Beschreibung von DNA-Sequenzen als Markov-Ketten

Eine Einführung

Uwe Menzel
Rudbeck Laboratory,
Uppsala University

Inhalt

1) Markov-Ketten für CpG-Islands
2) *Hidden Markov Models* (HMM) für CpG-Islands (Ausblick)

DNA-Sequenz

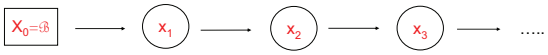
Die *Reihenfolge* der Basen (A,T,G,C) im DNA[1]-Molekül bestimmt den "Bauplan" eines Organismus.

[1]Desoxyribonukleinsäure

Markov-Ketten

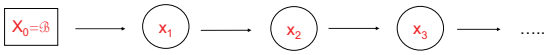
Modellierung von Sequenzen mit Hilfe von stochastischen Modellen – man nimmt an, dass die Sequenz durch einen Zufallsprozess "erzeugt" worden ist.

X: Stochastische Sequenz



- zeitliche oder örtliche Folge (Sprache, DNA)
- Durchnummerieren: x_i
- An jeder Stelle x_i kann die Sequenz einen von mehreren möglichen Werten annehmen = Alphabet
- DNA: x_i = {A,C,G,T}
- (A. Markov untersuchte die russische Literatur …)

P(X): Wahrscheinlichkeit der Sequenz

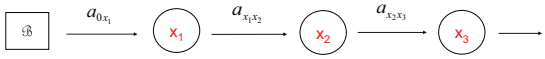


- Gesucht: Wahrscheinlichkeit des Auftauchens einer *bestimmten* Sequenz (Buchstabenfolge)
- $P(X) = P(X_1=x_1, X_2=x_2, X_3=x_3, \ldots) = P(x_1, x_2, x_3, \ldots)$
- DNA: P(C,C,C,C,C) oder P(C,G,C,G,C,G)

Beispiel: Stochastische Sequenz, aber keine Markov-Kette

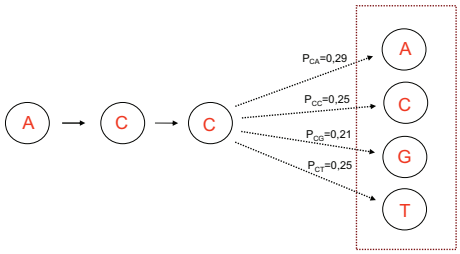
- Würfel: X=(3, 4, 2)
- P(*X*) = P(X_1=3, X_2=4, X_3=2) = P(3, 4, 2)
- P(3, 4, 2) = P(3) · P(4) · P(2) = 1/6 · 1/6 · 1/6
- Ereignisse unabhängig ⇨ keine Markov-Kette

Markov- Kette



- Welcher der Werte (des Alphabets) bei x_i vorliegt, hängt vom Wert beim unmittelbaren Vorgänger x_{i-1} ab,
- … jedoch ***nicht*** von noch weiter "davor" liegenden Werten x_{i-2}, x_{i-3}, …
- ***Markov-Eigenschaft*** - Kopplung (mit dem Nachbarn)

Markov-Kette für DNA



?

Für dieses Beispiel: Hat man viele Kettenglieder, so wird die Folge "CA" häufiger darin enthalten sein als "CG".

Wahrscheinlichkeit einer Konfiguration der Kette

$$P(\vec{x}) = P(x_1, x_2, x_3, \ldots, x_{N-1}, x_N)$$

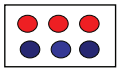
Durch multiples Anwenden von : $P(x,y) = P(x \mid y) \cdot P(y)$

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2, x_1) \cdot P(x_4 \mid x_3, x_2, x_1) \cdot \ldots$$

Bei Markov - Ketten vereinfacht sich dies :

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2) \cdot P(x_4 \mid x_3) \cdot \ldots \cdot P(x_N \mid x_{N-1})$$

Non-Markov Process



Ziehe "blind" drei Kugeln aus dem Behälter … wie groß ist die Wahrscheinlichkeit, drei rote Kugeln zu ziehen ?

$$P(rot; rot; rot) = P(x_1 = rot, x_2 = rot, x_3 = rot)$$

$$P(\vec{x}) = P(x_1 = rot) \cdot P(x_2 = rot \mid x_1 = rot) \cdot P(x_3 = rot \mid x_2 = rot, x_1 = rot) \quad = \tfrac{1}{2} \cdot \tfrac{2}{5} \cdot \tfrac{1}{4}$$

Vorhandene Kugeln vor der Ziehung	P, "rot" zu ziehen
● ● ● ● ● ●	P(x_1=rot) = ½ = 0,5
● ● ● ● ●	P(x_2=rot | x_1=rot) = 2/5 = 0,4
● ● ● ●	P(x_3=rot | x_1=rot; x_2=rot) = ¼ = 0,25

Dagegen:

● ● ● ●

P(x_3=rot | x_1=blau; x_2=rot) = ½ = 0,5

"Wahrscheinlichkeit" der Markov-Kette[1]

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2) \cdot P(x_4 \mid x_3) \cdot \ldots \cdot P(x_N \mid x_{N-1})$$

$$sei \quad a_{x_{i-1}x_i} = P(x_i \mid x_{i-1})$$

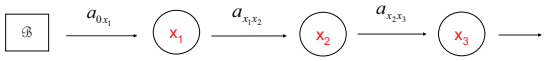
$$P(\vec{x}) = P(x_1) \cdot a_{x_1x_2} \cdot a_{x_2x_3} \cdot \ \ldots \ \cdot a_{x_{N-2}x_{N-1}} \cdot a_{x_{N-1}x_N}$$

$$P(\vec{x}) = P(x_1) \cdot \prod_{i=2}^{N} a_{x_{i-1}x_i} \quad mit \quad P(x_1) = a_{x_0x_1}$$

$$P(\vec{x}) = \prod_{i=1}^{N} a_{x_{i-1}x_i}$$

[1]Wir betrachten nur homogene Markov-Ketten

Wahrscheinlichkeit der Markov-Kette folgt aus den Übergangswahrscheinlichkeiten



$$P(x) = a_{0x_1} \cdot a_{x_1x_2} \cdot a_{x_2x_3} \cdot \quad \ldots \quad = \prod_{i=0}^{L} a_{x_i x_{i+1}}$$

Übergangswahrscheinlichkeiten

$P(x_i \mid x_{i-1})$	A	C	G	T
A	0,300	0,205	0,285	0,210
C	0,322	0,298	0,078	0,302
G	0,248	0,246	0,298	0,208
T	0,177	0,239	0,292	0,292

Stochastische Matrix

$$P(C,A,A,G) = a_{0C} \cdot a_{CA} \cdot a_{AA} \cdot a_{AG}$$

$$= 0{,}25 \cdot 0{,}322 \cdot 0{,}300 \cdot 0{,}285 = 0{,}00688$$

Übergangswahrscheinlichkeiten

- Ausgehend von genomischen Daten werden die Häufigkeiten der 16 möglichen Dinukleotide gezählt.
- Diese Häufigkeiten werden zu Wahrscheinlichkeiten normiert (a = Wahrscheinlichkeiten; c = "counts"):

$$a_{st} = \frac{c_{st}}{\sum_i c_{si}} \qquad s,t \in \{A,C,G,T\}$$

Übergangswahrscheinlichkeiten *für Di-Nukleotide*

$$c_{CG} = 100 \qquad c_{CA} = 150 \qquad c_{CT} = 50 \qquad c_{CC} = 100$$

$$a_{CG} = \frac{c_{CG}}{c_{CG} + c_{CA} + c_{CT} + c_{CC}} = \frac{100}{100 + 150 + 50 + 100} = 0{,}25$$

$$a_{CA} = \frac{150}{400} = 0{,}375$$

$$a_{CT} = \frac{50}{400} = 0{,}125$$

$$a_{CC} = \frac{100}{400} = 0{,}25$$

Stochastische Matrix

$$a_{CG} + a_{CA} + a_{CT} + a_{CC} = 1 \quad \textit{Zeilensumme}$$

CpG - Dinukleotide

- Statistisch gesehen: Häufigkeit ca. 4 - 6 %
  - P(C,G) ≈ ¼ · ¼ = 1/16 ; genauer 0,21 · 0,21 ≈ 4,4%
- Tatsächliche Häufigkeit: ≈ 0,8 % (Mammalia)
- Cytosin (C) ist chemisch instabil:
  - Methylierung, Desaminierung: CG → $C^{meth}G$ → TG
- In CpG-Islands[1] ist die Häufigkeit von CG-Dinukleotiden jedoch deutlich höher als im Rest des Genoms

[1]Das "p" in CpG bezieht sich auf die Binding zwischen Cytosin und Guanin (engl. phosphodiester)

Markov-Modell für CpG-Islands

- Da es sich um Dinukleotide handelt, also die Häufigkeit von CG (in dieser Reihefolge!), bietet sich ein Markov-Modell geradezu an !
- Wie häufig ist die Folge CG im Vergleich zu den anderen Dinukleotiden CA, CT, GT, usw …. ?
- Wie häufig ist die Folge ”CG” in CpG-Islands im Vergleich zur Häufigkeit dieser Folge in Nicht-Islands ?

”Training”

CpG-Islands

	A	C	G	T
A	0,180	0,274	0,426	0,120
C	0,171	0,368	**0,274**	0,188
G	0,161	0,339	0,375	0,125
T	0,079	0,355	0,384	0,182

Non-Islands

	A	C	G	T
A	0,300	0,205	0,285	0,210
C	0,322	0,298	**0,078**	0,302
G	0,248	0,246	0,298	0,208
T	0,177	0,239	0,292	0,292

$a^{+}_{CG} = 0{,}274$ $\qquad$ $a^{-}_{CG} = 0{,}078$

Unterscheidung: Insel - ”Rest”

X = (ATCGCGCGGC)

$$P(X \mid \mathrm{mod}+) = \prod_i a^{+}_{x_{i-1}x_i} = a^{+}_{0A} \cdot a^{+}_{AT} \cdot a^{+}_{TC} \cdot a^{+}_{CG} \cdot a^{+}_{GC} \cdot a^{+}_{CG} \cdot a^{+}_{GC} \cdot a^{+}_{CG} \cdot a^{+}_{GG} \cdot a^{+}_{GC}$$

$$= 0{,}25 \cdot 0{,}120 \cdot 0{,}355 \cdot 0{,}274 \cdot 0{,}339 \cdot 0{,}274 \cdot 0{,}339 \cdot 0{,}274 \cdot 0{,}375 \cdot 0{,}339 = 3{,}125 \cdot 10^{-6}$$

$$P(X \mid \mathrm{mod}-) = \prod_i a^{-}_{x_{i-1}x_i} = a^{-}_{0A} \cdot a^{-}_{AT} \cdot a^{-}_{TC} \cdot a^{-}_{CG} \cdot a^{-}_{GC} \cdot a^{-}_{CG} \cdot a^{-}_{GC} \cdot a^{-}_{CG} \cdot a^{-}_{GG} \cdot a^{-}_{GC}$$

$$= 0{,}25 \cdot 0{,}210 \cdot 0{,}239 \cdot 0{,}078 \cdot 0{,}246 \cdot 0{,}078 \cdot 0{,}246 \cdot 0{,}078 \cdot 0{,}298 \cdot 0{,}246 = 2{,}65 \cdot 10^{-8}$$

Ergebnis: Offenbar ist es wahrscheinlicher, dass es sich bei X um ein CpG-Island handelt.

Log-Odds Ratio als Score für Diskriminanzanalyse

- Eine Sequenz X ist nach diesem Modell ein CpG-Island, wenn:

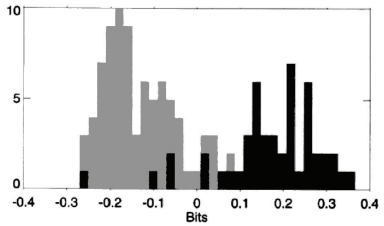
$$P(X \mid \mathrm{mod}+) > P(X \mid \mathrm{mod}-)$$

$$\frac{P(X \mid \mathrm{mod}+)}{P(X \mid \mathrm{mod}-)} > 1$$

$$S = \log\left[\frac{P(X \mid \mathrm{mod}+)}{P(X \mid \mathrm{mod}-)}\right] = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} > 0$$

”Probe”

- Wir berechnen den ”Score” S wieder für das Trainings-Set:



Abweichungen durch: Inkorrekte Labels im Trainings-Set, inkorrekte Bestimmung der Grenzen zwischen Island/Non-Island

Bildquelle: Durbin et al. (Ed): Biological Sequence Analysis, Cambridge University Press, 1998

Nachteile Scoring-Modell

- Lange Sequenz (BAC) ?
- Übergang zwischen CpG-Island und Umgebung ?

Finden von CpG-Inseln

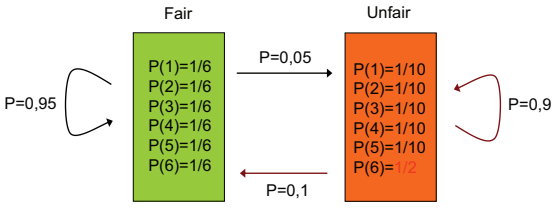
- Berechne Score S in jedem Fenster
- Probleme:
  - Laufzeit
  - Grösse der Insel nicht bekannt

Bildquelle: Sven Schuirer

Hidden Markov Model - HMM

- Detektion von CpG-Inseln in langen DNA-Sequenzen
- Auffinden der Grenze zwischen CpG-Insel und dem Rest-Genom
- Übergangswahrscheinlichkeiten a^+ und a^- in einem Modell

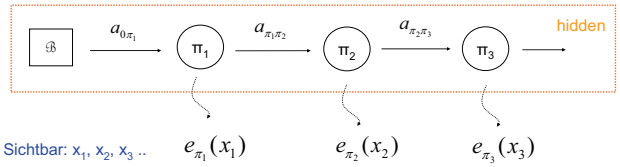
Beispiel: Casino mit 2 Würfeln



Gast sieht nur Augenzahl (Emissionen): 3 4 2 4 6 4 6 3 4 6 6 3 6 6 3 4 6 6

Der benutzte Würfel ist verborgen (state): F F F F F F F F U U U U U U U U

Hidden Markov Model



$$P(x,\pi) = a_{0\pi_1} \cdot \; e_{\pi_1}(x_1) \cdot a_{\pi_1\pi_2} \cdot \; e_{\pi_2}(x_2) \cdot a_{\pi_2\pi_3} \cdot \; \ldots$$

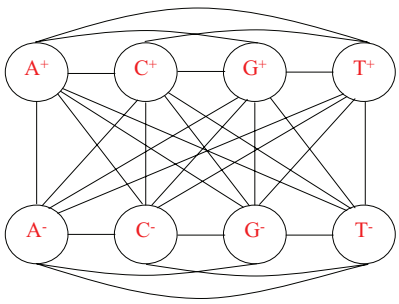
$$P(x,\pi) = a_{0\pi_1} \cdot \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i\pi_{i+1}}$$

HMM für CpG-Islands

- States: A^+, C^+, G^+, T^+, A^-, C^-, G^-, T^-
- Symbole: A, C, G, T

A^+	C^+	G^+	T^+	A^-	C^-	G^+	G^+	G^-	T^-
A	C	G	T	A	C	G	G	G	T

HMM für CpG Inseln in DNA-Sequenz



Bildquelle: Sven Schuirer

Übergangswahrscheinlichkeiten

π_i/π_{i+1}	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
A^+	0.180p	0.274p	0.426p	0.120p	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
C^+	0.171p	0.368p	0.274p	0.188p	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
G^+	0.161p	0.339p	0.375p	0.125p	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
T^+	0.079p	0.355p	0.384p	0.182p	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
A^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	0.300q	0.205q	0.285q	0.210q
C^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	0.322q	0.298q	**0.078q**	0.302q
G^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	0.248q	0.246q	0.298q	0.208q
T^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	0.177q	0.239q	0.292q	0.292q

Tabelle: Sven Schuirer

p = P (*bleibt in CpG-Insel*), q = *P (bleibt in Nicht-Insel)* $\approx 1\text{-}10^{-4}$

Emissionswahrscheinlichkeiten

$$e_{C^+}(C)=1;\ e_{C^-}(C)=1 \qquad e_{\pi_i}(C)=0\ \ \pi_i\ sonst$$

$$e_{A^+}(A)=1;\ e_{A^-}(A)=1 \qquad e_{\pi_i}(A)=0\ \ \pi_i\ sonst$$

$$e_{G^+}(G)=1;\ e_{G^-}(G)=1 \qquad e_{\pi_i}(G)=0\ \ \pi_i\ sonst$$

$$e_{T^+}(T)=1;\ e_{T^-}(T)=1 \qquad e_{\pi_i}(T)=0\ \ \pi_i\ sonst$$

Dekodierung

- Beobachtete Sequenz (”*emissions*”):
  - C G C G
- kann produziert werden von der ”*state*”-Sequenz:
  - $C^+ G^+ C^+ G^+$
  - $C^- G^- C^- G^-$
  - $C^+ G^- C^+ G^-$
  - … und vielen anderen …
- Welche ”*state*”-Sequenz ist am wahrscheinlichsten ?
  - Berechne:
  - P(X={ C, G, C, G }, π={ $C^+ G^+ C^+ G^+$ })
  - P(X={ C, G, C, G }, π={ $C^- G^- C^- G^-$ })
  - … und alle anderen

HMM: Übergänge

Aus der Tabelle Übergangswahrscheinlichkeiten:

$$p=0{,}999 \quad q=0{,}999$$

$$a_{C^+G^+}=0{,}274\cdot 0{,}999=0{,}2737$$

$$a_{G^+C^+}=0{,}339\cdot 0{,}999=0{,}3386$$

$$a_{C^-G^-}=0{,}078\cdot 0{,}999=0{,}0779$$

$$a_{G^-C^-}=0{,}246\cdot 0{,}999=0{,}2457$$

$$a_{C^+G^-}=(1-0{,}999)/4=0{,}00025 \quad klein$$

$$a_{G^-C^+}=(1-0{,}999)/4=0{,}00025 \quad klein$$

HMM:

$$P(x,\pi)=a_{0\pi_1}\cdot\prod_{i=1}^{L}e_{\pi_i}(x_i)\cdot a_{\pi_i\pi_{i+1}}$$

$$P(X=C,G,C,G;\ \ \pi=C^+,G^+,C^+,G^+)=$$
$$=a_{0C^+}\cdot e_{C^+}(C)\cdot\ a_{C^+G^+}\cdot e_{G^+}(G)\cdot\ a_{G^+C^+}\cdot\ e_{C^+}(C)\cdot\ a_{C^+G^+}\cdot\ e_{G^+}(G)\cdot a_{G^+0}$$
$$=0{,}5\cdot\ 1\cdot\ 0{,}2737\cdot\ 1\cdot\ 0{,}3386\cdot\ 1\cdot\ 0{,}2737\cdot\ 1\cdot\ 1=0{,}01268$$

$$P(X=C,G,C,G;\ \ \pi=C^-,G^-,C^-,G^-)=$$
$$=a_{0C^-}\cdot e_{C^-}(C)\cdot\ a_{C^-G^-}\cdot e_{G^-}(G)\cdot\ a_{G^-C^-}\cdot\ e_{C^-}(C)\cdot\ a_{C^-G^-}\cdot\ e_{G^-}(G)\cdot a_{G^-0}$$
$$=0{,}5\cdot\ 1\cdot\ 0{,}0779\cdot\ 1\cdot\ 0{,}2457\cdot\ 1\cdot\ 0{,}0779\cdot\ 1\cdot\ 1=0{,}000745$$

$$P(X=C,G,C,G;\ \ \pi=C^+,G^-,C^+,G^-)=$$
$$=a_{0C^+}\cdot e_{C^+}(C)\cdot\ a_{C^+G^-}\cdot\ e_{G^-}(G)\cdot\ a_{G^-C^+}\cdot\ e_{C^+}(C)\cdot\ a_{C^+G^-}\cdot\ e_{G^-}(G)\cdot\ a_{G^-0}$$
$$=0{,}5\cdot\ 1\cdot\ 0{,}00025\cdot\ 1\cdot\ 0{,}00025\cdot\ 1\cdot\ 0{,}00025\cdot\ 1\cdot\ 1=0{,}0000000000078125$$

Ergebnis: Wahrscheinlichster ”state path”

$$P(X=C,G,C,G;\ \ \pi=C^+,G^+,C^+,G^+)=0{,}01268$$

$$P(X=C,G,C,G;\ \ \pi=C^-,G^-,C^-,G^-)=0{,}000745$$

$$P(X=C,G,C,G;\ \ \pi=C^+,G^-,C^+,G^-)=0{,}0000000000078125$$

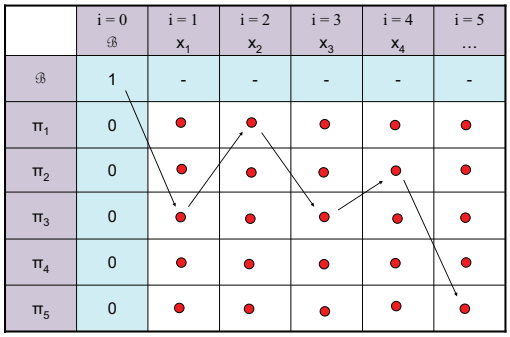
Es ist am wahrscheinlichsten, dass X=CGCG von einem CpG-Island stammt (unter diesen Dreien) – alle vier Symbole kommen von ”+” states.

Viterbi-Algorithmus: Berechnung des optimalen "state path"

– Natürlich kann man nicht alle möglichen Pfade durchrechnen (Anzahl = $|S|^N$)

→ Viterbi – Algorithmus ("dynamic programming")

Viterbi

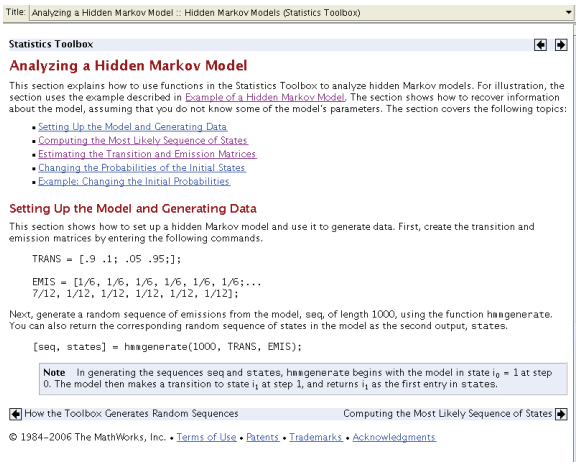


Literatur / Links

- Durbin et al (Ed.)., *Biological Sequence Analysis*, Cambridge University Press 1998
- A. Isaev, *Introduction to Mathematical Methods in Bioinformatics*, Universitext, ISBN 9783540219736
- Rabiner, L. R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, Vol. 77, No. 2, 1989
- http://www.stat.uni-muenchen.de/~semwiso/stochastische-prozesse/
- http://www.itu.dk/~sestoft/bsa.html

Software

- R – Scripte:
  - CRAN: Bioconductor
  - http://www.stat.uni-muenchen.de/~semwiso/stochastische-prozesse/ (*Ludwig Fahrmeir / Christiane Belitz*)
- MATLAB
  - stats package



Title: Analyzing a Hidden Markov Model :: Hidden Markov Models (Statistics Toolbox)

Statistics Toolbox

Analyzing a Hidden Markov Model

This section explains how to use functions in the Statistics Toolbox to analyze hidden Markov models. For illustration, the section uses the example described in Example of a Hidden Markov Model. The section shows how to recover information about the model, assuming that you do not know some of the model's parameters. The section covers the following topics:

- Setting Up the Model and Generating Data
- Computing the Most Likely Sequence of States
- Estimating the Transition and Emission Matrices
- Changing the Probabilities of the Initial States
- Example: Changing the Initial Probabilities

Setting Up the Model and Generating Data

This section shows how to set up a hidden Markov model and use it to generate data. First, create the transition and emission matrices by entering the following commands.

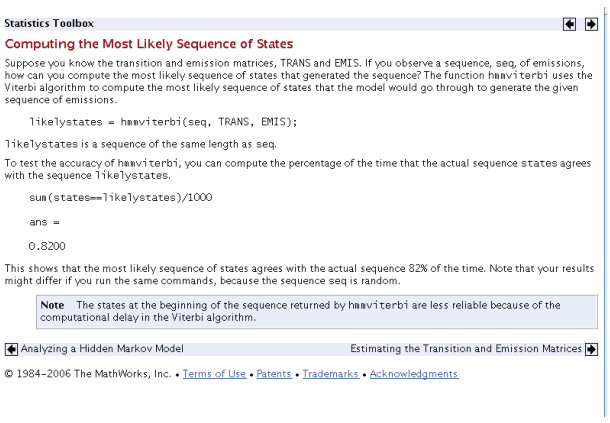
```
TRANS = [.9 .1; .05 .95];

EMIS = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6;...
7/12, 1/12, 1/12, 1/12, 1/12, 1/12];
```

Next, generate a random sequence of emissions from the model, seq, of length 1000, using the function hmmgenerate. You can also return the corresponding random sequence of states in the model as the second output, states.

```
[seq, states] = hmmgenerate(1000, TRANS, EMIS);
```

Note In generating the sequences seq and states, hmmgenerate begins with the model in state $i_0 = 1$ at step 0. The model then makes a transition to state i_1 at step 1, and returns i_1 as the first entry in states.

How the Toolbox Generates Random Sequences | Computing the Most Likely Sequence of States



Statistics Toolbox

Computing the Most Likely Sequence of States

Suppose you know the transition and emission matrices, TRANS and EMIS. If you observe a sequence, seq, of emissions, how can you compute the most likely sequence of states that generated the sequence? The function hmmviterbi uses the Viterbi algorithm to compute the most likely sequence of states that the model would go through to generate the given sequence of emissions.

```
likelystates = hmmviterbi(seq, TRANS, EMIS);
```

likelystates is a sequence of the same length as seq.

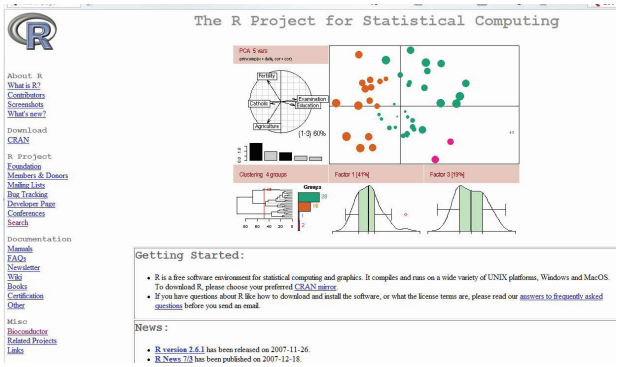
To test the accuracy of hmmviterbi, you can compute the percentage of the time that the actual sequence states agrees with the sequence likelystates.

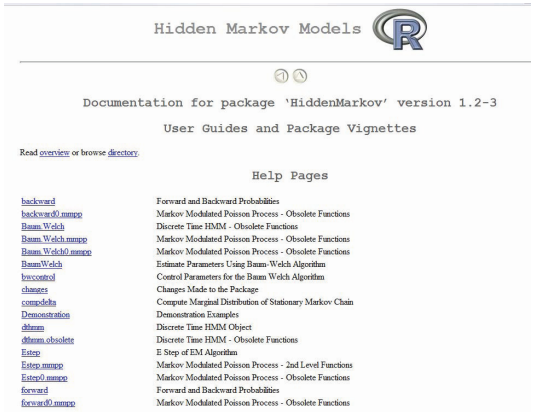
```
sum(states==likelystates)/1000

ans =

0.8200
```

This shows that the most likely sequence of states agrees with the actual sequence 82% of the time. Note that your results might differ if you run the same commands, because the sequence seq is random.

Note The states at the beginning of the sequence returned by hmmviterbi are less reliable because of the computational delay in the Viterbi algorithm.

Analyzing a Hidden Markov Model | Estimating the Transition and Emission Matrices





Vielen Dank für Ihre Aufmerksamkeit

http://puffer.genpat.uu.se/LECTURES/
uwe.menzel@genpat.uu.se

Anhang

Fragen

- Wieviele unterschiedliche DNA-Sequenzen der Länge L gibt es ?
- Ist das Würfeln mit einem (ehrlichen) Würfel ein Markow-Prozess ? $P(x_1=3, x_2=2, x_3=5)$ = ….
- Übergangsmatrix für den fairen/unfairen Würfel (S. 27) ?
- Ein HMM habe L Symbole und k Zustände: Wieviel mögliche Zustands-Pfade gibt es in diesem HMM ?
- Wie viele mögliche Zustandspfade gibt es für die Sequenz CGCG für das CpG-HMM (S. 29) ?
- Schwer: Unehrliches Casino: Erkläre die Bedeutung von
  - $P(X, \pi_3 = F)$
  - $P(X \mid \pi_3 = F)$
  - $P(\pi_3 = F \mid X)$ (a posteriori - Wahrscheinlichkeit)

”Aufspalten” von kombinierten Wahrscheinlichkeiten

$$P(x,y) = P(x \mid y) \cdot P(y)$$

$$P(A,B,C) = P(A \mid B,C) \cdot P(B,C)$$

$$P(\vec{x}) = P(x_N, x_{N-1}, x_{N-2}, \ldots, x_2, x_1) =$$

$$P(x_N \mid x_{N-1}, x_{N-2}, \ldots, x_2, x_1) \cdot P(x_{N-1}, x_{N-2}, \ldots, x_2, x_1) =$$

.... jetzt den letzten Faktor aufspalten ....

$$P(x_N \mid x_{N-1}, x_{N-2}, \ldots, x_2, x_1) \cdot P(x_{N-1} \mid x_{N-2}, \ldots, x_2, x_1) \cdot P(x_{N-2}, \ldots, x_2, x_1) =$$

.... jetzt wieder den letzten Faktor aufspalten ....

$$P(x_N \mid x_{N-1}, x_{N-2}, \ldots, x_2, x_1) \cdot P(x_{N-1} \mid x_{N-2}, \ldots, x_2, x_1) \cdot P(x_{N-2} \mid, x_{N-3} \ldots, x_2, x_1) \cdot \ldots \cdot P(x_2 \mid x_1) \cdot P(x_1)$$

Homogene Markov-Kette

Definition 2.1 (Homogeneous Markov chain) *A discrete-time stochastic process $\{X_k\}_{k\in\mathbb{N}}$ on a countable state space $\mathbf{S}$ is called a* ***homogeneous Markov chain****, if the so–called* ***Markov property***

$$\mathbb{P}[X_{k+1} = z|X_k = y, X_{k-1} = x_{k-1}, \ldots, X_0 = x_0] = \mathbb{P}[X_{k+1} = z|X_k = y] \quad (1)$$

holds for every $k \in \mathbb{N}$, $x_0, \ldots, x_{k-1}, y, z \in \mathbf{S}$, implicitly assuming that both sides of equation (1) are defined[1] *and, moreover, the right hand side of (1) does not depend on k, hence*

$$\mathbb{P}[X_{k+1} = z|X_k = y] \;=\; \ldots \;=\; \mathbb{P}[X_1 = z|X_0 = y]. \quad (2)$$

Wilhelm Huisinga, &
Eike Meerbach

Stochastische Matrix

Definition 2.2 *A matrix $P = (p_{xy})_{x,y\in\mathbf{S}}$ is called* ***stochastic****, if*

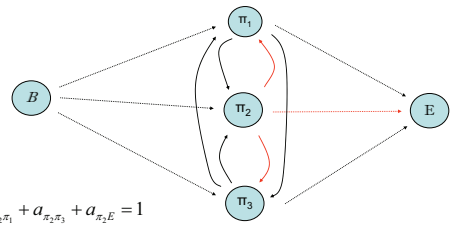
$$p_{xy} \geq 0, \;\; and \;\; \sum_{y\in\mathbf{S}} p_{xy} = 1 \quad (3)$$

for all $x, y \in \mathbf{S}$. Hence, all entries are non–negative and the row-sums are normalized to one.

Wilhelm Huisinga, &
Eike Meerbach

HMM - Grafik

Derselbe "state" kann wiederholt durchlaufen werden, daher zeichnen wir das Modell zweckmäßiger so:



$$a_{\pi_2\pi_1} + a_{\pi_2\pi_3} + a_{\pi_2 E} = 1$$

Entropie einer DNA-Sequenz

$$Sei \quad x_i \; ein \; Alphabet, z.B. \quad x_i = \{A, C, G, T\}$$

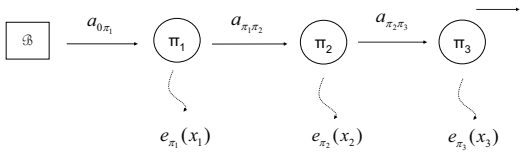
$$H(X) = -\sum_i p(x_i) \cdot \log(p(x_i)) = -\sum_i p_i \cdot \log(p_i)$$

$$p(A) = p(C) = p(G) = p(T) = 1/4$$

$$H = -\sum_{i=1}^{4} \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) = 2 \quad 2 \; bit; \quad 2 \; Ja/Nein - Fragen$$

Durbin et al, Chapter 11.2

HMM: combined probability of state path and observed symbols



$$P(x,\pi) = a_{0\pi_1} \cdot \;\; e_{\pi_1}(x_1) \cdot a_{\pi_1\pi_2} \cdot \;\; e_{\pi_2}(x_2) \cdot a_{\pi_2\pi_3} \cdot \;\; \ldots$$

$$P(x,\pi) = a_{0\pi_1} \cdot \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i\pi_{i+1}}$$

CpG - Inseln

- Genomische Region mit relativ großer Häufigkeit von CpG – Dinukleotiden[1]
  - Observed/Expected Ratio ≥ 0,65
  - GC-Gehalt > 55%
  - Länge ≥ 500 bp
- Epigenetische Regulation der Genexpression
- Rett - Syndrom

[1]Takai D, Jones PA (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22.". Proc Natl Acad Sci USA 99 (6): 3740-5