Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Cadeias de Markov

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

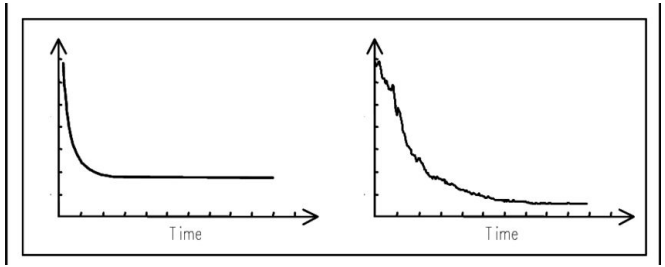
```
> cd /path/to/IMD0601
```

```
> git pull
```

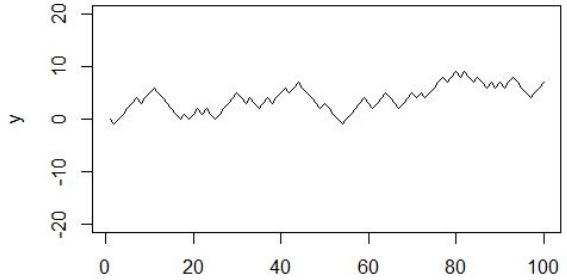
Revisão

- Complexidade em obter uma distribuição posteriori
- MCMC!
  - Modelos estocástico;
  - Processo estocástico;
  - Processo de Markov;
  - Cadeia de Markov;
- Processo de Markov;
  - Random Walk;
- Cadeia de Markov
  - Operações básicas de matriz;
  - O sapo e a vitória régia.

$$f(\mu,\phi \mid X_1,...,X_n) \propto \phi^{n/2} e^{-\frac{\phi}{2}\sum_{i=1}^{n}(X_i-\mu)^2} \frac{1}{\tau} e^{-\frac{1}{2\tau^2}\mu^2} \phi^{\alpha-1} e^{-(\phi/\beta)}$$



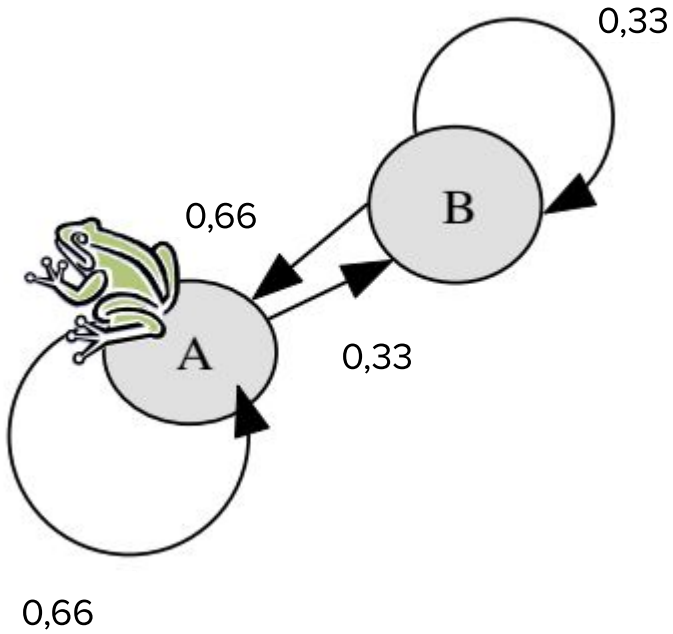
$$P(X_{n+1}=x_{n+1} | X_n=x_n, X_{n-1}=x_{n-1}, \ldots X_0=x_0) = P(X_{n+1}=x_{n+1} | X_n=x_n)$$



Um modelo simples de Cadeia de Markov

Conceitos importantes:

- Cadeia de Markov como modelo probabilístico;
- Estados da cadeia de Markov (folha A e B em um dado tempo);
- Matriz de transição de um estado para o outro;
- Computar as probabilidades de transição em intervalos k>1;
- Distribuição estacionária ➔ Convergência.



Sequência de DNA e Cadeia de Markov

Como modelar sequência de DNA com Cadeia de Markov?

- Cada posição do DNA pode ser considerada uma variável aleatória que pode assumir um dos valores {A, T, C, G};
- Tomando a direção 5’ ➔ 3’, X_0=A e X_1=T;
- Assumir que a probabilidade de estado em uma posição depende apenas do valor da posição anterior.

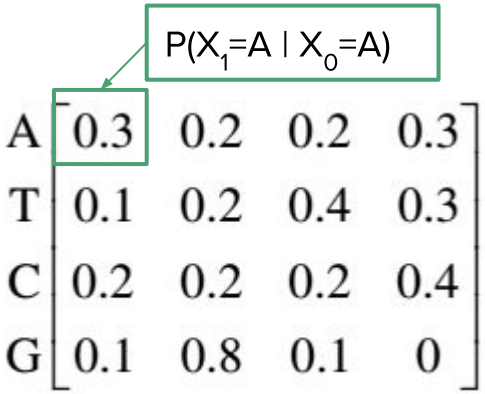
$$P(X5=C|X4=C,X3=G,X2=T,X1=T,X0=A)=P(X5=C|X4=C)$$

Matriz de transição

Sequência de DNA e Cadeia de Markov

Gerando uma matriz de transição

- Obter as probabilidades de transição através dos dados;
- Frequência do nucleotídeo adjacente;




```
TCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCAT
CATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATC
CGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGC
CCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCG
GGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTG
GCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGA
CCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTC
TTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACC
CTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCA
TACCGGTCAAGGGCATTTCCAATCTGAATAACATGGCAATGTT
GATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGC
CAATCATCTTCCGAATACAGCATCAGTTTCTGCGTTCCACAAA
TGCAGGAAGAGTTCTACCTGGAACTGAAAGAAGGCTTACTGGA
CATTATCTCGGTGGTAGGTGATGGTATGCGCACCTTGCGTGGG
GCCCGCGCCAATATCAACATTGTCGCCATTGCTCAGGGATCTT
ATAACGATGATGCGACCACTGGCGTGCGCGTTACTCATCAGAT
```

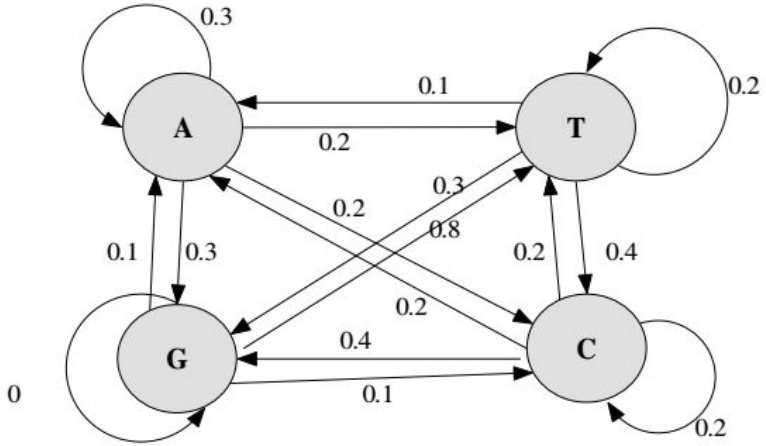
Sequência de DNA e Cadeia de Markov

Gerando uma matriz de transição

- Obter as probabilidades de transição através dos dados;
- Frequência do nucleotídeo adjacente;

$P(X_1=A \mid X_0=A)$

A	0.3	0.2	0.2	0.3
T	0.1	0.2	0.4	0.3
C	0.2	0.2	0.2	0.4
G	0.1	0.8	0.1	0



Sequência de DNA e Cadeia de Markov

Determinando a Convergência e a distribuição estacionária

- Multiplicar a matriz de transição com ela mesma (fazer no R)

Sequência de DNA e Cadeia de Markov

Determinando a Convergência e a distribuição estacionária

- Multiplicar a matriz de transição com ela mesma (fazer no R)

Convergência:

$$\begin{bmatrix} 0.155624 & 0.3497689 & 0.2449923 & 0.24961 \\ 0.155624 & 0.3497689 & 0.2449923 & 0.24961 \\ 0.155624 & 0.3497689 & 0.2449923 & 0.24961 \\ 0.155624 & 0.3497689 & 0.2449923 & 0.24961 \end{bmatrix}$$

Distribuição estacionária: $\pi = \begin{bmatrix} 0.15 & 0.35 & 0.25 & 0.25 \end{bmatrix}$

Distribuição estacionária

$$\pi = \begin{bmatrix} 0.15 & 0.35 & 0.25 & 0.25 \end{bmatrix}$$

Representa o ouro no processo de garimpagem dos modelos de Markov.

No contexto da Bayesiana ➜ **Distribuição posteriori**!

Para que a matriz de transição convirja, a cadeia deve ter as seguintes propriedades:

- Finito;
- Aperiódico;
- Irredutível

Finito

Características de uma Cadeia de Markov para obtenção de uma distribuição posteriori

Número finito de estados possíveis;

- Sapo → 2 estados (folha A e B);
- DNA → 4 estados (A, T, C, G);

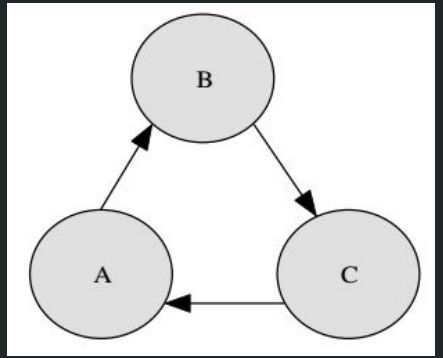
Não necessariamente que é pequeno, mas que o número possível de estado que é finito.

Periodicidade

Características de uma Cadeia de Markov para obtenção de uma distribuição posteriori

A cadeia não deve ser periódica ➜ a cadeia não se comporta como a função seno.

Exemplo:



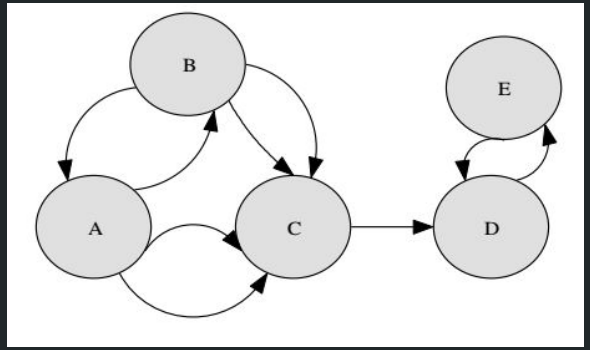
Cadeia de 3 períodos.

Irredutibilidade

Características de uma Cadeia de Markov para obtenção de uma distribuição posteriori

Todos os estados podem ser alcançados por todos os estados.

Exemplo de cadeia redutível:



Cadeia é redutível ➜ não se pode ir do estado E para A, por exemplo.

Não ocorre uma única distribuição estacionária

Ergodicidade

Características de uma Cadeia de Markov para obtenção de uma distribuição posteriori

Cadeias de Markov que sejam tanto aperiódico quanto irredutível.

Cadeias ergódicas possuem uma distribuição estacionária única que é positiva para todos os estados quando este é finito.

Reversibilidade

Características de uma Cadeia de Markov para obtenção de uma distribuição posteriori

Uma cadeia ergódica é reversível se:

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

Não é uma propriedade necessária para obter uma distribuição estacionária única e estável

Outro exemplo de Cadeias de Markov

Ilhas CpG

- Regiões regulatórias importantes encontrados em vários promotores de genes;
- Sofre metilação com facilidade;
  - Metil-C pode mutar para T nas sequências CpG → CpG é raro!
  - Mas essa característica não ocorre em regiões próximas a promotores → CpG mais frequente!
    - Ilha CpG
    - Região de poucas centenas ou milhares de pares de base
- Pergunta: dada uma sequência de DNA, ela veio de uma ilha CpG?

Outro exemplo de Cadeias de Markov

Sequência veio de uma ilha CpG?

- Treinar duas cadeias de Markov:
  - Modelo + ➜ dados de sequências de ilhas CpG;
  - Modelo - ➜ dados de sequências que não são de ilhas CpG;

Outro exemplo de Cadeias de Markov

Sequência veio de uma ilha CpG?

- Treinar duas cadeias de Markov:
  - Modelo + ➔ dados de sequências de ilhas CpG;
  - Modelo - ➔ dados de sequências que não são de ilhas CpG;
- Determinar as matrizes de transição para cada modelo;

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{bmatrix} 0.18 & 0.27 & 0.43 & 0.12 \\ 0.17 & 0.37 & 0.27 & 0.19 \\ 0.16 & 0.34 & 0.38 & 0.12 \\ 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix}
$$

CpG (+)

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{bmatrix} 0.30 & 0.21 & 0.28 & 0.21 \\ 0.32 & 0.30 & 0.08 & 0.30 \\ 0.25 & 0.24 & 0.30 & 0.21 \\ 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix}
$$

Não CpG (-)

Outro exemplo de Cadeias de Markov

Esta sequência veio de uma ilha CpG?

- `cgttcgacgta`

Outro exemplo de Cadeias de Markov

Esta sequência veio de uma ilha CpG?

- `cgttcgacgta`

  P(G|C)*P(T|G)*P(T|T)*P(C|T)*P(G|C)*P(A|G)*P(C|A)*P(G|C)*P(T|G)*P(A|T)

Outro exemplo de Cadeias de Markov

Esta sequência veio de uma ilha CpG?

- `cgttcgacgta`

P(G|C)*P(T|G)*P(T|T)*P(C|T)*P(G|C)*P(A|G)*P(C|A)*P(G|C)*P(T|G)*P(A|T)

A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.38	0.12
T	0.08	0.36	0.38	0.18

CpG (+)

A	0.30	0.21	0.28	0.21
C	0.32	0.30	0.08	0.30
G	0.25	0.24	0.30	0.21
T	0.18	0.24	0.29	0.29

Não CpG (-)

Outro exemplo de Cadeias de Markov

Esta sequência veio de uma ilha CpG?

- `cgttcgacgta`

P(G|C)*P(T|G)*P(T|T)*P(C|T)*P(G|C)*P(A|G)*P(C|A)*P(G|C)*P(T|G)*P(A|T)

A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.38	0.12
T	0.08	0.36	0.38	0.18

CpG (+)

A	0.30	0.21	0.28	0.21
C	0.32	0.30	0.08	0.30
G	0.25	0.24	0.30	0.21
T	0.18	0.24	0.29	0.29

Não CpG (-)

- Modelo + ➔ 0.27*0.12*0.18*0.36*0.27*0.16*0.27*0.27*0.12*0.08 = 6.347497e-08
- Modelo - ➔ 0.08*0.21*0.29*0.24*0.08*0.25*0.21*0.08*0.21*0.18 = 1.485079e-08

Outro exemplo de Cadeias de Markov

Sequência veio de uma ilha CpG?

- `cgttcgacgta`
  - P(G|C)*P(T|G)*P(T|T)*P(C|T)*P(G|C)*P(A|G)*P(C|A)*P(G|C)*P(T|G)*P(A|T)
  - Modelo + ➔ 0.27*0.12*0.18*0.36*0.27*0.16*0.27*0.27*0.12*0.08 = 6.347497e-08
  - Modelo - ➔ 0.08*0.21*0.29*0.24*0.08*0.25*0.21*0.08*0.21*0.18 = 1.485079e-08
- Log odds score

$$log_2 \frac{P(x|modelo+)}{P(x|modelo-)} = log_2 \frac{6.347e-08}{1.485e-08} = 2.095$$

Log odds score > 0, então modelo + é o mais provável.

Outro exemplo de Cadeias de Markov

Gerador de sequências aleatórias

Utilizando a matriz de transição, você pode gerar sequências aleatórias mais “realistas”;

Veja no script em R desta aula como você pode proceder.

Cadeias de Markov

Distribuição estacionária ➔ Distribuição posteriori

Cadeia de Markov deve ser:

- Finita;
- Ergódica ➔ Aperiódico e Irredutível;

Outras aplicações da cadeia de Markov:

- Padrões no DNA;
- Gerador de sequências aleatórias;