Markov chains, part 1

Ben Langmead

Markov chain

Assigning a probability to a sequence using Markov assumption:

$$P(x) \approx$$

Markov assumption

Say x is a length-k string of nucleotides (k-mer)

$P(x_i \mid x_{i-1})$ probability of seeing nucleotide x_i in i^{th} position given that previous nucleotide is x_{i-1}

Markov chain

Assigning a probability to a sequence using Markov assumption:

$$P(x) \underset{\text{assumption}}{\approx} P(x_k \mid x_{k-1})\, P(x_{k-1} \mid x_{k-2}) \ldots P(x_2 \mid x_1)\, P(x_1)$$

Say x is a length-k string of nucleotides (k-mer)

$P(x_i \mid x_{i-1})$ probability of seeing nucleotide x_i in i^{th} position given that previous nucleotide is x_{i-1}

Shorthand: P(G | C) = probability of G given previous is C

Markov chain

Say we are given sequences of several CpG islands. How do we estimate, say, P(G | C)?

P(**G** | **C**) = # times **CG** occurs / # times **C**X occurs

where X is any base

Markov chain

Given CpG island sequences from human chromosome 1, count dinucleotide occurrences and estimate all 16 possible $P(x_i \mid x_{i-1})$:

$$P(\mathbf{A} \mid \mathbf{A}) =$$

Markov chain

Given CpG island sequences from human chromosome 1, count dinucleotide occurrences and estimate all 16 possible $P(x_i \mid x_{i-1})$:

$P(\mathbf{A} \mid \mathbf{A}) =$ # times **AA** occurs / # times **A**X occurs

where X is any base

$P(\mathbf{C} \mid \mathbf{A}) =$ # times **AC** occurs / # times **A**X occurs

$P(\mathbf{G} \mid \mathbf{A}) =$ # times **AG** occurs / # times **A**X occurs

$P(\mathbf{T} \mid \mathbf{A}) =$ # times **AT** occurs / # times **A**X occurs

$P(\mathbf{A} \mid \mathbf{C}) =$ # times **CA** occurs / # times **C**X occurs

(etc)

Markov chain

Given example CpG island substrings we can estimate all P(*base* | *previous base*), P(x_i | x_{i-1}):

x_{i-1} (*previous*) \ x_i	A	C	G	T
A	P(A | A)	P(C | A)	P(G | A)	P(T | A)
C	P(A | C)	P(C | C)	P(G | C)	P(T | C)
G	P(A | G)	P(C | G)	P(G | G)	P(T | G)
T	P(A | T)	P(C | T)	P(G | T)	P(T | T)

Markov chain

Given example CpG island substrings we can estimate all P(*base* | *previous base*), P(X_i | X_{i-1}):

(real data from human chr1)

```
>>> iTab, nTab = islandTransitionTables(fn, ifn)
>>> print(iTab)
[[ 0.18427153, 0.27129525, 0.4055757 , 0.13885752],
 [ 0.19081672, 0.36113346, 0.24897947, 0.19907035],
 [ 0.17440554, 0.32764433, 0.35676759, 0.14118254],
 [ 0.09348595, 0.3474561 , 0.36885   , 0.19020795]]
```

X_{i-1} *(previous)*: **A** **C** **G** **T**

A **C** **G** **T**

X_i

P(T | G)

Rows sum to … 1

Markov chain

(real data from human chr1)

```
>>> iTab, oTab = islandTransitionTables(fn, ifn)
>>> print(iTab)
[[ 0.18427153, 0.27129525, 0.4055757 , 0.13885752],
 [ 0.19081672, 0.36113346, 0.24897947, 0.19907035],
 [ 0.17440554, 0.32764433, 0.35676759, 0.14118254],
 [ 0.09348595, 0.3474561 , 0.36885   , 0.19020795]]
>>> print(oTab)
 [[0.33804066, 0.17971034, 0.23104207, 0.25120694],
  [0.37777025, 0.25612117, 0.03987225, 0.32623633],
  [0.30257815, 0.20326794, 0.24910719, 0.24504672],
  [0.21790184, 0.20942905, 0.2642385 , 0.3084306 ]]
```

Inside: A C G T

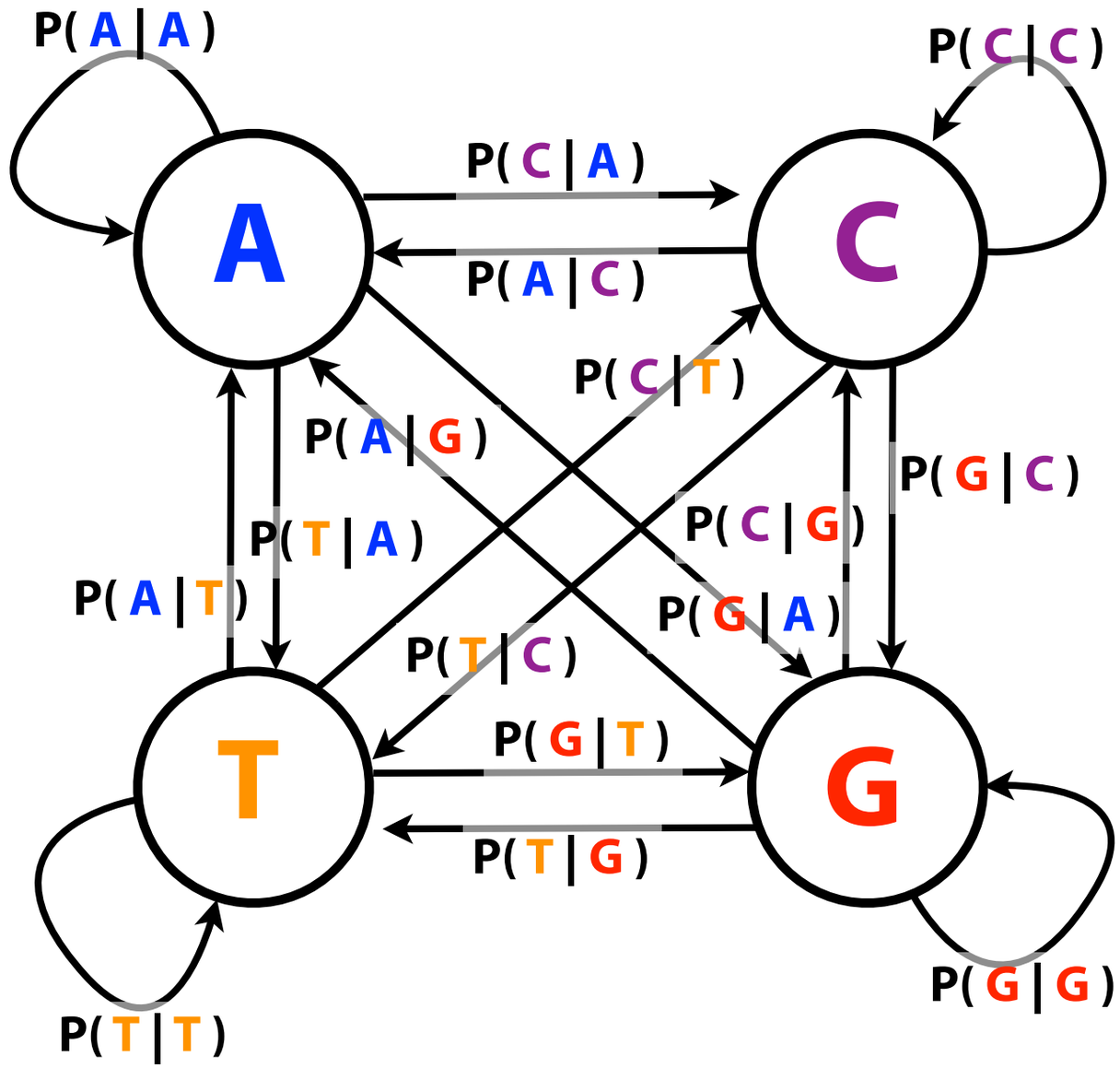
Outside: A C G T

A C G T

Notice anything about the ***outside*** conditional probabilities?

P(G | C) is low ; makes sense: outside CpG islands, C is rarely followed by G

Markov chain



Markov chain =
probabilistic automaton

Markov chain

```
>>> iTab, oTab = islandTransitionTables(fn, ifn)
>>> print(iTab)
[[ 0.18427153, 0.27129525, 0.4055757 , 0.13885752],
 [ 0.19081672, 0.36113346, 0.24897947, 0.19907035],
 [ 0.17440554, 0.32764433, 0.35676759, 0.14118254],
 [ 0.09348595, 0.3474561 , 0.36885   , 0.19020795]]
```

X_{i-1} **A C G T**

A C G T

X_i

$x = \text{GATC}$

Markov chain

```
>>> iTab, oTab = islandTransitionTables(fn, ifn)
>>> print(iTab)
[[ 0.18427153, 0.27129525, 0.4055757 , 0.13885752],
 [ 0.19081672, 0.36113346, 0.24897947, 0.19907035],
 [ 0.17440554, 0.32764433, 0.35676759, 0.14118254],
 [ 0.09348595, 0.3474561 , 0.36885   , 0.19020795]]
```

X_{i-1}: A, C, G, T

A C G T

X_i

$x = \text{GATC}$

$$P(x) = P(x_4 \mid x_3)\, P(x_3 \mid x_2)\, P(x_2 \mid x_1)\, P(x_1)$$

$$P(x) = P(C \mid T)\, P(T \mid A)\, P(A \mid G)\, P(G)$$

$$= 0.347 \times 0.139 \times 0.174 \times 0.25 = 0.0021$$

← (Assuming A, C, G, T equally likely)

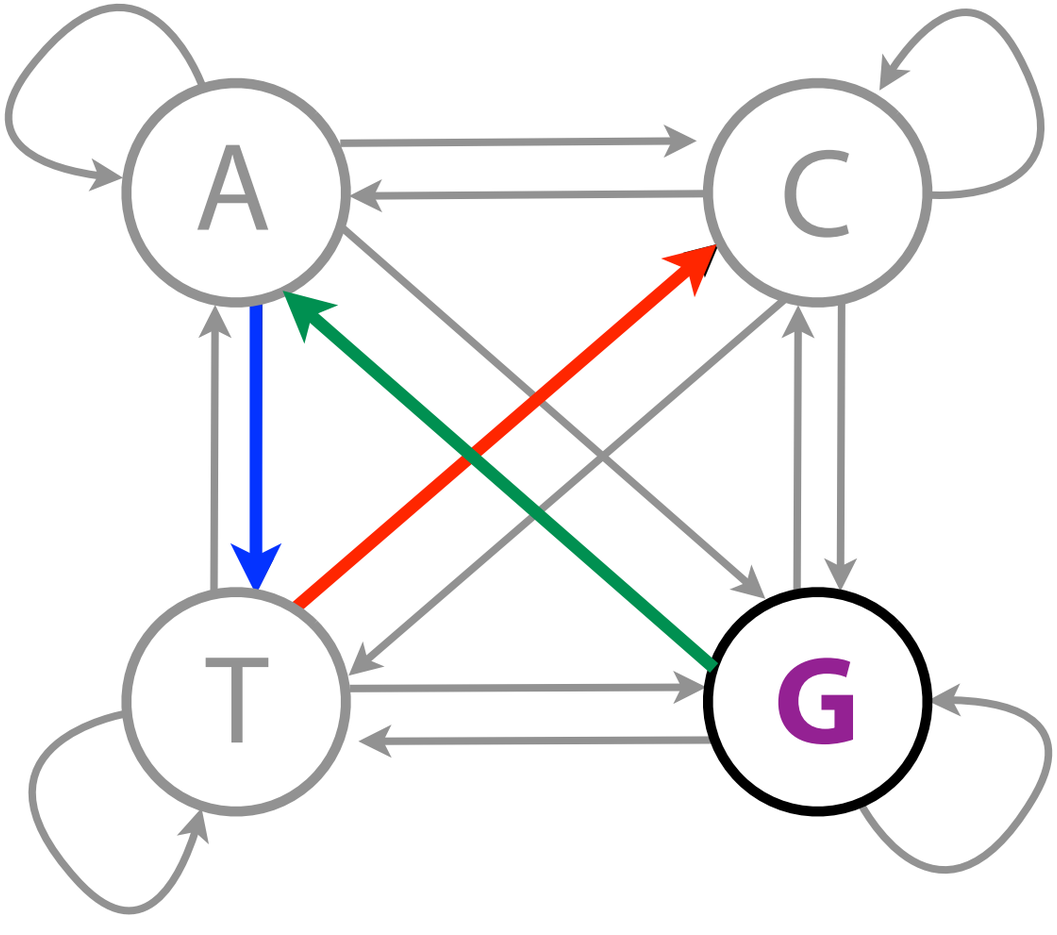
Markov chain

```
>>> iTab, oTab = islandTransitionTables(fn, ifn)
>>> print(iTab)
[[ 0.18427153, 0.27129525, 0.4055757 , 0.13885752],
 [ 0.19081672, 0.36113346, 0.24897947, 0.19907035],
 [ 0.17440554, 0.32764433, 0.35676759, 0.14118254],
 [ 0.09348595, 0.3474561 , 0.36885   , 0.19020795]]
```

X_{i-1} A C G T

A C G T

X_i



$x = \text{GATC}$

$P(x) = P(x_4 \mid x_3)\, P(x_3 \mid x_2)\, P(x_2 \mid x_1)\, P(x_1)$

$P(x) = P(\text{C} \mid \text{T})\, P(\text{T} \mid \text{A})\, P(\text{A} \mid \text{G})\, P(\text{G})$

$= 0.347 * 0.139 * 0.174 * 0.25$

$= 0.0021$