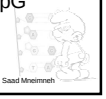
Computational Biology

Lecture 9

Saad Mneimneh

CG islands

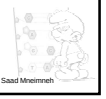
- Whenever CG occurs (CpG to distinguish it from C-G pair across the two strands), there is a high chance that C mutates to T by methylation.
- The CpG dinucleotide is rarer than would be expected by the independent probabilities of C and G.
- The methylation process is suppressed in areas around genes. Such regions are called CpG islands (100-1000 bases long).



Saad Mneimneh

Questions

- Q1: Given a short sequence, is it from a CpG island or not?
- Q2: Given a long sequence, does it contain a CpG island or not?
- Start with Q1.



Saad Mneimneh

Modeling Dinucleotides

- **Di**nucleotides are important here. We need to model them in a sequence.
- Build a model that generates sequences in which the probability of a symbol depends on the previous symbol (why?).
- Markov Chain!

Saad Mneimneh

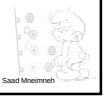
Markov Chain

A Markov Chain is defined as:

– A set of states

– For each pair of states i and j, a transition probability a_{ij}.

– $\Sigma_j a_{ij} = 1$

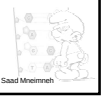


Saad Mneimneh

Markov Chain

- We transition from one state to another in discrete time steps $n = 1, 2, 3, \ldots$
- If we are at state i in time step n, we go to state j in time step $n+1$ with probability a_{ij}.
- The state at time n, x_n, depends on the states $x_{n-1}, x_{n-2}, \ldots$ only through the most recent state x_{n-1}.

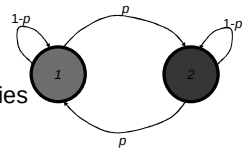
$$p(x_n = j \mid x_0, x_1, \ldots, x_{n-2}, x_{n-1} = i) = p(x_n = j \mid x_{n-1} = i) = a_{ij}$$



Saad Mneimneh

Example

- States: 1, 2
- Transition probabilities
  - a_{11}=1-p
  - a_{12}=p
  - a_{21}=p
  - a_{22}=1-p



$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

$p(x_n = j \mid x_0 = i)$

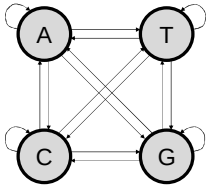
The probability of being in state j at time n given that state i is the starting state:

$p(x_n = j \mid x_0 = i) = P^n_{ij}$

Modeling sequences

We will model
dinucleotides
using a Markov Chain

states = DNA symbols

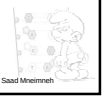


Probability of a path

(or a sequence)

The probability of a given sequence of states $x_1 \ldots x_n$ is:

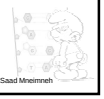
$p(x_1 \ldots x_n) =$
$p(x_1 \ldots x_{n-1}, x_n) =$
$p(x_n, x_1 \ldots x_{n-1}) =$
$\boldsymbol{p(x_n \mid x_1 \ldots x_{n-1})} p(x_1 \ldots x_{n-1}) =$
$\boldsymbol{p(x_n \mid x_{n-1})} p(x_1 \ldots x_{n-1}) =$
$\boldsymbol{a}_{x_{n-1} x_n} p(x_1 \ldots x_{n-1}) =$
$p(x_1) \Pi_{i=2 \ldots n} \boldsymbol{a}_{x_{i-1} x_i}$



Saad Mneimneh

Modeling the beginning and end of the sequence

- $p(x_1 \ldots x_n) = p(x_1) \Pi_{i=2 \ldots n} \boldsymbol{a}_{x_{i-1} x_i}$
- what is $p(x_1)$? Depends on how we start.
- Add a distinct start state $x_0 = S$.

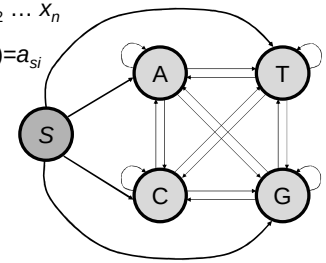


Saad Mneimneh

Adding a start state

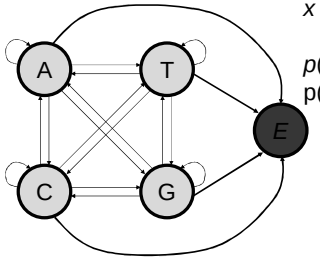
$x = x_1 x_2 \ldots x_n$

$p(x_1 = i) = a_{si}$



Saad Mneimneh

Adding an end state

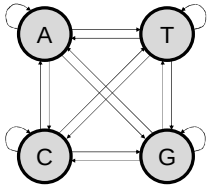


$x = x_1 x_2 \ldots x_n$

$p(x_n = i) =$
$p(E \mid x_n = i) = a_{iE}$

Modeling sequences

- Build the model
- Obtain transitional probabilities
  from statistical data
  - For CpG islands
  - For non CpG islands



CpG island v.s. non CpG island

- Bring a set of DNA sequences labeled '+' for CpG island and '-' for non CpG island
- In each case, estimate a_{ij}^+ and a_{ij}^-
- For instance, $a_{ij}^+ = c_{ij}^+ / \Sigma_k c_{ik}^+$ where c_{ij}^+ = number of times j followed i in regions labeled '+'
- Use log-odds ratio to decide whether a sequence came from a CG-island or not

Example

a_{ij}^{+}	A	C	G	T
A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.37	0.13
T	0.08	0.36	0.38	0.18

a_{ij}^{-}	A	C	G	T
A	0.30	0.20	0.29	0.21
C	0.32	0.30	0.08	0.30
G	0.25	0.25	0.29	0.21
T	0.18	0.24	0.29	0.29

$$S(x) = \log \frac{p(x|+)}{p(x|-)} = \log \frac{\prod_{i=1..n} a^{+}_{x_{i-1}x_i}}{\prod_{i=1..n} a^{-}_{x_{i-1}x_i}} = \sum_{i=1..n} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

Saad Mneimneh

Questions

- Done with Q1.
- Q2: Given a long sequence, does it contain a CpG island or not?
- How can we answer Q2?

Saad Mneimneh

Markov Chain model

- The Markov Chain model that we have just build can be used
- Calculate the log-odds score for windows of size, say 100, in the sequence
- CpG islands will stand out with positive values

Saad Mneimneh

Problems?

The previous approach is unsatisfactory

– CpG islands have variable length

– Why use a window of size 100? Why not 10 or 50 or 200? (no way to tell best size, could be average length of CpG island, but still unsatisfactory)

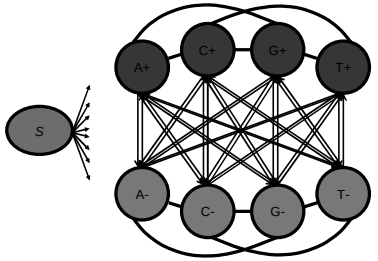
Saad Mneimneh

Better solution

- Represent CpG islands and non CpG islands in one model
- Both Markov chains build earlier put together with small transition probability between them
- We will have two states for each nucleotides ➔ rename them A+, C+, G+, T+ and A-, C-, G-, T-

Saad Mneimneh

New Model



Saad Mneimneh

What is the big difference now?

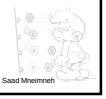
- There is not a one-to-one correspondence between the states and the symbols.
  - Given a symbol C, it could have been generated by state C+ or state C-
- Before, a sequence uniquely determines the path
- Now, for a given sequence, we want to find the most likely path

Saad Mneimneh

Hidden Markov Model

A Hidden Markov model HMM is defined as:
[state is hidden, decouple states from symbols]

- A set of *hidden* states
- For each pair of states i and j, a transition probability a_{ij}.
- $\Sigma_j a_{ij} = 1$
- For each state k, emission probabilities
  $e_k(b) = p(x_i = b \mid \pi_i = k)$
  [now we use variable π for states and variable x for symbols]
- $\Sigma_b e_k(b) = 1$ for each state k
- Markov property: $p(\pi_n = j \mid x_0 \ldots x_{n-1}, \pi_0 \ldots \pi_{n-2}, \pi_{n-1} = i) = p(\pi_n = j \mid \pi_{n-1} = i) = a_{ij}$



Saad Mneimneh

Questions with HMMs

- *Evaluation*: given x, what is the probability $p(x)$ that it was produced by the model?
- *Decoding*: given x, what is the most probable path that produces x in the model?
- *Learning*: given x, what are the most probable parameters (transitional probabilities and emission probabilities) of the model?

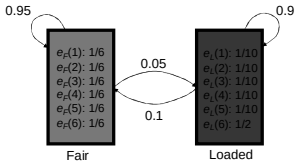
Saad Mneimneh

The dishonest Casino

- Casino uses a fair die most of the time
- Loaded die has p(1) = p(2) = p(3) = p(4) = p(5) = 0.1, p(6) = 0.5
- Casino switches from fair to loaded with a probability of 0.05
- Switches back with probability 0.1

[think about similarities with CpG island]

HMM for dishonest casino



We just see the sequence of rolls $x = x_1 \ldots x_n$. We do not know the path $\pi = \pi_1 \ldots \pi_n$ that generated the sequence x. This is why the state is hidden.

We need to find the most probable path π.

Most probable path

- Label start state and end state by 0.
- The joint probability of observing a sequence of symbols $x = x_1 \ldots x_n$ emitted by a sequence of states $\pi = \pi_1 \ldots \pi_n$:

  $p(x,\pi) = a_{0\pi 1}.e_{\pi 1}(x_1) \ldots a_{\pi n-1\ \pi n} e_{\pi n}(x_n).a_{\pi n 0}$

  $= a_{0\pi 1} \Pi_{i=1..n}\ e_{\pi i}(x_i) a_{\pi i\ \pi i+1}$

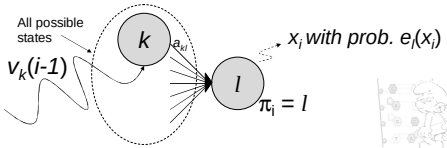
  where $\pi_{n+1} = 0$
- We want to find $\pi^* = \text{argmax}_\pi\ p(\pi \mid x) = \text{argmax}_\pi\ p(x,\pi)$

Why?

- Try all possible π: EXPONENTIAL!

DP… again ☹

- $v_l(i)$ = probability of most probable path $\pi = \pi_1 \ldots \pi_i$ for $x_1 \ldots x_i$ ending in state l ($\pi_i = l$)
- $v_l(i) = e_l(x_i).\max_k (v_k(i-1).a_{kl})$



Saad Mneimneh

Derivation

$v_l(i)$

$= \max_{\pi_1 \ldots \pi_{i-1}} p(x_1 \ldots x_i, \pi_1 \ldots \pi_{i-1}, \pi_i = l)$

$= \max_{\pi_1 \ldots \pi_{i-1}} p(x_i, \pi_i = l, x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-1})$

$= \max_{\pi_1 \ldots \pi_{i-1}} p(x_i, \pi_i = l \mid x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-1}).p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-1})$

$= \max_{\pi_1 \ldots \pi_{i-1}} p(x_i, \pi_i = l \mid \pi_{i-1}).p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-1})$

$= \max_{\pi_1 \ldots \pi_{i-2}, k} p(x_i, \pi_i = l \mid \pi_{i-1} = k).p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-2}, \pi_{i-1} = k)$

$= \max_{\pi_1 \ldots \pi_{i-2}, k} e_l(x_i) a_{kl} p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-2}, \pi_{i-1} = k)$

$= \max_k \max_{\pi_1 \ldots \pi_{i-2}} e_l(x_i) a_{kl} p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-2}, \pi_{i-1} = k)$

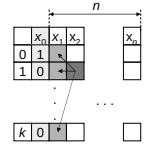
$= \max_k \max_{\pi_1 \ldots \pi_{i-2}} p(x_1 \ldots x_{i-1}, \pi_1 \ldots \pi_{i-2}, \pi_{i-1} = k)\ e_l(x_i) a_{kl}$

$= \max_k\ v_k(i\text{-}1) e_l(x_i) a_{kl} = e_l(x_i) \max_k\ v_k(i\text{-}1) a_{kl}$

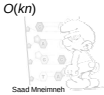
Saad Mneimneh

Viterbi *decoding* algorithm

- Initialization
  $v_0(0) = 1$, $v_k(0) = 0$ for $k > 0$
- Main iteration
  for $i = 1 \ldots n$
  $v_l(i) = e_l(x_i).\max_k (v_k(i-1).a_{kl})$
  $\text{ptr}_l(i) = \text{argmax}_k(v_k(i-1)a_{kl})$
- Termination
  $p(x,\pi^*) = \max_k(v_k(n)a_{k0})$



Time = $O(k^2 n)$
Space = $O(kn)$



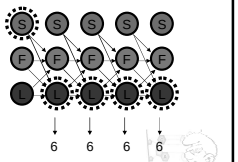
Saad Mneimneh

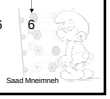
Example

What is the most probable path for the following outcome of the casino?

6666

		6	6	6	6
S	1	0	0	0	0
F	0	.0833	.0132	.0021	.0008
L	0	.25	.1125	.0506	.0228





Saad Mneimneh