Lecture 3
Probabilistic Sequence Models

Burr Settles

IBS Summer Research Program 2008

bsettles@cs.wisc.edu

www.cs.wisc.edu/~bsettles/ibs08/

Probability 101

- *frequentist* interpretation: the probability of an event is the proportion of the time events of same kind will occur in the long run

- examples
  - the probability my flight to Chicago will be on time
  - the probability this ticket will win the lottery
  - the probability it will rain tomorrow

- always a number in the interval [0,1]

  0 means "never occurs"

  1 means "always occurs"

Sample Spaces

- *sample space*: a set of possible outcomes for some event

- examples
  - flight to Chicago: {on time, late}
  - lottery:{ticket 1 wins, ticket 2 wins,…,ticket *n* wins}
  - weather tomorrow:

    {rain, not rain} or

    {sun, rain, snow} or

    {sun, clouds, rain, snow, sleet} or…

Random Variables

- *random variable*: a variable representing the outcome of an experiment
- example:
  - X represents the outcome of my flight to Chicago
  - we write the probability of my flight being on time as $\Pr(X = \text{on-time})$
  - or when it's clear which variable we're referring to, we may use the shorthand $\Pr(\text{on-time})$

Notation

- uppercase letters and capitalized words denote random variables
- lowercase letters and uncapitalized words denote values
- we'll denote a particular value for a variable as follows

$$\Pr(X = x) \qquad \Pr(Fever = true)$$

- we'll also use the shorthand form

$$\Pr(x) \quad \text{for} \quad \Pr(X = x)$$

- for Boolean random variables, we'll use the shorthand

$$\Pr(fever) \ \text{ for } \ \Pr(Fever = true)$$

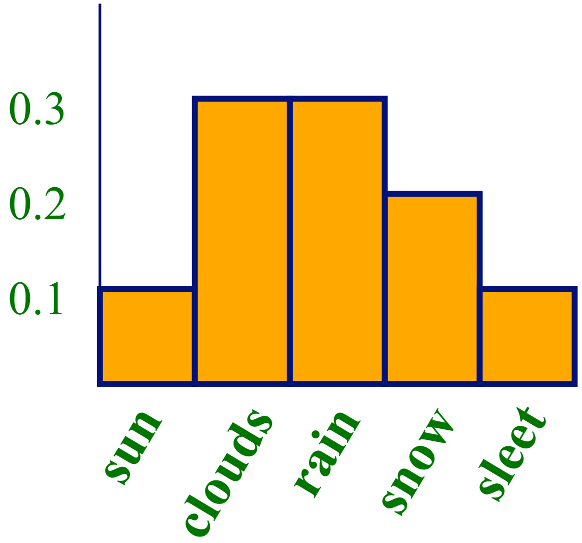
$$\Pr(\neg fever) \ \text{ for } \ \Pr(Fever = false)$$

Probability Distributions

- if X is a random variable, the function given by $\Pr(X = x)$ for each x is the *probability distribution* of X
- requirements:

$$\Pr(x) \geq 0 \quad \text{for every } x$$

$$\sum_x \Pr(x) = 1$$



Joint Distributions

- *joint probability distribution*: the function given by $\Pr(X = x, Y = y)$
- read "X equals x and Y equals y"
- example

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal Distributions

- the *marginal distribution* of *X* is defined by

$$\Pr(x) = \sum_{y} \Pr(x, y)$$

"the distribution of *X* ignoring other variables"

- this definition generalizes to more than two variables, e.g.

$$\Pr(x) = \sum_{y} \sum_{z} \Pr(x, y, z)$$

Marginal Distribution Example

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for X

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional Distributions

- the *conditional distribution* of *X* given *Y* is defined as:

$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{P(Y = y)}$$

"the distribution of *X* given that we know *Y*"

Conditional Distribution Example

joint distribution

x, y	$\Pr(X = x,\ Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$\Pr(X = x \mid Y = \textit{on-time})$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

Independence

- two random variables, *X* and *Y*, are *independent* if

$$\Pr(x, y) = \Pr(x) \times \Pr(y) \quad \text{for all } x \text{ and } y$$

Independence Example #1

joint distribution

x, y	$\Pr(X = x,\ Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
on-time	0.45
late	0.55

Are *X* and *Y* independent here? NO.

Independence Example #2

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
fly-United	0.9
fly-Northwest	0.1

Are X and Y independent here? YES.

Conditional Independence

- two random variables *X* and *Y* are *conditionally independent* given *Z* if

$$\Pr(X \mid Y, Z) = \Pr(X \mid Z)$$

“once you know the value of *Z*, knowing *Y* doesn’t tell you anything about *X*”

- alternatively

$$\Pr(x, y \mid z) = \Pr(x \mid z) \times \Pr(y \mid z) \quad \text{for all } x, y, z$$

Conditional Independence Example

Flu	Fever	Vomit	Pr
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Fever and Vomit are not independent: e.g. $\Pr(fever, vomit) \neq \Pr(fever) \times \Pr(vomit)$

Fever and Vomit are conditionally independent given Flu:

$$\Pr(fever, vomit \mid flu) = \Pr(fever \mid flu) \times \Pr(vomit \mid flu)$$

$$\Pr(fever, vomit \mid \neg flu) = \Pr(fever \mid \neg flu) \times \Pr(vomit \mid \neg flu)$$

etc.

Bayes Theorem

$$\Pr(x \mid y) = \frac{\Pr(y \mid x)\Pr(x)}{\Pr(y)} = \frac{\Pr(y \mid x)\Pr(x)}{\sum_x \Pr(y \mid x)\Pr(x)}$$

- this theorem is extremely useful
- there are many cases when it is hard to estimate Pr(x | y) directly, but it's not too hard to estimate Pr($y \mid x$) and Pr(x)

Bayes Theorem Example

- MDs usually aren't good at estimating Pr(*Disorder* | *Symptom*)
- they're usually better at estimating Pr(*Symptom* | *Disorder*)
- if we can estimate Pr(*fever* | *flu*) and Pr(*flu*) we can use Bayes' Theorem to do diagnosis

$$\Pr(\mathit{flu} \mid \mathit{fever}) = \frac{\Pr(\mathit{fever} \mid \mathit{flu})\Pr(\mathit{flu})}{\Pr(\mathit{fever} \mid \mathit{flu})\Pr(\mathit{flu}) + \Pr(\mathit{fever} \mid \neg \mathit{flu})\Pr(\neg \mathit{flu})}$$

Expected Values

- the *expected value* of a random variable that takes on numerical values is defined as:

$$E[X] = \sum_x x \times \Pr(x)$$

this is the same thing as the *mean*

- we can also talk about the expected value of a function of a random variable

$$E[g(X)] = \sum_x g(x) \times \Pr(x)$$

Expected Value Example

- Suppose each lottery ticket costs \$1 and the winning ticket pays out \$100. The probability that a particular ticket is the winning ticket is 0.001.

$$E\left[gain(Lottery)\right] =$$

$$gain(\text{winning})\Pr(\text{winning}) + gain(\text{losing})\Pr(\text{losing}) =$$

$$(\$100 - \$1)\times 0.001 - \$1\times 0.999 =$$

$$-\$0.90$$

Probabilistic Sequence Models in Computational Biology

- there are many cases in which we would like to represent the statistical regularities of some class of sequences
  - genes
  - various regulatory sites in DNA (e.g. where RNA polymerase and transcription factors bind)
  - proteins in a given family

Probability Of A Sequence

- given some sequence *x* of length *L*, we want to compute its probability (likelihood)
- one way to compute this is the joint probability of all the characters in the sequence:

$$\begin{aligned}\Pr(x) &= \Pr(x_1, x_2, \ldots, x_L) \\ &= \Pr(x_1)\Pr(x_2 \mid x_1)\ldots\Pr(x_L \mid x_1, \ldots, x_{L-1})\end{aligned}$$

- for example:

$$\Pr(cggt) = \Pr(c)\Pr(g \mid c)\Pr(g \mid cg)\Pr(t/cgg)$$

- *problem*: biological sequences tend to be very long; that's too many conditional probabilities to estimate!

The Markov Assumption

- trick: assume the probability of a character is only dependent on the *previous character*, not the entire prefix

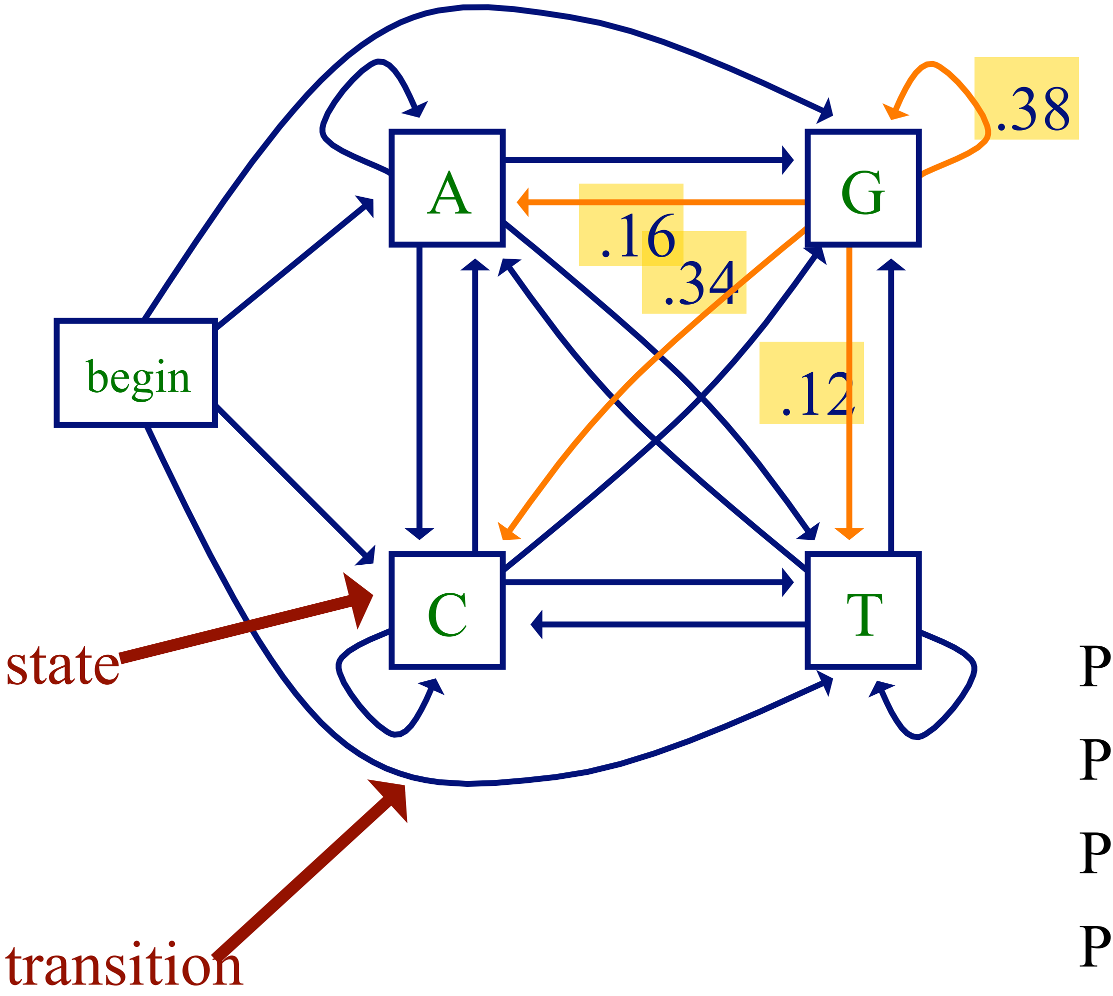
$$\begin{aligned}\Pr(x) &= \Pr(x_1, x_2, \ldots, x_L) \\ &\approx \Pr(x_1)\Pr(x_2 \mid x_1)\ldots\Pr(x_{L-1} \mid x_{L-2})\Pr(x_L / x_{L-1}) \\ &= \Pr(x_1)\prod_{i=2}^{L}\Pr(x_i \mid x_{i-1})\end{aligned}$$

- now our probabilities are easier to estimate:

$$\Pr(cggt) = \Pr(c)\Pr(g \mid c)\Pr(g \mid g)\Pr(t/g)$$

- this trick is called the *Markov assumption*, and a statistical process that uses it is called a *Markov chain*

Markov Chain Models



state

transition

transition probabilities

$$\Pr(x_i = a \mid x_{i-1} = g) = 0.16$$

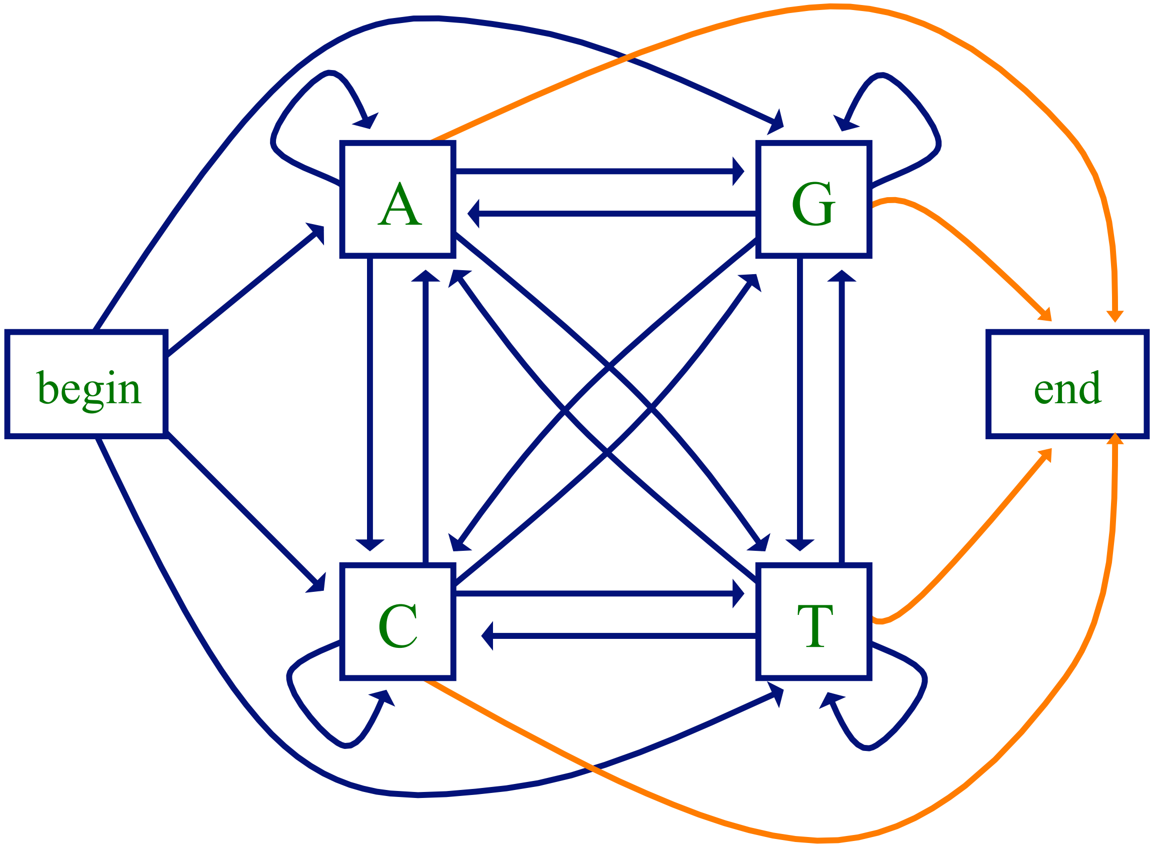
$$\Pr(x_i = c \mid x_{i-1} = g) = 0.34$$

$$\Pr(x_i = g \mid x_{i-1} = g) = 0.38$$

$$\Pr(x_i = t \mid x_{i-1} = g) = 0.12$$

Markov Chain Models

- can also have an *end* state; allows the model to represent
  - a distribution over sequences of different lengths
  - preferences for ending sequences with certain symbols



Markov Chain Models

- a Markov chain model is defined by
  - a set of states
    - some states *emit* symbols
    - other states (e.g. the *begin* and *end* states) are *silent*
  - a set of transitions with associated probabilities
    - the transitions emanating from a given state define a distribution over the possible next states

Markov Chain Notation

- the transition parameters can be denoted by $a_{x_{i-1}x_i}$ where

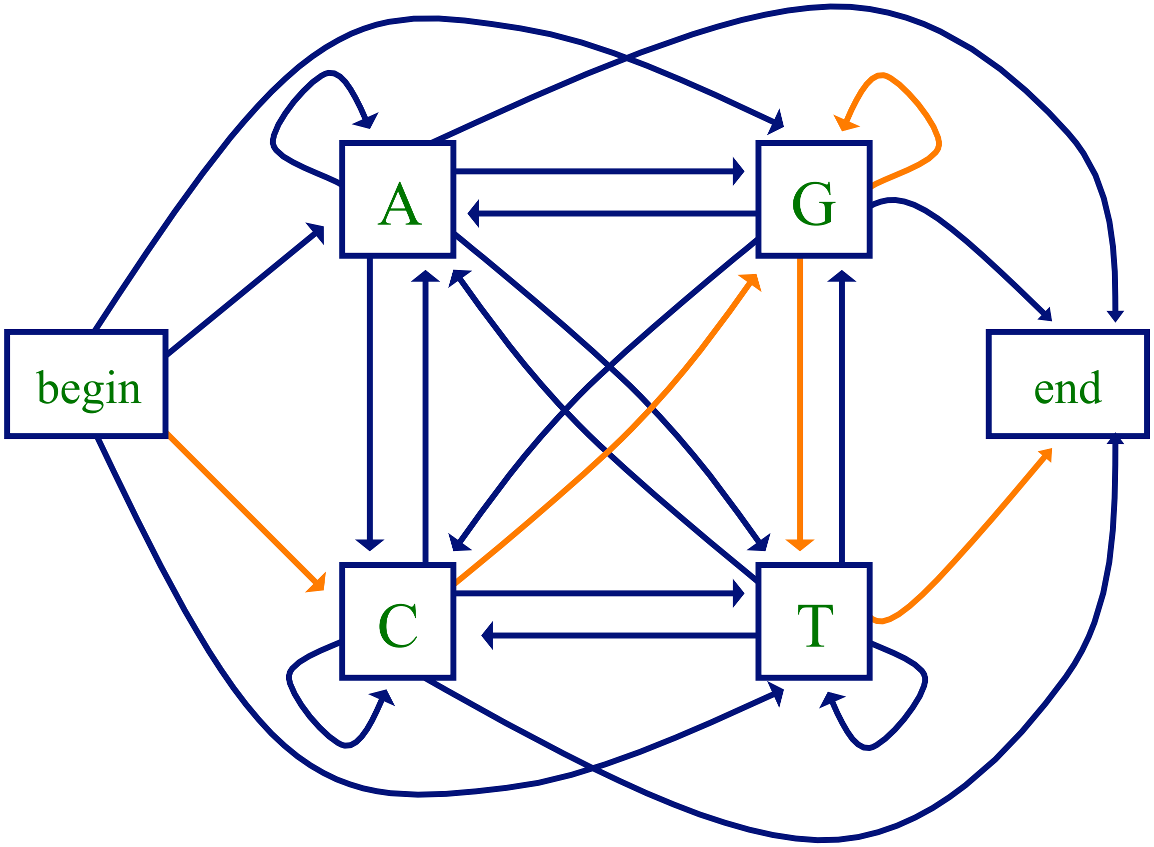
$$a_{x_{i-1}x_i} = \Pr(x_i \mid x_{i-1})$$

- similarly we can denote the probability of a sequence *x* as

$$a_{\mathrm{B}x_1} \prod_{i=2}^{L} a_{x_{i-1}x_i} = \Pr(x_1) \prod_{i=2}^{L} \Pr(x_i \mid x_{i-1})$$

where $a_{\mathrm{B}x_1}$ represents the transition from the *begin* state

The Probability of a Sequence for a Given Markov Chain Model



$$\Pr(cggt) = \Pr(c)\Pr(g \mid c)\Pr(g \mid g)\Pr(t/g)\Pr(\text{end} \mid t)$$

Estimating the Model Parameters

- given some data (e.g. a set of sequences), how can we determine the probability parameters of our model?

- one approach: *maximum likelihood estimation*
  - given a set of data D
  - set the parameters θ to maximize

$$\Pr(D \mid \theta)$$

  - i.e. make the data D look <u>as likely as possible</u> under the model θ

Maximum Likelihood Estimation

- suppose we want to estimate the parameters:
  Pr(a), Pr(c), Pr(g), Pr(t)
- and we're given the sequences

  accgcgctta

  gcttagtgac

  tagccgttac

$$\Pr(a) = \frac{n_a}{\sum_i n_i}$$

- then the maximum likelihood estimates are

$$\Pr(a) = \frac{6}{30} = 0.2 \qquad \Pr(g) = \frac{7}{30} = 0.233$$

$$\Pr(c) = \frac{9}{30} = 0.3 \qquad \Pr(t) = \frac{8}{30} = 0.267$$

Maximum Likelihood Estimation

- suppose instead we saw the following sequences

  gccgcgcttg

  gcttggtggc

  tggccgttgc

- then the maximum likelihood estimates are

$$\boxed{\Pr(a) = \frac{0}{30} = 0} \qquad \Pr(g) = \frac{13}{30} = 0.433$$

$$\Pr(c) = \frac{9}{30} = 0.3 \qquad \Pr(t) = \frac{8}{30} = 0.267$$

do we really want to set this to 0?

A Bayesian Approach

- instead of estimating parameters strictly from the data, we could start with some prior belief for each
- for example, we could use *Laplace estimates*

$$\Pr(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}$$

pseudocount

- where n_i represents the number of occurrences of character i
- using Laplace estimates with the sequences

gccgcgcttg

gcttggtggc

tggccgttgc

$$\Pr(a) = \frac{0 + 1}{34}$$

$$\Pr(c) = \frac{9 + 1}{34}$$

A Bayesian Approach

- a more general form: *m-estimates*

$$\Pr(a) = \frac{n_a + p_a m}{\left(\sum_i n_i\right) + m}$$

prior probability of *a*

number of “virtual” instances

- with *m*=8 and uniform priors

gccgcgcttg

gcttggtggc

tggccgttgc

$$\Pr(c) = \frac{9 + 0.25 \times 8}{30 + 8} = \frac{11}{38}$$

Estimation for 1st Order Probabilities

- to estimate a 1st order parameter (where each character depends on 1 previous character), such as Pr(*c*|*g*), we count the number of times that *c* follows the history *g* in our given sequences
- using Laplace estimates with the sequences:

gccgcgcttg

gcttggtggc

tggccgttgc

$$\Pr(a \mid g) = \frac{0+1}{12+4} \qquad \Pr(a \mid c) = \frac{0+1}{7+4}$$

$$\vdots$$

$$\Pr(c \mid g) = \frac{7+1}{12+4}$$

$$\Pr(g \mid g) = \frac{3+1}{12+4}$$

$$\Pr(t \mid g) = \frac{2+1}{12+4}$$

Higher Order Markov Chains

- the Markov property specifies that the probability of a state depends only on the probability of the previous state
- but we can build more “memory” into our states by using a higher order Markov model
- in an *n*th order Markov model

$$\Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_1) = \Pr(x_i \mid x_{i-1}, \ldots, x_{i-n})$$

Selecting the Order of a Markov Chain Model

- higher order models remember more “history”
- additional history can have predictive value
- example:
  - predict the next word in this sentence fragment
    “…finish ______” (up, it, first, last, …?)

  - now predict it given more history
    “nice guys finish ______”

Selecting the Order of a Markov Chain Model

- but the number of parameters we need to estimate grows exponentially with the order
  - for modeling DNA we need $O(4^{n+1})$ parameters for an *n*th order model

- the higher the order, the less reliable we can expect our parameter estimates to be
  - estimating the parameters of a 2nd order Markov chain from the complete genome of E. Coli, we'd see each "word" 72,000+ times on average
  - estimating the parameters of an 8th order chain, we'd see each "word" about 5 times on average

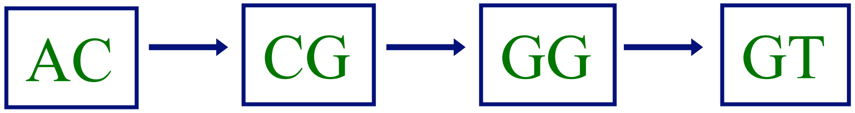
Higher Order Markov Chains

- an *n*th order Markov chain over some alphabet A is equivalent to a first order Markov chain over the alphabet of *n*-tuples A^n
- example: a 2nd order Markov model for DNA can be treated as a 1st order Markov model over alphabet

  AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT

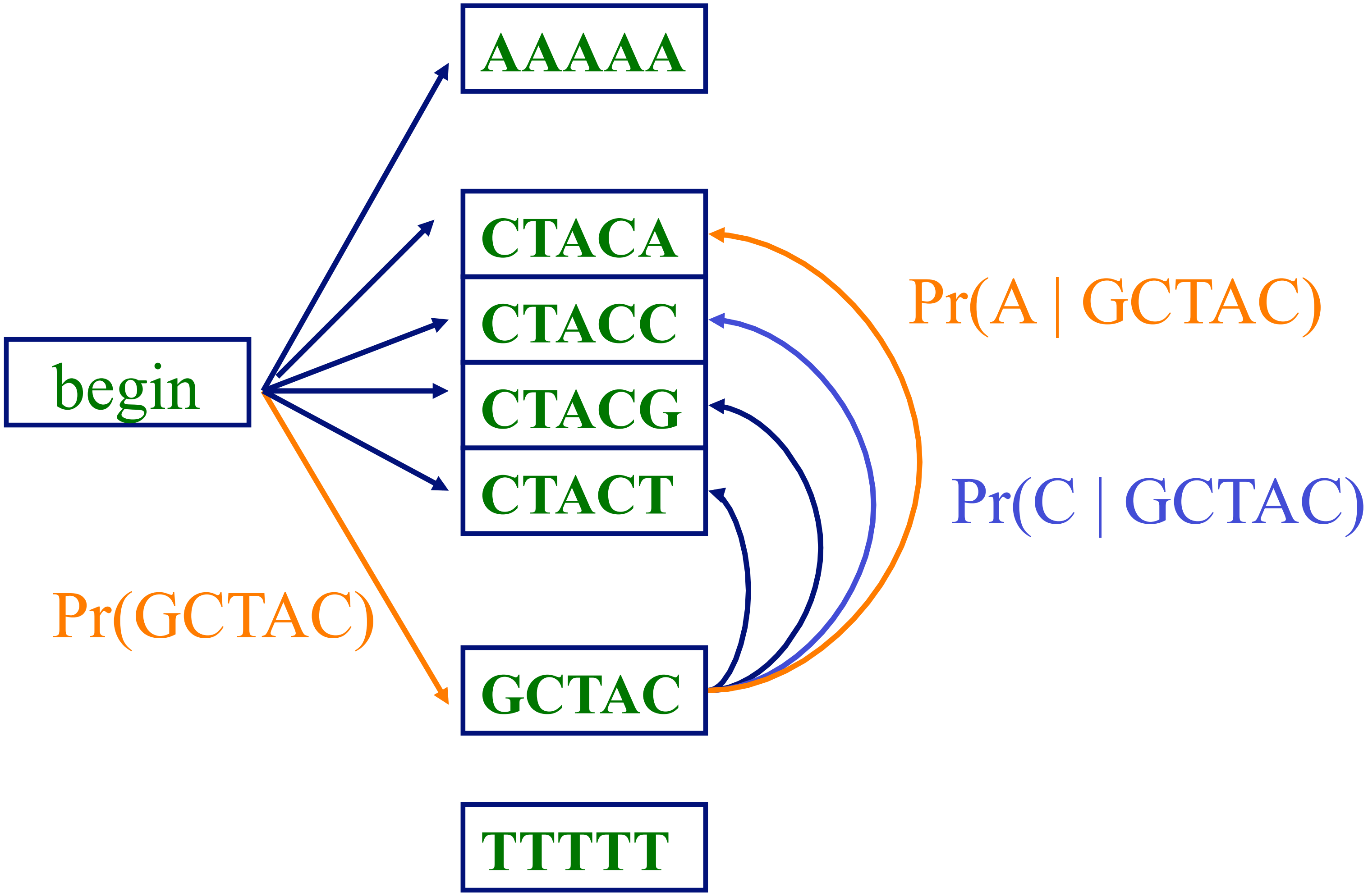
- caveat: we process a sequence one character at a time

  A C G G T

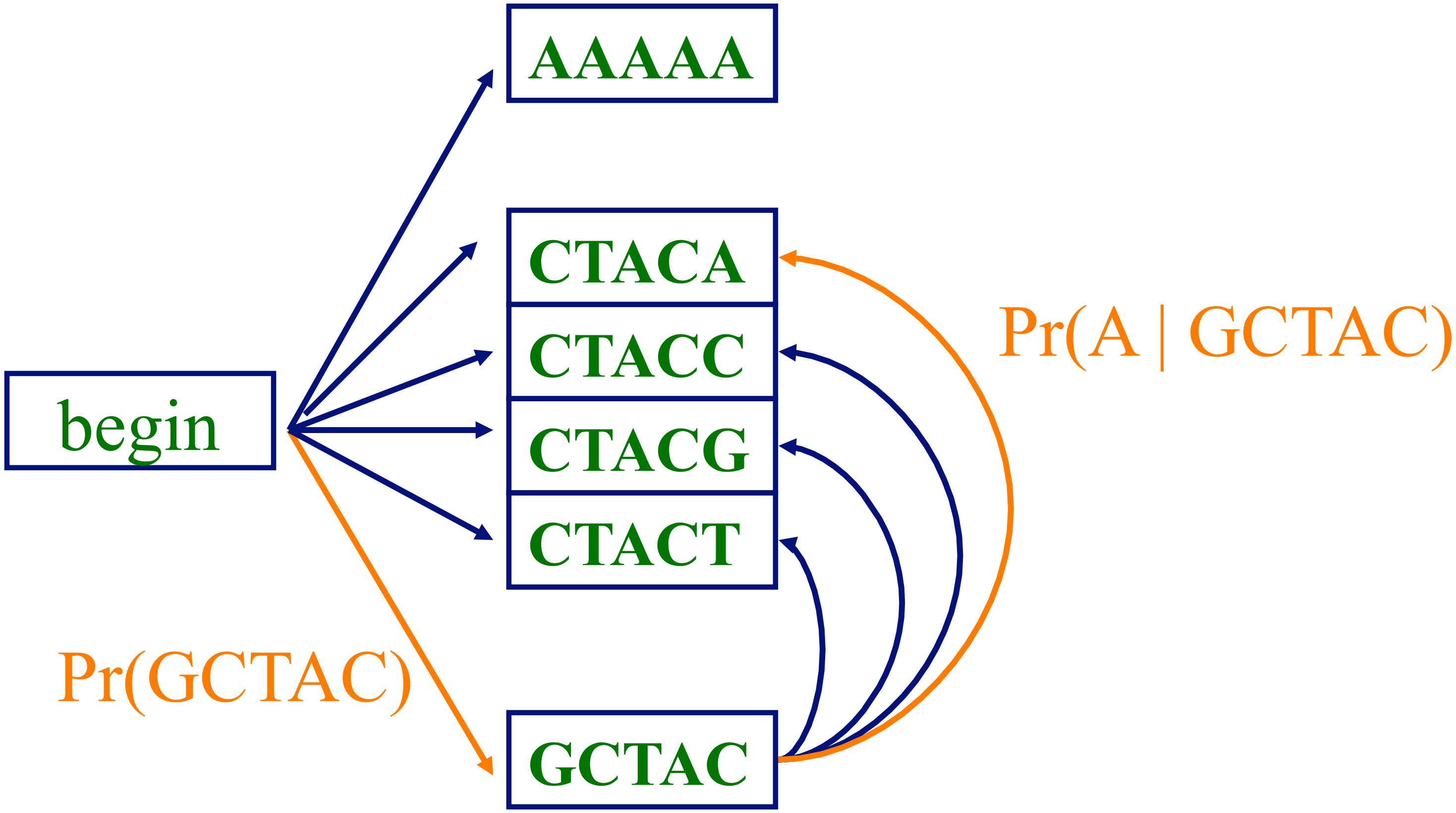


AC → CG → GG → GT

A Fifth Order Markov Chain



A Fifth Order Markov Chain



$$\Pr(gctaca) = \Pr(gctac)\Pr(a \mid gctac)$$

Example Application

- language classification
- given:
  - passages of text from different languages
  - e.g. newspaper articles written in English, French, Spanish, German, and Italian
- do:
  - learn a Markov chain model for each language
  - use these models to determine the most likely language for some new passage of text

- **http://pages.cs.wisc.edu/~bsettles/webtoys/polyglot/**

Example Biological Application

- CpG islands
  - CG dinucleotides are rarer in eukaryotic genomes than expected given the marginal probabilities of C and G
  - but the regions upstream of genes are richer in CG dinucleotides than elsewhere – *CpG islands*
  - useful evidence for finding genes

Example Biological Application

- given sequences from CpG islands, and sequences from other regions, we can construct
  - a model to represent CpG islands
  - a *null model* to represent the other regions

- can then score a test sequence by:

$$score(x) = \log \frac{\Pr(x \mid \text{CpG model})}{\Pr(x \mid \text{null model})}$$

Example Biological Application

- parameters estimated for CpG and null models
  - human sequences containing 48 CpG islands
  - 60,000 nucleotides

$\Pr(c \mid a)$

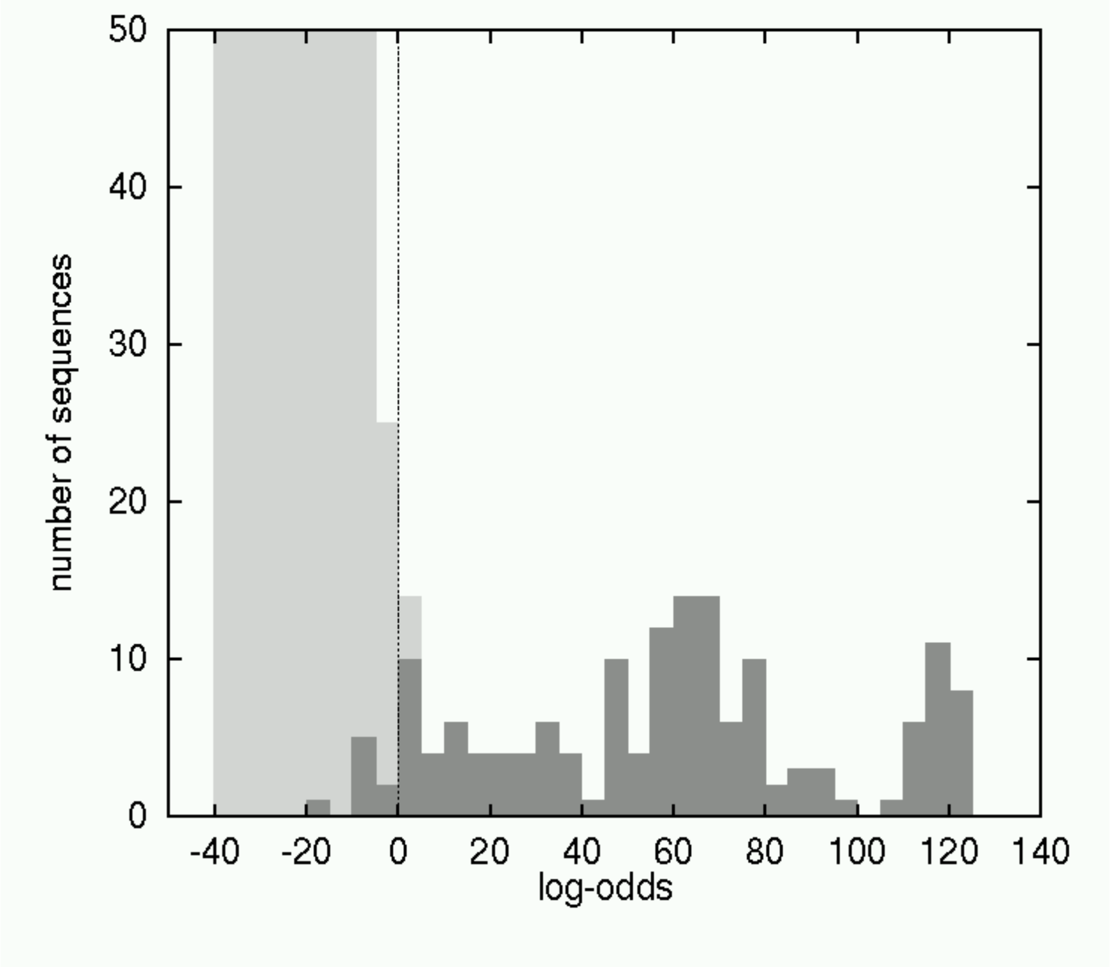
+	*a*	*c*	*g*	*t*
a	.18	.27	.43	.12
c	.17	.37	.27	.19
g	.16	.34	.38	.12
t	.08	.36	.38	.18

CpG

-	*a*	*c*	*g*	*t*
a	.30	.21	.28	.21
c	.32	.30	.08	.30
g	.25	.24	.30	.21
t	.18	.24	.29	.29

null

Example Biological Application



- light bars represent negative sequences
- dark bars represent positive sequences
- the actual figure here is not from a CpG island discrimination task, however

Figure from A. Krogh, “An Introduction to Hidden Markov Models for Biological Sequences” in Computational Methods in Molecular Biology, Salzberg et al. editors, 1998.

Example Biological Application

- why use

$$score(x) = \log \frac{\Pr(x \mid CpG)}{\Pr(x \mid null)}$$

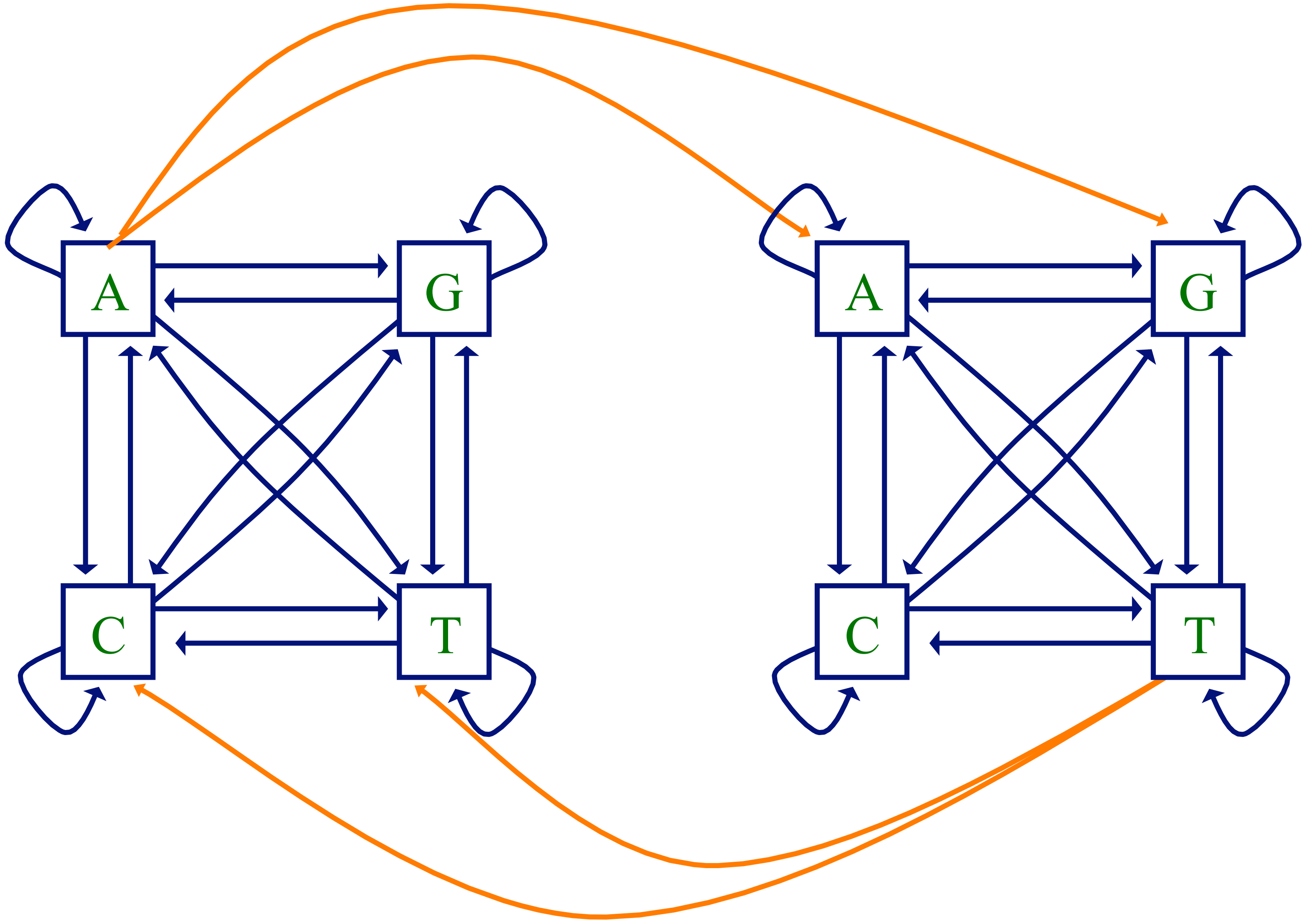
- Bayes' rule tells us

$$\Pr(CpG \mid x) = \frac{\Pr(x \mid CpG)\Pr(CpG)}{\Pr(x)}$$

$$\Pr(null \mid x) = \frac{\Pr(x \mid null)\Pr(null)}{\Pr(x)}$$

- if we're not taking into account prior probabilities of two classes ($\Pr(CpG)$ and $\Pr(null)$) then we just need to compare $\Pr(x \mid CpG)$ and $\Pr(x \mid null)$

Hidden Markov Models



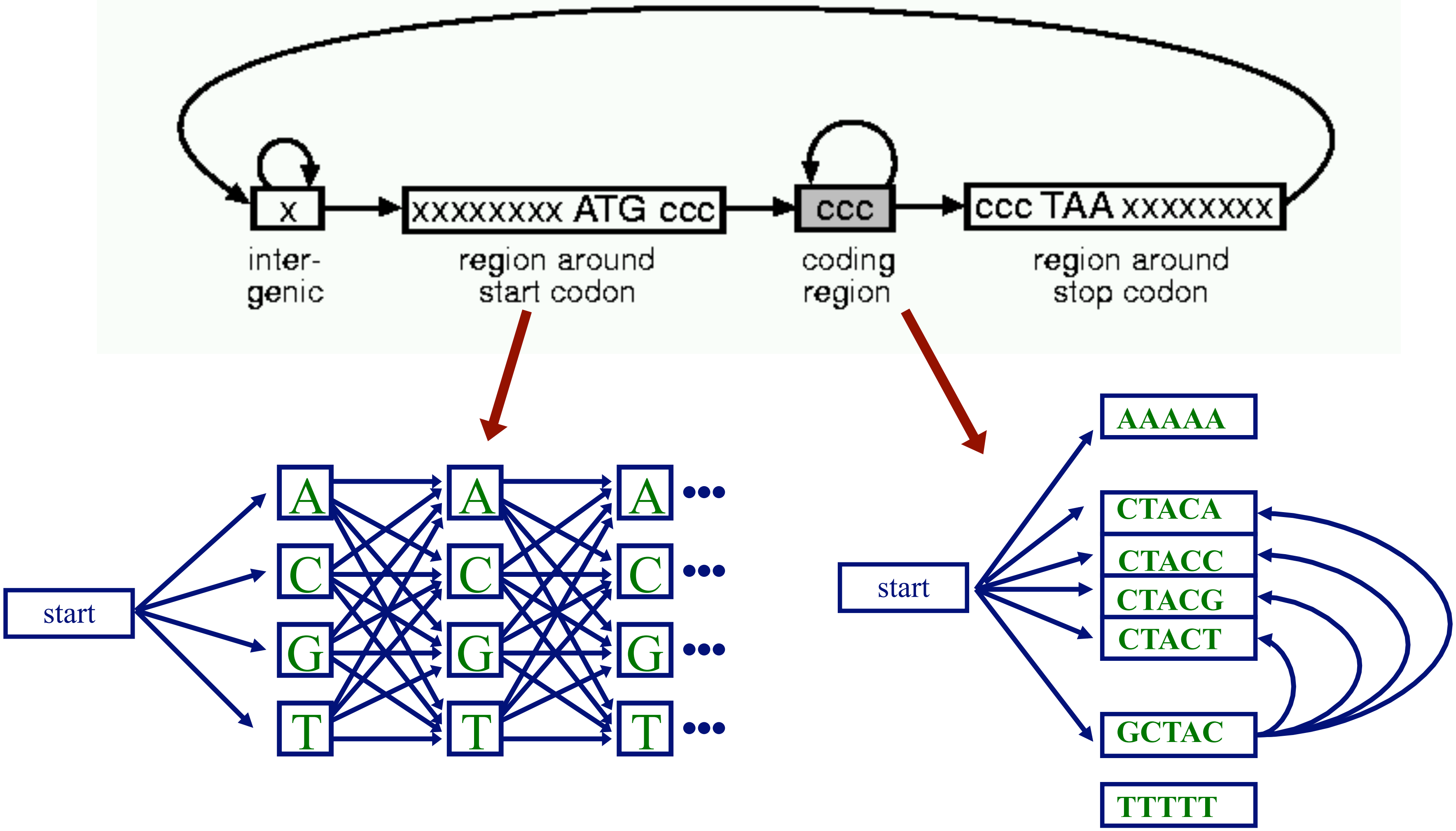
- given say a *T* in our input sequence, which state emitted it?

Hidden State

- we’ll distinguish between the *observed* parts of a problem and the *hidden* parts
- in the Markov models we’ve considered previously, it is clear which state accounts for each part of the observed sequence
- in this example, there are multiple states that could account for each part of the observed sequence
  - this is the *hidden* part of the problem
  - *hidden Markov models* (HMMs) are Markov chain models with hidden state

Simple HMM for Gene Finding

Figure from A. Krogh, An Introduction to Hidden Markov Models for Biological Sequences



HMM Applications

- *classification*
  - *given*: a set of models representing different sequence classes (e.g. protein families), and a test sequence
  - *do*: determine which model/class best explains the sequence
  - use Forward algorithm to calculate probability of sequence under each each model

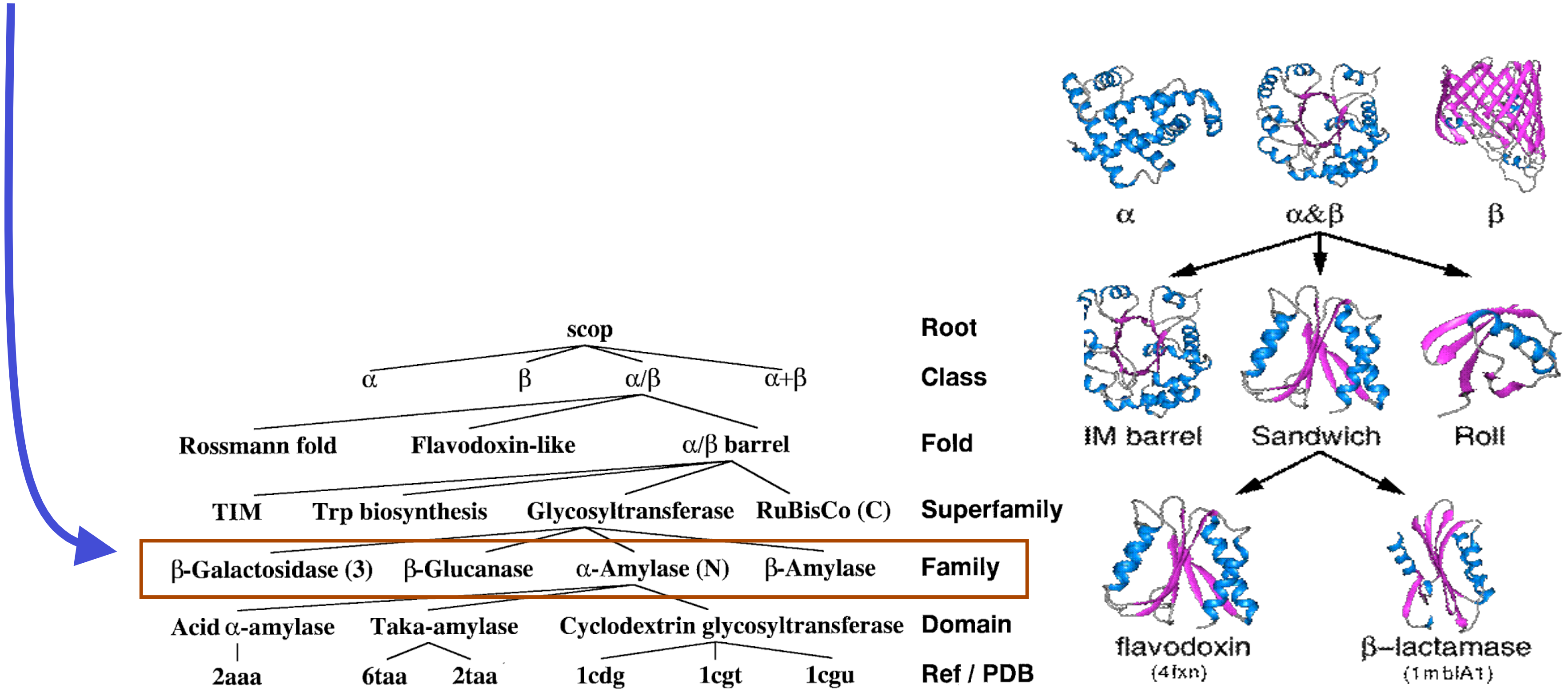
- *segmentation*
  - *given*: a model representing different sequence classes, a test sequence
  - *do*: segment the sequence into subsequences, predicting the state labels for each subsequence
  - use Viterbi algorithm to find most probable path for sequence

Example: Protein Classification

given: amino-acid sequence of a protein

do: predict the *family* to which it belongs

GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVCVLAHHFGKEFTPPVQAAYAKVVAGVANALAHKYH



scop — Root

α, β, α/β, α+β — Class

Rossmann fold, Flavodoxin-like, α/β barrel — Fold

TIM, Trp biosynthesis, Glycosyltransferase, RuBisCo (C) — Superfamily

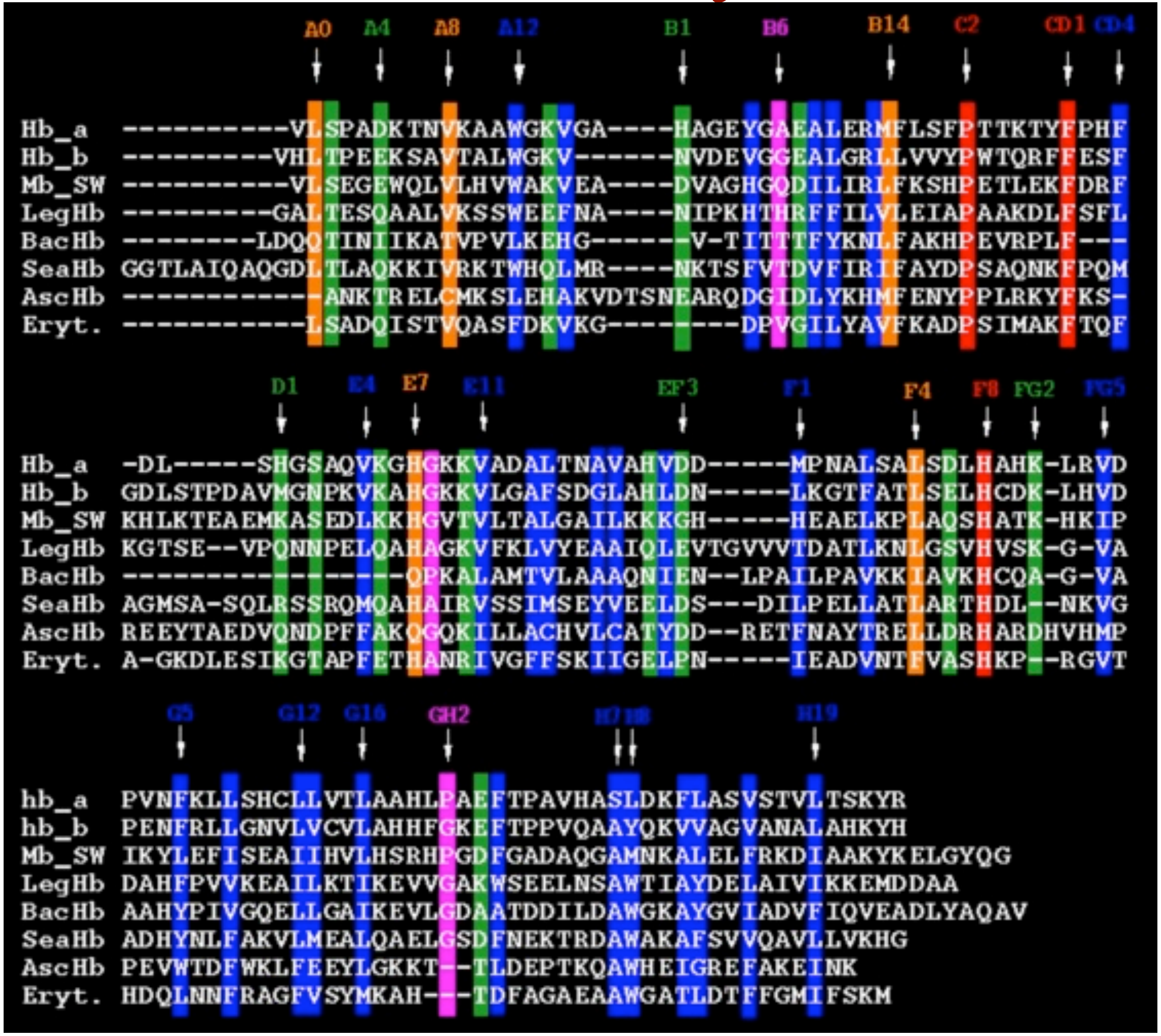
β-Galactosidase (3), β-Glucanase, α-Amylase (N), β-Amylase — Family

Acid α-amylase, Taka-amylase, Cyclodextrin glycosyltransferase — Domain

2aaa, 6taa, 2taa, 1cdg, 1cgt, 1cgu — Ref / PDB

Alignment of Globin Family Proteins

- The sequences in a family may vary in length
- Some positions are more conserved than others



```
Hb_a   ----------VLSPADKTNVKAAWGKVGA----HAGEYGAEALERMFLSFPTTKTYFPHF
Hb_b   ---------VHLTPEEKSAVTALWGKV------NVDEVGGEALGRLLVVYPWTQRFFESF
Mb_SW  ----------VLSEGEWQLVLHVWAKVEA----DVAGHGQDILIRLFKSHPETLEKFDRF
LegHb  ---------GALTESQAALVKSSWEEFNA----NIPKHTHRFFILVLEIAPAAKDLFSFL
BacHb  --------LDQQTINIIKATVPVLKEHG------V-TITTTFYKNLFAKHPEVRPLF---
SeaHb  GGTLAIQAQGDLTLAQKKIVRKTWHQLMR----NKTSFVTDVFIRIFAYDPSAQNKFPQM
AscHb  ------------ANKTRELCMKSLEHAKVDTSNEARQDGIDLYKHMFENYPPLRKYFKS-
Eryt.  -----------LSADQISTVQASFDKVKG--------DPVGILYAVFKADPSIMAKFTQF

Hb_a   -DL-----SHGSAQVKGHGKKVADALTNAVAHVDD-----MPNALSALSDLHAHK-LRVD
Hb_b   GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN-----LKGTFATLSELHCDK-LHVD
Mb_SW  KHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH-----HEAELKPLAQSHATK-HKIP
LegHb  KGTSE--VPQNNPELQAHAGKVFKLVYEAAIQLEVTGVVVTDATLKNLGSVHVSK-G-VA
BacHb  -----------------QPKALAMTVLAAAQNIEN--LPAILPAVKKIAVKHCQA-G-VA
SeaHb  AGMSA-SQLRSSRQMQAHAIRVSSIMSEYVEELDS---DILPELLATLARTHDL--NKVG
AscHb  REEYTAEDVQNDPFFAKQGQKILLACHVLCATYDD--RETFNAYTRELLDRHARDHVHMP
Eryt.  A-GKDLESIKGTAPFETHANRIVGFFSKIIGELPN-----IEADVNTFVASHKP--RGVT

hb_a   PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
hb_b   PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
Mb_SW  IKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LegHb  DAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
BacHb  AAHYPIVGQELLGAIKEVLGDAATDDILDAWGKAYGVIADVFIQVEADLYAQAV
SeaHb  ADHYNLFAKVLMEALQAELGSDFNEKTRDAWAKAFSVVQAVLLVKHG
AscHb  PEVWTDFWKLFEEYLGKKT--TLDEPTKQAWHEIGREFAKEINK
Eryt.  HDQLNNFRAGFVSYMKAH---TDFAGAEAWGATLDTFFGMIFSKM
```

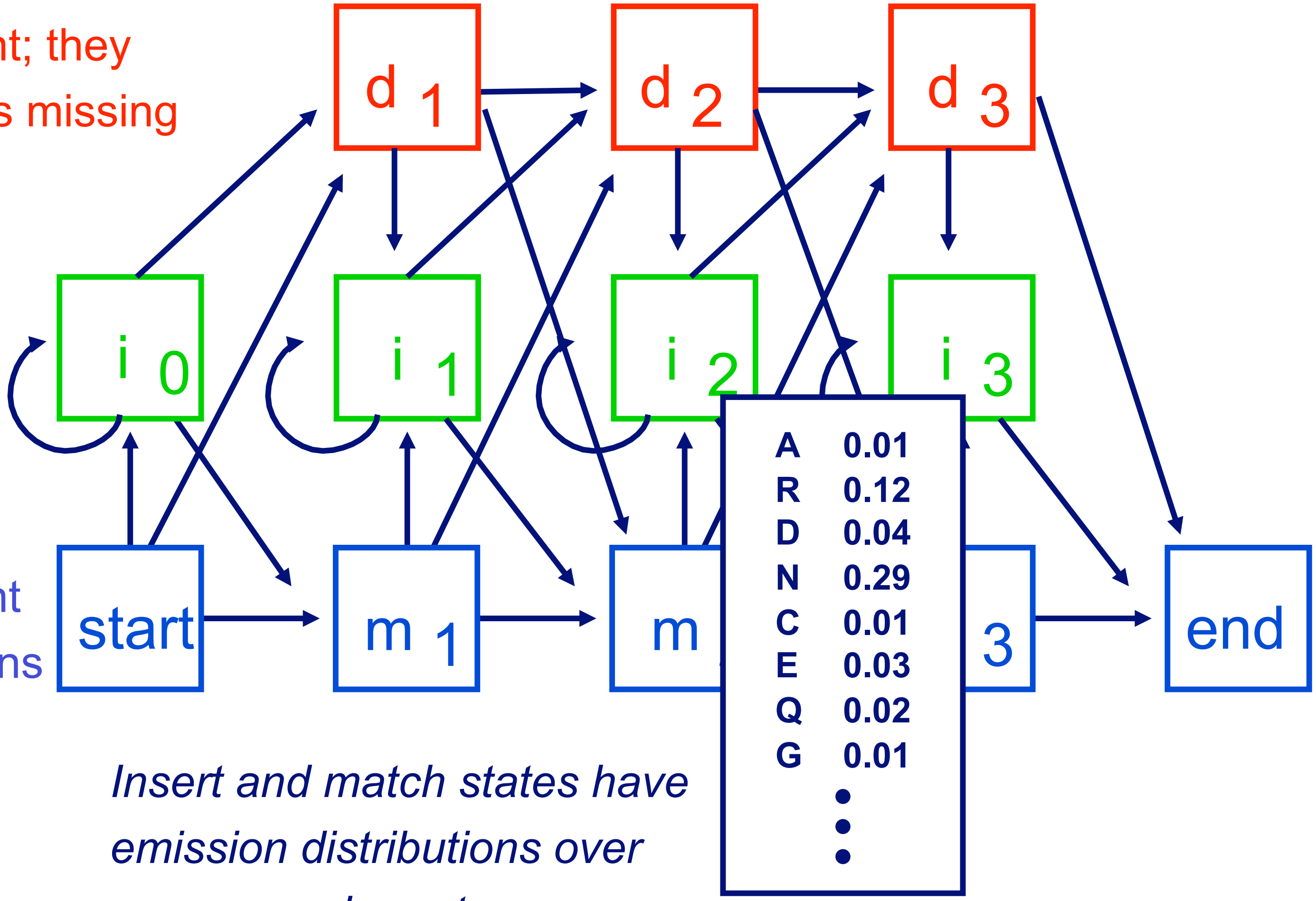
Profile HMMs

- profile HMMs are commonly used to model families of sequences

Delete states are silent; they Account for characters missing in some sequences

Insert states account for extra characters in some sequences

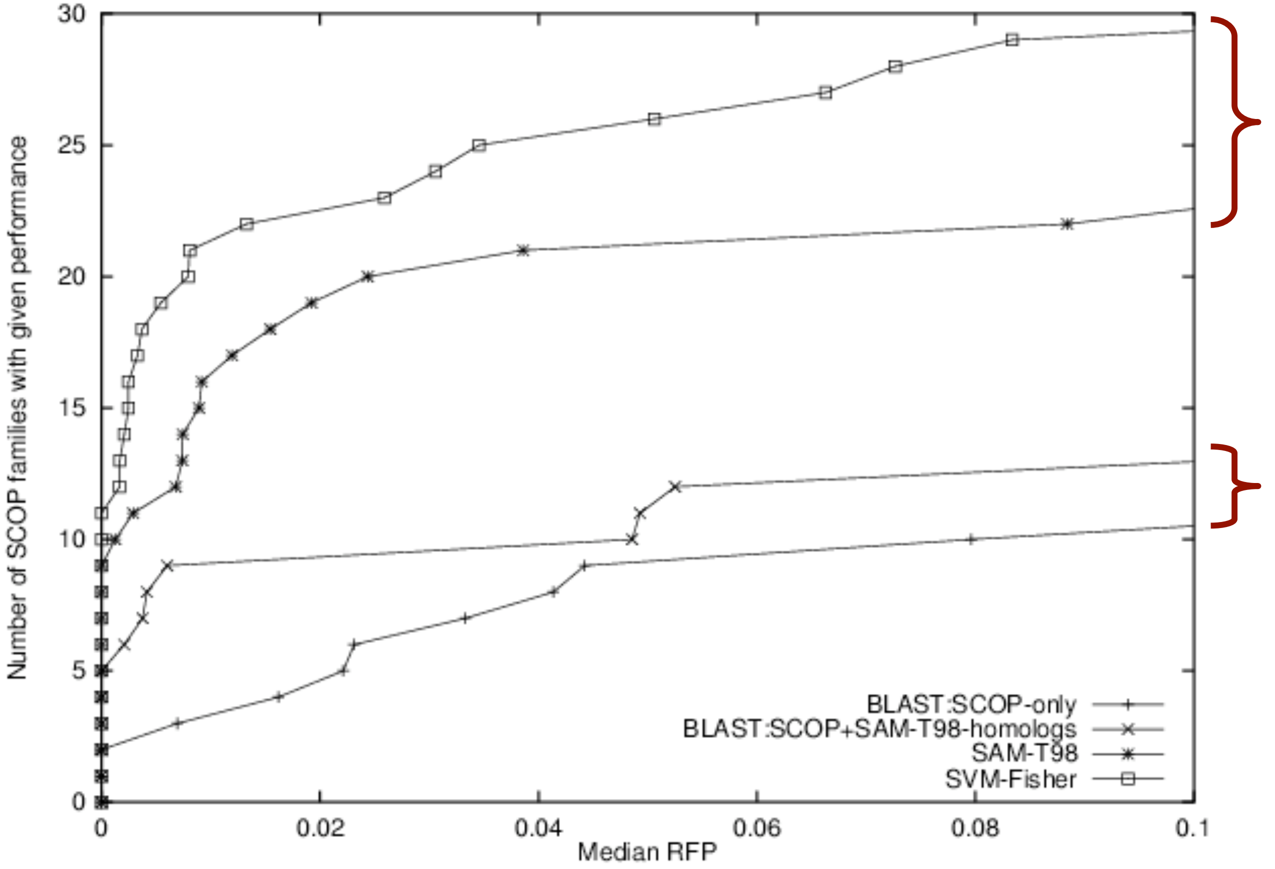
Match states represent key conserved positions



A	0.01
R	0.12
D	0.04
N	0.29
C	0.01
E	0.03
Q	0.02
G	0.01
•	
•	
•	

Insert and match states have emission distributions over sequence characters

Profile HMM Accuracy



profile HMM-based methods

BLAST-based methods

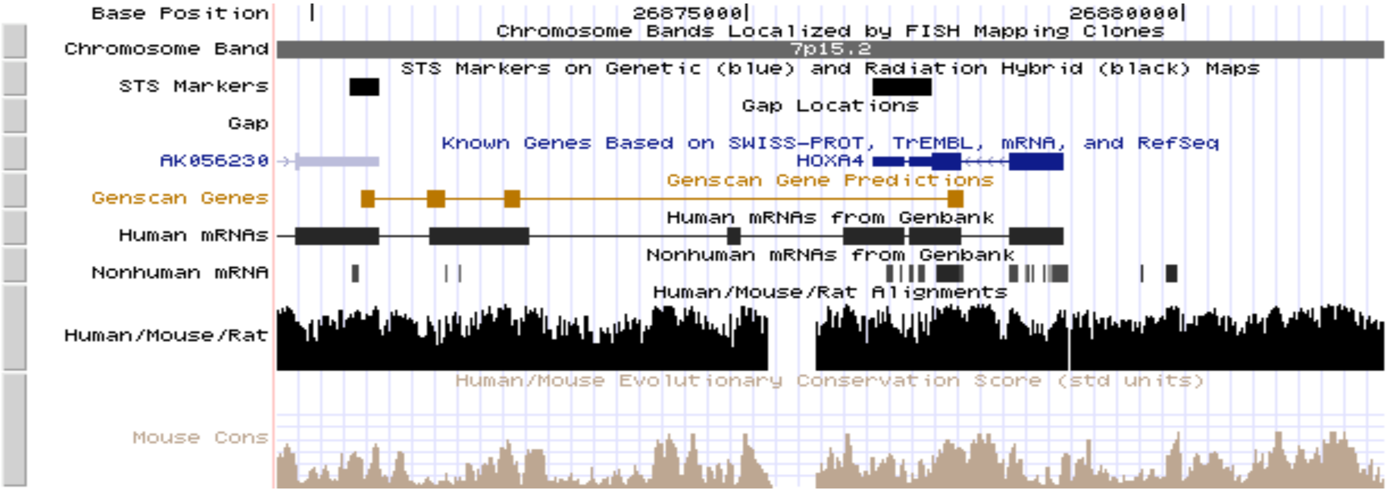
Figure from Jaakola et al., ISMB 1999

- classifying 2447proteins into 33 families
- *x*-axis represents the median # of negative sequences that score as high as a positive sequence for a given family's model

Example: Gene Finding

given: an uncharacterized DNA sequence

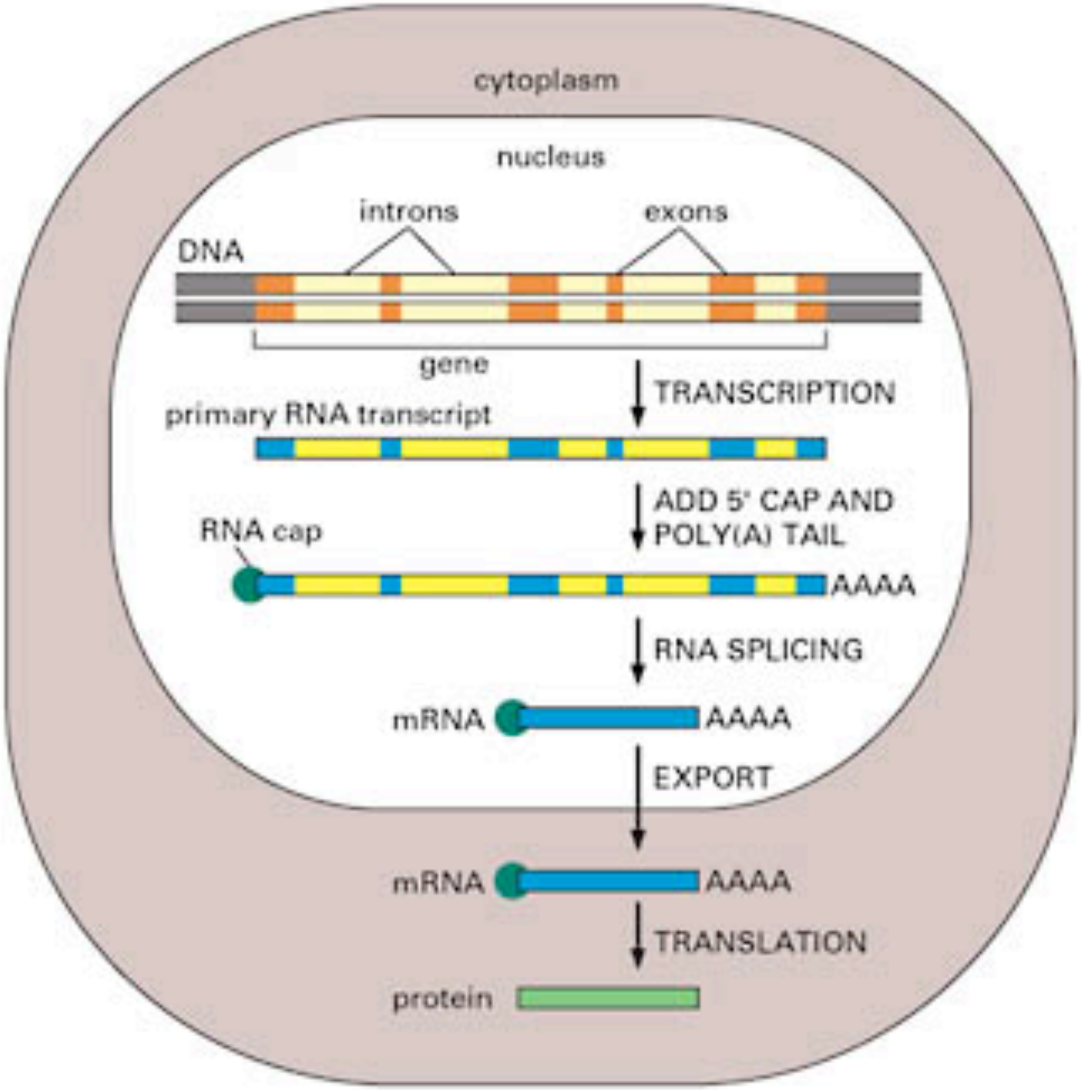
do: locate the genes in the sequence, including the coordinates of individual *exons* and *introns*



Base Position
26875000
26880000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7p15.2
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
AK056230
HOXA4
Genscan Gene Predictions
Genscan Genes
Human mRNAs from Genbank
Human mRNAs
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Human/Mouse/Rat Alignments
Human/Mouse/Rat
Human/Mouse Evolutionary Conservation Score (std units)
Mouse Cons

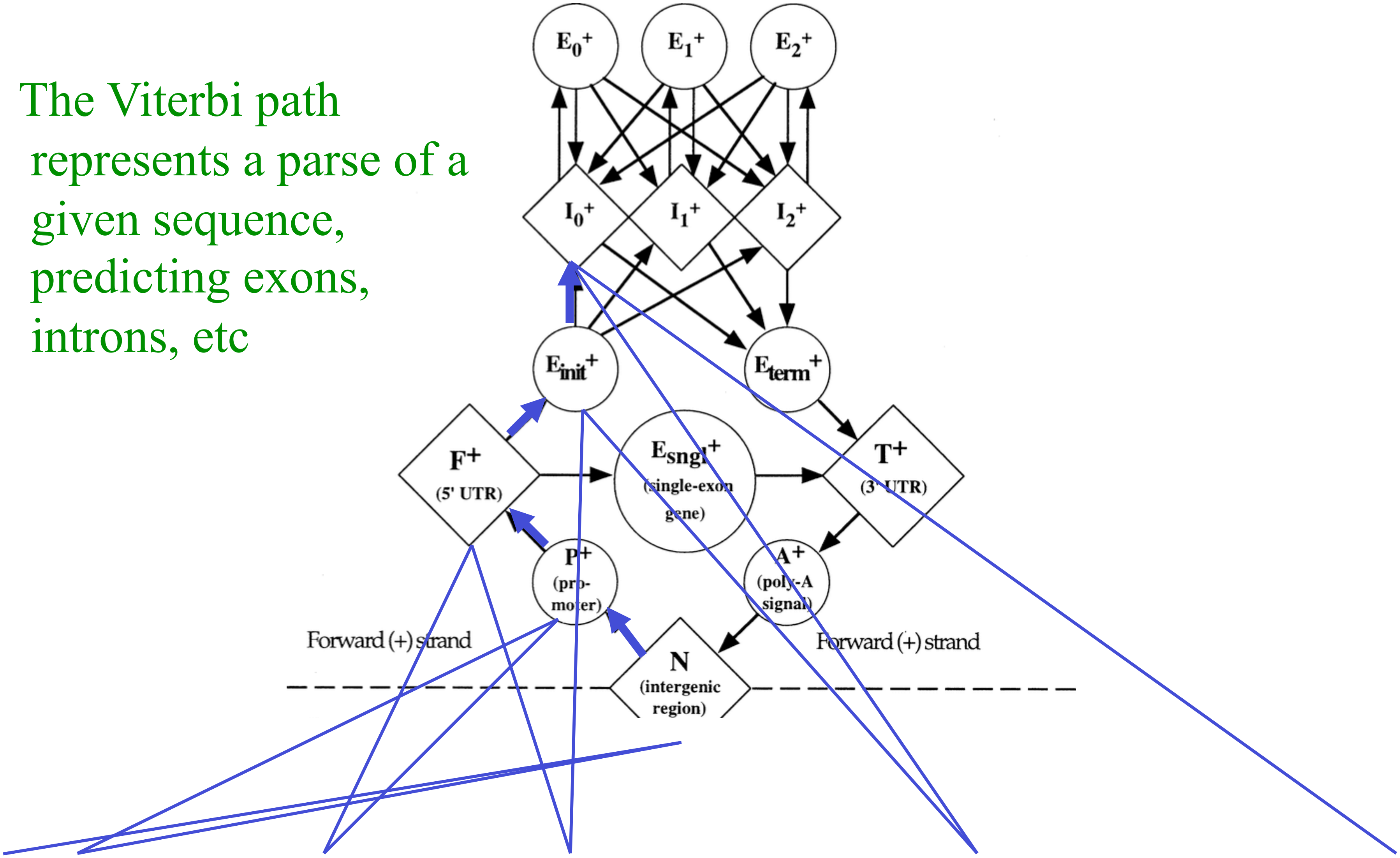
image from the UCSC Genome Browser
http://genome.ucsc.edu/

Eukaryotic Gene Structure



Parsing a DNA Sequence

The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc

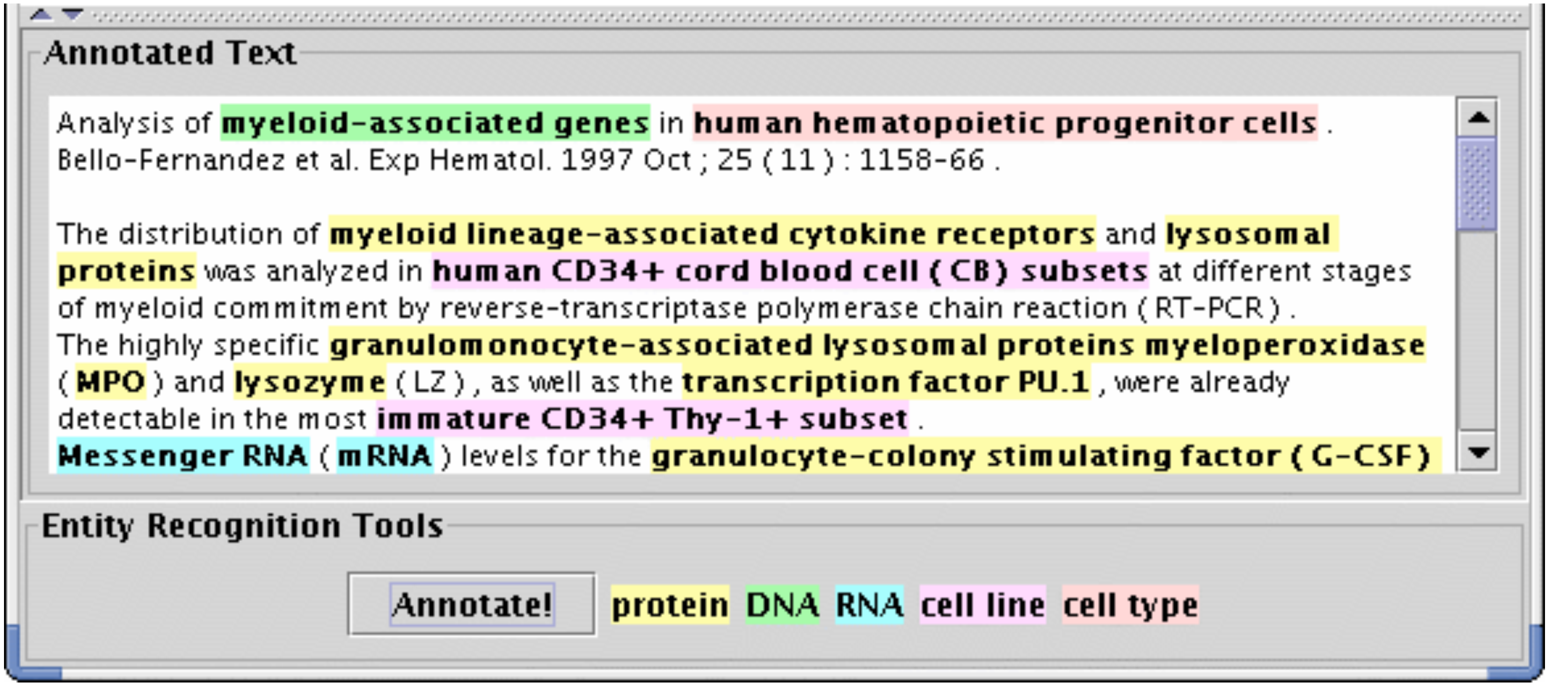


ACCGTTACGTGTCATTCTACGTGATCATCGGATCCTAGAATCATCGATCCGTGCGATCGATCGGATTAGCTAGCTTAGCTAGGAGAGCATCGATCGGATCGAGGAGGAGCCTATATAAATCAA

Example: Information Extraction From Biomedical Literature

given: a passage of text from a scientific article

do: identify mentions of genes or proteins, annotate the article with this information in a database



Analysis of **myeloid-associated genes** in **human hematopoietic progenitor cells** .
Bello-Fernandez et al. Exp Hematol. 1997 Oct ; 25 (11) : 1158-66 .

The distribution of **myeloid lineage-associated cytokine receptors** and **lysosomal proteins** was analyzed in **human CD34+ cord blood cell (CB) subsets** at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR) . The highly specific **granulomonocyte-associated lysosomal proteins myeloperoxidase** (**MPO**) and **lysozyme** (LZ) , as well as the **transcription factor PU.1** , were already detectable in the most **immature CD34+ Thy-1+ subset** . **Messenger RNA** (**mRNA**) levels for the **granulocyte-colony stimulating factor (G-CSF)**

Next Time…

- basic molecular biology
- sequence alignment
- probabilistic sequence models
- **gene expression analysis**
- protein structure prediction
  - by Ameet Soni