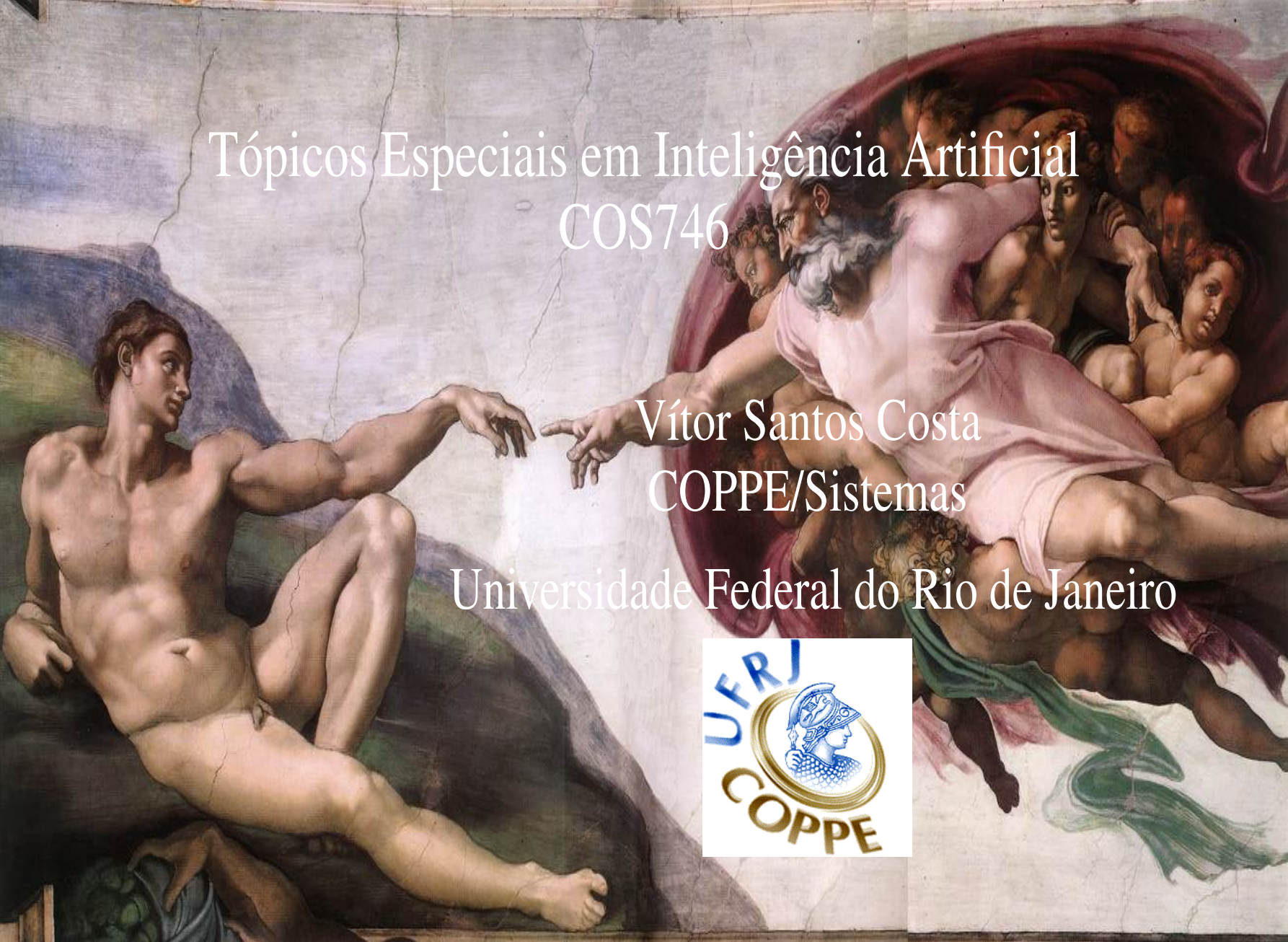


Tópicos Especiais em Inteligência Artificial
COS746

Vítor Santos Costa
COPPE/Sistemas
Universidade Federal do Rio de Janeiro

Agradecimento

- Copiado dos slides de Mark Craven/C. David Page para BMI/CS 576, UW-Madison

Definição de Probabilidade

- *Interpretação Frequentista:* a probabilidade de um evento é a proporção de eventos temporais desse tipo que vão ocorrer passado muito tempo.
- Exemplos:
  - ⋆ a probabilidade que o meu voo para o Porto saia a tempo
  - ⋆ a probabilidade que o bilhete ganhe a lotaria
  - ⋆ a probabilidade que chova amanhã
- Sempre um número entre $[0, 1]$
  - ⋆ 0 significa que nunca vai acontecer
  - ⋆ 1 significa que acontece sempre

Espaço de Amostragem

- *Espaço de Amostragem*: o conjunto de resultados possíveis para o evento
- exemplos:
  - ⋆ voo para o Porto: {certo, atrasado }
  - ⋆ lotaria: {bilhete 1 ganha, bilhete 2 ganha,. . .,bilhete n ganha}
  - ⋆ tempo amanhã:
    - ∗ {chove,não chove}
    - ∗ {sol, chuva}
    - ∗ {sol, nublado, chuvisco, chuva, tempestade}

Variáveis Aleatórias

- *Variável Aleatória:* a variável representa o resultado de uma experiência
- exemplo:
  - X representa o resultado do meu voo para o Porto
  - escrevemos a probabilidade do voo estar certo como $Pr(X = acerto)$
  - quando for claro no que estamos a pensar, podemos escrever $Pr(acerto)$

Notação

- Letras maiúsculas e palavras capitalizadas denotam variáveis aleatórias
- Letras minúsculas e palavras não-capitalizadas denotam valores
- vamos escrever um valor para uma variável como sendo:

$$Pr(X = x) \qquad Pr(Febre = \texttt{verdade})$$

- Podemos usar uma notação compacta

$$Pr(x) \text{ em vez de } Pr(X = x)$$

- para variáveis booleanas, podemos usar:
  - $Pr(\texttt{febre})$ para $Pr(Febre = \texttt{verdade})$
  - $Pr(\neg\texttt{febre})$ para $Pr(Febre = \texttt{falso})$

Distribuição de Probabilidade

- Se X for uma variável aleatória, a função dada por $Pr(X = x)$ para cada x é a distribuição de probabilidade de X
- Requisitos:
  - $Pr(X = x) \geq 0$ para qualquer x
  - $\sum_x Pr(x) = 1$

Distribuição Conjunta

- *Probabilidade Conjunta* a função dada por $Pr(X = x, Y = y)$
- ler como "X vale x e Y vale y"
- exemplo:

x, y	$Pr(X = x, Y = y)$
sol, certo	0.20
chuva, certo	0.30
sol, atrasado	0.40
chuva, atrasado	0.10

Distribuição Marginal

- a *distribuição marginal* de X é definida como:

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

  "a distribuição de X ignorando outras variáveis"

- Esta distribuição generaliza para mais do que 2 variáveis, e.g.

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

Exemplo de Distribuição Marginal

Distribuição Conjunta

x, y	$Pr(X = x, Y = y)$
sol, certo	0.20
chuva, certo	0.30
sol, atrasado	0.40
chuva, atrasado	0.10

Distribuição Marginal para X

x	$Pr(X = x)$
sol	0.60
chuva	0.40

Distribuições Condicionais

- A *distribuição condicional* de X dado Y é definida como:

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

"A distribuição de X dado que sabemos Y"

Exemplo de Distribuição Condicional

Distribuição Conjunta

x, y	$Pr(X = x, Y = y)$
sol, certo	0.20
chuva, certo	0.30
sol, atrasado	0.40
chuva, atrasado	0.10

Distribuição Condicional para X dado que $Y =$ certo

x	$Pr(X = x \mid Y = \texttt{certo})$
sol	0.20/0.50 = 0.40
chuva	0.30/0.50 = 0.60

Independência

Duas variáveis aleatórias X e Y são *independentes* se

$$Pr(x, y) = Pr(x) \times Pr(y) \quad \text{para qualquer } x \text{ e } y$$

Exemplo de Distribuição Marginal

Distribuição Conjunta

x, y	$Pr(X = x, Y = y)$
sol, certo	0.20
chuva, certo	0.30
sol, atrasado	0.40
chuva, atrasado	0.10

Distribuição Marginal para X

x	$Pr(X = x)$
sol	0.60
chuva	0.40

y	$Pr(Y = y)$
sol	0.50
chuva	0.50

Independência Condicional

- Duas variáveis X e Y são condicionalmente independentes dado Z se

$$Pr(X|Y,Z) = Pr(X|Z)$$

"se soubermos o valor de Z, saber Y não faz diferença para saber X"

- Alternativamente:

$$Pr(X,Y|Z) = Pr(X|Z) \times Pr(Y|Z) \text{ para qualquer } `x, y, z$$

Exemplo de Independência Condicional

Gripe	Febre	Vomitar	**Pr**
V	V	V	0.04
V	V	F	0.04
V	F	V	0.01
V	F	F	0.01
F	V	V	0.009
F	V	F	0.081
F	F	V	0.081
F	F	F	0.729

- Febre e Vomitar não são independentes: $Pr(\texttt{febre}, \texttt{vomitar}) \neq Pr(\texttt{febre}) \times Pr(\texttt{vomitar})$
- Febre e vomitar são condicionalmente independentes dado gripe:
  - $\star$ $Pr(\texttt{febre}, \texttt{vomitar}|\texttt{gripe}) \neq Pr(\texttt{febre}|\texttt{gripe}) \times Pr(\texttt{vomitar}|\texttt{gripe})$
  - $\star$ $Pr(\texttt{febre}, \texttt{vomitar}|\neg\texttt{gripe}) \neq Pr(\texttt{febre}|\neg\texttt{gripe}) \times Pr(\texttt{vomitar}|\neg\texttt{gripe})$

Teorema de Bayes

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

- Este teorema é extremamente útil
- Há muitos casos em que é difícil estimar $Pr(x|y)$, mas não é difícil estimar $Pr(y|x)$ e $Pr(x)$.

Exemplo do Teorema de Bayes

- Habitualmente médicos não conseguem estimar facilmente $Pr(Doenca|Sintoma)$
- É mais fácil para eles estimar $Pr(Sintoma|Doenca)$
- Se pudermos estimar $Pr(\texttt{febre}|\texttt{gripe})$ e $Pr(\texttt{gripe})$ podemos usar o teorema de Bayes para fazer o diagnóstico:

$$Pr(\texttt{gripe}|\texttt{febre}) = \frac{Pr(\texttt{febre}|\texttt{gripe})Pr(\texttt{gripe})}{Pr(\texttt{febre}|\texttt{gripe})Pr(\texttt{gripe}) + Pr(\texttt{febre}|\neg\texttt{gripe})Pr(\neg\texttt{gripe})}$$

Valores Esperados

- O *valor esperado* de uma variável aleatória que toma valores numéricos é definido como:

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

  É o mesmo que a média

- Podemos também falar do valor esperado de um função de variável aleatória:

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

Exemplo de Valor Esperado

- Clientes de Sapataria:

$$E[TamanhoDoPe] = 38 \times Pr(TamanhoDoPe = 38) + \ldots$$

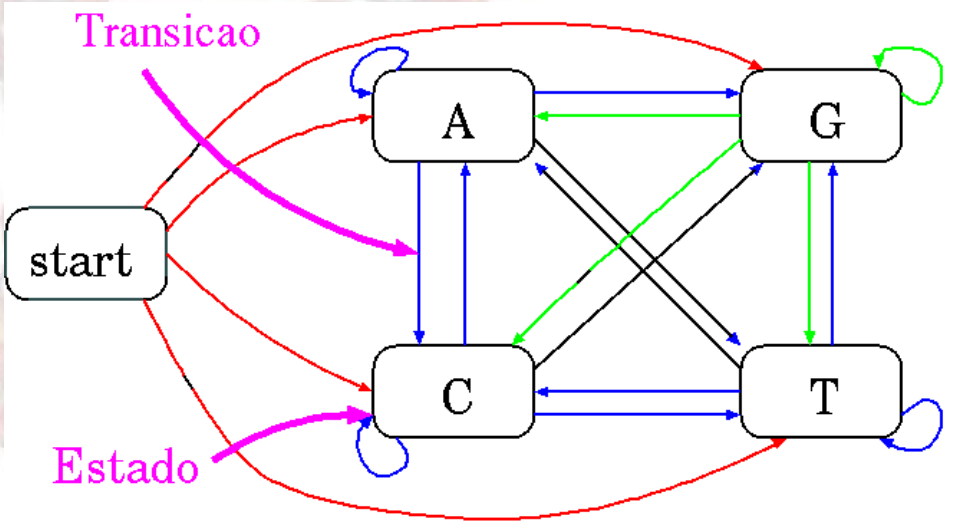
- Lotaria Simples:

$$E[lucro(Lotaria)] =$$
$$ganho(\texttt{premiado})Pr(\texttt{premiado}) + ganho(\neg\texttt{premiado})Pr(\neg\texttt{premiado}) =$$
$$\$100 \times 0.001 - \$1 \times 0.999 =$$
$$-\$0.8999$$

Cadeias de Markov

- Há muitos casos em que queremos representar regularidades estatísticas de certos tipos de sequências:
  - ⋆ genes
  - ⋆ locais reguladores em DNA (eg, onde a polimerase do RNA e os factores de transcrição associam)
  - ⋆ proteínas de uma família
- Modelos de Markov são muito apropriados para este tipo de tarefa.

Cadeias de Markov



probabilidades de transição

$$Pr(x_i = a|x_{i-1} = g) = 0.16$$
$$Pr(x_i = c|x_{i-1} = g) = 0.34$$
$$Pr(x_i = g|x_{i-1} = g) = 0.38$$
$$Pr(x_i = t|x_{i-1} = g) = 0.12$$

Modelos de Cadeia de Markov

- Um modelo de cadeia de Markov é definido como:
  - ⋆ um conjunto de estados
  - ⋆ alguns estados emitem símbolos
  - ⋆ outros estados (eg, estado `begin`) são silenciosos
  - ⋆ um conjunto de transições com probabilidades associadas
- as transições saindo de um estado definem uma distribuição sobre os estados seguintes possíveis.

Modelos de Cadeia de Markov

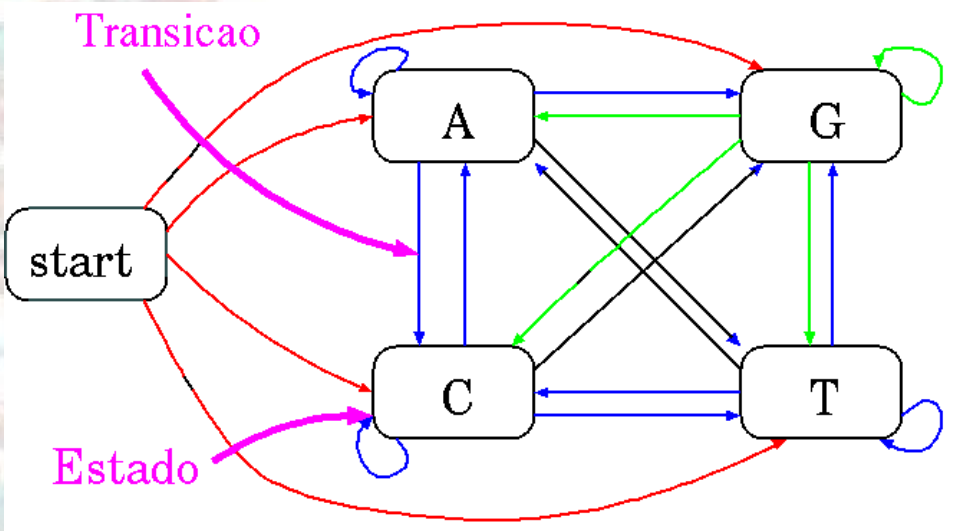
- Dada uma sequência x de comprimento L, queremos saber qual a probabilidade dado o nosso modelo
- Para qualquer modelo probabilístico de sequências, podemos escrever esta probabilidade como:

$$\begin{aligned} Pr(x) = \; & Pr(x_L, x_{L-1}, \ldots, x_1) = \\ & Pr(x_L|x_{L_1}, \ldots, x_1) Pr(x_{L-1}|x_{L-2}, \ldots, x_1) \ldots Pr(x_1) \end{aligned}$$

- propriedade chave das cadeias de Markov da primeira ordem: a probabilidade de cada x_i depende apenas de x_{i-1}:

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

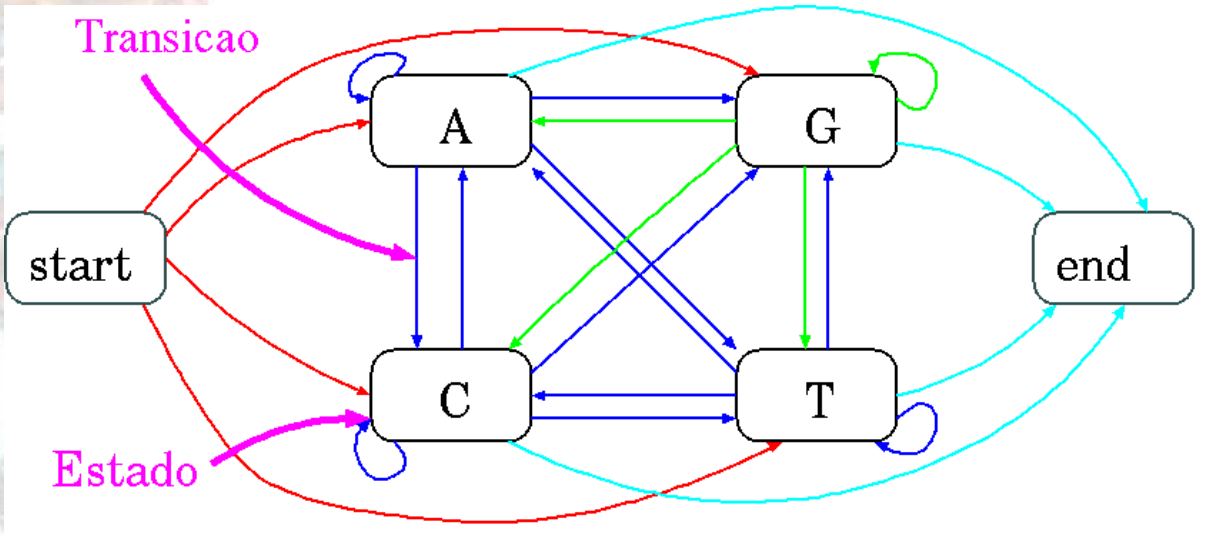
Probabilidade de uma Sequência



$$Pr(cggt) = Pr(c)Pr(g|c)Pr(g|g)Pr(t|g)$$

Modelos de Cadeia de Markov

- Podem ter um estado final: permitem ao modelo representar:
  - ⋆ uma distribuição sobre sequências de tamanhos diferentes
  - ⋆ preferências para terminar sequências com certos tamanhos



Notação de Cadeias de Markov

- Os parâmetros de transição podem ser denotados como $a_{x_{i-1}x_i}$ onde:

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

- podemos denotar a probabilidade de uma sequência x como sendo

$$a_{Bx_1} \prod_{i=2}^{L} a_{x_{i-1}x_i} = Pr(x_1) \prod_{i=2}^{L} Pr(x_i|x_{i-1})$$

onde a_{Bx_1} representa a transição do estado `begin`.

Exemplo de Aplicação

- Ilhas CpG:
  - ⋆ dinucleotídeos CG são mais raros em genomas eucarióticos do que esperado dada as probabilidade marginais de C e de G
  - ⋆ mas as regiões acima de um gene são mais ricas em dinucleotídeos CG do que em qualquer outro lado: *ilhas CpG*
  - ⋆ informação útil para encontrar genes
- podemos prever ilhas CpG com cadeias de Markov:
  - ⋆ uma para representar ilhas CpG
  - ⋆ outra para representar o resto do genoma.

Estimando os Parâmetros do Modelo

- Obtido um conjunto de dados, (eg, um conjunto de sequências de ilhas CpG), como determinar os parâmetros do nosso modelo?
- Uma solução: `estimação com parecença máxima`:
  - ⋆ Dado um conjunto de dados D
  - ⋆ obter os parâmetros θ para maximizar

$$Pr(D|\theta)$$

  - ⋆ ie, fazer com que os dados pareçam o mais possível de acordo com o modelo.

Estimação de Máxima Semelhança (Likelihood)

- suponhamos que queremos estimar os parâmetros $Pr(a)$, $Pr(c)$, $Pr(g)$, e $Pr(t)$
- e recebemos as sequências:
  - `accgcgctta`
  - `gcttagtgac`
  - `tagccgttac`
- A estimação de máxima semelhança será:

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

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

O Método Bayesiano

- Em vez de estimarmos os parâmetros estreitamentos dos dados, podemos começar com alguma crença prévia
- por exemplo, podemos usar estimativas de Laplace:

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

- $n_a + 1$ é uma *pseudo-contagem*
- e n_i representa o número de ocorrências do caracter i
- usando estimativas de Laplace com as sequências:
  - ⋆ `gccgcgcttg`
  - ⋆ `gcttggtggc`
  - ⋆ `tggccgttgc`

O Método Bayesiano

- uma forma mais geral: *m-estimates*

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

- p_a é a probabilidade prévia para a
- m é o número de instâncias virtuais
- com $m = 8$ e prévias uniformes:

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

Estimação de Parâmetros de Primeira Ordem

- Para estimar um parâmetro de primeira ordem, como $Pr(c|g)$, contamos o número de vezes que g segue a história de c nas nossas sequências
- Usando estimadores de Laplace:

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

$$Pr(a|g) = \frac{7+1}{12+4} \quad \ldots$$

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

$$Pr(a|g) = \frac{2+1}{12+4}$$

Cadeias de Markov para Discriminação

- queremos distinguir ilhas CpG de outras regiões
- dadas sequências de ilhas CpG, e sequências de outras regiões, podemos construir:
  - ⋆ um modelo para representar ilhas CpG
  - ⋆ um *modelo nulo* para representar as outras regiões
- Podemos avaliar uma sequência de teste por:

$$score(x) = log \frac{Pr(x|\text{modelo CpG})}{Pr(x|\text{modelo nulo})}$$

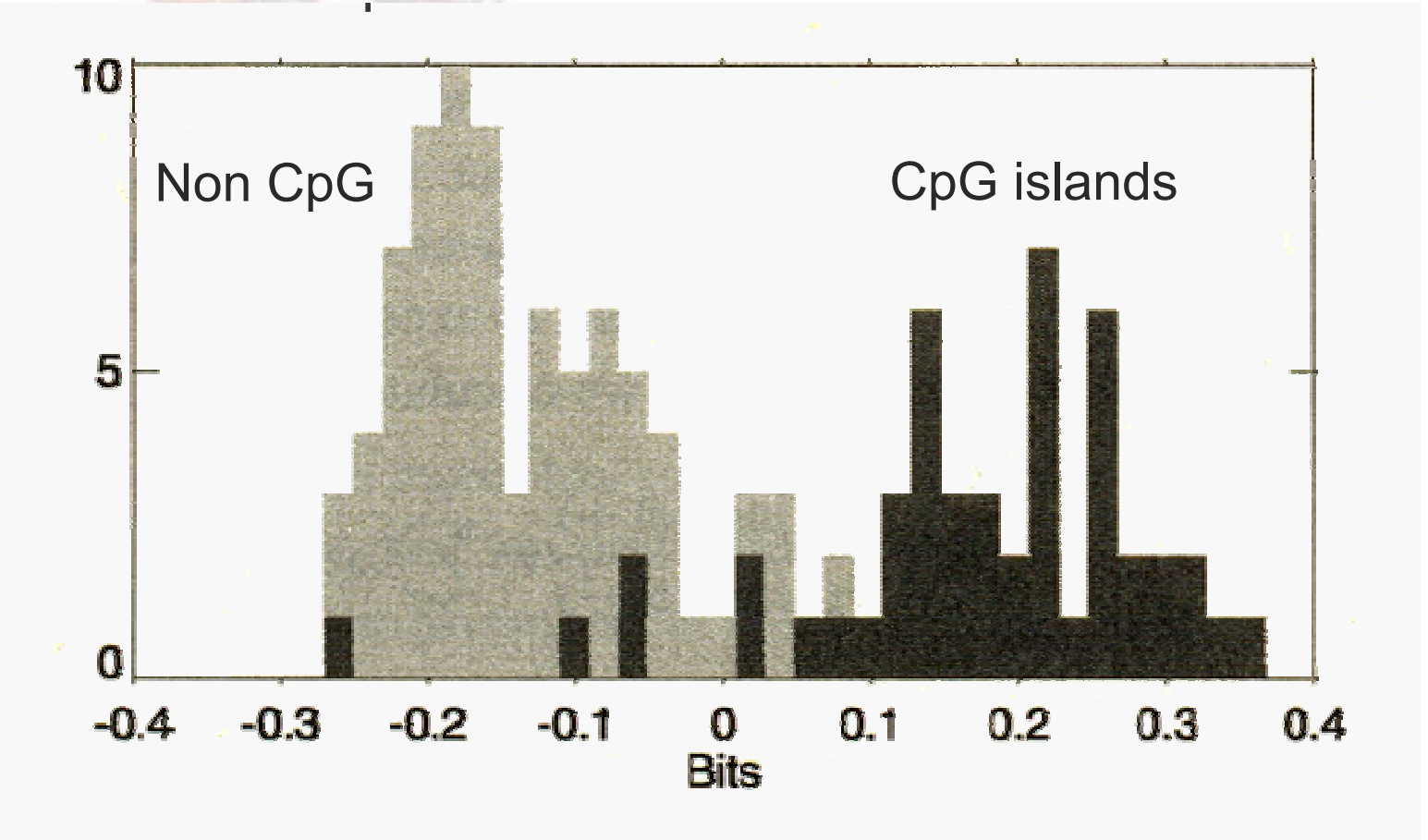
Cadeias de Markov para Discriminação

- Parâmetros estimados para CpG e para modelos nulos:
  - ⋆ genoma humano contém 48 ilhas CpGs
  - ⋆ 60000 nucleotídeos

$+$	A	C	G	T
A	.18	.27	.43	.12
C	.17	.37	.27	.19
G	.16	.34	.38	.12
T	.08	.36	.38	.18

$-$	A	C	G	T
A	.30	.21	.28	.21
C	.32	.30	.08	.30
G	.25	.24	.30	.21
T	.18	.24	.29	.29

Cadeias de Markov para Discriminação



Normalizado por tamanho de sequências.

Cadeias de Markov para Discriminação

- Porque usar

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

- A regra de Bayes diz-nos que:

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

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

- se não estivermos a considerar probabilidades prévias das duas classes ($Pr(CpG)$) e ($Pr(null)$) então basta comparar $Pr(x|CpG)$ e $Pr(x|null)$.

Cadeias de Markov de Mais Alta Ordem

- A propriedade de Markov especifica que a probabilidade de um estado depende apenas da probabilidade do estado prévio
- mas podemos construir mais memória nos nossos estados usando um modelo de mais alta ordem
- num modelo de Markov de ordem n:

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

Selecção da Ordem de um Modelo

- Modelos de ordem mais alta lembram-se de mais *história*
- mais história pode ser útil para previsões
- Por exemplo
  - ⋆ Preveja a próxima palavra neste fragmento de frase:
    - ∗ . . . quer ____
  - ⋆ Agora com mais história
    - ∗ . . . quem desdenha quer ____

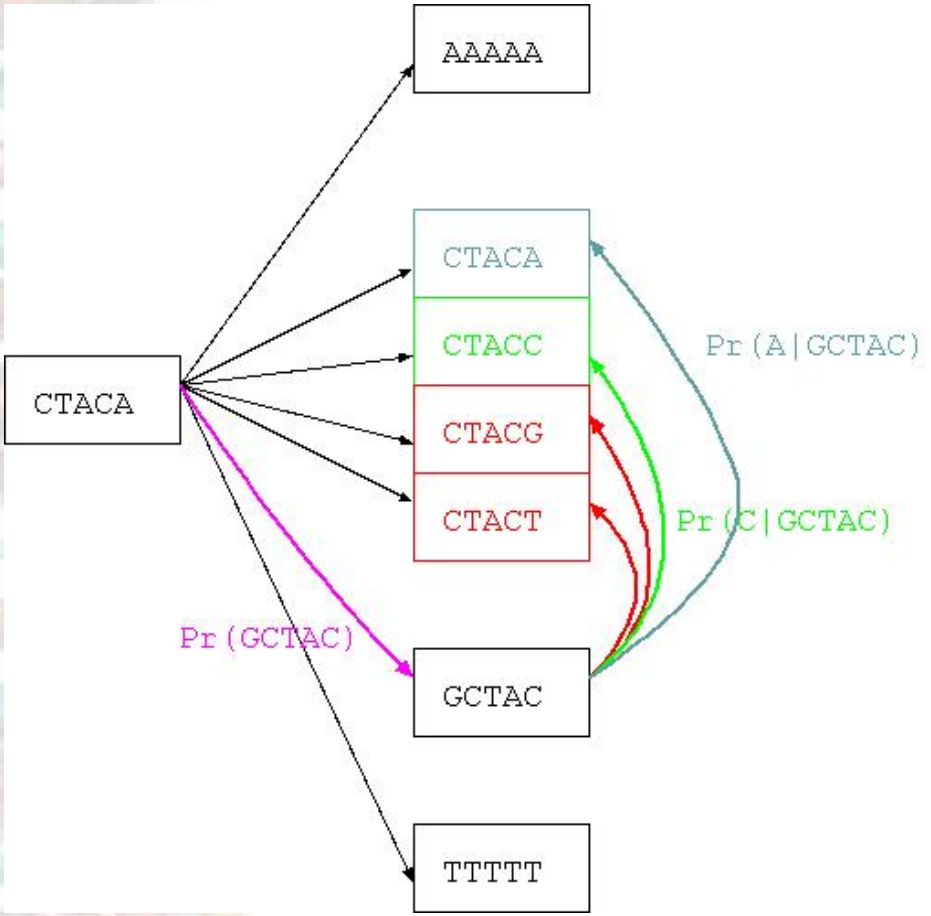
Estimando a ordem de um Modelo de Markov

- O problema é que o número de parâmetros que precisamos de estimar cresce exponencialmente com a ordem
  - ⋆ Para modelar DNA precisamos de $O(4^{n+1})$ para um modelo de ordem n
- quanto maior a ordem, quanto menos confiáveis serão as nossas estimas de parâmetros:
  - ⋆ para estimar os parâmetros de uma cadeia de segunda ordem de Markov do genoma completo de E. coli, precisamos de ver cada palavra > 72000 em média
  - ⋆ para estimar os parâmetros de uma cadeia de oitava ordem, precisamos de ver cada palavra 5 vezes em média

Cadeias de Markov de Mais Alta Ordem

- Uma cadeia de Markov de ordem n sobre um alfabeto A é equivalente a uma cadeia de primeira ordem sobre o alfabeto dos tuplos-n A^n
- Exemplo, uma cadeia de segunda ordem pode ser tratada como uma cadeia de primeira ordem sobre os pares:
  - ⋆ AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT

Uma Cadeia de Markov de Quinta Ordem

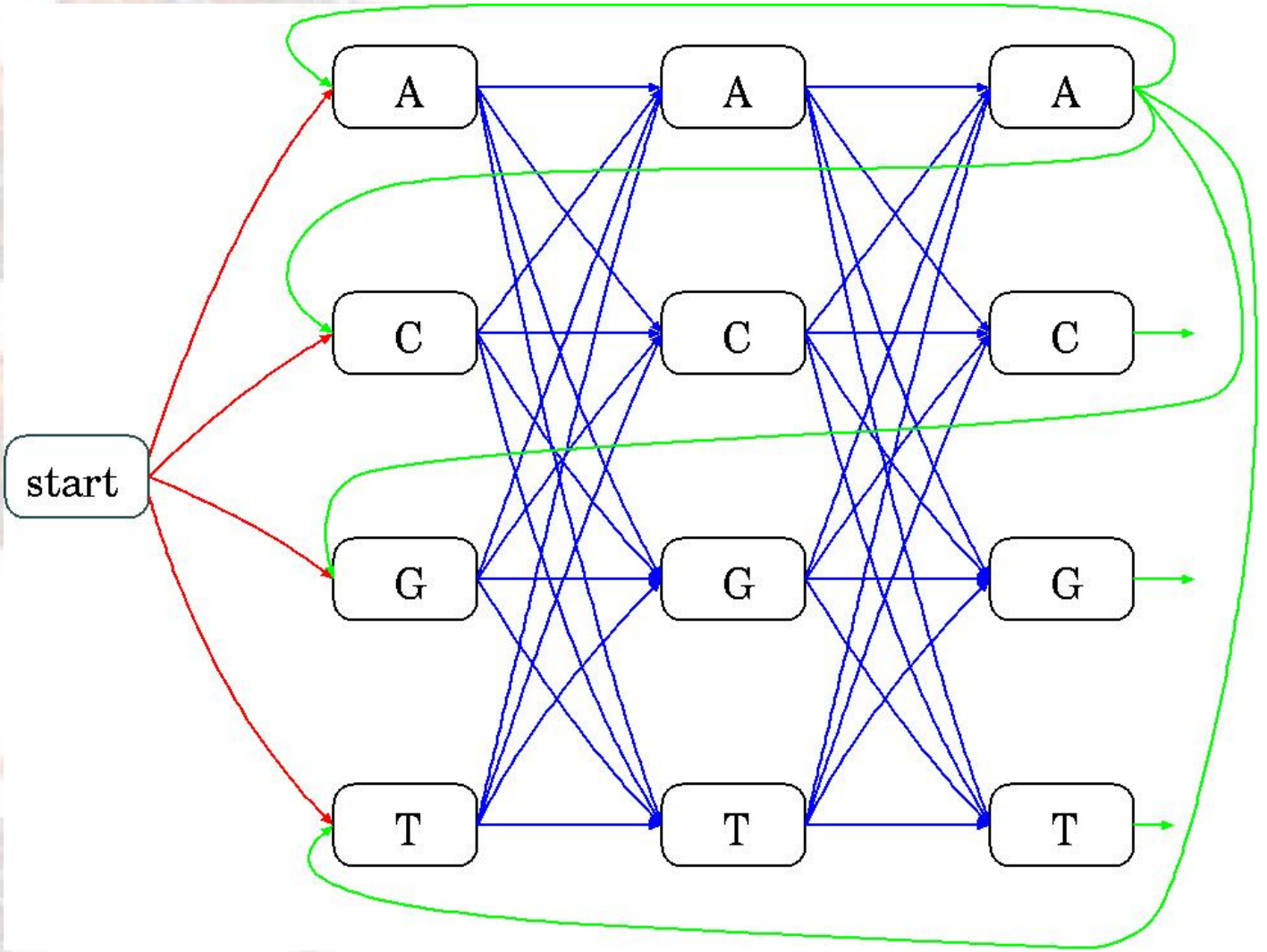


$$Pr(gctaca) = Pr(a|gctac)Pr(gctac)$$

Cadeias Não Homogéneas de Markov

- Nos modelos de Markov que temos considerado até agora, as probabilidades não dependem de onde estamos numa dada sequência
- Num modelo não homogéneo de Markov, podemos ter distribuições diferentes em posições diferentes da sequência
- Considere a modelação de codões em regiões de codificação de proteínas

Uma Cadeia Não Homogénea de Markov



Aplicação de Cadeia Não Homogénea

- construir modelos de Markov para regiões que codificam ou não
- aplicar modelos para ORFs (quadros de leitura abertos) ou para janelas de tamanho fixo da sequências
- Exemplo: GeneMark de Borodovsky et al:
  - ⋆ sistema popular para identificar genes em genomas de bactérias
  - ⋆ usa cadeias de quinta ordem não homogéneas de Markov
  - ⋆ `http://opal.biology.gatech.edu/GeneMark/`

ORF: Quadro de Leitura Aberta

- Um quadro de leitura aberta é uma sequência que pode potencialmente codificar uma proteína:
  - ⋆ começa com um potencial codon de inicio
  - ⋆ acaba com um potencial codon de fim
  - ⋆ satisfaz um tamanho mínimo
  - ⋆ em procariotes, o codon de inicio e de estão na mesma janela

```
ATG CCG AAA TCG CAT GCA TTT GTG ... TAG
```

Técnicas para Encontrar Genes

Procurar por:

- *semelhança de sequência*: encontrar genes procurando sequências semelhantes a sequências que se sabem ser relacionadas a genes
- *sinal*: encontrar genes identificando os sinais de sequenciamento envolvidos em expressão de genes
- *conteúdo*: usar propriedades estatísticas que distinguem DNA que codifica para proteínas do resto do DNA
- *combinado*: métodos mais avançados combinam estas estratégias.

Procura Por Conteúdo

- codificar uma proteína afecta as propriedades estatísticas de uma sequência de DNA:
  - ⋆ alguns amino-ácidos são ácidos mais frequentemente que outros (Leu é mais popular do que Trp)
  - ⋆ número diferente de codões para amino-ácidos diferentes (Leu tem 6, Trp tem 1)
  - ⋆ para um certo amino-ácido, habitualmente um codon é usado mais frequentemente que outros:
    - ∗ chamado de *preferência de codon*
    - ∗ essas preferências variam com a espécie

Exemplo: E. Coli

AA	codon	freq. por 1000
Gly	GGG	1.89
Gly	GGA	0.44
Gly	GGU	52.99
Gly	GGC	34.55
Glu	GAG	15.68
Glu	GAA	57.20
Asp	GAU	21.63
Asp	GAC	43.26

Quadros de Leitura

- 6 quadros para uma fita:

Fita de DNA	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A

Modelos de Markov e Quadros de Leitura

- Imagine modelar uma sequência de codificação
- para cada palavra que avaliamos, queremos considerar a sua posição em relação ao quadro de leitura que estamos assumindo

Fita de DNA	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A
	A	C	G	C	A	G	A	T	A	T	C	A	T	G	A

Cadeia de Quinta Ordem Não Homogénea