

Bioinformatics II

Hidden Markov Models

Part 2

Dr. Konrad Schlude
schlude@inf.ethz.ch

Institute of Theoretical Computer Science
ETH Zürich

Contents

1. Last week
2. CpG Islands
3. Most probable state: Backward algorithm
4. HMM parameter estimation
5. Conclusion

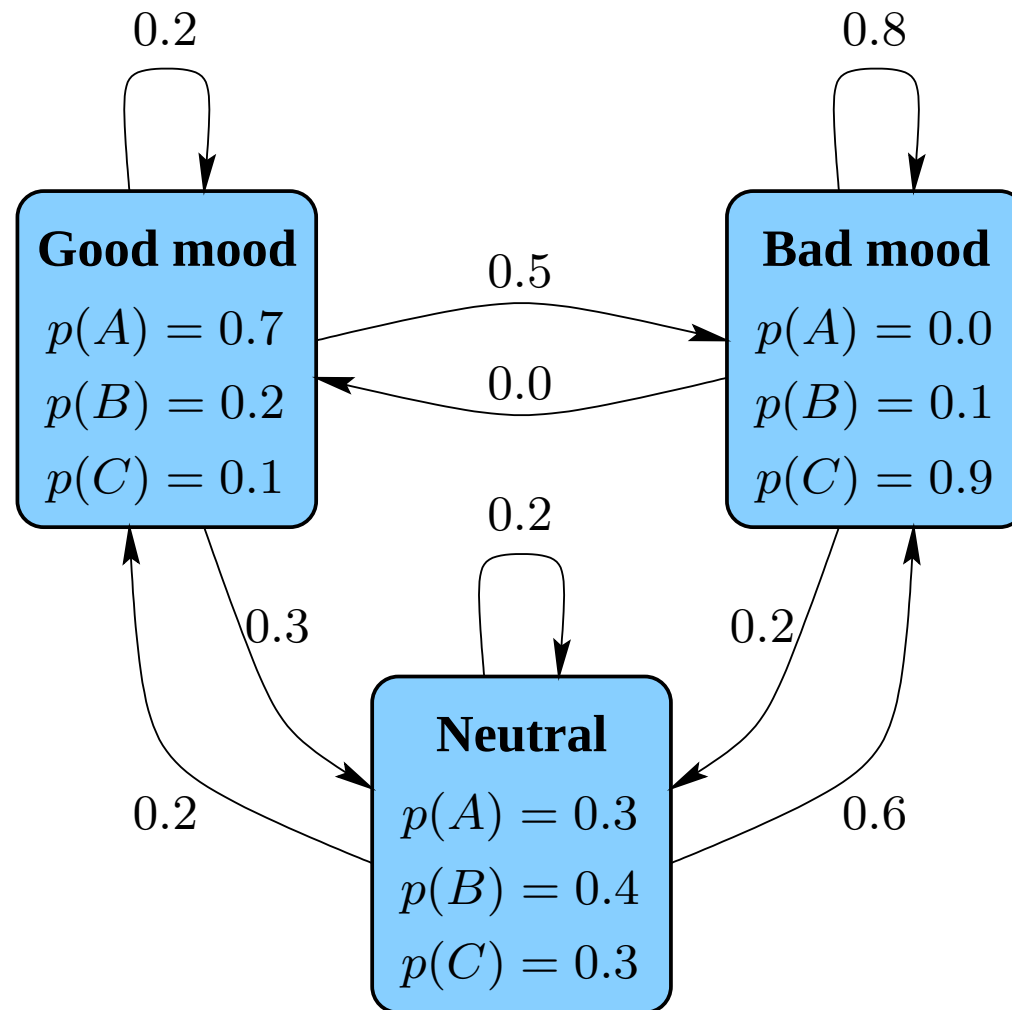
Teacher

Situation:

- Your school teacher gave three different types of daily homework assignments:
 - A: took about 5 minutes to complete
 - B: took about 1 hour to complete
 - C: took about 3 hours to complete
- Mood state was not known: Good, neutral, or bad.
- Mood changes occurred only overnight.

Question: How were his moods related to the homework type assigned that day?

Teacher



Questions Teacher

One week, your teacher gave the following homework assignments:

Monday:	Tuesday:	Wednesday:	Thursday:	Friday:
A	C	B	A	C

Questions:

- What did his mood curve look like most likely that week?
- What is the probability that he would assign this order of homework assignments?
- What is the probability that he was in a good mood on Thursday?

Hidden Markov Model

Definition

- States $S = \{s_1, s_2, \dots, s_m\}$
- a_{kl} = probability that state l follows state k (Markov chain)
- Alphabet Σ
- $e_k(x)$ = probability that state k emits symbol $x \in \Sigma$

Questions:

Most probable state path: Viterbi algorithm

Probability of emission: Forward algorithm

Most probable state: Backward algorithm

Complexity

Number of paths: $|S|^n$

Viterbi, Forward, Backward algorithm:

- Space: $O(|S|n)$
- Time: $O(|S|^2n)$

Example: $|S| = 4, n = 100$

- $|S|n = 400$
- $|S|^2n = 1600$
- $|S|^n \approx 1.61 * 10^{60}$

Contents

1. Last week
2. CpG Islands
3. Most probable state: Backward algorithm
4. HMM parameter estimation
5. Conclusion

CpG Islands in human genome

Wherever the dinucleotide CpG occurs (not the C–G base pair across the two strands) in the human genome, the cytosine C is typically chemically modified by methylation and there is a high chance that this methyl–C mutates into a T.

CpG dinucleotides are rarer in the genome than could be expected from the independent probabilities of C and G. However, the methylation process is suppressed around promoter regions of genes. CpG dinucleotides occur at a higher rate around promoter regions than elsewhere. These regions are called *CpG islands*.

CpG Islands in human genome

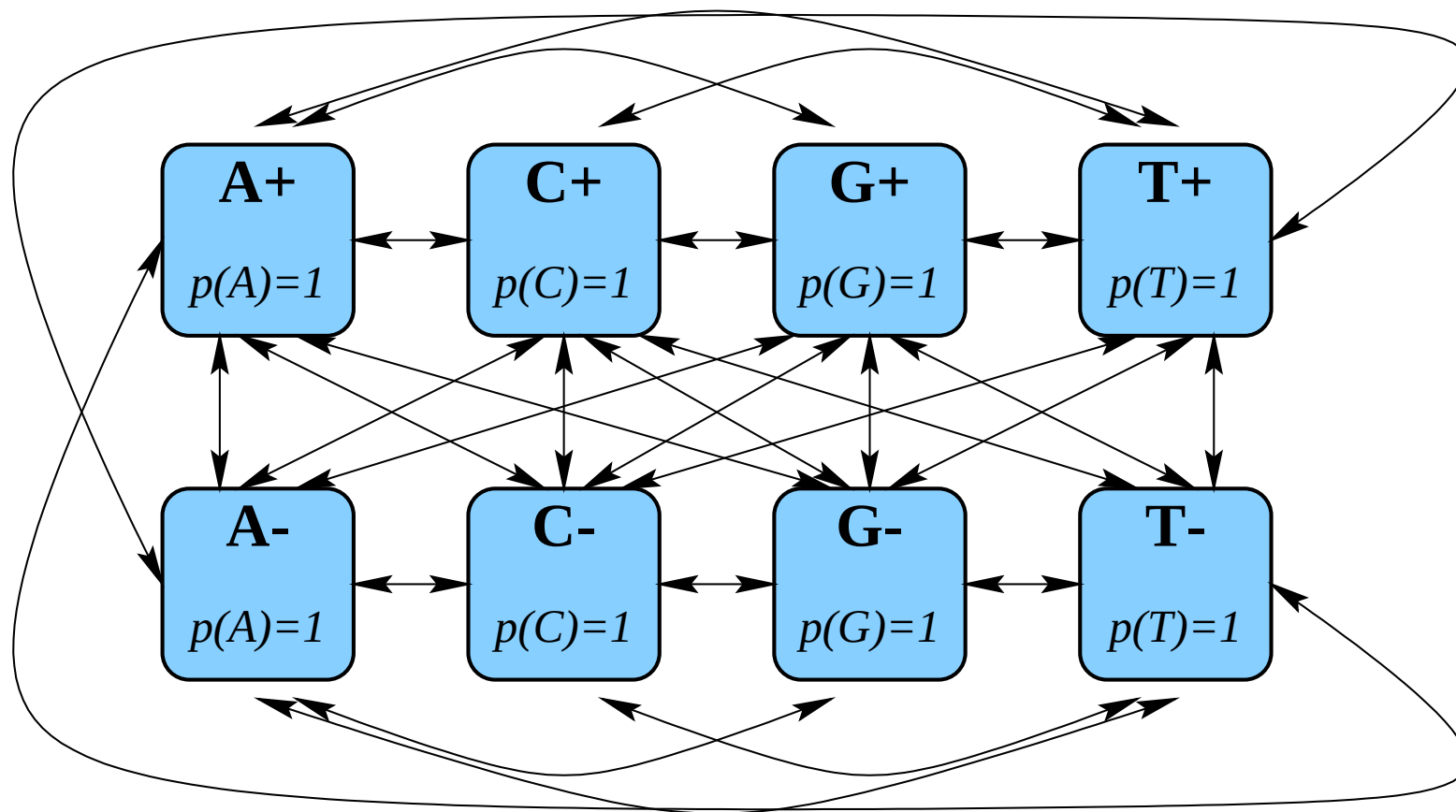
Question:

Given a long sequence, how can we find the CpG islands in it?

States $\Sigma = \{\underbrace{A+, C+, G+, T+}_{\text{CpG islands}}, A-, C-, G-, T-\}$.

Markov Chain CpG islands

+ $\equiv$ island



Transition Probabilities CpG islands

	A+	C+	G+	T+	A-	C-	G-	T-
A+	.16	.25	.41	.10	.02	.02	.02	.02
C+	.15	.35	.25	.17	.02	.02	.02	.02
G+	.14	.32	.36	.10	.02	.02	.02	.02
T+	.06	.34	.36	.16	.02	.02	.02	.02
A-	.01	.01	.01	.01	.29	.20	.27	.20
C-	.01	.01	.01	.01	.31	.29	.07	.29
G-	.01	.01	.01	.01	.24	.23	.29	.20
T-	.01	.01	.01	.01	.17	.23	.28	.28

Example CpG islands

Emitted symbol sequence: T A C G C G

State path: + - - + + +

Transition probability: .13 .02 .20 .01 .36 .25

Emission probability: 1 1 1 1 1 1

Overall probability = 0.00004

From a given symbol sequence and a corresponding state path, we can compute the probability of the path by multiplying the probability of the state changes.

Contents

1. Last week
2. CpG Islands
3. Most probable state: Backward algorithm
4. HMM parameter estimation
5. Conclusion

Most probable state

Question:

What is the probability that the teacher was in a good mood on Thursday?

Type of homework: A

Without further informations:

$$\Pr(\text{good}|\text{A}) = \frac{0.05 * 0.7}{0.05 * 0.7 + 0.20 * 0.3 + 0.75 * 0.0} = 0.368$$

But, we know the whole sequence $E = x_1, \dots, x_n$.

Most probable state

Given:

- Hidden Markov model: $S, a_{kl}, \Sigma, e_l(x)$.
- Observed symbol sequence $E = x_1, \dots, x_n$.

Questions:

Probability $\Pr(s_i = l|E)$ that observation x_i came from state l ?

State $s_i \in S$ that maximizes $\Pr(s_i = l|E)$?

Observations

$$\Pr(s_i = l|E) = \frac{\Pr(s_i = l, E)}{\Pr(E)}$$

$$\Pr(E, s_i = l) =$$

$$\Pr(x_1, x_2, \dots, x_i, s_i = l) \Pr(x_{i+1}, \dots, x_n | x_1, \dots, x_i, s_i = l) =$$

$$\underbrace{\Pr(x_1, x_2, \dots, x_i, s_i = l)}_{f_l(i)} \underbrace{\Pr(x_{i+1}, \dots, x_n | s_i = l)}_{b_l(i)}$$

Simple Algorithm

Step 1, Forward algorithm

	A	C	B	A	C
good	0.23	0.0066	.001332	.0023604	.0000857
neutral	0.1	0.0267	.015528	.0019235	.0003278
bad	0.0	0.1575	.014532	0	.0021009

Σ .0025144

Simple Algorithm

Step 2, initialization:

	A	C	B	A	C
good				1	
neutral				0	
bad				0	

Simple Algorithm

Step 2, Forward algorithm

	A	C	B	A	C
good				1	0.02
neutral				0	0.09
bad				0	0.45

Simple Algorithm

Step 2, sum of entries in last column.

	A	C	B	A	C
good				1	0.02
neutral				0	0.09
bad				0	0.45

Σ 0.56

Most probable state

What is the probability that the teacher was in a good mood on Thursday?

$$\begin{aligned}\Pr(\text{good}|E) &= \frac{\Pr(s_i = l, E)}{\Pr(E)} = \frac{f_l(i) * b_l(i)}{\Pr(E)} \\ &= \frac{0.0023604 * 0.56}{0.0025144} = 0.526\end{aligned}$$

Problem:

l, i are assumed to be constant.

Backward Algorithm

$b_l(i)$ has the following property:

$$b_l(i) = \sum_k a_{lk} * e_k(x_{i+1}) * b_k(i+1) \quad i < n$$

Backward Algorithm:

- Initialization: $b_l(n) = 1$ for all states $l \in S$.
- $b_l(i) = \sum_k a_{lk} * e_k(x_{i+1}) * b_k(i+1)$ for all $l \in S, i < n$.

Backward Computation

Initialization: $b_l(n) = 1$

	A	C	B	A	C
good					1
neutral					1
bad					1

Backward Computation

Backward Iteration: $b_l(i - 1) = \sum_k a_{lk} * e_k(x_i) * b_k(i)$

	A	C	B	A	C
good	.00743912	.021100	0.1342	0.56	1
neutral	.00803384	.016848	0.1156	0.62	1
bad	.00981216	.012224	0.0372	0.78	1

Most probable state on Wednesday

The type of homework was B.

$$\Pr(\text{good}|E) = \frac{0.00133 * 0.1342}{0.0025144} = 0.071$$

$$\Pr(\text{neutral}|E) = \frac{0.015528 * 0.1156}{0.0025144} = 0.714$$

$$\Pr(\text{bad}|E) = \frac{0.014532 * 0.0372}{0.0025144} = 0.215$$

Contents

1. Last week
2. CpG Islands
3. Most probable state: Backward algorithm
4. HMM parameter estimation
5. Conclusion

Parameter Estimation

Question:

- How do we know the transition probabilities a_{kl} and the emission probabilities $e_k(x)$?

Answer:

- We use training sequences and construct maximum likelihood estimators.

Parameter Estimation

Case 1: State sequences of training sequences are known.

We construct maximum likelihood estimators.

Let A_{kl} be the number of transitions from state k into state l in training data (+ constant).

Let $E_k(x)$ be the number of emissions of symbol x from state k in training data (+ constant).

Estimators:

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad e_k(x) = \frac{E_k(x)}{\sum_{x'} E_k(x')}$$

Parameter Estimation

Case 2: State sequences of training sequences are not known.

Let $E^1, \dots, E^\nu$ be training sequences

$$\Pr(s_i = k, s_{i+1} = l | E^j) = \frac{f_k^j(i) * a_{kl} * e_l(x_i^j) * b_l^j(i+1)}{\Pr(E^j)}$$

Expected number of times a_{kl} is used:

$$A_{kl} = \sum_j \frac{1}{\Pr(E^j)} \sum_i f_k^j(i) * a_{kl} * e_l(x_i^j) * b_l^j(i+1)$$

Parameter Estimation

Similarly

$$E_k(x) = \sum_j \frac{1}{\Pr(E^j)} \sum_{i, x_i^j = x} f_k^j(i) b_k^j(i)$$

Baum-Welch algorithm:

Calculate A_{kl} and $E_k(x)$ as the expected number of times each transition or emission is used, given the training sequences.

Baum-Welch algorithm

Algorithm sketch:

Initialization: Pick arbitrary model parameters.

REPEAT

Set all A_{kl} and $E_k(x)$ to their constant value (zero)

FOR each training sequence E^j

Compute $f_k(i)$ for E^j (Forward algorithm)

Compute $b_k(i)$ for E^j (Backward algorithm)

Add contribution to A_{kl} and $E_k(x)$.

Update $a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$ and $e_k(x) = \frac{E_k(x)}{\sum_{x'} E_k(x')}$.

UNTIL stopping criterion is reached.

Contents

1. Last week
2. CpG Islands
3. Most probable state: Backward algorithm
4. HMM parameter estimation
5. Conclusion

Further reading and reference

A. Krogh:

An Introduction to Hidden Markov Models for Biological Sequences

In *Computational Methods in Molecular Biology*, edited by S.L. Salzberg, D.B. Searls and S. Kasif, pages 45-63. Elsevier, 1998.

R. Durbin , S.R. Eddy, A. Krogh, G. Mitchison:

Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids

Cambridge Univ. Pr., 1999