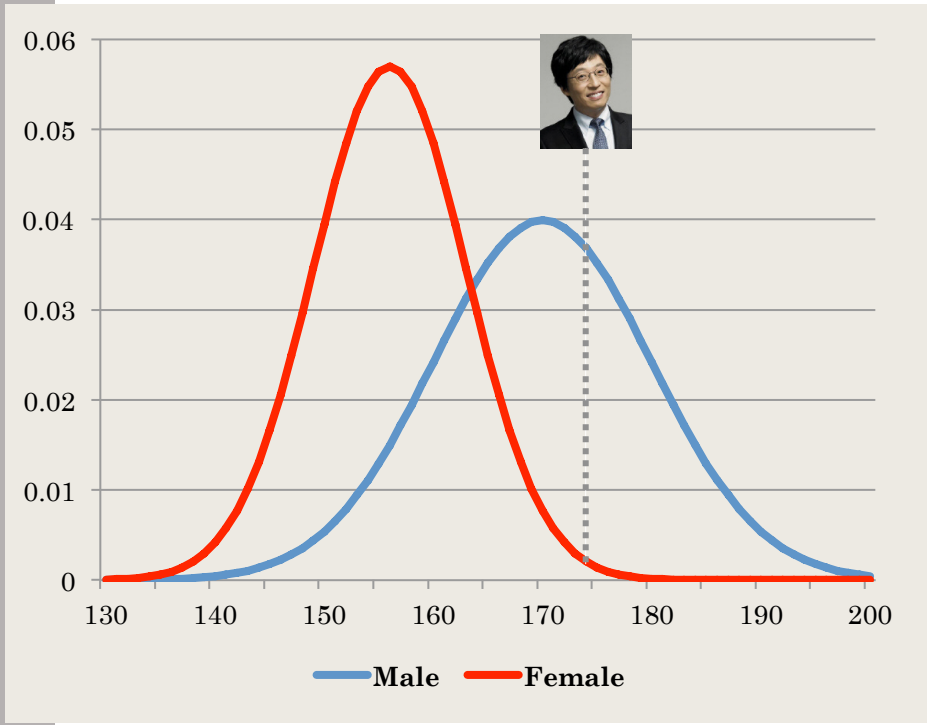
Gender-conditional Height Distributions

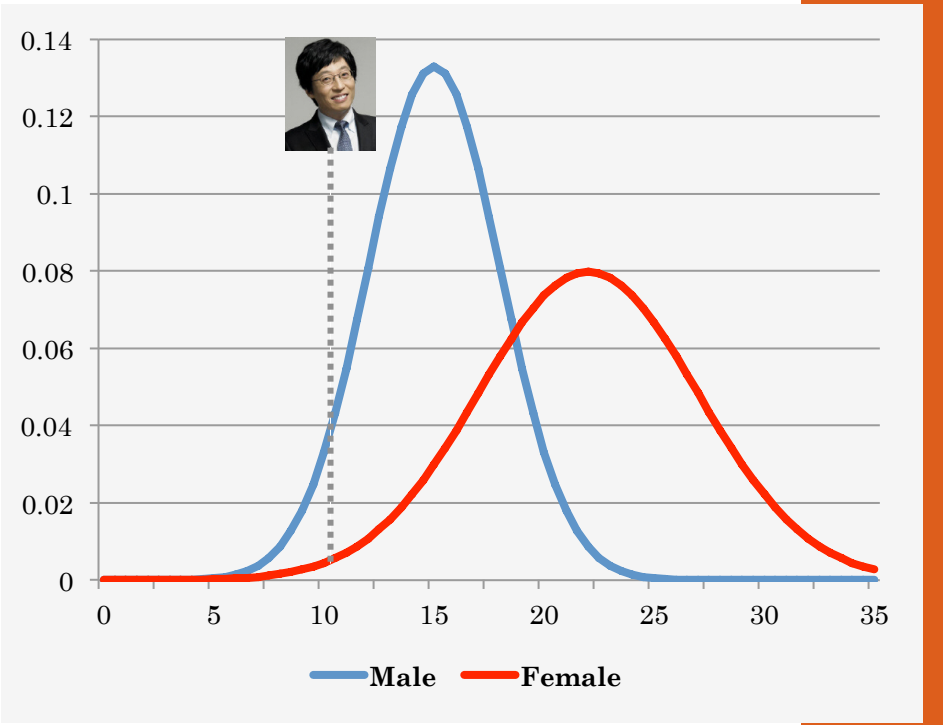


Gender-conditional BFP distributions



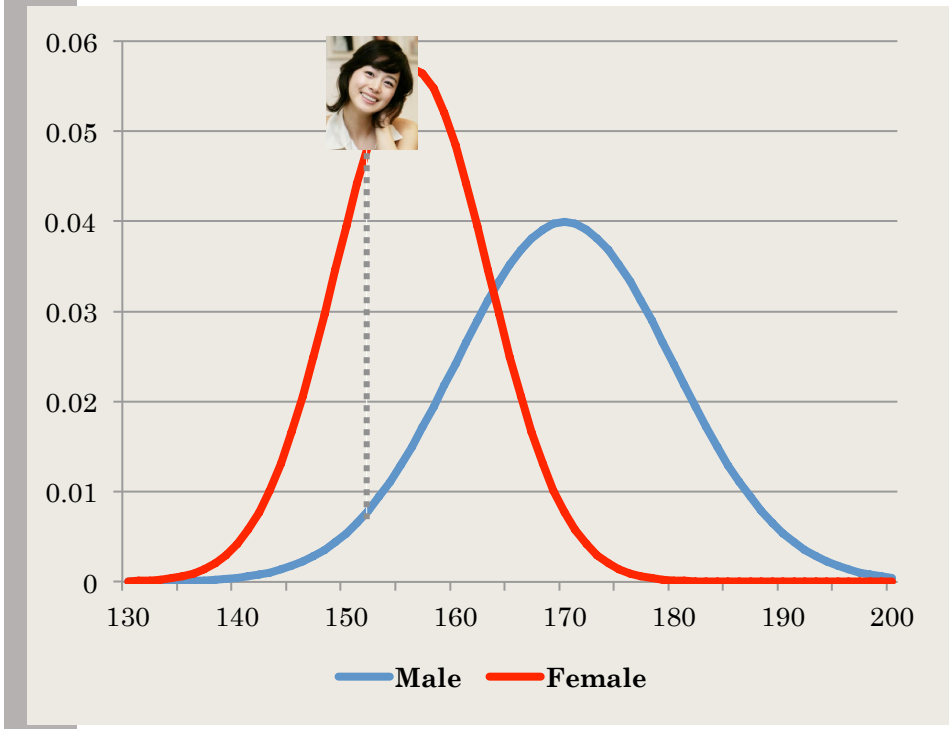
- In case of 

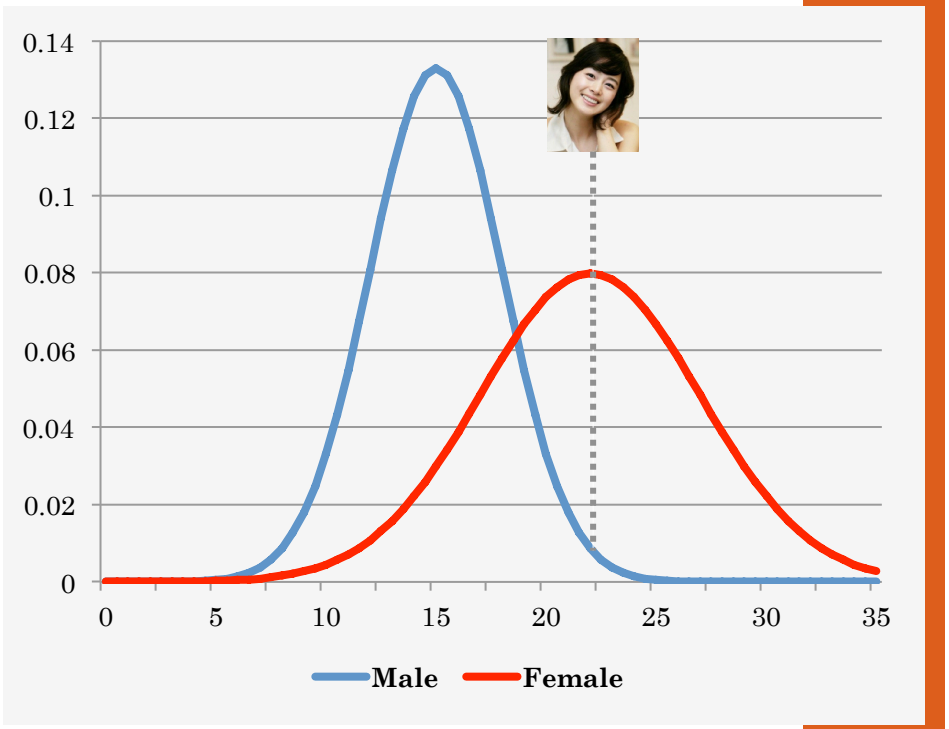




➔ Classify him as male

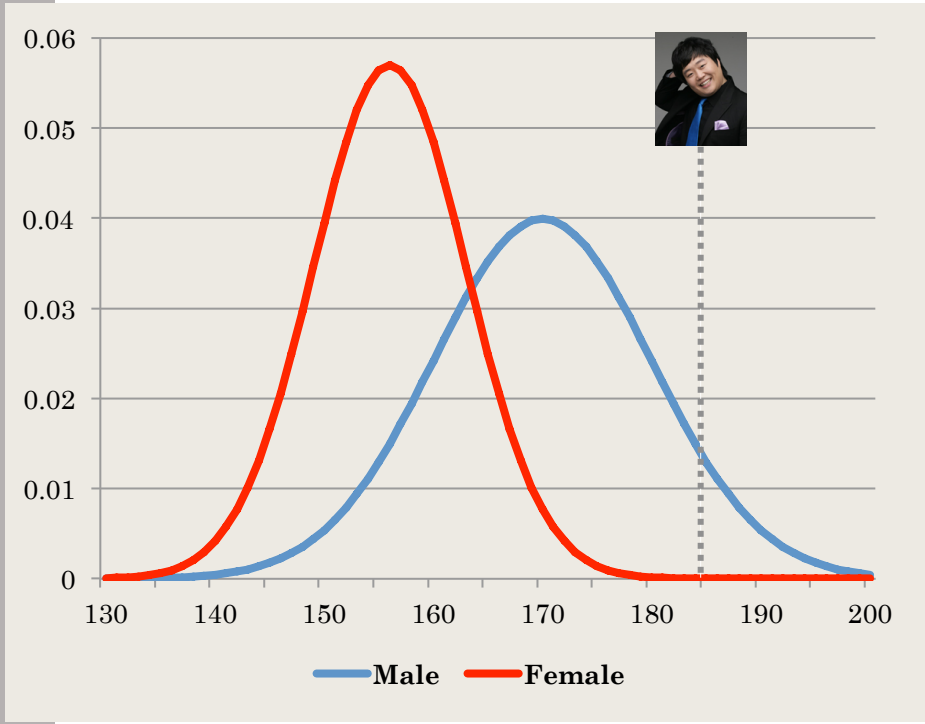
- In case of 

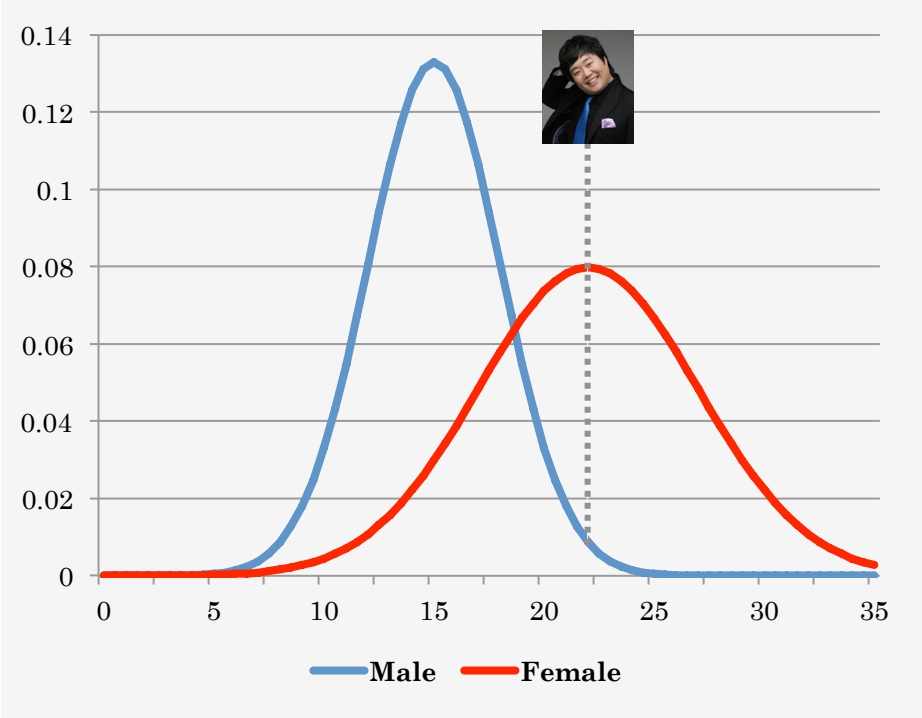




➔ Classify her as female

- In case of 





➔ Classify him as male or female ???

Exact Bayesian Classifier

$$P(Male)P(H = 186, BFP = 11|Male)$$

vs.

$$P(Female)P(H = 186, BFP = 11|Female)$$

Find all the other records just like it

- Find all the other people with the same height and BFP.

Person	Height	BFP	Class
홍길동	186	11	M
김영희	186	11	F
김철수	186	11	M
김가네	186	11	M

Exact Bayesian Classifier

In general, how to find

$$P(H = x, BFP = y|Male), P(H = x, BFP = y|Female)$$

Find all the other records just like it

Person	Height	BFS	Class
홍길동	x	y	M
김영희	x	y	F
김철수	x	y	M
김가네	x	y	M

Difficult to find exactly the same records when the there are many attributes(features) with small number of training data.

Naïve Bayes Classifier

- Assume each variable (attribute) is independent:

P(A & B)=P(A)P(B) if A,B independent

$$P(Height = 186, BFP = 11|Male) = P(Height = 186|Male) * P(BFP = 11|Male)$$

$$P(Height = 186, BFP = 11|Female) = P(Height = 186|Female) * P(BFP = 11|Female)$$

Naïve Bayes Classifier

- Assuming that the height and BFP are independent.

P(height| male) = 0.015

P(BFP | male) = 0.01

P(height,BFP | male)

=P(height| male)*P(BFP | male) = 0.00015

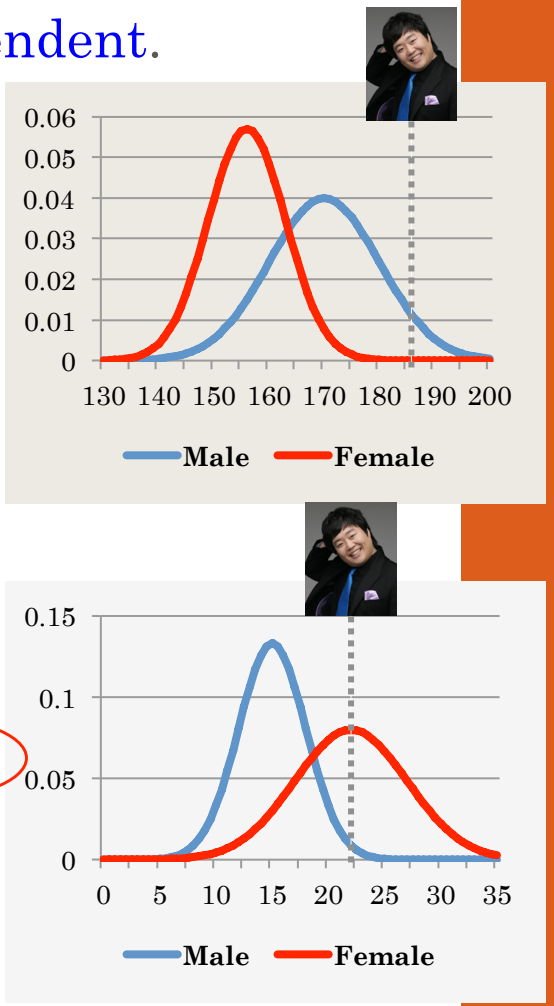
P(height| female) = 0.001

P(BFP | female) = 0.08

P(height,BFP | female)

=P(height| female)* P(BFP | female) = 0.00008

Likelihood of the observed data



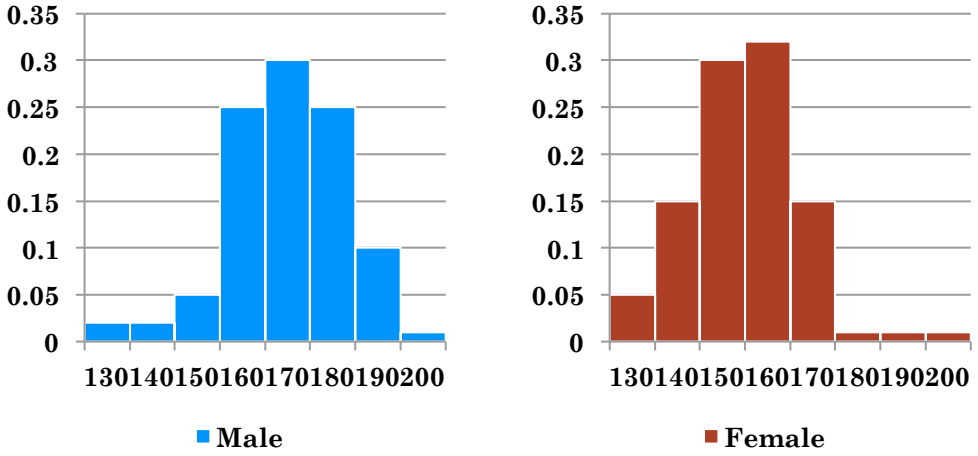
Naïve Bayes Classification

1. Prepare the training data

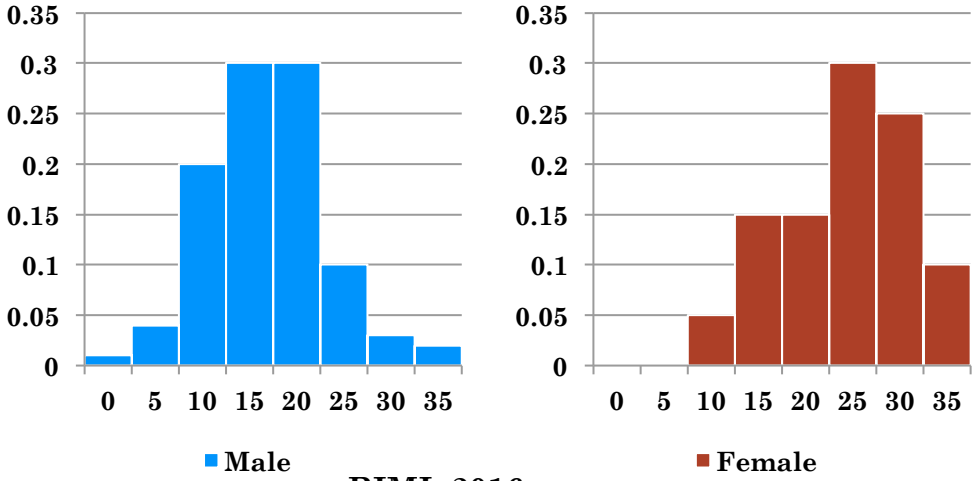
	Height	BFP	Gender
1	187	15	**M**
2	165	25	**F**
3	174	14	**M**
4	156	29	**F**
…	…	…	**…**
N	168	12	**M**

2. Estimate the probability distribution of the attributes for each class

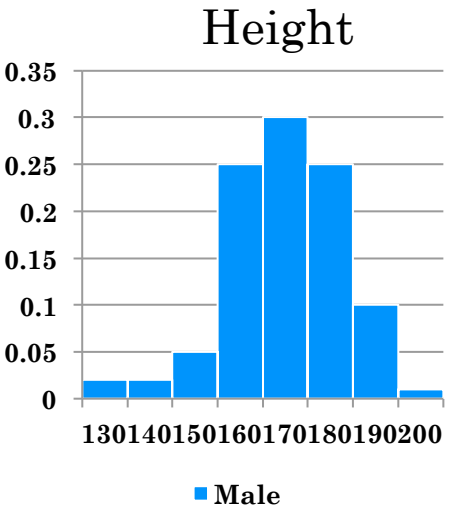
Height

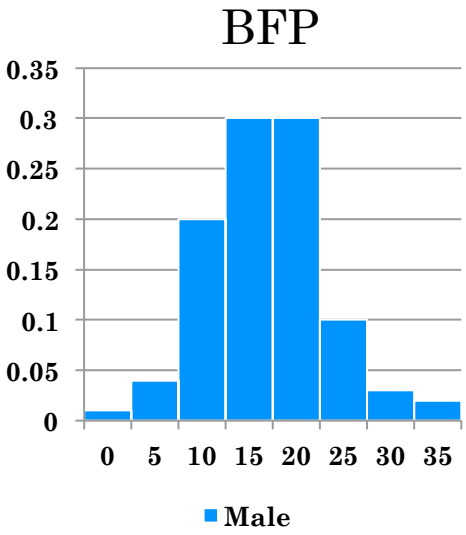


BFP



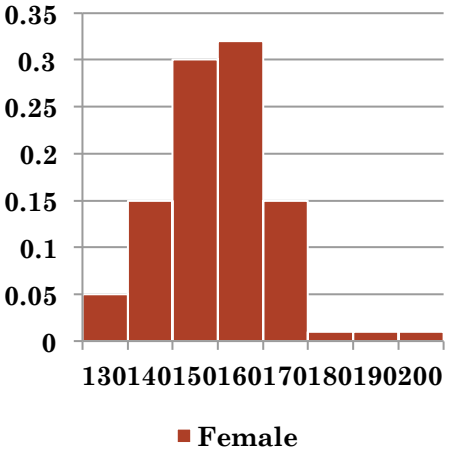
3. For a given input  (h=186, BFP=11), compute the conditional probability for each attribute

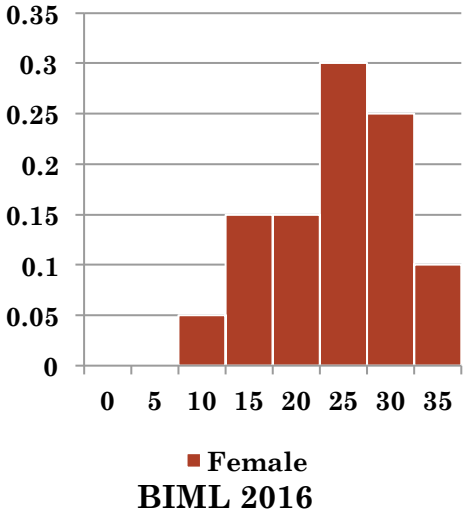




P(h=186 | Male) =

P(BFP=11 | Male) =





P(h=186 | Female) = 0.01

P(BFP=11 | Female) = 0.05

4. Compute the posterior probability

- Compute the likelihood for each class

  P(h = 186, BFP = 11 | Male)

  = P(h = 186 | Male)* P(BFP = 11 | Male)

  = 0.25*0.2 = 0.05

  P(h = 186, BFP = 11 | Female)

  = P(h = 186 | Female)* P(BFP = 11 | Female)

  = 0.01*0.05 = 0.0005

- P(Height=186, BFS=11 | Male) > P(Height=186, BFS=11 | Female)

- What if there are 400 males and 100 females in the training data?

  Compute the prior probability P(Male) & P(Female):

  ✓ P(Height=186, BFS=11 | Male)*P(Male) = 0.05*0.8 = 0.04

  ✓ P(Height=186, BFS=11 | Female)*P(Female) = 0.0005*0.2 = 0.0001

Naïve Bayes Classification: Theory

- Conditional Probability

$$P(C_i \mid x) = \frac{P(C_i, x)}{P(x)} = \frac{\frac{P(x, C_i)}{P(C_i)} \cdot P(C_i)}{P(x)} = \frac{P(x \mid C_i) P(C_i)}{P(x)}$$

- Baye's rule

$$P(C_i \mid x_1, x_2, \ldots, x_d) = \frac{P(x_1, x_2, \ldots, x_d \mid C_i) P(C_i)}{P(x)}$$

$$= \frac{\left(P(x_1 \mid C_i) \cdot P(x_2 \mid C_i) \cdot \ldots \cdot P(x_n \mid C_i)\right) P(C_i)}{P(x)}$$

Bayesian inference

- A method of statistical inference in which Bayes theorem is used to update **the probability for a hypothesis** as more **evidence or information** becomes available
- The prior distribution is the distribution of the parameter(s) before any data is observed, i.e. $p(\theta)$
  - θ is the parameter of the data distribution, i.e. $x \sim p(x \mid \theta)$
- Likelihood is the distribution of the observed data conditional on its parameters = $p(X \mid \theta)$
- Posterior is the distribution of the parameter after taking into account the observed data

$$p(\theta \mid X) = p(\theta)\, p(X \mid \theta) \;/\; p(X) \sim p(\theta)\, p(X \mid \theta)$$

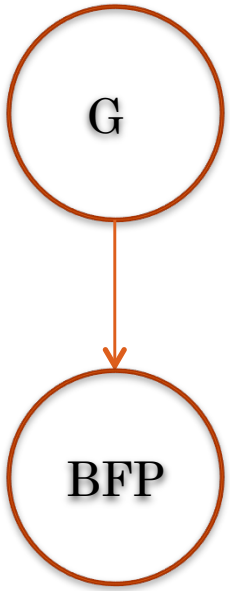
Posterior ~ Prior * Likelihood

Graphical models

- They are *diagrammatic representations* of probability distributions
  - marriage between probability theory and graph theory
- Also called *probabilistic graphical models (PGM)*

- What is a *graph*?
  - Consists of random nodes (also called vertices) and links (also called edges or arcs)
- In a probabilistic graphical model
  - Each node represents a random variable (or group of random variables)
  - Links express probabilistic relationships between variables

Graphical Model for Naïve Bayes

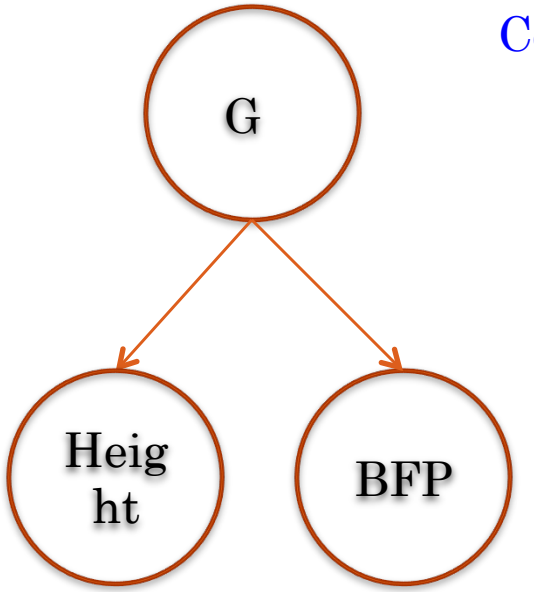


P(Gender)

P(BFP | Gender)

$$P(G, BFP) = P(G) * P(BFP \mid G)$$

Graphical Model for Naïve Bayes



Conditional Parameters

P(Gender)

P(BFP | Gender)

P(Height | Gender)

Joint distribution

$$P(G, BFP, Height) = P(G) * P(BFP, Height \mid G)$$
$$= P(G) * P(BFP \mid G) * P(Height \mid G)$$

Key questions in PGM

- How do we specify distributions that satisfy particular independence properties?

  → **Representation**

- How can we exploit independence properties for efficient computation?

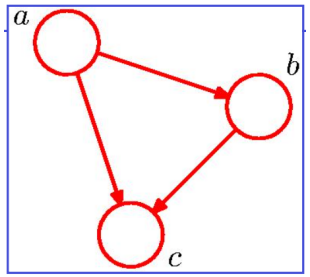
  → **Inference**

- How can we identify independence properties present in data?

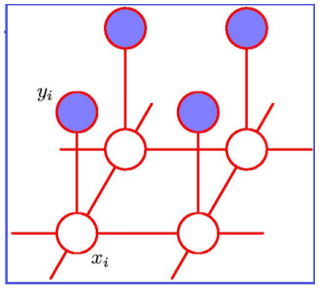
  → **Learning**

Types of GMs

- **Directed edges** give causality relationships
  - Bayesian network or Directed graphical models
  - More popular with AI and statistics

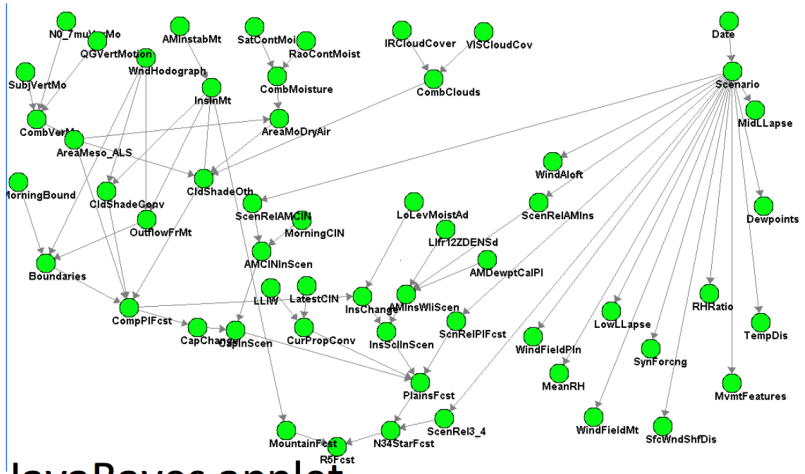


- **Undirected edges** simply give correlations between variables
  - Markov Random Field or Undirected Graphical models
  - More popular in Vision and Physics



Bayesian networks

- Compact representation of distributions over large number of variables
- (Often) allows efficient exact inference (computing marginals, etc.)



JavaBayes applet

HailFinder
56 vars
~ 3 states each

➔~10^{26} terms
> **10.000 years**
on Top
supercomputers

Bayesian networks

- A Bayesian network structure is a directed, acyclic graph G, where each vertex *s* of G is interpreted as a random variable X_s (with unspecified distribution)

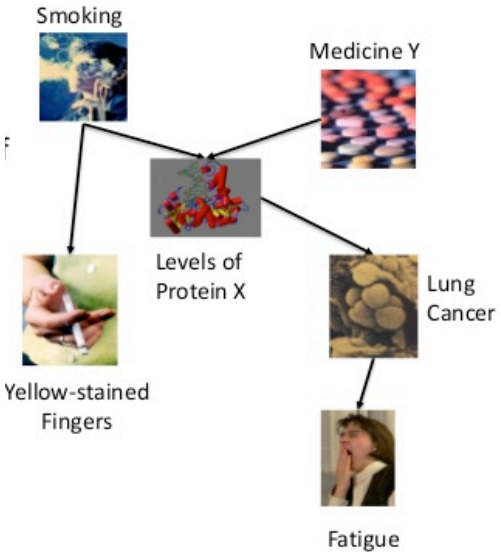
- A Bayesian network (G,P) consists of
  - A BN structure G and ..
  - ..a set of conditional probability distributions (CPTs) $P(X_s \mid Pa_{Xs})$, where Pa_{Xs} are the parents of node X_s such that
  - (G,P) defines joint distribution

$$P(X_1, ..., X_n) = \prod_i P(X_i \mid \mathbf{Pa}_{X_i})$$

Example: using a Bayesian network

1. Factorize the joint probability distribution

2. Answer questions like
   1. P(Lung Cancer | Levels of Protein X) = ?
   2. Ind(Smoking, Fatigue | Levels of Protein X)?



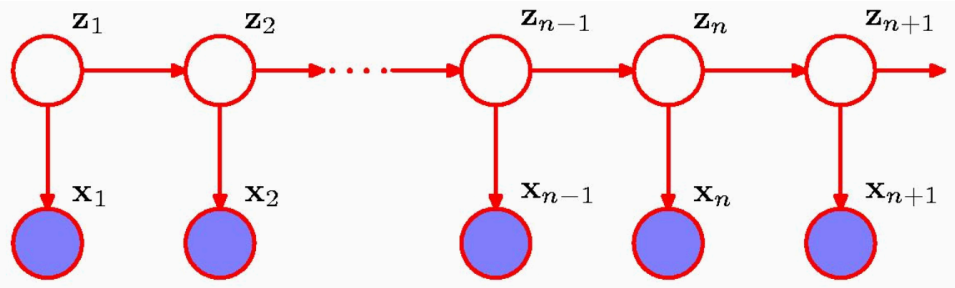
Hidden Markov Model

What is an HMM?

- Ubiquitous tool for modeling time series data
- Used in
  - Almost all speech recognition systems
  - Handwritten word recognition
  - Computational molecular biology
- A tool for representing probability distribution over sequences of observations
- HMM gets its name from two defining properties
  - Observation x_t at time t was generated by some process whose state z_t is hidden
  - Assumes that state at z_t is dependent only on state z_{t-1} and independent of all prior states

Graphical model of HMM

- Has the graphical model shown below and latent variables are discrete



- Joint distribution has the form

$$p(x_1,..x_N,z_1,..z_n) = p(z_1)\left[\prod_{n=2}^{N} p(z_n \mid z_{n-1})\right]\prod_{n=1}^{N} p(x_n \mid z_n)$$

Andrei A Markov (1856 ~ 1922)

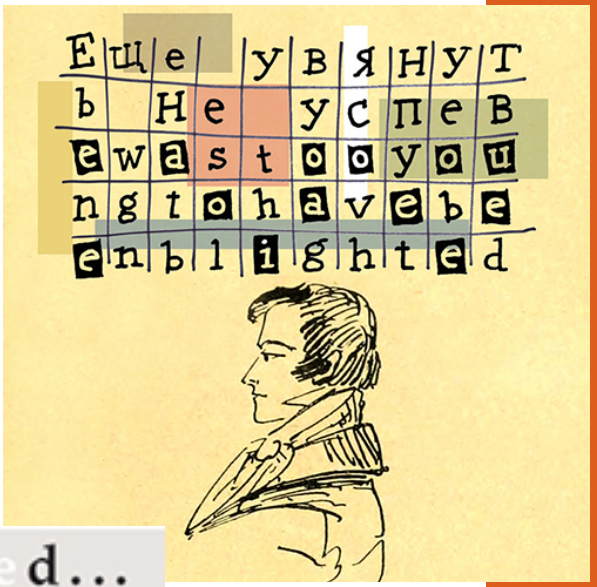
- Russian Mathematician
- Best known for stochastic process

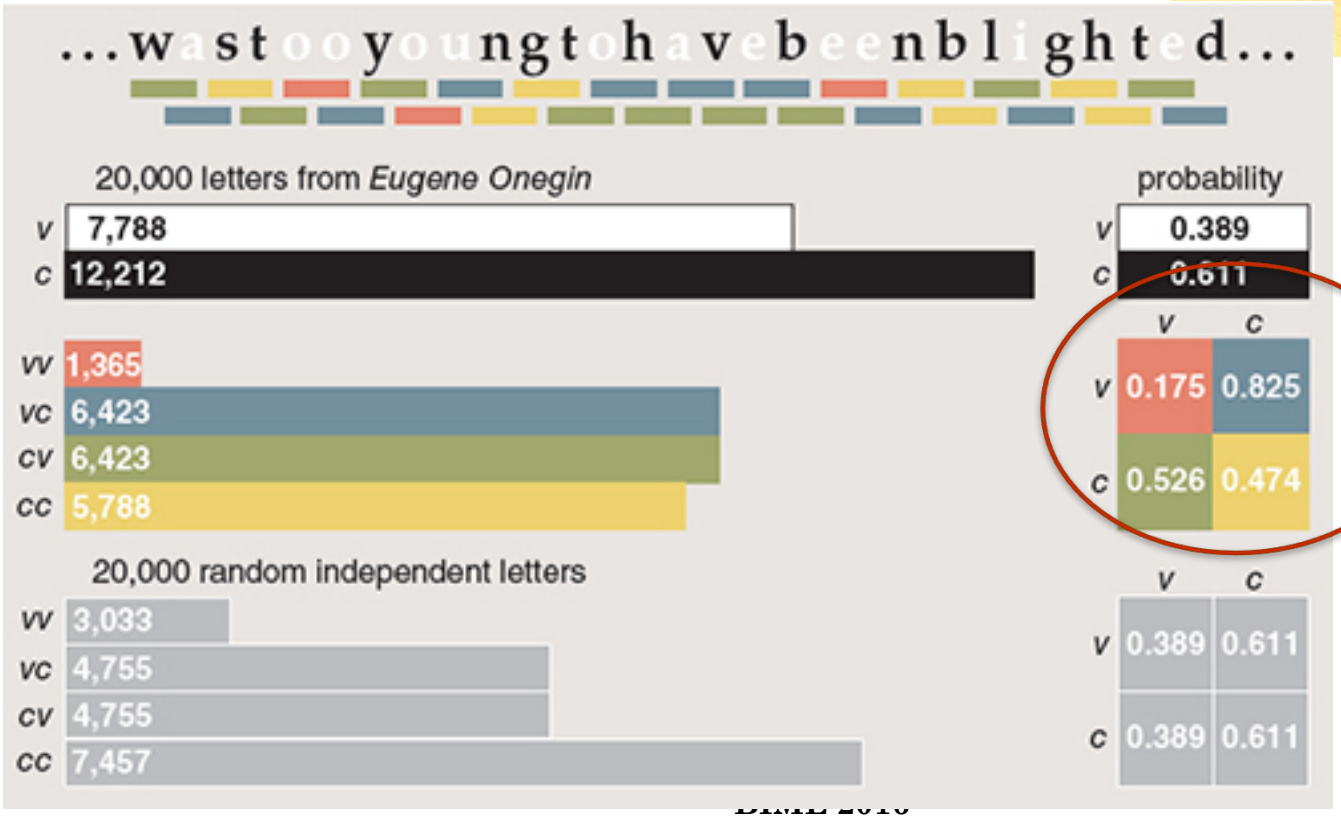
1913

- Russian government celebrates the 300th anniversary of the House of Romanov
- AA Markov organizes a counter-celebration-the 200th anniversary of Bernoulli's Law of Large Numbers

101 years of Markov Chains

- First Links in the Markov chain
- **Probability and poetry were unlikely partners in the creation of a computational tool**





Analysis of Eug ene Onegin by Pushkin (1913)

Transition matrix

http://www.americanscientist.org/issues/pub/first-links-in-the-markov-chain

Markov models

- Used to model sequences of events that occur one after another

- An event might be followed by one of several subsequent events, each with a different probability
  - Sequences of vowels and consonants in *Eugene Onegin*
  - Sequences of words in sentences
  - Daily changes in the weather

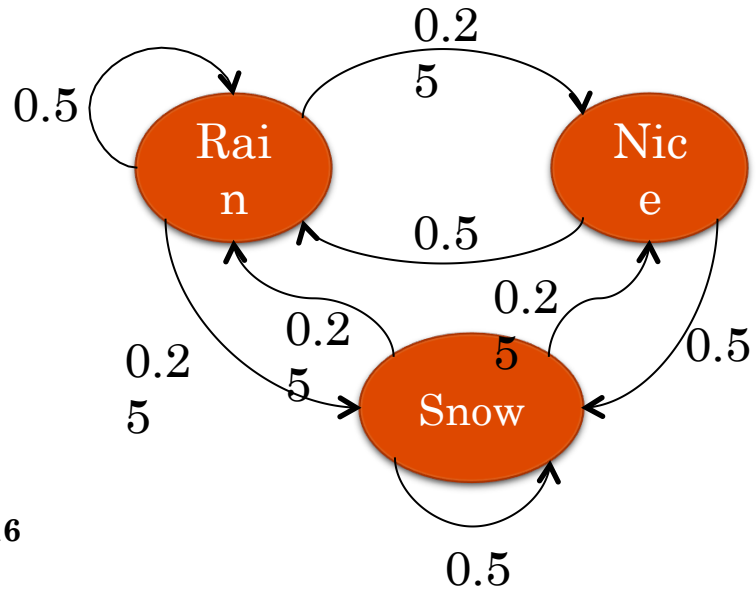
Markov Process: simple example

Weather in the Land of Oz:

- **raining today** ⇒ **50% rain tomorrow**
  ⇒ **25% nice tomorrow**
  ⇒ **25% snow tomorrow**
- **nice today** ⇒ **50% rain tomorrow**
  ⇒ **50% snow tomorrow**
- **snowy today** … …

The transition matrix:

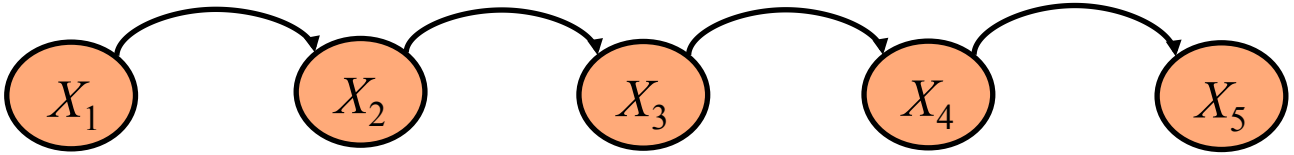
$$P = \begin{array}{c} \\ R \\ N \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & N & S \end{array} \\ \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{array}$$



Markov Process Property

- Markov Property: the state of the system at time t+1 depends only on the state of the system at time t

$$\Pr\left[X_{t+1} = x_{t+1} \mid X_1 \cdots X_t = x_1 \cdots x_t\right] = \Pr\left[X_{t+1} = x_{t+1} \mid X_t = x_t\right]$$



- Stationary Assumption: Transition probabilities are independent of time (t)

$$\Pr\left[X_{t+1} = b \mid X_t = a\right] = p_{ab}$$

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}$$

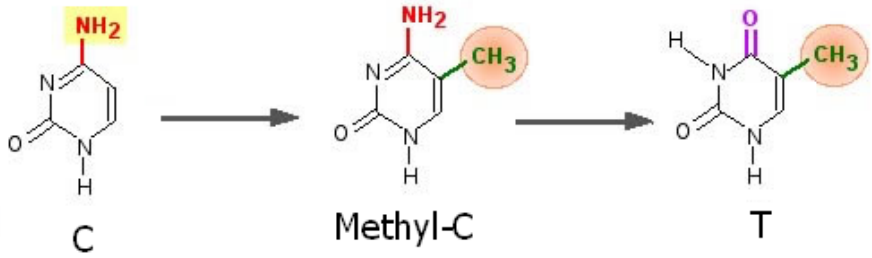
Markov Chain

- A set of *states* S = $\{s_1,\ldots,s_r\}$
- The process starts in one of these states and moves successively from one state to another
- If the chain is in state s_i, then it moves to states s_j at the next step with probability P_{ij} (transition matrix)
- This probability does not depend on which states the chain was in before the current state (Markov property)

$$P(x) = P(x_1)\cdot P(x_2 \mid x_1)\ldots P(x_n \mid x_1, x_2, \ldots, x_{n-1})$$
$$= P(x_1)\cdot P(x_2 \mid x_1)\ldots P(x_n \mid x_{n-1})$$

Using Markov Chains in Genome Search: CpG Islands

- The CG island is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions. It is also called the CpG island, where "p" simply indicates that "C" and "G" are connected by a phosphodiester bond.
- Whenever the dinucleotide CpG occurs, the C nucleotide is typically chemically modified by methylation.
  - C of CpG is methylated into methyl-C.
  - methyl-C mutates into T relatively easily.



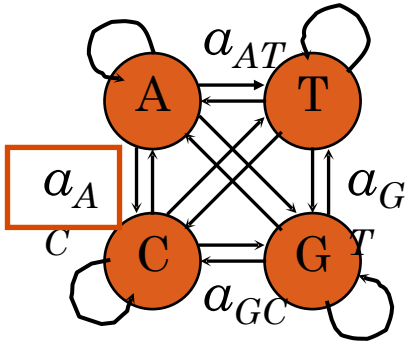
CpG Islands

- Hence the pair CG appears less than expected from what is expected from the independent frequencies of C and G alone. F(CpG) < F(C) * F(G)
- Due to biological reasons, methylation process is sometimes suppressed in short stretches of genomes such as in the start regions of many genes.
- These areas are called *CpG islands* (p denotes "pair").
- CpG islands have more CpG than elsewhere
- Identification of CpG islands is important for gene finidng

Modeling CpG Islands

- Questions:
  - Given a short stretch of genomic data, does it come from a CpG island ?
  - Given a long piece of genomic data, does it contain CpG islands in it, where, what length ?
- We "solve" the first question by modeling strings with and without CpG islands as Markov Chains over the same states {A,C,G,T} but different transition probabilities.

Markov Chains for CpG discrimination



- A state for each of the four letters A,C, G, and T in the DNA alphabet
- Arrow <-> probability of a residue following another residue

+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

- Training Set:
  - set of DNA sequences w/ known CpG islands
- Derive two Markov chain models:
  - **'+' model**: from the CpG islands
  - **'-' model**: from the remainder of sequence
- Transition probabilities for each model:

$$a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{t'} c_{st'}^{+}}$$

c_{st}^{+} is the number of times letter t followed letter s in the Cp G islands

- To use these models for discrimination, calculate the log-odds ratio:

$$S(x) = \log \frac{P(x|\text{model}+)}{P(x|\text{model}-)} = \sum_{i=1}^{L} \log \frac{a_{x_{i-1}x_i}^{+}}{a_{x_{i-1}x_i}^{-}}$$

Markov Chains for CpG discrimination

P^+ (For *CpG* islands):

+	A	C	G	T
A	0.18	0.27	0.43	0.12
C	0.17	0.23	**0.27**	$p_+(T\|C)$
G	0.16	0.25	0.45	$p_+(T\|G)$
T	0.08	$p_+(C\|T)$	$p_+(G\|T)$	$p_+(T\|T)$

P^- (For non *CpG* islands):

-	A	C	G	T
A	0.3	0.2	0.29	0.21
C	0.32	0.24	**0.08**	$p_-(T\|C)$
G	0.25	0.27	0.41	$p_-(T\|G)$
T	0.18	$p_-(C\|T)$	$p_-(G\|T)$	$p_-(T\|T)$

- x = CGCGG

$P(CGCGG|+) = P(C)P(G|C,+)P(C|G,+)P(G|C,+)P(G|G,+)=P(C)$ x 0.27 x 0.25 x 0.27 x 0.45 = 0.00820125

$P(CGCGG|-) = P(C)P(G|C,-)P(C|G,-)P(G|C,-)P(G|G,-)=P(C)$ x 0.08 x 0.27 x 0.08 x 0.41 = 0.00070848

→ $P(CGCGG|+)/P(CGCGG|-) = 11.5758384 > 1$, or

$\log(P(CGCGG|+) / P(CGCGG|-)) = \log(P(G|C,+)/P(G|C,-)) + \log(P(C|G,+)/P(C|G,-)) +$

$\log(P(G|C,+)/P(G|C,-)) + \log(P(G|G,+)/P(G|G,-))$

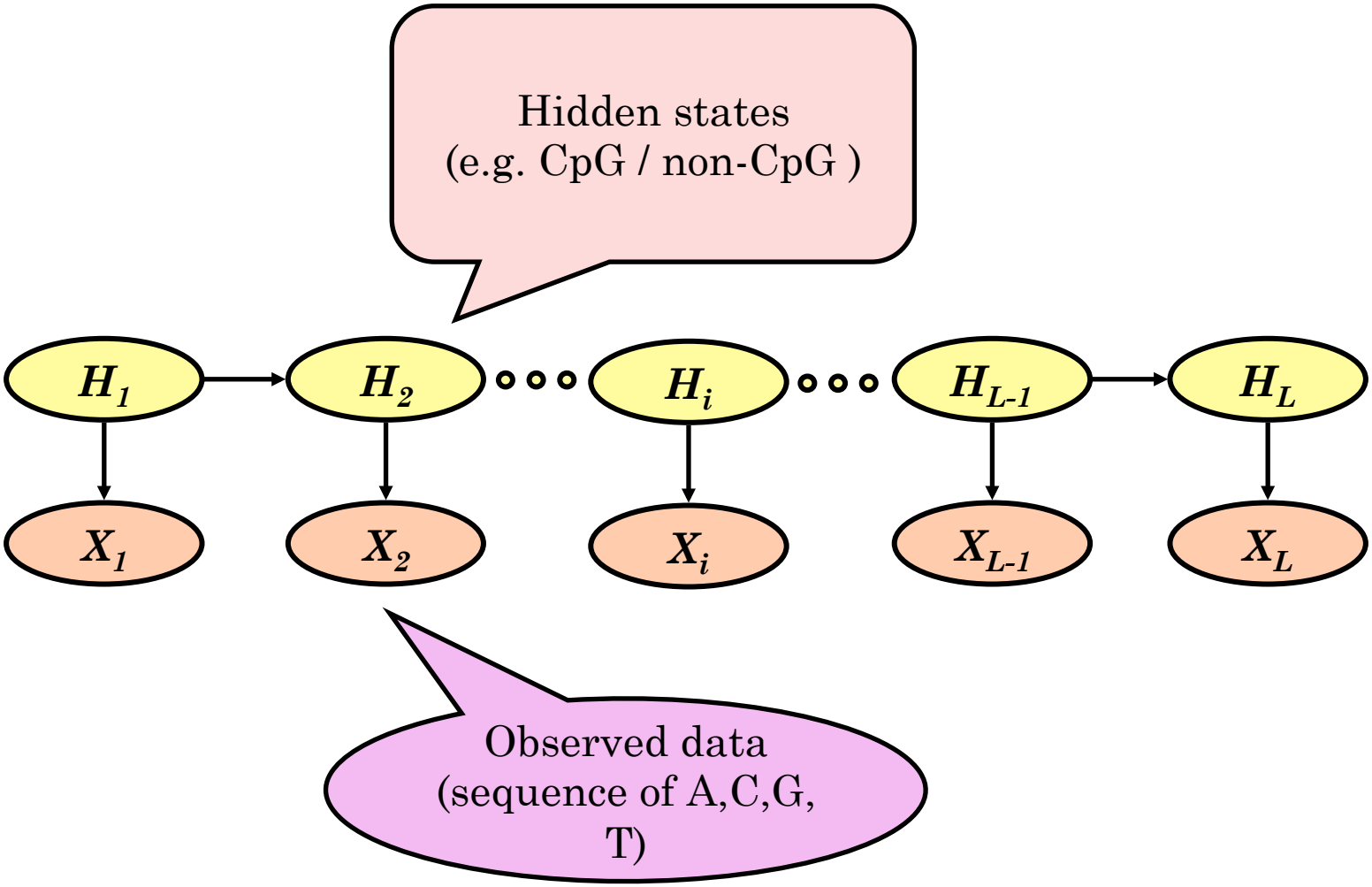
$= \log(0.27/0.08) + \log(0.25/0.27) + \log(0.27/0.08) + \log(0.45/0.41)$

$= 2.44892 > 0$

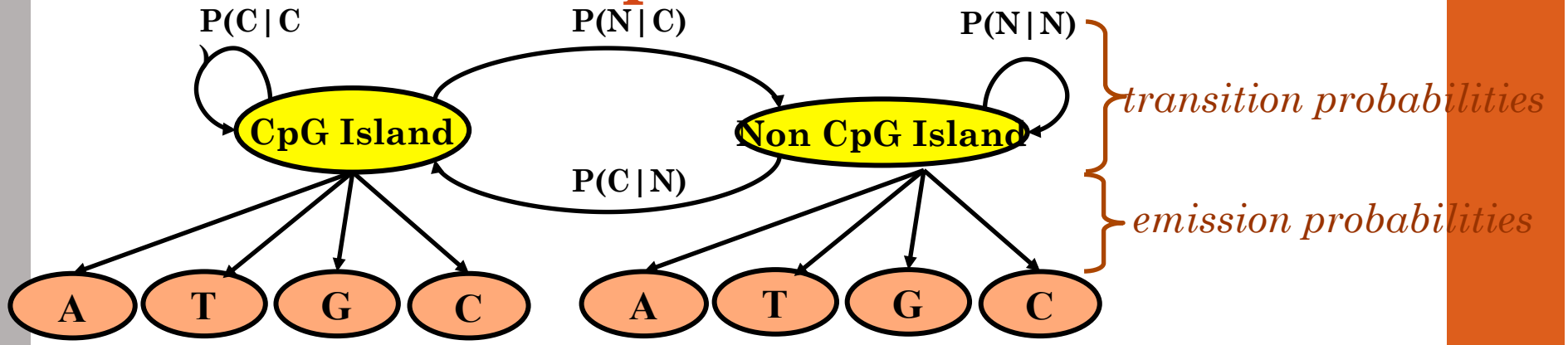
Modeling CpG Islands

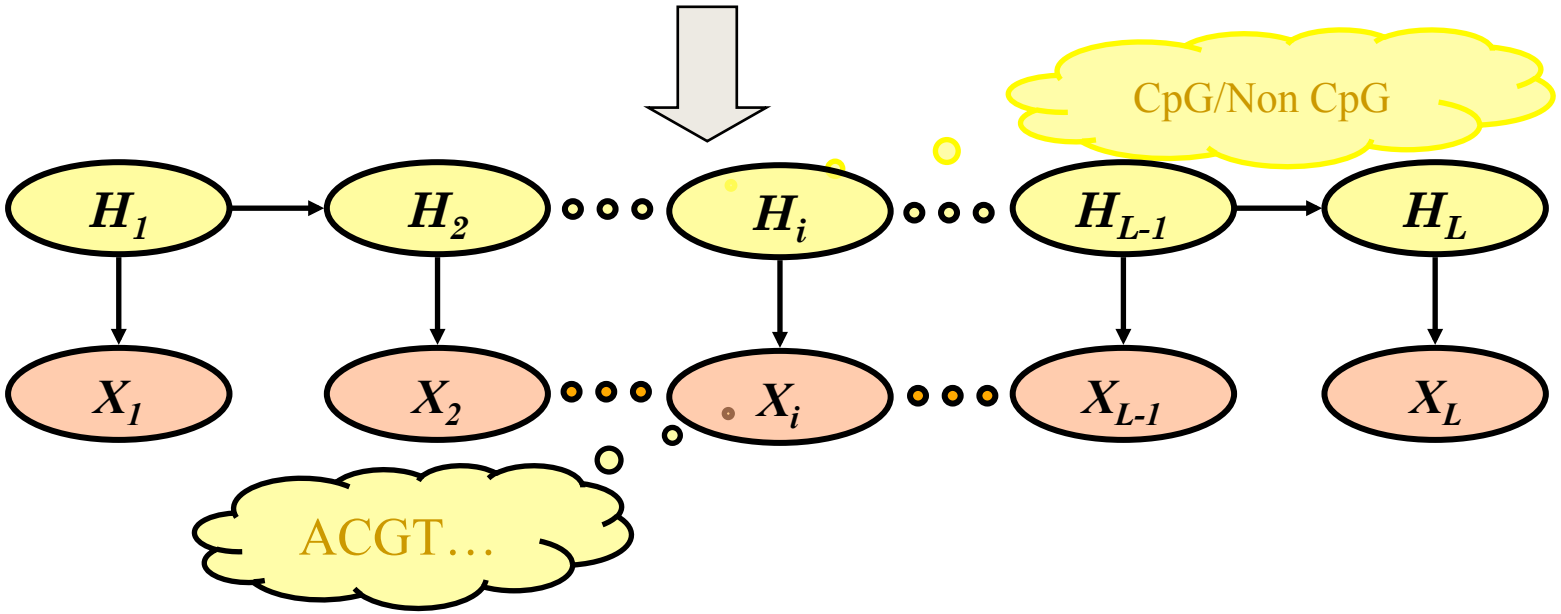
- Questions:
  - Given a short stretch of genomic data, does it come from a CpG island ?
  - Given a long piece of genomic data, does it contain CpG islands in it, where, what length ?
- We "solve" the second question by modeling strings with the so called Hidden Markov Model

Hidden Markov Model – HMM



Hidden Markov Model: CpG Islands example





HMM Definition

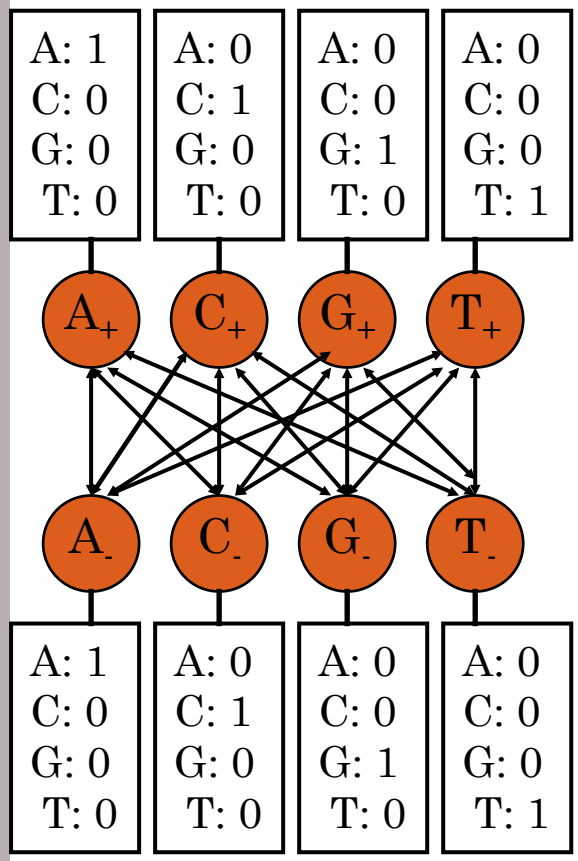
How do we find CpG islands in a long sequence?

- An HMM is a 5-tuple (S,V,p,A,E) where:
  - S is a finite set of states, $|S|=K$
  - V is a finite set of observation symbols per state, $|V|=M$
  - p is the initial state probabilities
  - A is the state transition probabilities, denoted by a_{jk} for each j, k in S
  - For each j, k in S, the transition probability is
  - $a_{jk} = P(s_t=k \mid s_{t-1}=j)$
  - E is a probability emission matrix, $e_{ki}=P(x_t=v_i \mid s_t = k)$

- Only emitted symbols are observable by the system but not the underlying random walk between states -> "hidden"

HMM for CpG islands

How do we find CpG islands in a long sequence?



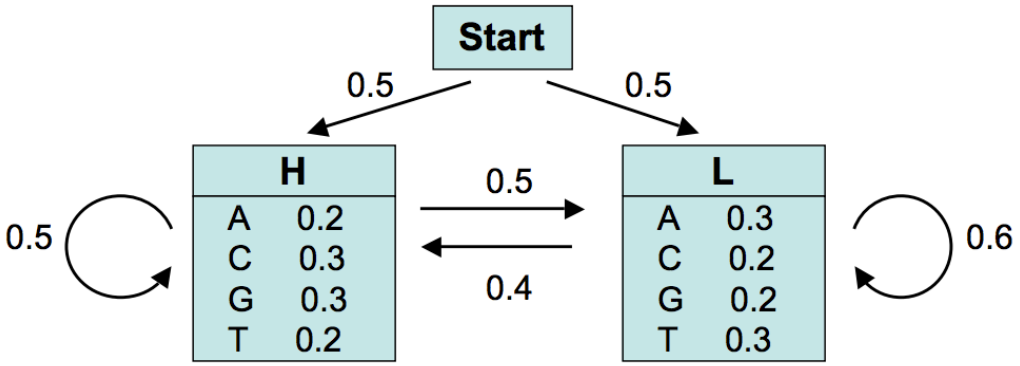
Note: Each set ('+' or '-') has an addition al set of transitions as in previous Mark ov chain

- Build a single model that combines both Markov chains:
  - **'+' states**: A_+, C_+, G_+, T_+
    - Emit symbols: A, C, G, T in CpG islands
  - **'-' states**: A_-, C_-, G_-, T_-
    - Emit symbols: A, C, G, T in non-islands
- If a sequence CGCG is emitted by states (C_+,G_-,C_-,G_+), then:

$$P(CGCG) = a_{0,C_+} \times a_{C_+,G_-} \times a_{G_-,C_-} \times a_{C_-,G_+}$$

- In general, we DO NOT know the path. How to estimate the path?

HMM with 2 hidden states

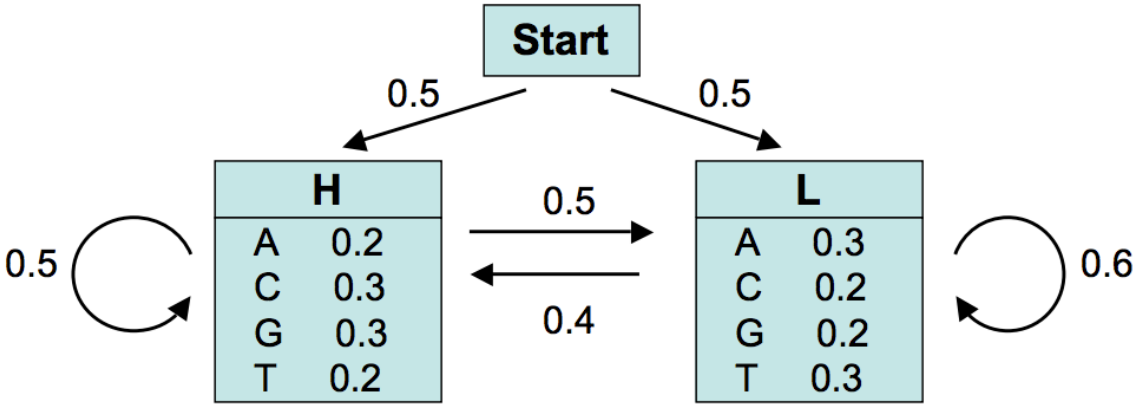


- This model is composed of 2 hidden states, H (high GC content: +) and L (low GC content: -).
- The model can then be used to predict the region of coding DNA from a given sequence.

Sources: For the theory, see Durbin et al (1998);

- For the example, see Borodovsky & Ekisheva (2006), pp 80-81

HMM with 2 hidden states



Consider the sequence S= **GGCACTGAA**

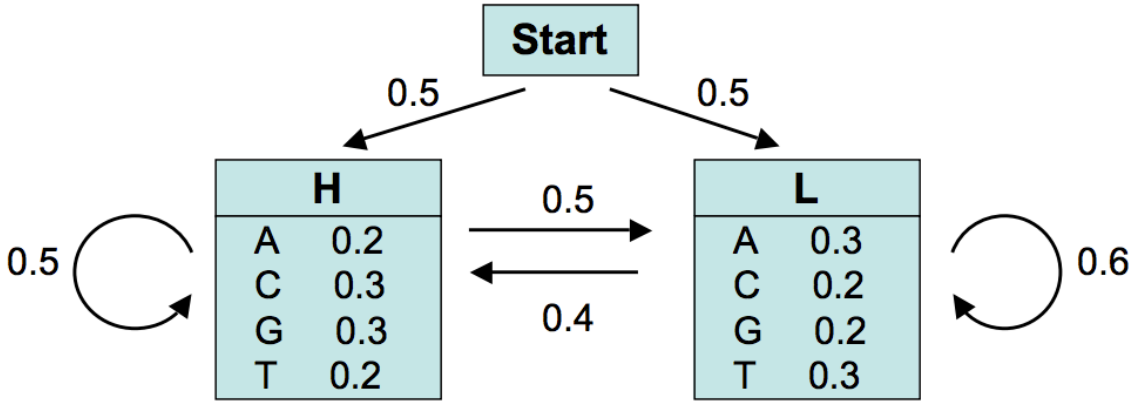
There are several paths through the hidden states (H and L) that lead to the given sequence S.

Example: P = **LLHHHHLLL**

The probability of the HMM to produce sequence S through the path P is:

$$p = p_L(0) * p_L(G) * p_{LL} * p_L(G) * p_{LH} * p_H(C) * \ldots$$
$$= 0.5 * 0.2 * 0.6 * 0.2 * 0.4 * 0.3 * \ldots$$
$$= \ldots$$

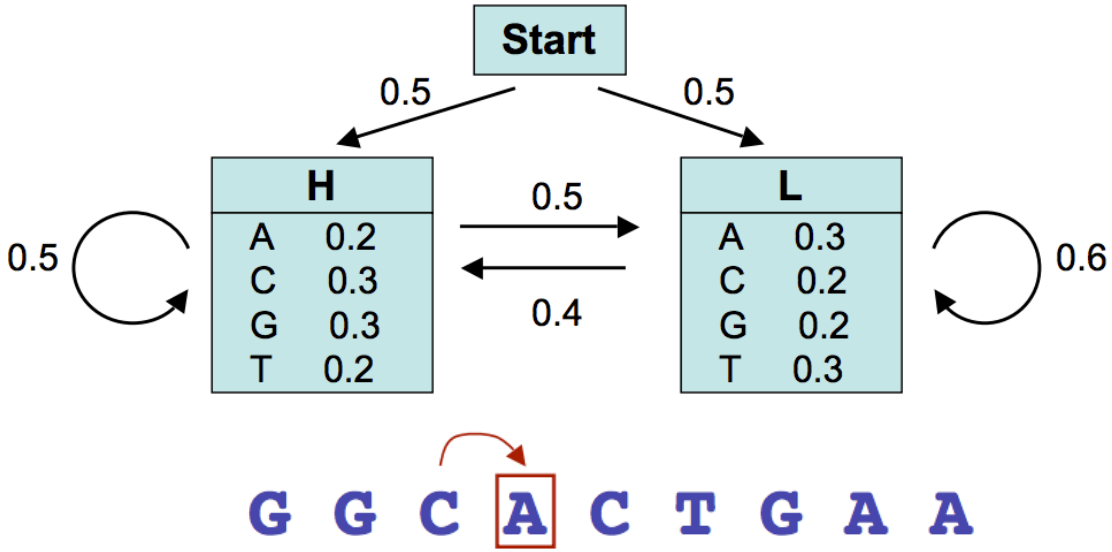
HMM: Decoding



- There are many possible paths through the hidden states (H & L), but they do not have the same probability

- Decoding problem: compute the most probable path (w/ the highest probability) among all possible hidden state sequence
  - Viterbi algorithm

HMM: Viterbi algorithm



- The probability of the most probable path ending in state *l* with observation "*i*" is

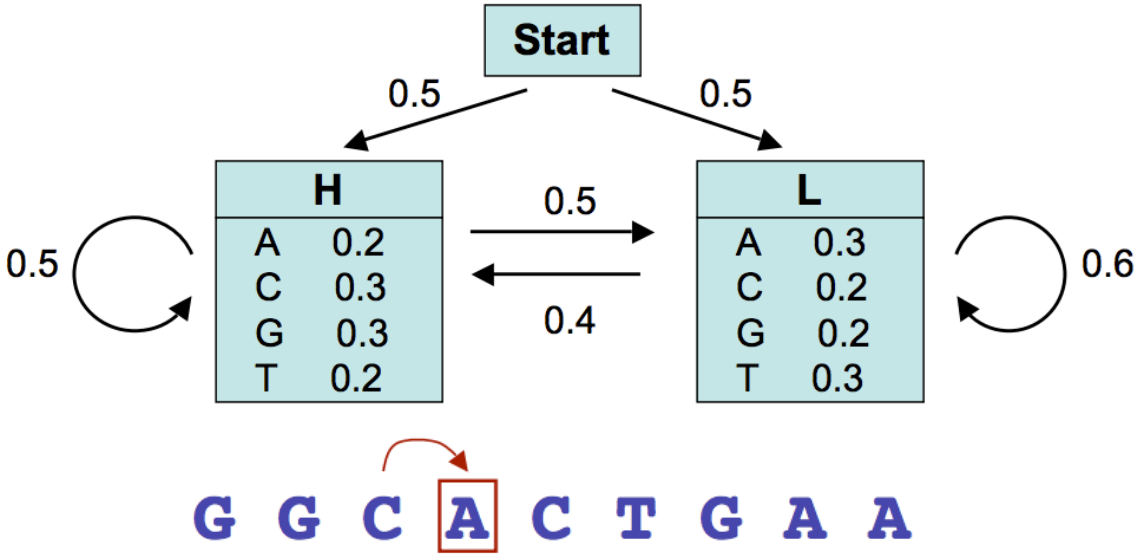
$$p_l(i,x) = e_l(i)\max_k\left(p_k(j,x-1).p_{kl}\right)$$

probability to observe element *i* in state *l*

probability of the most probable path ending at position x-1 in state *k* with element *j*

probability of the transition from state *l* to state *k*

HMM: Viterbi algorithm



The probability of the most probable path ending in state **k** with observation "i" is

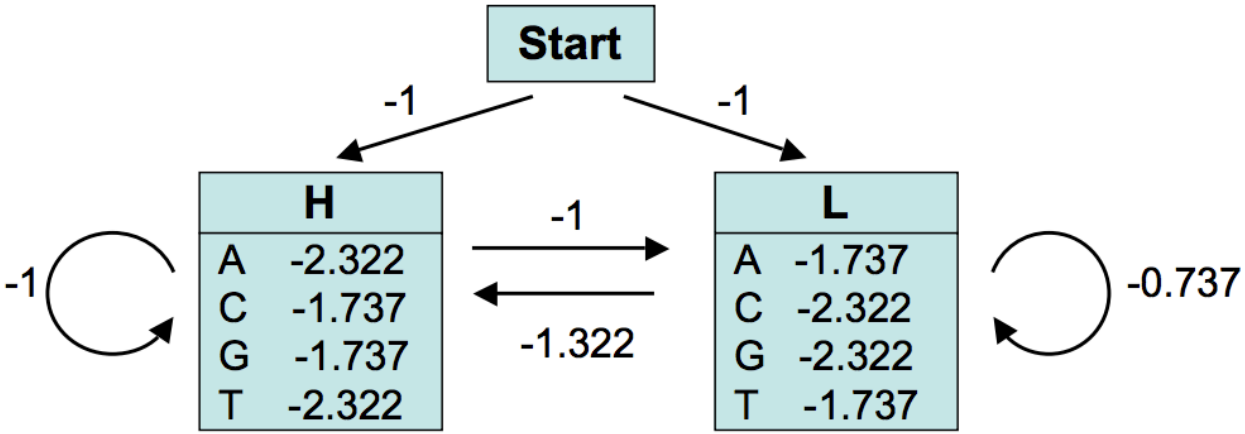
$$p_l(i,x) = e_l(i)\max_k\left(p_k(j,x-1).p_{kl}\right)$$

In our example, the probability of the most probable path ending in state **H** with observation "A" at the 4th position is:

$$p_H(A,4) = e_H(A)\max\left(p_L(C,3)p_{LH}, p_H(C,3)p_{HH}\right)$$

We can thus compute recursively (from the first to the last element of our sequence) the probability of the most probable path.

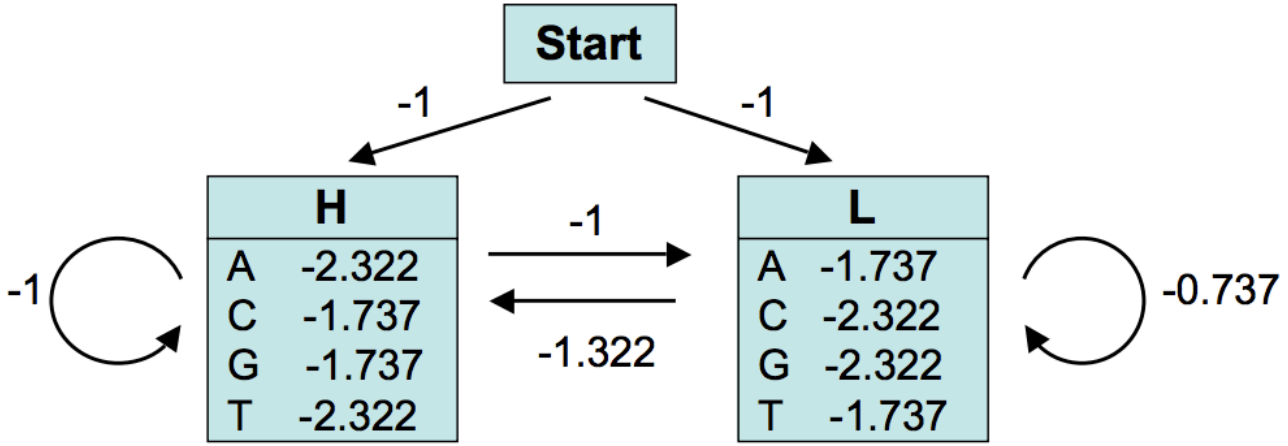
HMM: Viterbi algorithm



GGCACTGAA

- It is convenient to use the log of the probabilities (rather than the probabilities themselves). This allows us to compute sums instead of products, which is more efficient and accurate. We used here log2(p).

HMM: Viterbi algorithm



GGCACTGAA

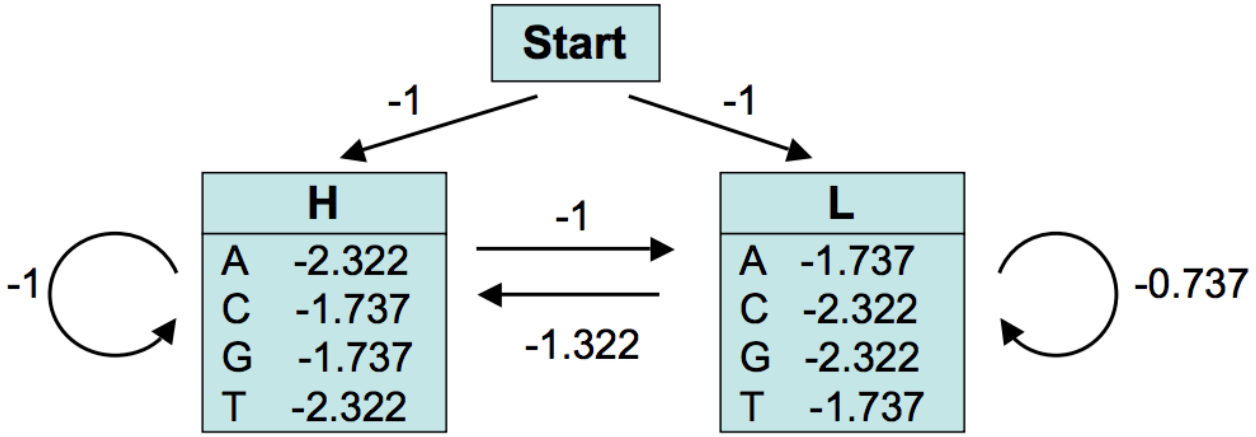
Probability (in $\log_2$) that G at the first position was emitted by state **H**

$p_H(G,1) = -1 - 1.737 = -2.737$

Probability (in $\log_2$) that G at the first position was emitted by state **L**

$p_L(G,1) = -1 - 2.322 = -3.322$

HMM: Viterbi algorithm



GGCACTGAA

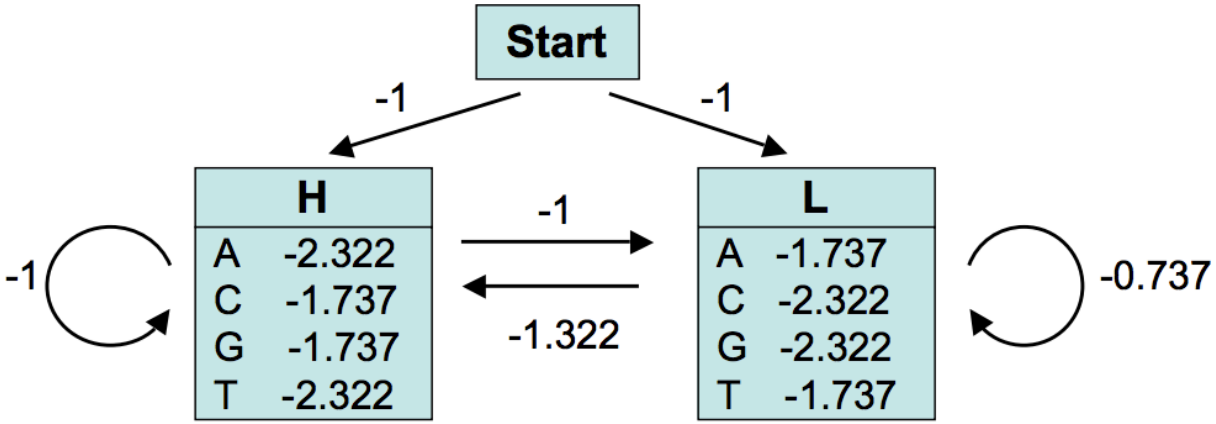
Probability (in $\log_2$) that G at the 2nd position was emitted by state **H**

$$p_H(G,2) = -1.737 + \max(p_H(G,1)+p_{HH},\ p_L(G,1)+p_{LH})$$
$$= -1.737 + \max(-2.737 - 1,\ -3.322 - 1.322)$$
$$= -5.474 \text{ (obtained from } p_H(G,1))$$

Probability (in $\log_2$) that G at the 2nd position was emitted by state **L**

$$p_L(G,2) = -2.322 + \max(p_H(G,1)+p_{HL},\ p_L(G,1)+p_{LL})$$
$$= -2.322 + \max(-2.737 - 1.322,\ -3.322 - 0.737)$$
$$= -6.059 \text{ (obtained from } p_H(G,1))$$

HMM: Viterbi algorithm



GGCACTGAA

	G	G	C	A	C	T	G	A	A
H	**-2.73**	**-5.47**	**-8.21**	-11.53	-14.01	...			-25.65
L	-3.32	-6.06	-8.79	**-10.94**	**-14.01**	...			**-24.49**

The most probable path is: **HHHLLLLLLL**

Its probability is $2^{-24.49}$ = 4.25E-8
(remember that we used $\log_2(p)$)

Summary: Viterbi Algorithm

- Initialization (i=0): $p_0(0) = 1, p_k(0) = 0 \text{ for } k > 0$
- Recursion (i=1…L):

$$p_l(i+1) = e_{l,x_{i+1}} \max_k (p_k(i) a_{kl})$$

$$ptr_i(l) = \arg\max_k (p_k(i-1) a_{kl})$$

- Termination:

$$P(x,\pi^*) = \max_k (p_k(L) a_{k0})$$

$$\pi^*_L = \arg\max_k (p_k(L) a_{k0})$$

- Traceback (i=L…1):

$$\pi^*_{i-1} = ptr_i(\pi^*_i)$$

Computational Complexity:

- Brute Force: $O(N^L)$
- Viterbi: $O(L*N^2)$
- N – number of states
- L – sequence length

Likelihood of evidence

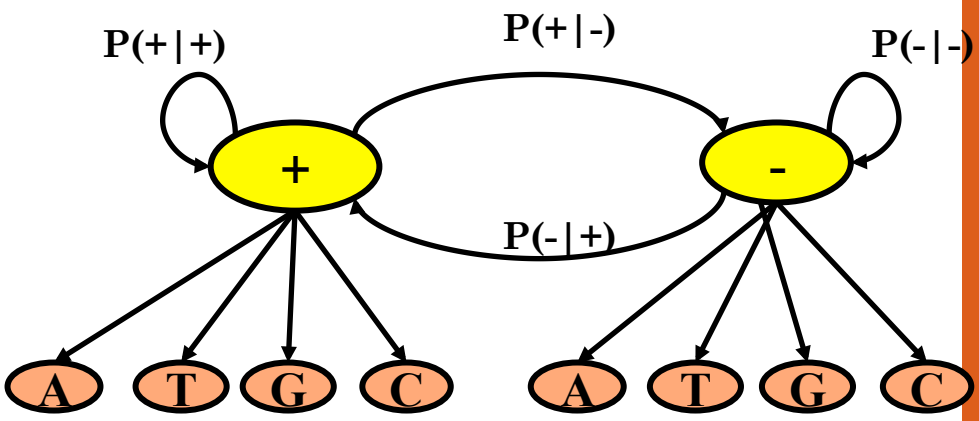
- What is the probability of getting the observation sequence under the current HMM model (transition & emission probabilities)?

- Can be used to evaluate which HMM model is more likely to have produced the observation sequence

HMM – Training problem

From raw sequence data ....

To Parameters



Parameter Estimation for HMMs (Case 1)

- Case 1: All the paths/hidden state labels in the set of training sequences are known:
  - Use the Maximum Likelihood (ML) estimators for:

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \text{ and } e_{kx} = \frac{E_k(x)}{\sum_{x'} E_k(x')}$$

  - Where A_{kl} and $E_k(x)$ are the number of times each transition or emission is used in training sequences

  - Drawbacks of ML estimators:
    - Vulnerable to overfitting if not enough data
    - Estimations can be undefined if never used in training set (add pseudocounts to reflect a prior biases about probability values)

Parameter Estimation for HMMs (Case 2)

- Case 2: The paths/labels in the set of training sequences are UNknown:
  - Use Iterative methods (e.g., Baum-Welch):
    1. Initialize a_{kl} and e_{kx} (e.g., randomly)
    2. Estimate A_{kl} and $E_k(x)$ using current values of a_{kl} and e_{kx}
    3. Derive new values for a_{kl} and e_{kx}
    4. Iterate Steps 2-3 until some stopping criterion is met (e.g., change in the total log-likelihood is small)
  - Drawbacks of Iterative methods:
    - Converge to local optimum
    - Sensitive to initial values of a_{kl} and e_{kx} (Step 1)
    - Convergence problem is getting worse for large HMMs

HMM- Baum-Welch algorithm

- Given the observation sequence, learn the HMM parameters iteratively
  - Using the current HMM parameter, estimate the hidden state sequence probability
  - Using the new hidden state assignment, update the HMM parameters
  - Repeat this until convergence

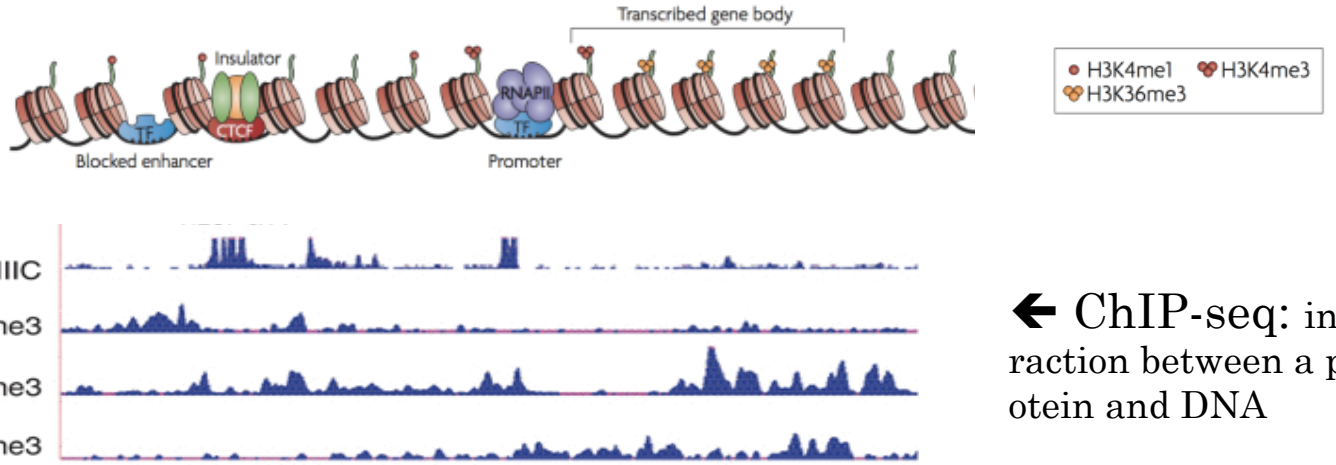
HMM Applications in Bioinformatics

- HMMs can be applied efficiently to well known biological problems
  - Gene finding
  - Protein secondary structure recognition
  - Multiple sequence alignment
  - Splicing signals prediction
  - Chromatin state segmentation

Next-generation genomics: an integrative approach (2010). *Nature Reviews Genetics* 11, 476-486

Annotating the genome

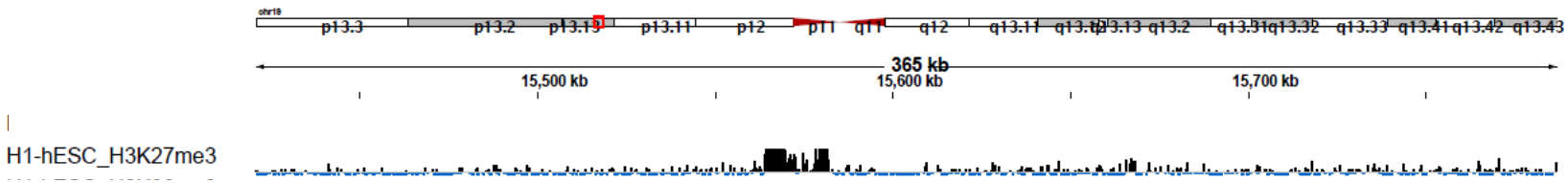
through detecting TFBS and histone-modification states



← ChIP-seq: interaction between a protein and DNA

Promoter	By TFBSs, or localization of ***H3K4me3***
Transcribed gene body	*By localization of* ***H3K36me3***
Enhancer	By distal TF binding sites, or by ***H3K4me1***

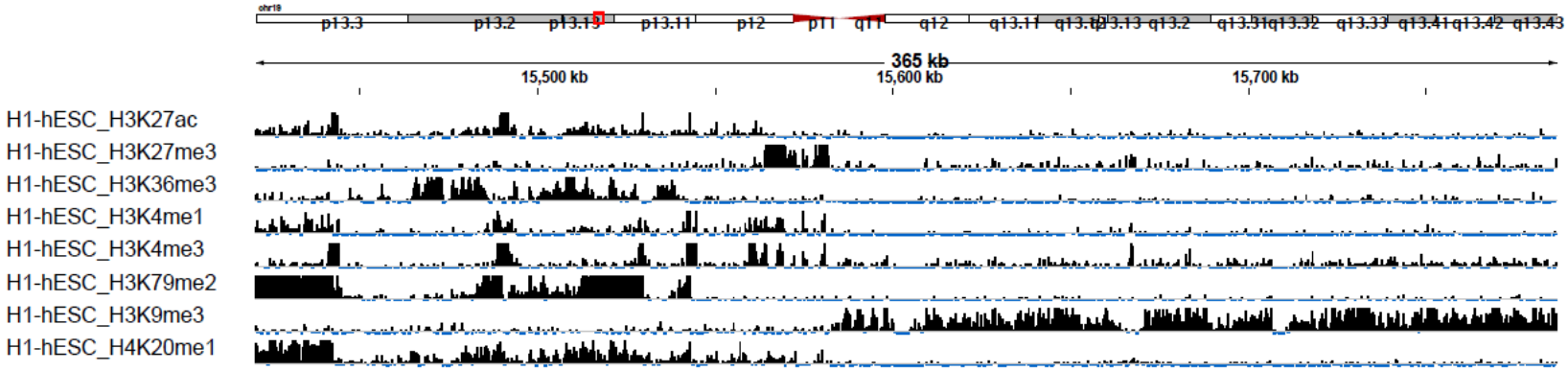
Analysis of histone modification state: single track



Modification state

Either ON or OFF ➔ binary state (K=2)

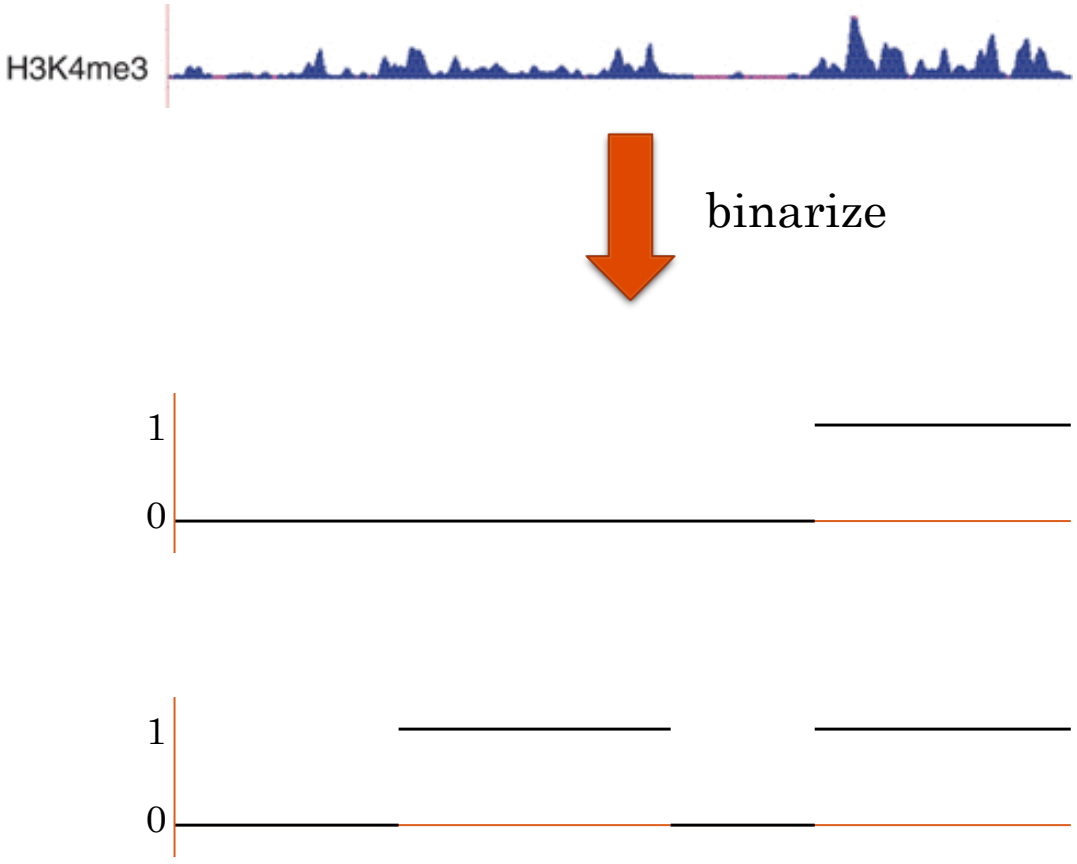
Joint analysis of multiple signal tracks



m signal tracks, each with 2 possible states

How many possible states in total?

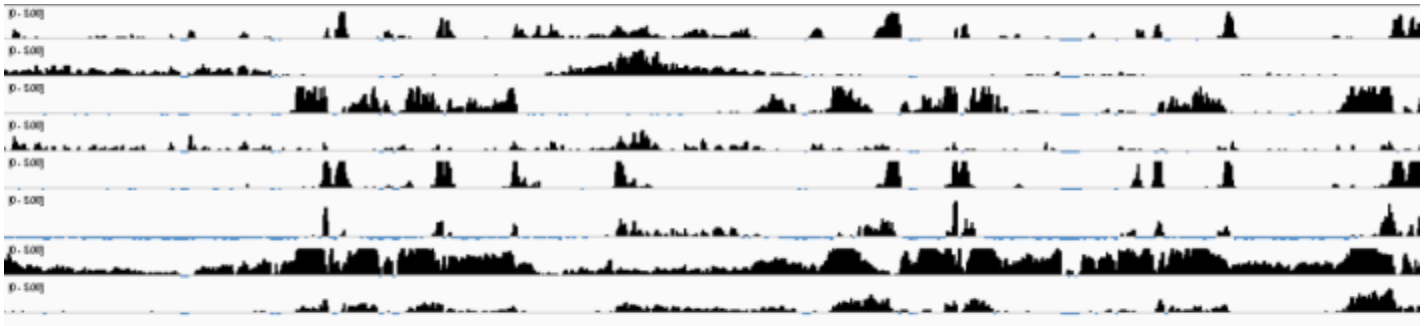
Analysis of histone modification state



Chromatin state segmentation using ChIP-seq data

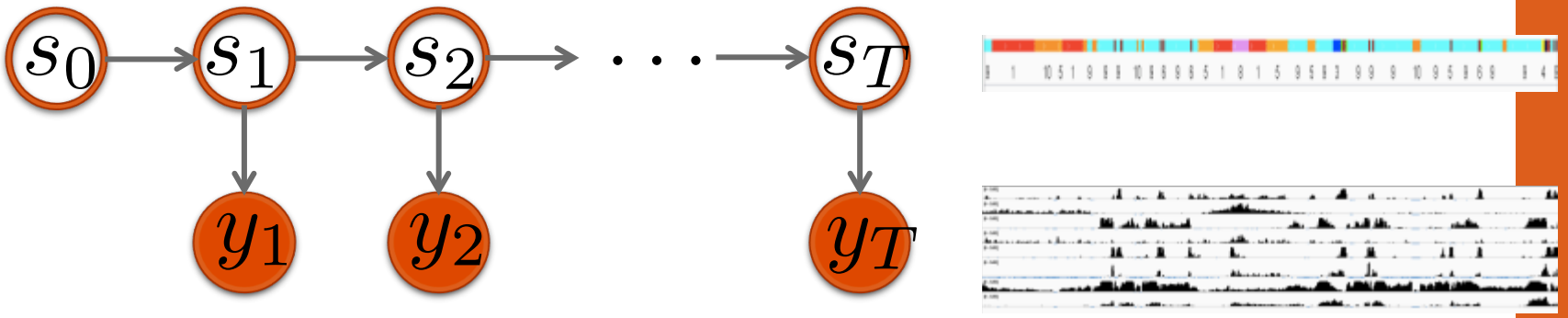
Joint analysis of multiple signal tracks

Input



Output

HMM for chromatin state inference



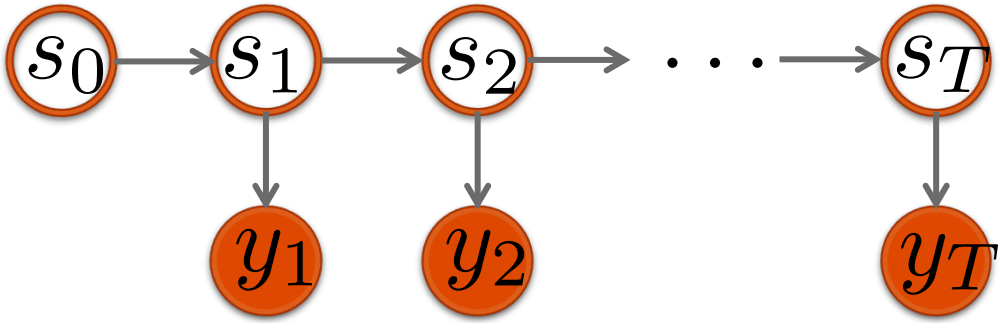
- Observation sequence: multiple signal tracks
  - $y_t \in \mathbb{R}^m$
  - e.g.

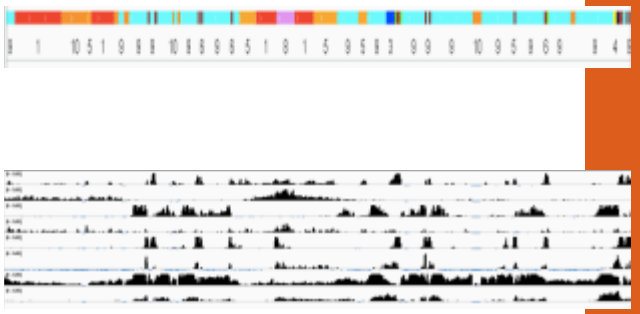
$$y_t = (0.01, 0.05, 0.02, 2.0, 004, 0.9, 0.1)^T, or$$

$$y_t = (0, 0, 0, 1, 0, 1, 0)^T$$

- Hidden state: functional element
  - e.g. 1:promoter, 2:enhancer, 3:transcribed gene body, 4:repressive
    ➔ 2 2 1 1 3 3 3 3 3 4 4 4 4 4 4 4 4 4

HMM for chromatin state inference



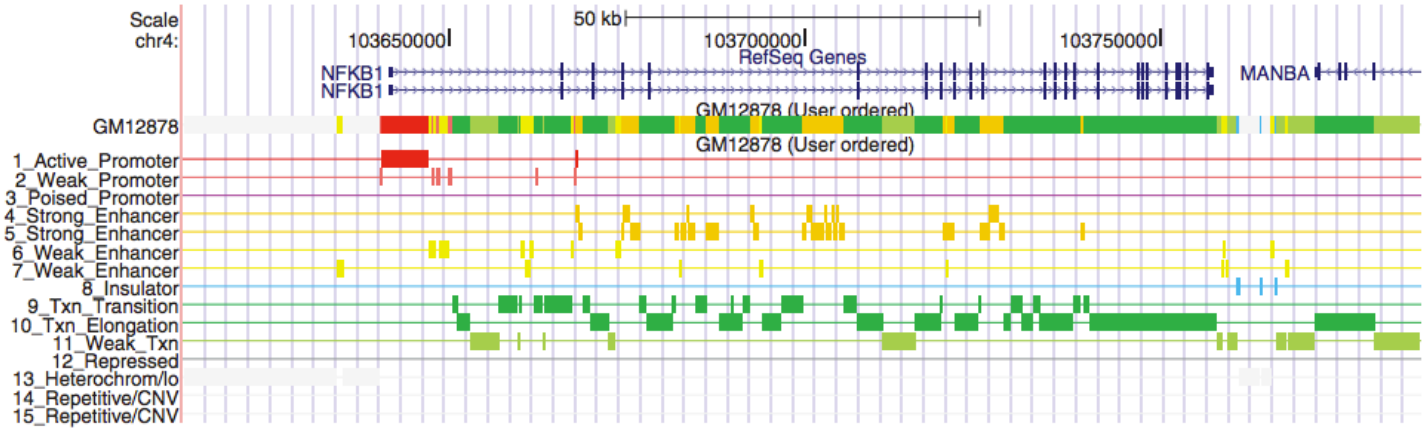


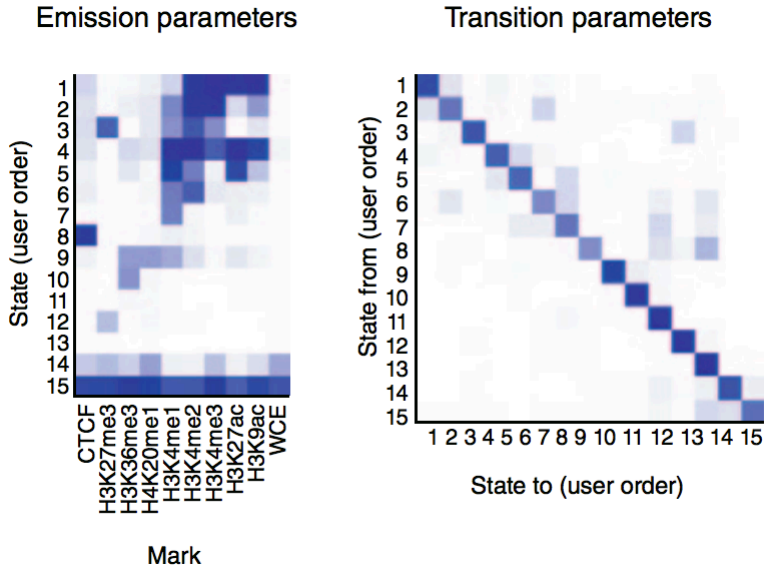
- Transition model
  - Transition probability between functional elements
  - Can be estimated by Baum-Welch algorithm

$$y_t \mid s_t = k \sim N(\mu_k, \Sigma_k)$$

- Emission model
  - Multi-variable Gaussian model
  - μ_k Average signal strength of m tracks on hidden state k
  - Σ_k Covariance matrix

Chromatin state segmentation by ChromHMM





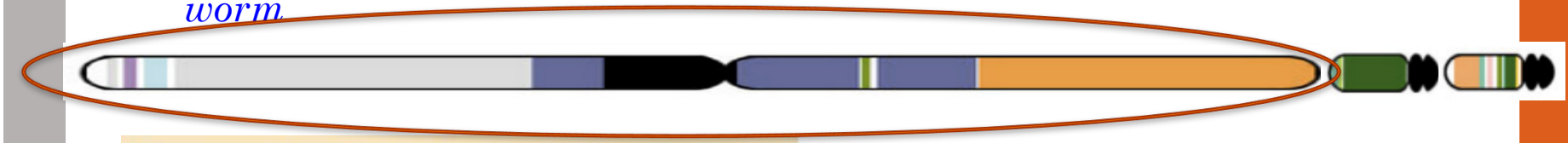
BIML 2016

Software

- ChromHMM (Ernst and Kellis, 2012)
  - Hidden Markov Model on binary tracks
  - Relatively fast, but tend to over-segment
- Segway (Hoffman et al., 2012)
  - Dynamic Bayesian Network on continuous tracks
  - Slow
- hiHMM (Sohn et al., 2015)
  - For joint inference from multiple related genomes

Cross-species chromatin state comparison

- Motivated by a recent modENCODE (model organism encyclopedia of DNA elements) project
  - Aims to *systematically compare chromatin organization in human, fly, and worm*



Genome size comparison

Species	Chromosomes	Genes	Base pairs
Human *(Homo sapiens)*	46 (23 pairs)	28-35,000	3.1 billion
Mouse *(Mus musculus)*	40	22.5-30,000	2.7 billion
Puffer fish *(Fugu rubripes)*	44	31,000	365 million
Malaria mosquito *(Anopheles gambiae)*	6	14,000	289 million
Fruit fly *(Drosophila melanogaster)*	8	14,000	137 million
Roundworm *(C. elegans)*	12	19,000	97 million
Bacterium* *(E. coli)*	1	5,000	4.1 million

*Bacterial chromosomes are chromonemes, not true chromosomes

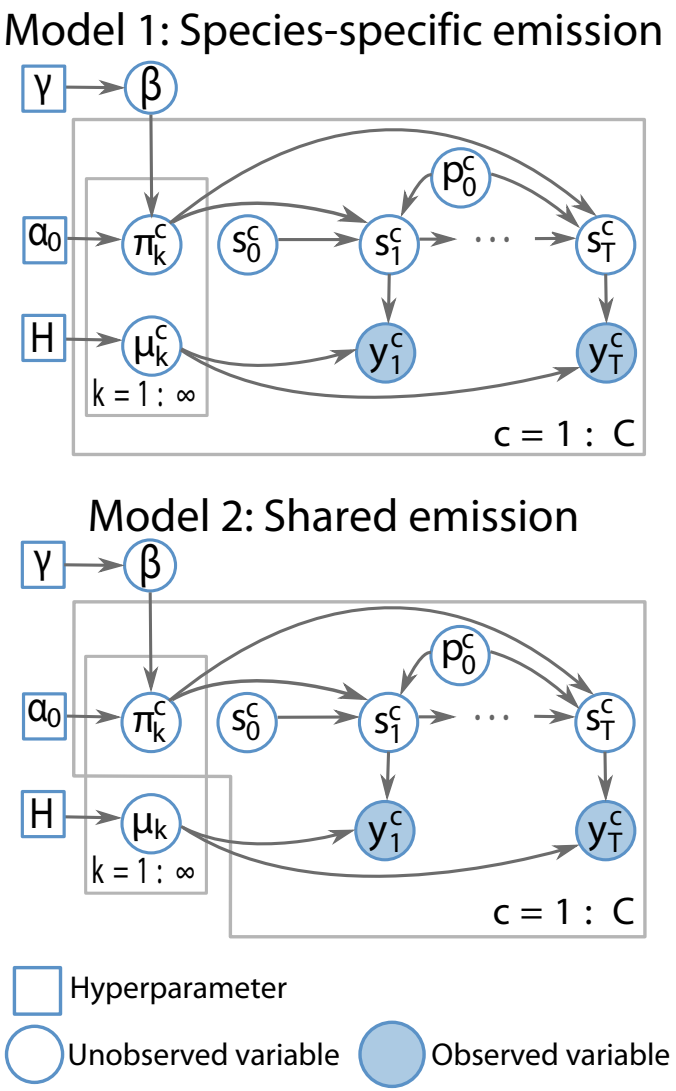
JOHN BLANCHARD / *The Chronicle*

- Challenges caused by
  - different genome sizes
  - different dynamic ranges in ChIP signal
  - the need to capture common chromatin states while allowing for species-specific patterns of histone modifications

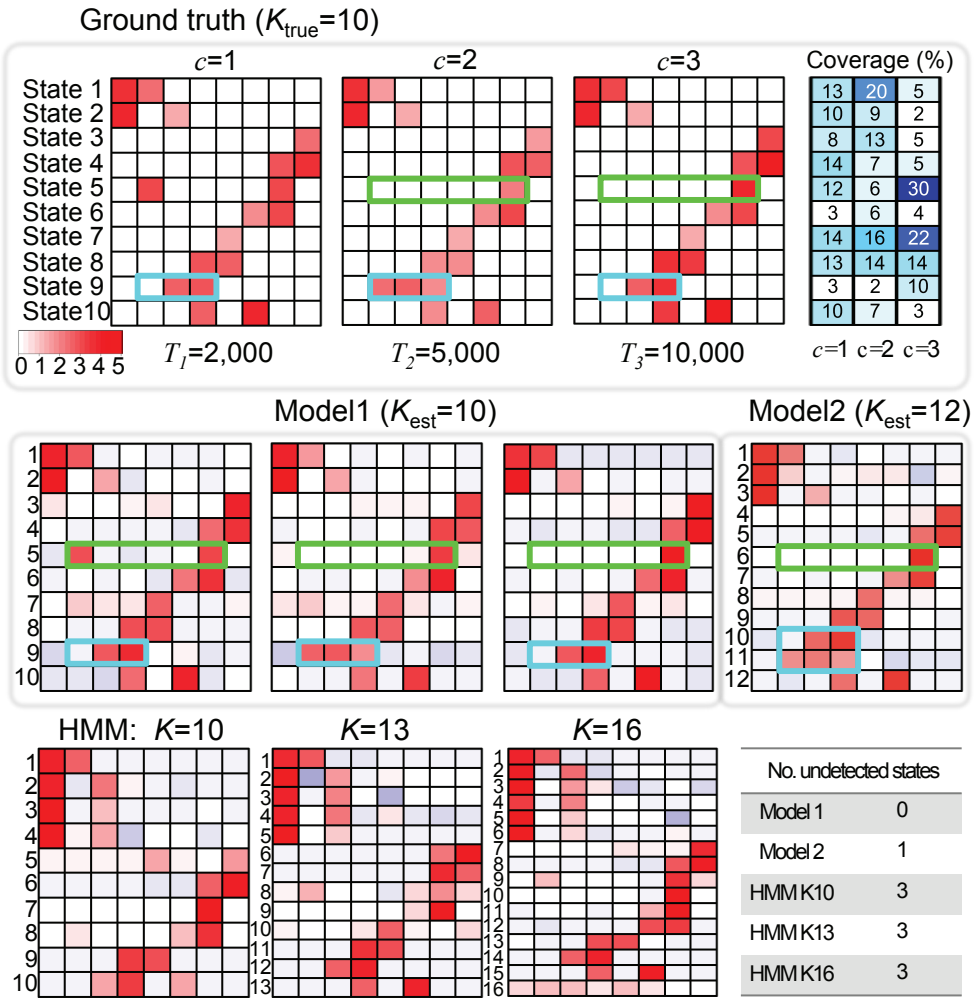
hiHMM

: hierarchically linked HMM

Graphical model representation of hiHMM model 1 and 2

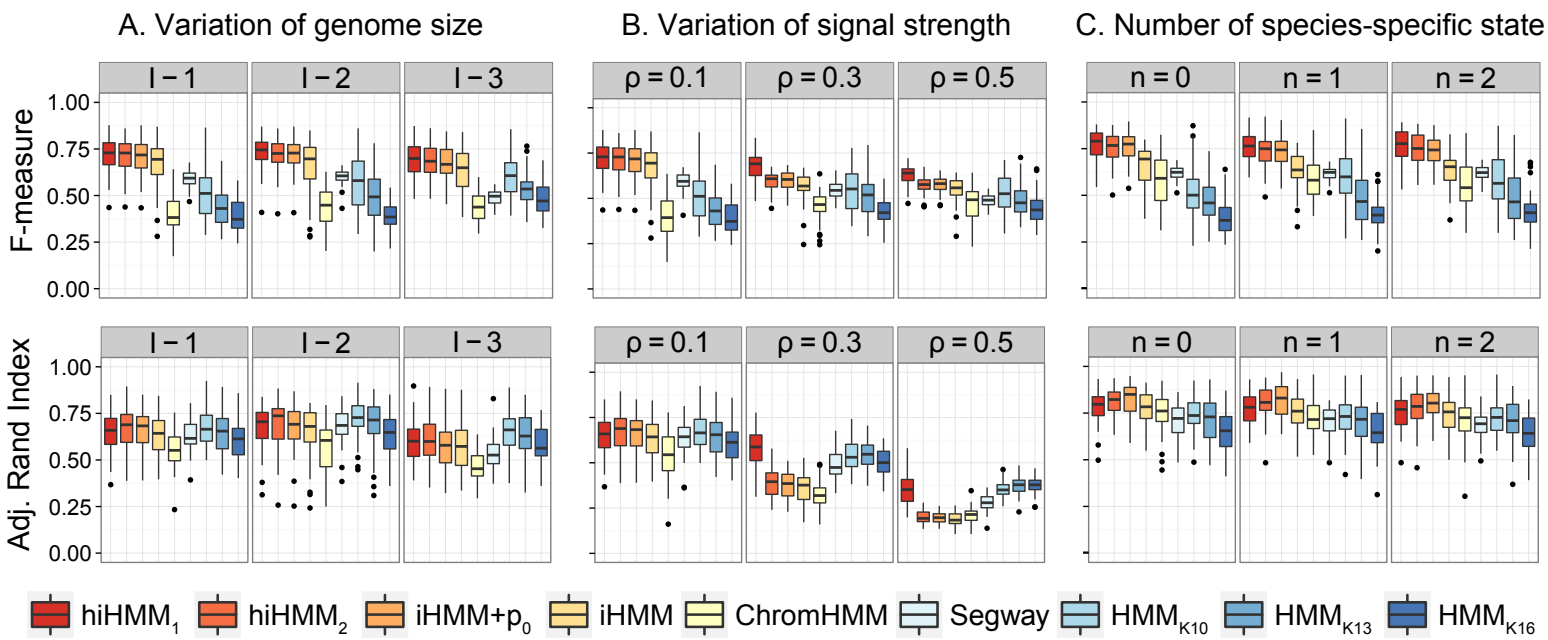


Simulation study



No. undetected states	
Model 1	0
Model 2	1
HMM K10	3
HMM K13	3
HMM K16	3

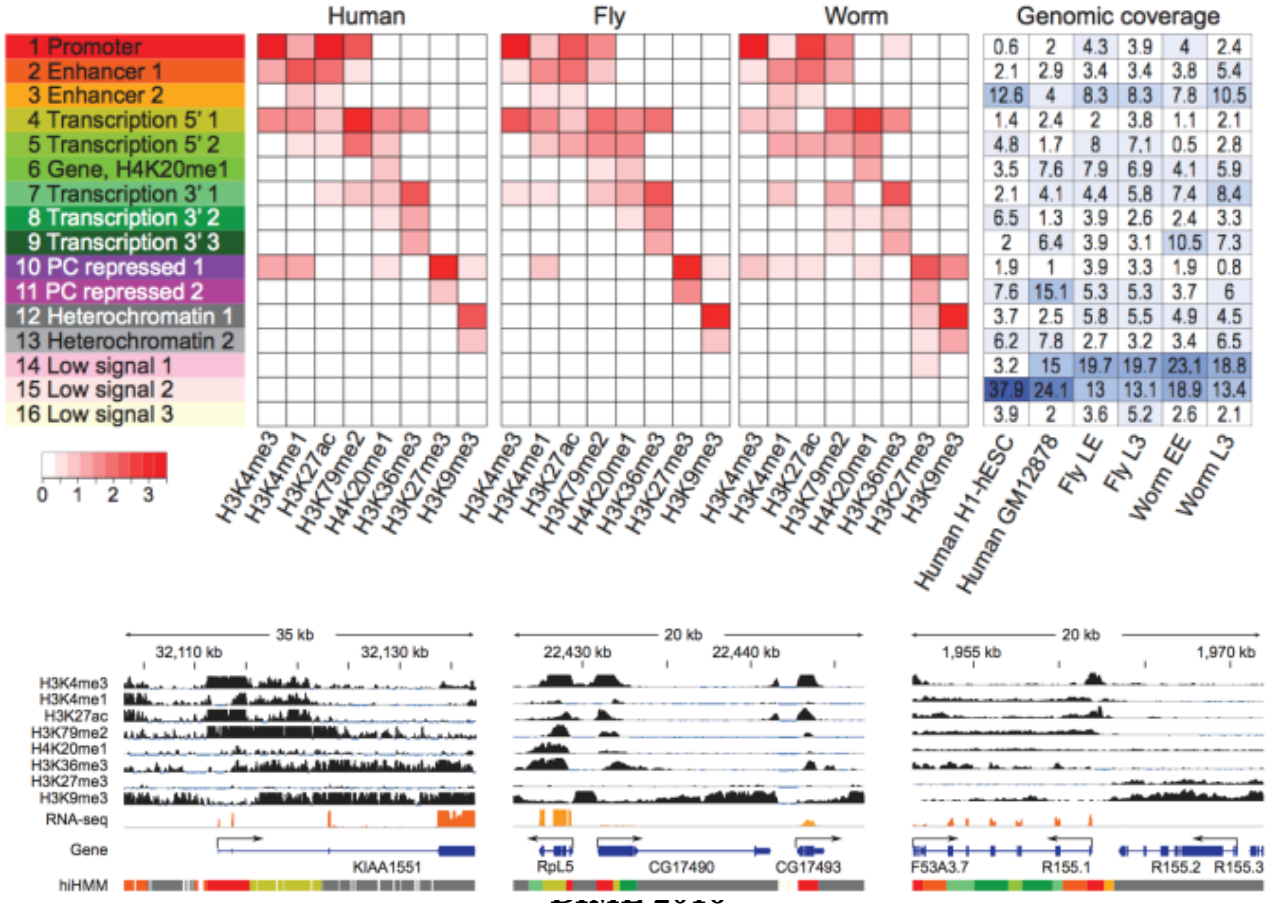
Simulation study



Ho, et al. Nature 2014

Comparative analysis of metazoan chromatin organization

- hiHMM applied to modENCODE data analysis



HMM Summary

- In general, HMM states and transitions are designed based on the knowledge of the problem under study

- The Viterbi algorithm is used to compute the most probable path (as well as its probability). It requires knowledge of the parameters of the HMM model and a particular output sequence

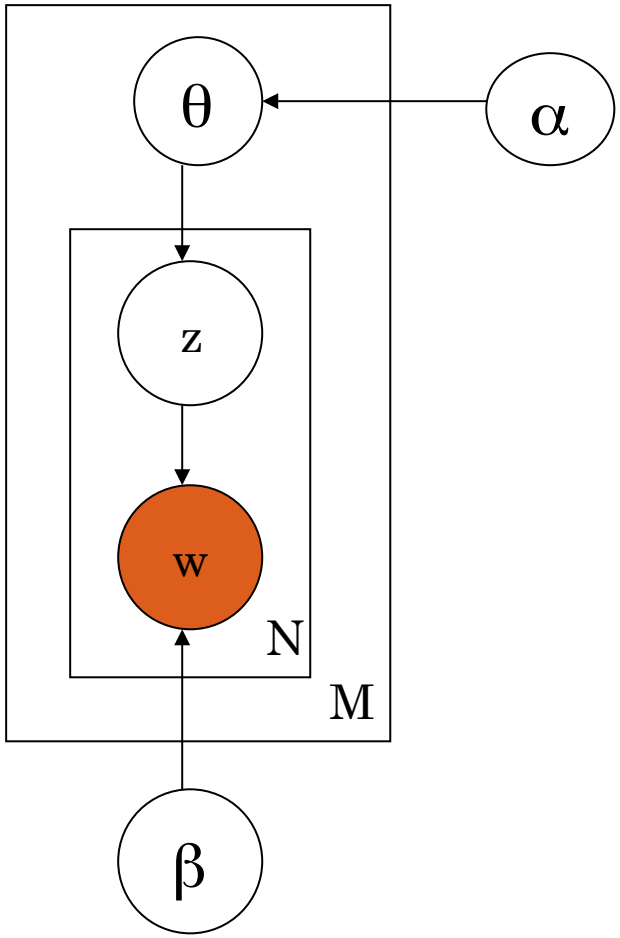
- To create a HMM model, we need a set of (training) sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the Baum-Welch algorithm.

Topic model: Latent Dirichlet Allocation

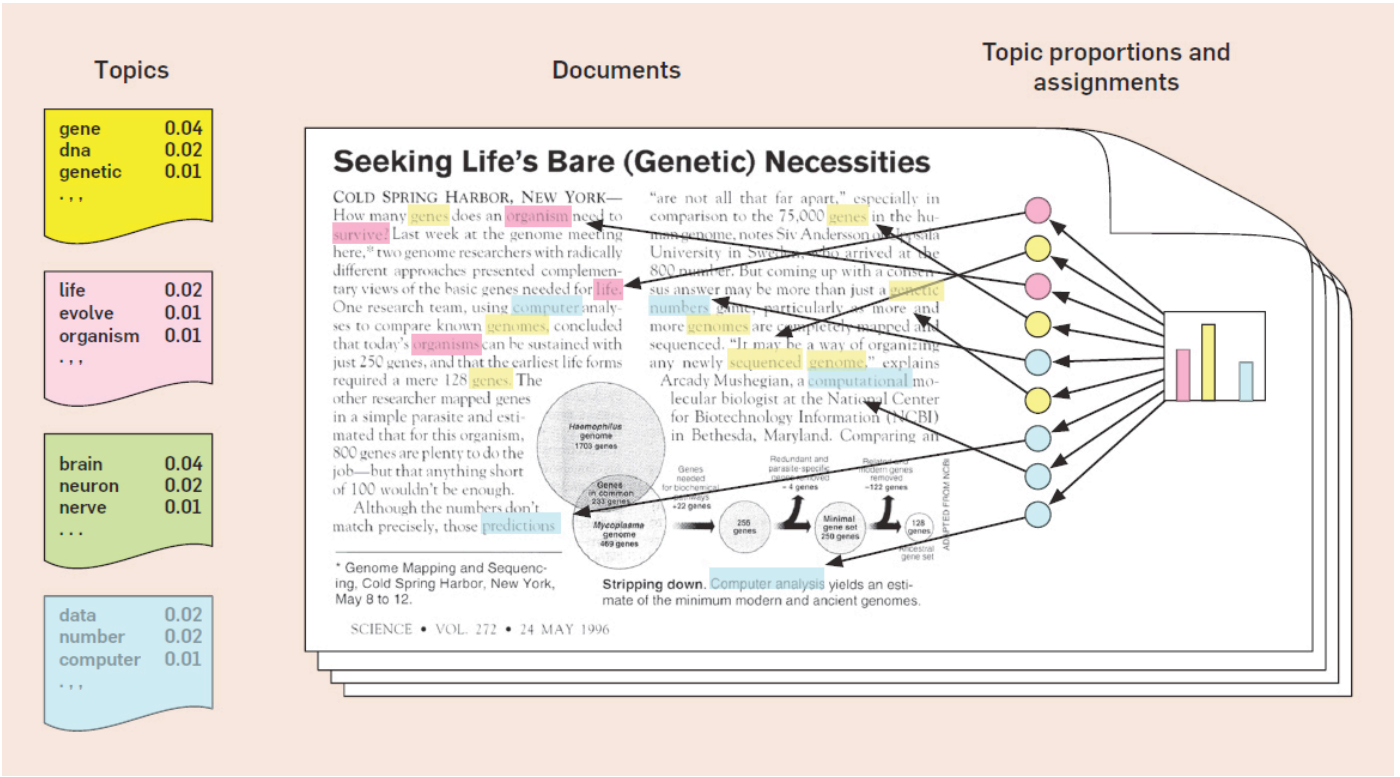
- For each document d = 1,⋯,M
  - Generate $\theta_d \sim \mathrm{Dir}(\cdot \mid \alpha)$
  - For each position n = 1,⋯, N_d
    - generate $z_n \sim \mathrm{Mult}(\cdot \mid \theta_d)$
    - generate $w_n \sim \mathrm{Mult}(\cdot \mid \beta_{z_n})$

$$\prod_{d=1}^{N_d} P(w_1, \cdots, w_{N_d} | \beta, \alpha)$$

$$= \prod_{d=1}^{N_d} \int_{\theta_d} P(\theta_d | \alpha) \left\{ \prod_{n=1}^{N_d} \left(\sum_k \theta_{dk} \beta_{k w_n} \right) \right\} d\theta_d$$



Document modeling using LDA



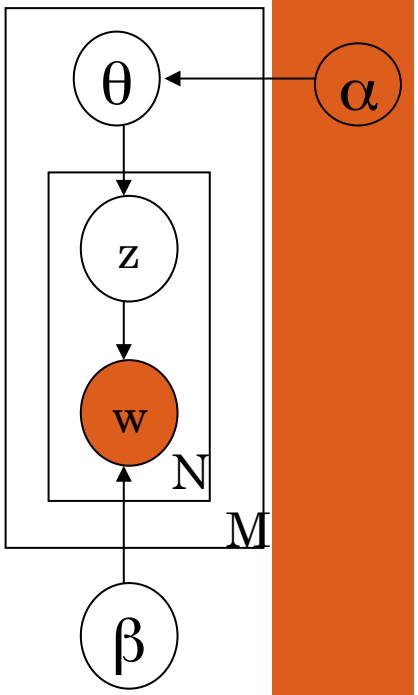


Illustration from Blei, D. 2012. "Probabilistic Topic Models."

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Summary

- Graphical models are useful to visualize structure of probabilistic models
- Joint distributions can be factored into conditional distributions
- Directed & undirected graphical models

Graphical model ~ multivariate statistics + structure