

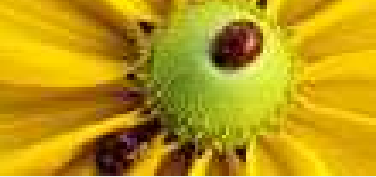


# Ein springendes Profil-HMM zur Erkennung von rekombinanten HI-Viren

Mario Stanke

Department of Bioinformatics, University of Göttingen

mstanke@gwdg.de



# Überblick

Verallg. HMMs

---

Profil-HMMs

---

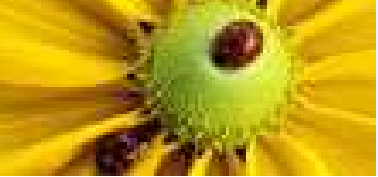
Rekombination bei HIV

---

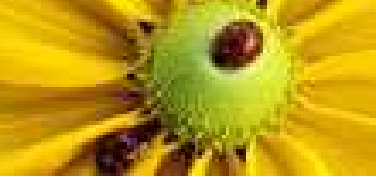
jpHMM

---

1. Verallgemeinerte Hidden Markov Modelle
2. Profil-HMMs zur Modellierung von Sequenzfamilien
3. Rekombination bei HIV
4. jpHMM - ein “springendes” Profil-HMM



# Verallgemeinerte Hidden Markov Modelle



# Erinnerung HMMs

$Z$ : Menge von Zuständen (z.B. 1="fairer Würfel", 2="unfairer Würfel")

$\Sigma$ : Emissionsalphabet (z.B.  $\Sigma = \{1, 2, 3, 4, 5, 6\}$ )

Ein HMM ist ein stochastischer Prozess

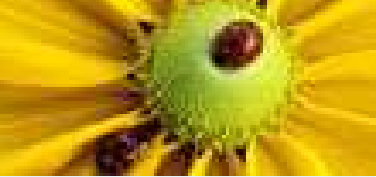
$$X_0, Y_0, X_1, Y_1, X_2, Y_2, \dots,$$

wobei  $X_i \in Z$  (Zustände) und  $Y_i \in \Sigma$  (Emissionen).

Die Folge  $X_0, X_1, X_2, \dots$  ist eine homogene Markov-Kette und

$$P(Y_i = y_i \mid X_0 = x_0, \dots, X_i = x_i, Y_0 = y_0, \dots, Y_{i-1} = y_{i-1}) = P(Y_i = y_i \mid X_i = x_i)$$

für alle  $i > 1$ ,  $x_0, \dots, x_i \in Z$ ,  $y_0, \dots, y_i \in \Sigma$ .



# Brauchen Verallgemeinerung von HMM ...

## Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

## Profil-HMMs

## Rekombination bei HIV

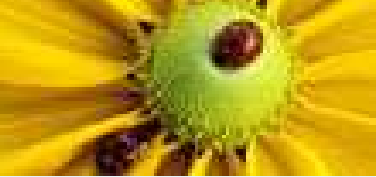
## jpHMM

... für viele Anwendungen.

Im HMM wird in jedem Zustand immer genau ein Zeichen emittiert. Aber z.B.

- Bei der Modellierung von sogenannten Sequenzalignments wird ein Modell benötigt, bei dem in einem Zustand auch **kein Zeichen** emittiert werden kann.
- In der Genvorhersage ist ein Modell sinnvoller, bei dem in den Zuständen **ganze Worte** (variabler Länge) emittiert werden können.

**Beispiel:** Unehrlisches Kasino, bei dem jeder Würfel 10-20 mal gewürfelt wird bis er gewechselt wird.



# Idee: Emission von Wörtern anstatt Zeichen

## Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

## Profil-HMMs

## Rekombination bei HIV

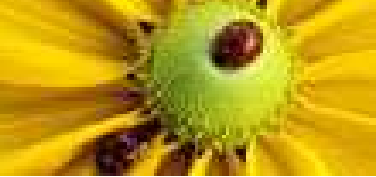
## jpHMM

Idee: Wir erlauben, dass eine Emission  $Y_i$  ein zufälliges **ganzes Wort** ist anstatt nur ein zufälliges Zeichen.

D.h.

$$Y_i \in \Sigma^* \quad (i = 0, 1, 2, \dots)$$

$\Sigma^*$  ist die Menge aller Wörter (beliebiger endlicher Länge), die aus Zeichen aus dem Alphabet  $\Sigma$  gebildet werden können.  $\Sigma^*$  enthält auch das **leere Wort**  $\varepsilon$ , d.h. es soll in einem Zustand auch nichts emittiert werden können.



# Verallgemeinertes HMM (GHMM)

Ein **verallgemeinertes HMM** ist ein stochastischer Prozess

$$X_0, Y_0, X_1, Y_1, X_2, Y_2, \dots,$$

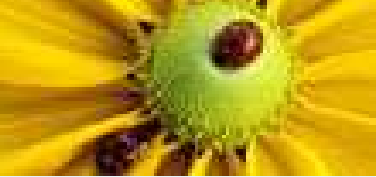
wobei  $X_i \in Z$  (Zustände) und  $Y_i \in \Sigma^*$  (Emissionen).

Die Folge  $X_0, X_1, X_2, \dots$  ist eine homogene Markov-Kette und

$$\begin{aligned} & P(Y_i = y_i \mid X_0 = x_0, \dots, X_i = x_i, Y_0 = y_0, \dots, Y_{i-1} = y_{i-1}) \\ &= P(Y_i = y_i \mid X_i = x_i) \end{aligned} \tag{1}$$

für alle  $i > 0$ ,  $x_0, \dots, x_i \in Z$ ,  $y_0, \dots, y_i \in \Sigma^*$ . (1) ist unabhängig von  $i$ .





# Bezeichnungen

## Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- **Bezeichnungen**
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

## Profil-HMMs

## Rekombination bei HIV

## jpHMM

Sei analog zum HMM

$$a_{q',q} := P(X_i = q \mid X_{i-1} = q').$$

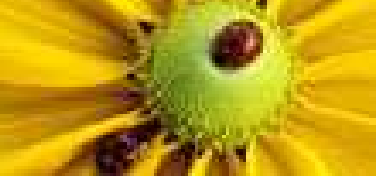
D.h.  $a_{q',q}$  ist die Übergangswkeit von Zustand  $q'$  zu Zustand  $q$ . Wir führen auch hier einen imaginären Zustand '-1' ein um die Startwkeiten als Übergangswkeiten zu modellieren:

$$P(X_0 = q) = a_{-1,q}.$$

Bezeichnung der Emissionswkeiten:

$$e_q(w) := P(Y_i = w \mid X_i = q) \quad (q \in Z, w \in \Sigma^*)$$

ist die Wkeit, dass im Zustand  $q$  das Wort  $w$  emittiert wird.



# Bezeichnungen

Sei  $\sigma = \sigma[0, \ell) \in \Sigma^+$  eine **Beobachtung** der Länge  $\ell$ .

Ein GHMM ist zwar unendliche Folge, aber wir betrachten in Anwendungen endliche Folgen bis zu einem Index  $n$ .

$$(X_0, Y_0, X_1, Y_1, \dots, X_n, Y_n)$$

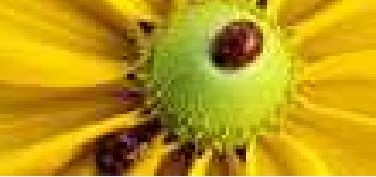
Sei

$$S = S_n = Y_0 Y_1 Y_2 \cdots Y_n$$

die Verkettung der ersten  $n + 1$  emittierten Wörter. Wir werden den Index  $n$  oft weglassen, wenn klar ist, auf welches  $n$  es sich bezieht. Wir nennen die Größe

$$\Phi = \Phi_n = ((X_0, U_0, V_0), (X_1, U_1, V_1), \dots, (X_n, U_n, V_n))$$

mit  $U_0 = 0$ ,  $V_i = U_i + |Y_i|$ ,  $U_{i+1} = V_i$  den **Pfad**. Die Segmentgrenzen  $U_i, V_i$  sind so gewählt, dass  $Y_i = S[U_i, V_i)$ . Zusätzlich zum Pfad eines normalen HMMs enthält der Pfad des GHMMs also auch die Info, wo die Zustände anfangen und enden.



# Problemstellung in Anwendungen

## Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- **Problemstellung**
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

## Profil-HMMs

## Rekombination bei HIV

## jpHMM

Suchen den Pfad  $\varphi$ , der der Beobachtung  $\sigma$  zugrunde liegt.  
Fassen  $\sigma$  als Realisierung von  $S$  auf.

Ein GHMM ist wie ein HMM ein **generatives Modell**.

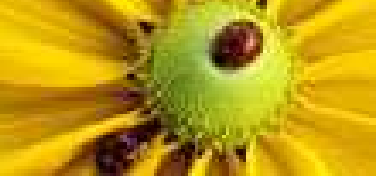
Dekodieren: Suchen wahrscheinlichsten Pfad, gegeben die Beobachtung  $\sigma$ :

$$\hat{\varphi} \in \operatorname{argmax}_{\varphi} P(\Phi = \varphi \mid S = \sigma)$$

Pfad  $\varphi$ ,

$$v_n = \ell$$

( $n$  definiert durch  $\varphi$ ). Nennen  $\hat{\varphi}$  einen **Viterbi-Pfad**.



# Viterbi-Algorithmus für GHMM

$$\begin{aligned}\hat{\varphi} &\in \operatorname{argmax}_{\text{Pfad } \varphi, v_n = \ell} P(\Phi = \varphi \mid S = \sigma) \\ &= \operatorname{argmax}_{\text{Pfad } \varphi, v_n = \ell} P(\Phi = \varphi, S = \sigma) \\ &= \operatorname{argmax}_{\text{Pfad } \varphi, v_n = \ell} \prod_{i=0}^n a_{x_{i-1}x_i} \cdot e_{x_i}(\sigma[u_i, v_i])\end{aligned}$$

Viterbi-Variable:

$$\begin{aligned}\gamma_{q,t} := & \max_{\substack{\text{Pfad } \varphi, \\ v_n = t+1, \\ x_n = q}} P(\Phi = \varphi, S = \sigma[0, t]) \quad (q \in Z, 0 \leq t < \ell)\end{aligned}$$

Angenommen,  $P(|Y_i| = \varepsilon) = 0$  für alle  $i$ , d.h. jedes emittierte Wort hat mindestens die Länge 1. Dann ist ganz analog zur Viterbi-Rekursion im normalen HMM

$$\begin{aligned}\gamma_{q,t} = & \max_{\substack{q' \in Z, \\ 0 \leq t' \leq t}} \gamma_{q',t'} \cdot a_{q',q} \cdot e_q(\sigma(t', t)).\end{aligned} \tag{2}$$



# Vorwärts-Algorithmus für GHMM

Vorwärts-Variable:

$$\alpha_{q,t} := \sum_{\substack{\text{Pfad } \varphi, \\ v_n=t+1, \\ x_n=q}} P(\Phi = \varphi, S = \sigma[0, t]) \quad (q \in Z, 0 \leq t < \ell)$$

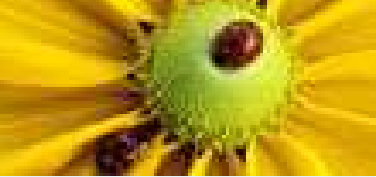
Auch der Vorwärts-Algorithmus funktioniert unter der Bedingung

$P(|Y_i| = \varepsilon) = 0$  ganz analog:

$$\alpha_{q,t} = \sum_{\substack{q' \in Z, \\ 0 \leq t' \leq t}} \alpha_{q',t'} \cdot a_{q',q} \cdot e_q(\sigma(t', t]). \quad (3)$$

Dasselbe gilt für den Rückwärts-Algorithmus.

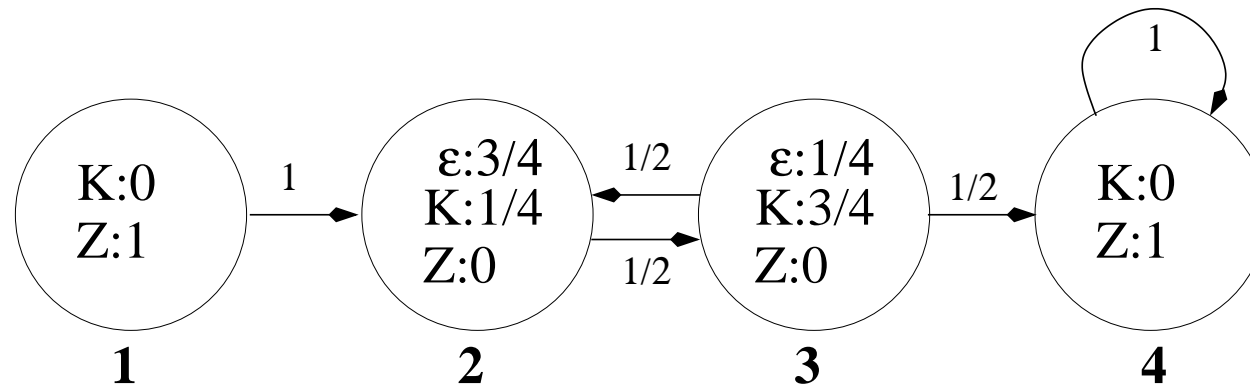
Erinnerung: Vorwärts- und Rückwärts-Algorithmus können dazu verwendet werden, die a-posteriori-Wkeiten der Zustände und die Wkeit der Beobachtung zu berechnen.



# Leere Emission problematisch

Die Viterbi-, Vorwärts- und Rückwärts-Algorithmen funktionieren im allgemeinen nicht, wenn leere Worte emittiert werden können.

Beispiel:



Sei  $\sigma = ZKZ$ . Hier ist  $\hat{\varphi} = ((1, 0, 1), (2, 1, 1), (3, 1, 2), (4, 2, 3))$  aber die Viterbirekursion funktioniert nicht, weil man etwa zur Berechnung von  $\gamma_{2,1}$  den Wert von  $\gamma_{3,1}$  benötigt und umgekehrt.

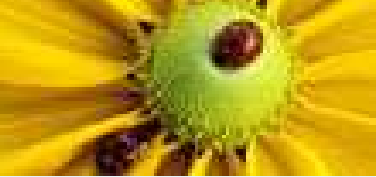
Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

Profil-HMMs

Rekombination bei HIV

jpHMM



## Weitere Forderung bzgl. Emission vom leeren Wort

Damit die Algorithmen ( Viterbi, Vorwärts, Rückwärts) analog übertragen werden können, fordern wir noch:

$$a_{q',q} \cdot e_q(\varepsilon) = 0 \quad \text{falls } q' \geq q.$$

Dies kann unter Umständen durch Umnummerierung der Zustände erreicht werden.

In Worten: Die Zustände sollten so sortiert werden können, dass die möglichen Vorgängerzustände eines Zustands  $q$ , in dem das leere Wort emittiert werden kann, vor  $q$  kommen.

Dann funktioniert die Viterbi-(Vorwärts, Rückwärts) Rekursion ohne Änderung genau wie z.B. in (2), da bei jedem möglichen Pfad zwei aufeinanderfolgende Zustände  $x_{i-1}, x_i$  mit leerer Emission  $y_i$  die Reihenfolge  $x_{i-1} < x_i$  ist, und die Rekursionsvariable  $\gamma_{x_{i-1},t}$  bei der Berechnung von  $\gamma_{x_i,t}$  bereits zur Verfügung steht.

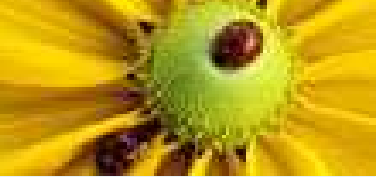
### Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- **Weitere Forderung**
- GHMM

### Profil-HMMs

### Rekombination bei HIV

### jpHMM



# Beobachtungen/Bemerkungen

## Verallg. HMMs

- Erinnerung HMMs
- Brauchen Verallgemeinerung
- Emission von Wörtern
- Verallgemeinertes HMM
- Bezeichnungen
- Problemstellung
- Viterbi
- Vorwärts
- Leere Emission problematisch
- Weitere Forderung
- GHMM

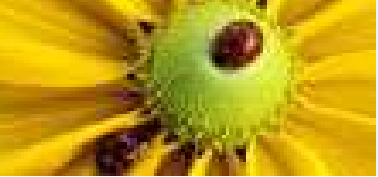
## Profil-HMMs

## Rekombination bei HIV

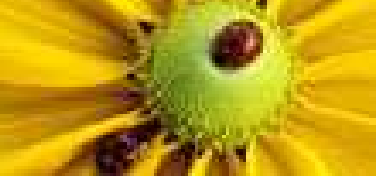
## jpHMM

- Jedes HMM ist offenbar ein GHMM.
- Ein in diesem Sinne verallgemeinertes HMM heisst manchmal auch **Hidden Semi-Markov Model (HSMM)**.
- Laufzeit von Viterbi/Vorwärts/Rückwärts:  $O(N^2 \cdot \ell \cdot M)$   
Speicherbedarf:  $O(N \cdot \ell)$   
Dabei:  
 $N = |Z|$ : Anzahl der Zustände  
 $\ell$ : Länge der Beobachtung  
 $M = \max\{|w| \mid e_q(w) > 0, q \in Z, w \in \Sigma^*\}$  maximale Länge einer Emission





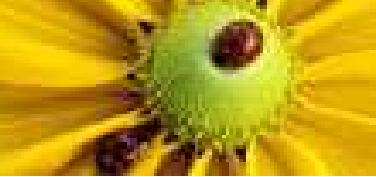
# Profil-HMMs zur Modellierung von Sequenzfamilien



# Multiples Sequenzalignment (MSA)

```
Subtyp A1 : GAGCAGAAAGACAGGGAACAGGCCCAACCCTTAGTT
Subtyp A2 : GAGAACAGGGAGCCGTCACCCCTGCAATT
Subtyp B  : GAGCCGATAGACAAGGAACTGTATCCTTTAACT
Subtyp C  : GAGACGATAGACAAGGAACTGCCCTTAACT
Subtyp D  : GAGCAGAAAGACAAGGAACTGTATCCTTTAACT
Subtyp F1 : GAGCAGAAGGACGAGGGACAGGGACTGTATCCTCCCTTAGCC
Subtyp G  : GAGCAGAAGGAAAAGGAACTATATCCTCTATCT
Subtyp H  : GAGCAGCTGAAGGACAAGGAACCTCCCTTAGCT
Subtyp J  : GAGCCGAAGGACAAGGAACTGTATCCTCTAACT
Subtyp K  : GAGACCAAAGACAAGGAACAGAGCCCTCCTTTAACT
```

einige Ausschnitte von HIV-Sequenzen aus einer bestimmten Region eines Gens



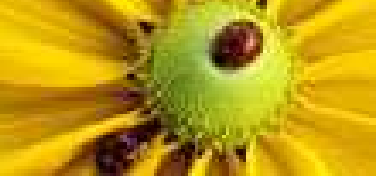
# Multiples Sequenzalignment (MSA)

```
Subtyp A1 : GAGCAGAAAGACAGGGAACAGGCCCAACCCTTAGTT
Subtyp A2 : GAGAACAGGGAGCCGTCACCCCTGCAATT
Subtyp B  : GAGCCGATAGACAAGGAACTGTATCCTTTAACT
Subtyp C  : GAGACGATAGACAAGGAACTGCCCTTAACT
Subtyp D  : GAGCAGAAAGACAAGGAACTGTATCCTTTAACT
Subtyp F1 : GAGCAGAAGGACGAGGGACAGGGACTGTATCCTCCCTTAGCC
Subtyp G  : GAGCAGAAGGAAAAGGAACTATATCCTCTATCT
Subtyp H  : GAGCAGCTGAAGGACAAGGAACCTCCCTTAGCT
Subtyp J  : GAGCCGAAGGACAAGGAACTGTATCCTCTAACT
Subtyp K  : GAGACCAAAGACAAGGAACAGAGCCCTCCTTTAACT
```

einige Ausschnitte von HIV-Sequenzen aus einer bestimmten Region eines Gens

## Annahmen:

- Alle Sequenzen stammen von einer einzigen Vorfahr-Sequenz ab
- Die Sequenz verändert sich im Laufe der Evolution durch **Mutationen, Insertionen und Deletionen**

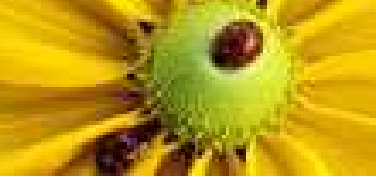


# Multiples Sequenzalignment (MSA)

```
Subtyp A1 : GAGCA---GAAAGACAG-----GGAACAGGCCCAACCCTTAGTT
Subtyp A2 : GAGAA-----CAG-----GGAGCCGTCCACCCCTGCAATT
Subtyp B  : GAGCC---GATAGACAA-----GGAACTGTA---TCCTTTAACT
Subtyp C  : GAGAC---GATAGACAA-----GGAACT-----GCCCTTAACT
Subtyp D  : GAGCA---GAAAGACAA-----GGAACTGTA---TCCTTTAACT
Subtyp F1 : GAGCA---GAAGGACGAGGGACAGGGACTGTATCCTCCCTTAGCC
Subtyp G  : GAGCA---GAAGGAAAA-----GGAACTATA---TCCTCTATCT
Subtyp H  : GAGCAGCTGAAGGACAA-----GGAACC-----TCCCTTAGCT
Subtyp J  : GAGCC---GAAGGACAA-----GGAACTGTA---TCCTCTAACT
Subtyp K  : GAGAC---CAAAGACAA-----GGAACAGAGCCCTCCTTTAACT
```

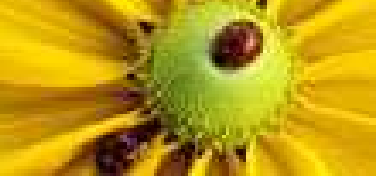
## Alignment der Sequenzen:

Es werden Lückenzeichen '-' so eingefügt, dass Positionen, die wahrscheinlich von derselben Position im gemeinsamen Vorfahren abstammen, untereinander stehen.



# Profil-HMMs (Krogh 1994, Eddy 1995)

|            |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|
| Subtyp A1: | C | A | - | - | - | G | A | A | A |
| Subtyp A2: | A | A | - | - | - | - | - | - | - |
| Subtyp B : | C | C | - | - | - | G | A | T | A |
| Subtyp C : | A | C | - | - | - | G | A | T | A |
| Subtyp D : | C | A | - | - | - | G | A | A | A |
| Subtyp F1: | C | A | - | - | - | G | A | A | G |
| Subtyp G : | C | A | - | - | - | G | A | A | G |
| Subtyp H : | C | A | G | C | T | G | A | A | G |
| Subtyp J : | C | C | - | - | - | G | A | A | G |
| Subtyp K : | A | C | - | - | - | C | A | A | A |

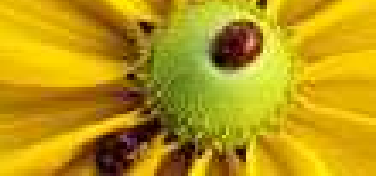


# Profil-HMMs (Krogh 1994, Eddy 1995)

|            |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|
| Subtyp A1: | C | A | - | - | - | G | A | A | A |
| Subtyp A2: | A | A | - | - | - | - | - | - | - |
| Subtyp B : | C | C | - | - | - | G | A | T | A |
| Subtyp C : | A | C | - | - | - | G | A | T | A |
| Subtyp D : | C | A | - | - | - | G | A | A | A |
| Subtyp F1: | C | A | - | - | - | G | A | A | G |
| Subtyp G : | C | A | - | - | - | G | A | A | G |
| Subtyp H : | C | A | G | C | T | G | A | A | G |
| Subtyp J : | C | C | - | - | - | G | A | A | G |
| Subtyp K : | A | C | - | - | - | C | A | A | A |

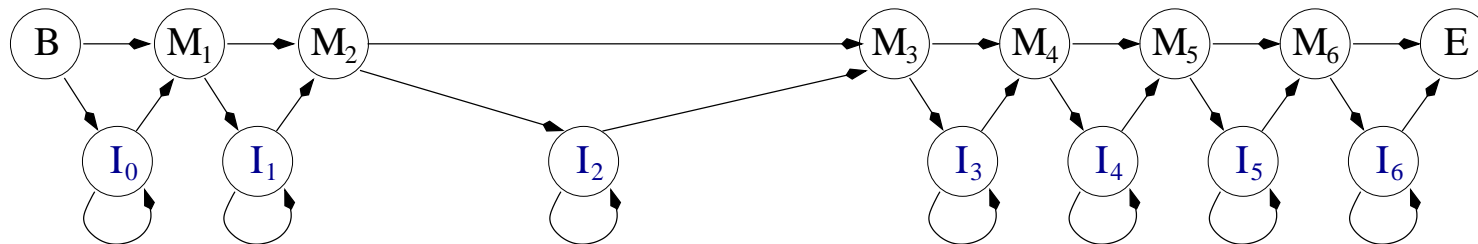


Match-Zustände  $M_i$  für jede hinreichend konservierte Pos. , emittiert 1 Zeichen

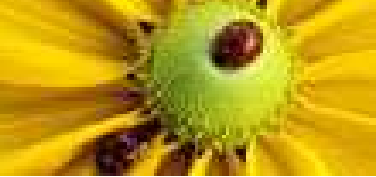


# Profil-HMMs (Krogh 1994, Eddy 1995)

|            |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|
| Subtyp A1: | C | A | - | - | - | G | A | A | A |
| Subtyp A2: | A | A | - | - | - | - | - | - | - |
| Subtyp B:  | C | C | - | - | - | G | A | T | A |
| Subtyp C:  | A | C | - | - | - | G | A | T | A |
| Subtyp D:  | C | A | - | - | - | G | A | A | A |
| Subtyp F1: | C | A | - | - | - | G | A | A | G |
| Subtyp G:  | C | A | - | - | - | G | A | A | G |
| Subtyp H:  | C | A | G | C | T | G | A | A | G |
| Subtyp J:  | C | C | - | - | - | G | A | A | G |
| Subtyp K:  | A | C | - | - | - | C | A | A | A |

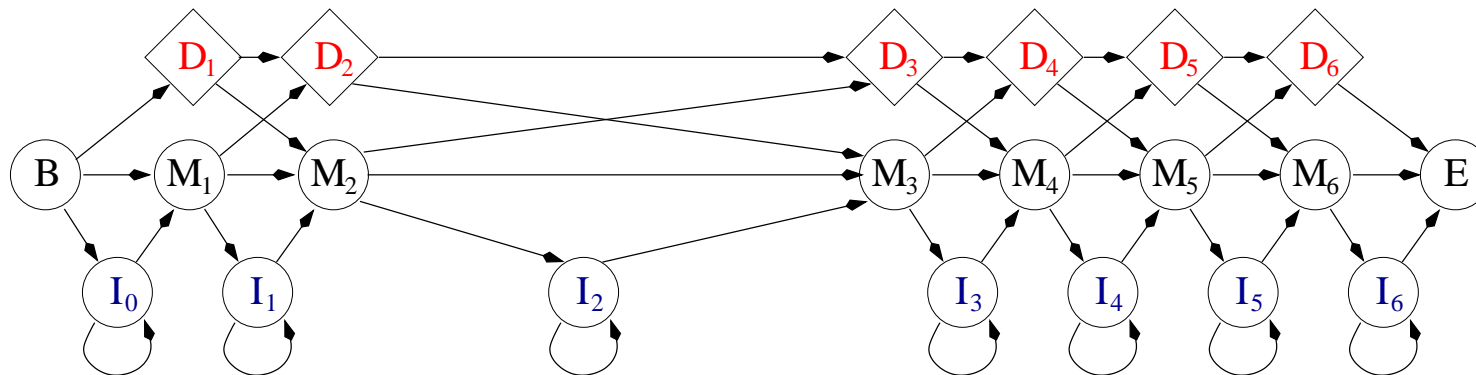


Match-Zustände  $M_i$  für jede hinreichend konservierte Pos. , emittiert 1 Zeichen  
Insert-Zustände  $I_i$  um Einfügungen zu erlauben, emittiert 1 Zeichen



# Profil-HMMs (Krogh 1994, Eddy 1995)

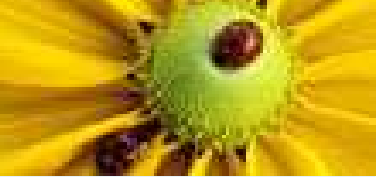
|            |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|
| Subtyp A1: | C | A | - | - | - | G | A | A | A |
| Subtyp A2: | A | A | - | - | - | - | - | - | - |
| Subtyp B:  | C | C | - | - | - | G | A | T | A |
| Subtyp C:  | A | C | - | - | - | G | A | T | A |
| Subtyp D:  | C | A | - | - | - | G | A | A | A |
| Subtyp F1: | C | A | - | - | - | G | A | A | G |
| Subtyp G:  | C | A | - | - | - | G | A | A | G |
| Subtyp H:  | C | A | G | C | T | G | A | A | G |
| Subtyp J:  | C | C | - | - | - | G | A | A | G |
| Subtyp K:  | A | C | - | - | - | C | A | A | A |



Match-Zustände  $M_i$  für jede hinreichend konservierte Pos. , emittiert 1 Zeichen  
 Insert-Zustände  $I_i$  um Einfügungen zu erlauben, emittiert 1 Zeichen  
 Delete-Zustände  $D_i$  um Löschungen zu erlauben, emittiert  $\varepsilon$



# Schätzen der Parameter



Verallg. HMMs

---

Profil-HMMs

---

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

---

jpHMM

---

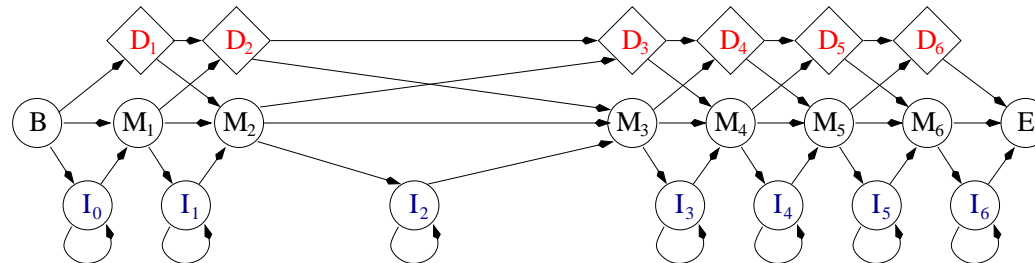
Müssen schätzen:

- Emissionswahrscheinlichkeiten  $e_q(w)$  für Match- und Insertzustände
- Übergangswahrscheinlichkeiten  $a_{q',q}$

# Schätzen der Parameter

Das MSA induziert **beobachtete** Häufigkeiten der Emissionen und Übergänge.

|            |   |   |   |   |   |   |   |   |   |
|------------|---|---|---|---|---|---|---|---|---|
| Subtyp A1: | C | A | - | - | - | G | A | A | A |
| Subtyp A2: | A | A | - | - | - | - | - | - | - |
| Subtyp B : | C | C | - | - | - | G | A | T | A |
| Subtyp C : | A | C | - | - | - | G | A | T | A |
| Subtyp D : | C | A | - | - | - | G | A | A | A |
| Subtyp F1: | C | A | - | - | - | G | A | A | G |
| Subtyp G : | C | A | - | - | - | G | A | A | G |
| Subtyp H : | C | A | G | C | T | G | A | A | G |
| Subtyp J : | C | C | - | - | - | G | A | A | G |
| Subtyp K : | A | C | - | - | - | C | A | A | A |

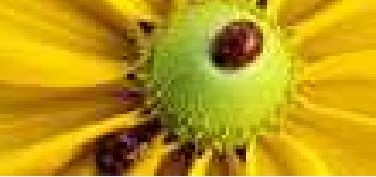


Z.B.:

In  $M_1$  werden Emissionen beobachtet: 3xA, 7xC

In  $I_2$  werden Emissionen beobachtet: 1xC, 1xG, 1xT

Übergänge aus  $M_2$ : 8x zu  $M_3$ , 1x zu  $D_3$ , 1x zu  $I_2$



# Schätzen der Parameter einer Multinomialverteilung

Allgemeines Problem:

Ein Zufallsexperiment hat  $k$  verschiedene Ausgänge, die Ereignisse  $1, 2, \dots, k$ .

Suchen Wkkeiten  $p_1, p_2, \dots, p_k$  der  $k$  Ausgänge.

Wir haben

- einen Häufigkeitsvektor  $\vec{n} = (n_1, n_2, \dots, n_k)$   
(Stichprobe vom Umfang  $n := n_1 + \dots + n_k$ )
- **a-priori Wissen** über die Verteilung von  $\vec{p} = (p_1, p_2, \dots, p_k)$

**Bsp 1:** Reisen in unbekanntest Land. Erste beiden Einwohner haben schwarze Haare. 1=schwarze Haare, 2=nicht schwarze Haare.  $\vec{n} = (2, 0)$ . Gesucht  $p_1 \sim$  Anteil Einwohner mit schwarzen Haaren.

**Bsp 2:** Würfel. Auf jeder Seite steht 1, 2 oder 3. Würfeln 7 mal.  $\vec{n} = (4, 3, 0)$ . Suchen Wkkeiten für 1,2,3.

Verallg. HMMs

Profil-HMMs

● Alignment

● Profile HMM

● **Schätzen der Parameter**

● Multinomialverteilung

● A-priori Wissen

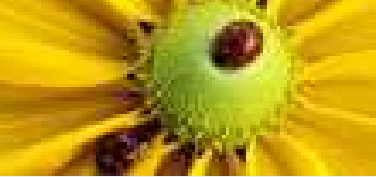
● Bayessches Modell

● Dirichlet-Verteilung

● Dirichlet-Mixtures

Rekombination bei HIV

jpHMM



# Multinomialverteilung

Verallg. HMMs

Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- **Multinomialverteilung**
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

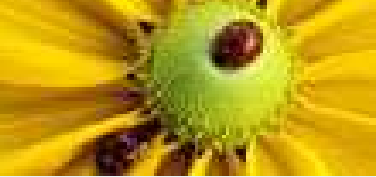
Rekombination bei HIV

jpHMM

$\vec{n}$  ist multinomialverteilt:

$$P(\vec{n} | \vec{p}) = \frac{n!}{n_1! \cdot \dots \cdot n_k!} p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_k^{n_k}$$

(elementare Kombinatorik)



# Bayessches Modell: A-priori Wissen

Verallg. HMMs

Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

jpHMM

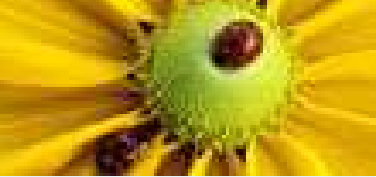
Wir betrachten den unbekanntem **Parametervektor** selbst als **zufällig**.

Sei  $D = \{(p_1, \dots, p_k) \mid p_1 + \dots + p_k = 1\}$  die Menge aller möglichen Parametervektoren.

A-priori Wissen kann als Verteilung auf  $D$  modelliert werden.

Sei  $f$  eine Dichte dieser **a-priori Verteilung**:

$$\int_D f(\vec{p}) d\vec{p} = 1$$



# Bayessches Modell

Verallg. HMMs

Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

jpHMM

Der Zufällige Prozess, der den Häufigkeitsvektor  $\vec{n}$  und den Parametervektor  $\vec{p}$  erzeugt, läuft wie folgt ab:

1. Der Parametervektor  $\vec{p} \in D$  wird zufällig gewählt gemäss der a-priori-Verteilung gegeben durch  $f$ .
2. Es werden  $n$  unabhängige Zufallsexperimente gemacht gemäss der Verteilung  $\vec{p}$ .  $\vec{n}$  ist dann der Vektor, der die Häufigkeiten der  $k$  mgl. Ereignisse zählt.

# Schätzung

Die Wkeit für das  $i$ -te Ereignis wird nun als

$$\begin{aligned}\hat{p}_i &:= P(i\text{-tes Ereignis} \mid \vec{n}) \\ &= \int_{D_{20}} p_i \cdot f(\vec{p} \mid \vec{n}) d\vec{p}.\end{aligned}\tag{4}$$

geschätzt.

Dabei bezeichnet  $f(\vec{p} \mid \vec{n}) = \frac{f(\vec{p}) \cdot P(\vec{n} \mid \vec{p})}{P(\vec{n})}$  die bedingte Dichte des Parametervektors  $\vec{p}$  gegeben den Häufigkeitsvektor  $\vec{n}$ .

$P(\vec{n})$  kann nach der Formel von der totalen Wkeit berechnet werden

$$P(\vec{n}) = \int_{D_{20}} P(\vec{n} \mid \vec{p}) \cdot f(\vec{p}) d\vec{p}.\tag{5}$$

Mit obigen Formeln können die Schätzungen  $\hat{p}_i$  berechnet werden, wenn die Integrale in (4) und (5) berechnet werden können.

Verallg. HMMs

Profil-HMMs

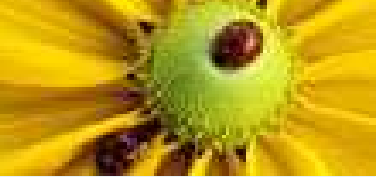
- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen

● Bayessches Modell

- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

jpHMM



# Dirichlet-Verteilung als a-priori-Verteilung

G. Lejeune Dirichlet *“Über die Lehre von den einfachen und mehrfachen bestimmten Integralen”* (G.Arendt, 1904) gehalten im Jahr 1852 in Göttingen:

$$\int_D \prod_{i=1}^k x_i^{\alpha_i - 1} d(x_1, \dots, x_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\alpha_1 + \dots + \alpha_k)} \quad (6)$$

$\alpha_1, \alpha_2, \dots, \alpha_k > 0$  Parameter

Dichte der Dirichlet-Verteilung:

$$f(\vec{p}) = \frac{\prod_{i=1}^k p_i^{\alpha_i - 1}}{\int_D \prod_{i=1}^k p_i^{\alpha_i - 1} d\vec{p}} \quad (7)$$

Erwartungswert:  $E p_i = \frac{\alpha_i}{\alpha_1 + \dots + \alpha_k}$

Verallg. HMMs

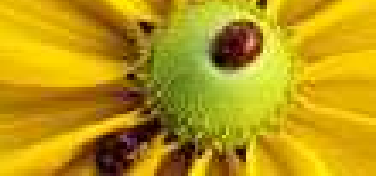
Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

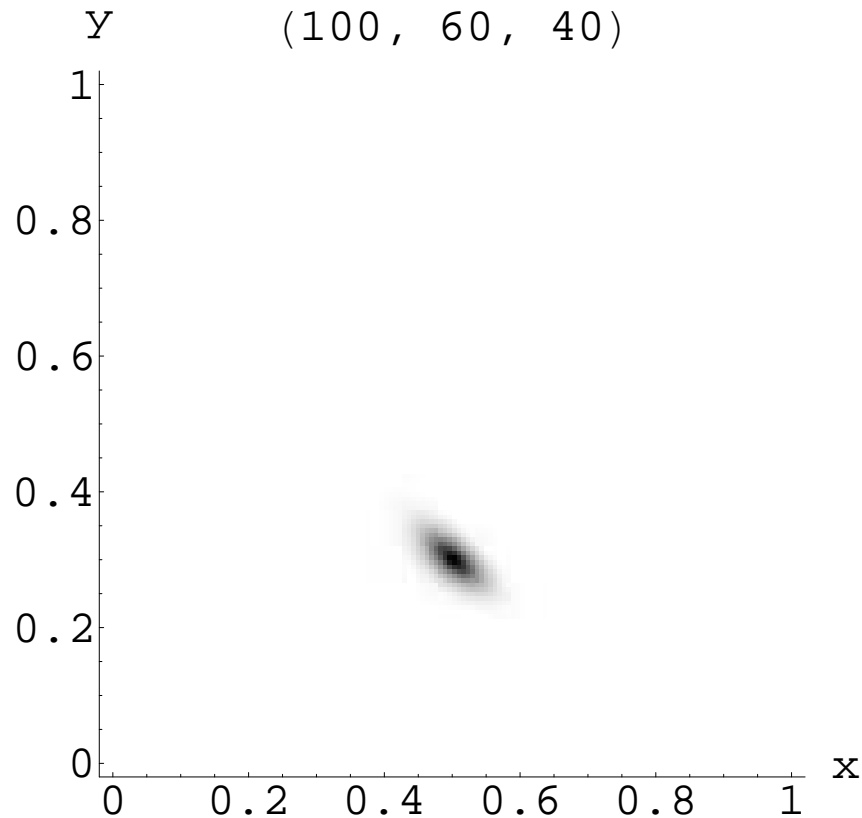
Rekombination bei HIV

jpHMM





# Dirichlet-Verteilung als a-priori-Verteilung

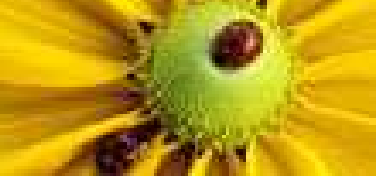


Einige Dirichlet-Verteilungen für  $k = 3$  Dimensionen.

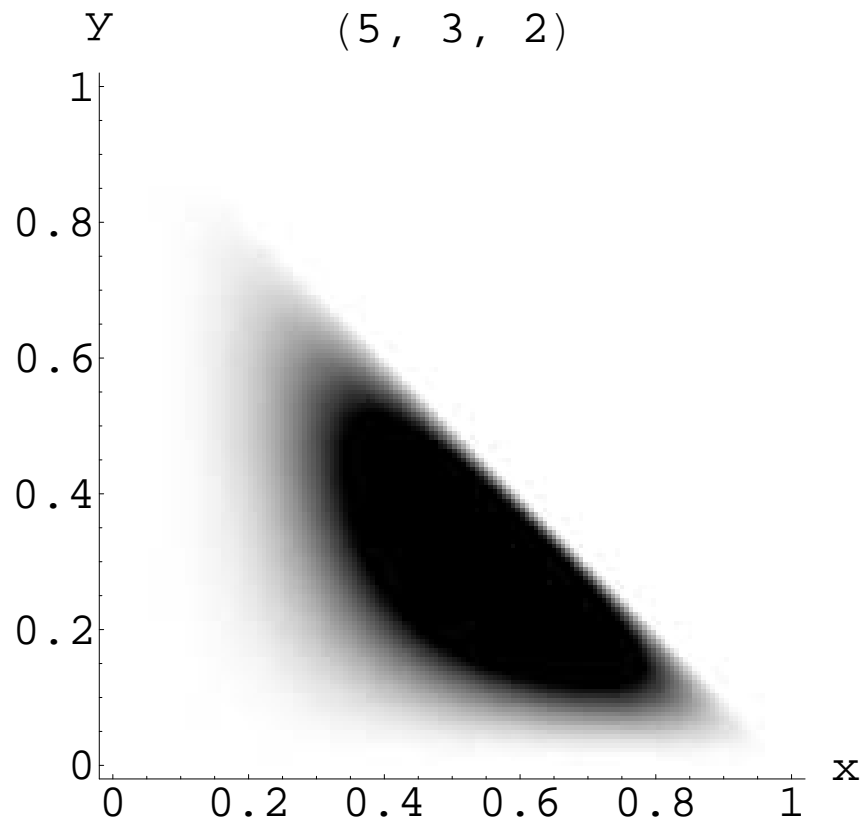
Oben stehen jeweils die Parameter  $(\alpha_1, \alpha_2, \alpha_3)$ .

Die Projektion auf die ersten beiden Dimensionen  $x$  und  $y$  wird gezeigt.

Die dritte Dimension  $z$  ist gleich  $1 - x - y$ .  
Je dunkler, desto grösser der Wert der Dichte.



# Dirichlet-Verteilung als a-priori-Verteilung

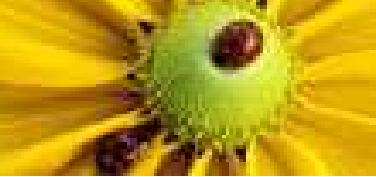


Einige Dirichlet-Verteilungen für  $k = 3$  Dimensionen.

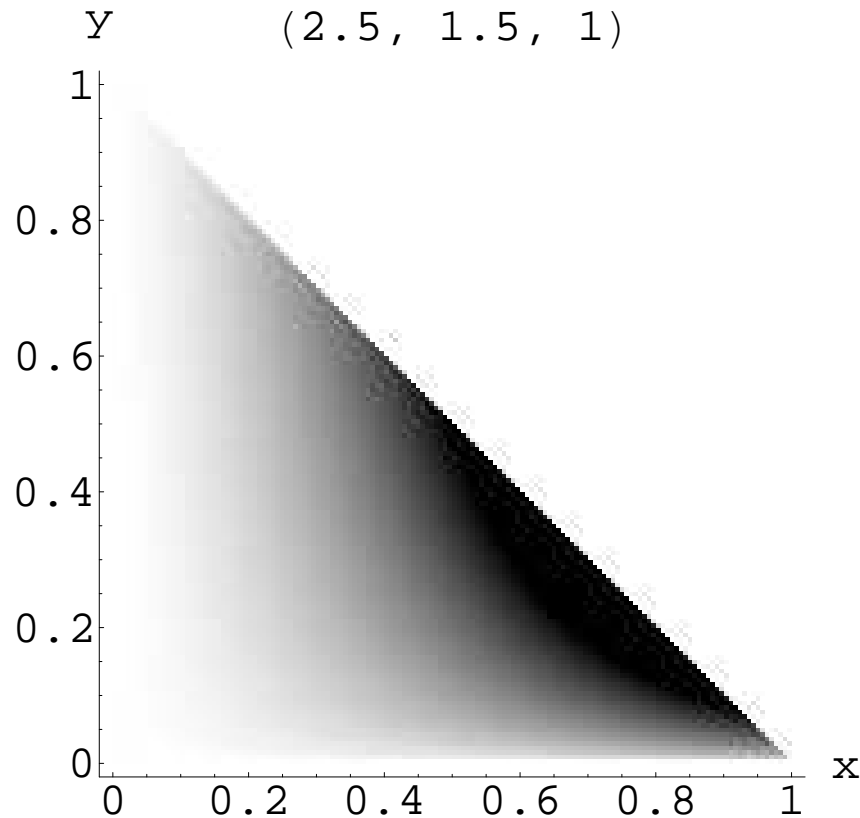
Oben stehen jeweils die Parameter  $(\alpha_1, \alpha_2, \alpha_3)$ .

Die Projektion auf die ersten beiden Dimensionen  $x$  und  $y$  wird gezeigt.

Die dritte Dimension  $z$  ist gleich  $1 - x - y$ .  
Je dunkler, desto grösser der Wert der Dichte.



# Dirichlet-Verteilung als a-priori-Verteilung

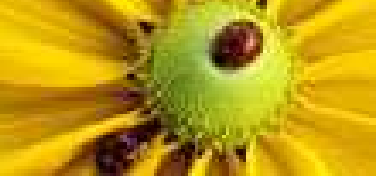


Einige Dirichlet-Verteilungen für  $k = 3$  Dimensionen.

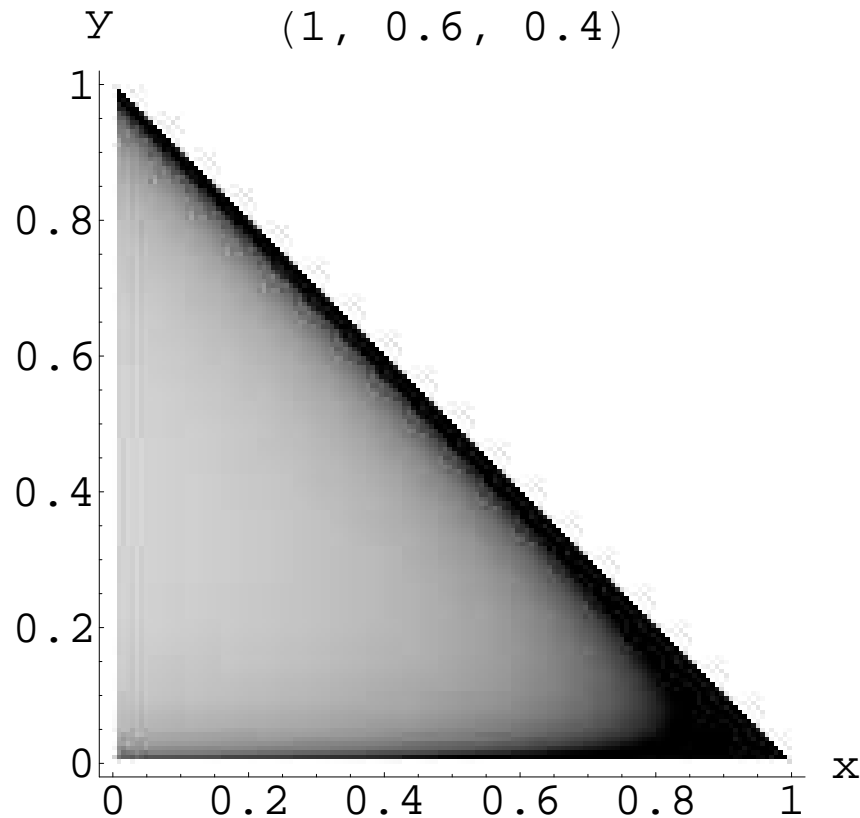
Oben stehen jeweils die Parameter  $(\alpha_1, \alpha_2, \alpha_3)$ .

Die Projektion auf die ersten beiden Dimensionen  $x$  und  $y$  wird gezeigt.

Die dritte Dimension  $z$  ist gleich  $1 - x - y$ .  
Je dunkler, desto grösser der Wert der Dichte.



# Dirichlet-Verteilung als a-priori-Verteilung

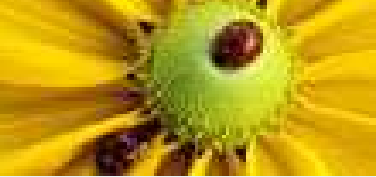


Einige Dirichlet-Verteilungen für  $k = 3$  Dimensionen.

Oben stehen jeweils die Parameter  $(\alpha_1, \alpha_2, \alpha_3)$ .

Die Projektion auf die ersten beiden Dimensionen  $x$  und  $y$  wird gezeigt.

Die dritte Dimension  $z$  ist gleich  $1 - x - y$ .  
Je dunkler, desto grösser der Wert der Dichte.



# Schätzung mit Dirichlet-A-Priori-Verteilung

Verallg. HMMs

Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

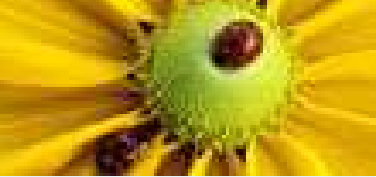
jpHMM

Setzt man dieses  $f$  in obige Formel ein, so erhält man

$$\hat{p}_i = \frac{n_i + \alpha_i}{\sum_j (n_j + \alpha_j)}$$

## Bemerkung:

Die  $\alpha_i$  nennt man in diesem Fall auch **Pseudocounts**, weil man auf dieses Ergebnis auch durch Schätzung mittels relativer Häufigkeiten kommt, wenn man so tut als beobachtete man  $n_i + \alpha_i$  Ereignisse  $i$ .



# Mischung von Dirichlet-Verteilungen

Nicht jedes a-priori Wissen kann hinreichend gut mit einer Dirichlet-Verteilung modelliert werden.

**Beispiel:** Haben Urne mit drei etwa gleich häufigen Typen von Würfeln: Stehen jeweils nur zwei Zahlen von 1, 2, 3 drauf. Entweder  $\vec{p} \approx (\frac{1}{2}, \frac{1}{2}, 0)$  oder  $\vec{p} \approx (\frac{1}{2}, 0, \frac{1}{2})$  oder  $\vec{p} \approx (0, \frac{1}{2}, \frac{1}{2})$ .

**Lösung:** Mischung von Dirichlet-Verteilungen

Sei  $f_j$  die Dichte einer Dirichlet-Verteilung ( $j = 1, \dots, \ell$ ) mit Parametern  $\vec{\alpha}_j = (\alpha_{j,1}, \dots, \alpha_{j,k})$  und  $q_1, \dots, q_k$ , so dass  $q_1 + \dots + q_k = 1$ .

Dann ist

$$f = q_1 f_1 + \dots + q_\ell f_\ell.$$

die **Mischung** der Dirichletverteilungen  $f_1, \dots, f_\ell$  mit **Mischungskoeffizienten**  $q_i$ .

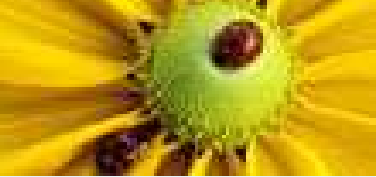
Verallg. HMMs

Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

jpHMM



# Mischung von Dirichlet-Verteilungen

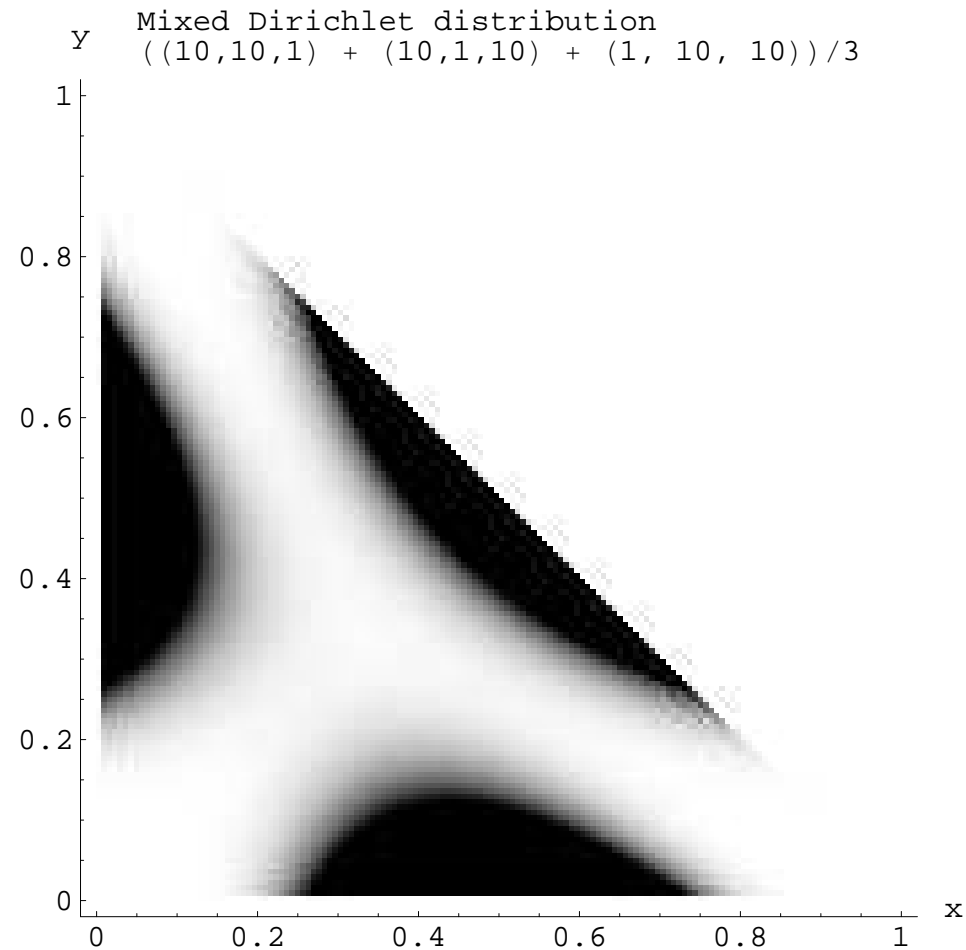
Verallg. HMMs

Profil-HMMs

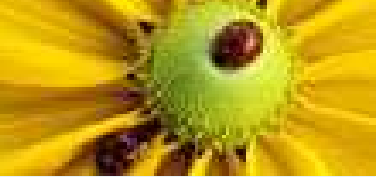
- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

Rekombination bei HIV

jpHMM



Eine Mischung aus 3 Dirichlet-Verteilungen, die in etwa dem a-priori-Wissen aus dem Würfelbeispiel entspricht.



# Schätzen der Parameter im Profil-HMM

Schätzen der **Übergangswkeiten**:

Bayessches Modell mit Dirichlet a-priori-Verteilung  
(Pseudocounts)

Schätzen der **Emissionswkeiten**:

- DNA: Bayessches Modell mit Dirichlet a-priori-Verteilung  
(Pseudocounts)
- Proteinsequenzen: Bayessches Modell mit Mischung von  
Dirichlet-Verteilungen als a-priori-Verteilung

Wahl der Parameter der a-priori-Verteilung: Z.B. nach  
Maximum-Likelihood-Prinzip. Seien  $\vec{n}_1, \vec{n}_2, \dots, \vec{n}_t$  die  
beobachteten Häufigkeitsvektoren aus  $t$  unabhängigen  
Stichproben (Trainingsdaten). Nach dem ML-Prinzip wird dann  
 $\alpha$  wie folgt geschätzt.

$$\hat{\alpha} = \operatorname{argmax}_{\alpha} \prod_{j=1}^t P(\vec{n}_j | \alpha)$$

Verallg. HMMs

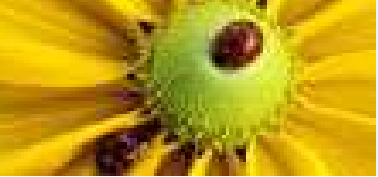
Profil-HMMs

- Alignment
- Profile HMM
- Schätzen der Parameter
- Multinomialverteilung
- A-priori Wissen
- Bayessches Modell
- Dirichlet-Verteilung
- Dirichlet-Mixtures

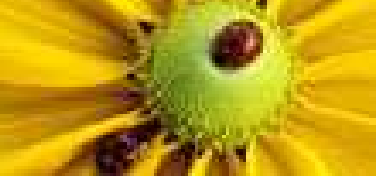
Rekombination bei HIV

jpHMM

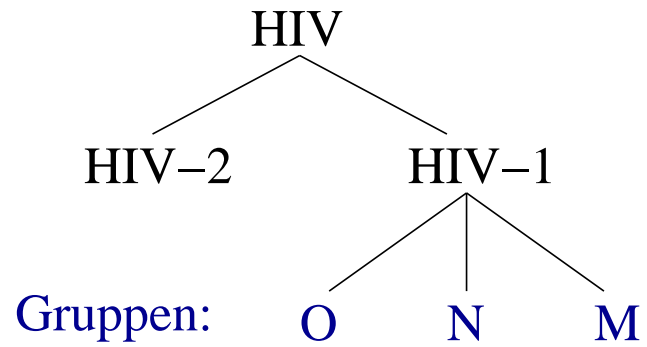


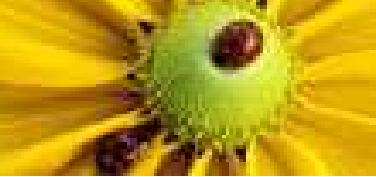


# Rekombination bei HIV

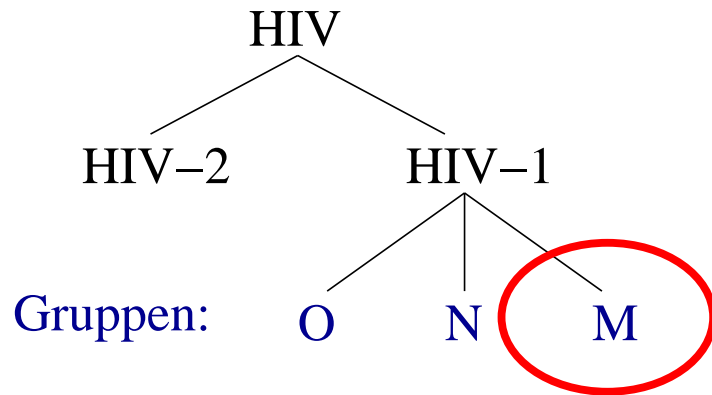


# Human Immunodeficiency Virus (HIV)



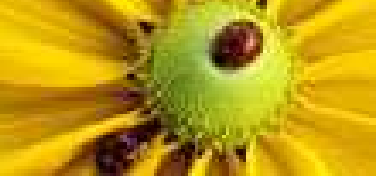


# Human Immunodeficiency Virus (HIV)

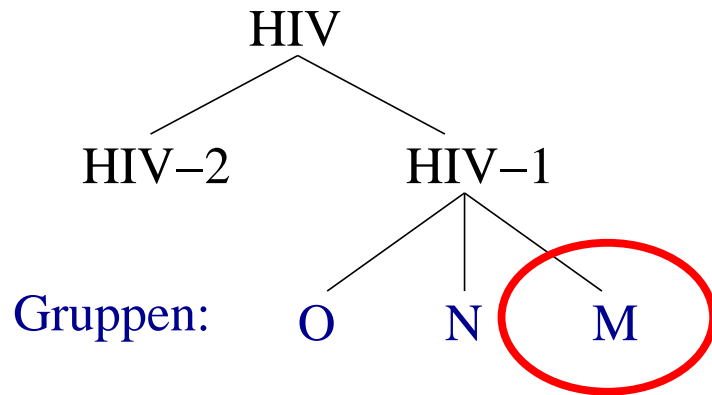


## Gruppe M:

- hauptverantwortlich für die globale Pandemie
- > 96% aller Infektionen



# Human Immunodeficiency Virus (HIV)

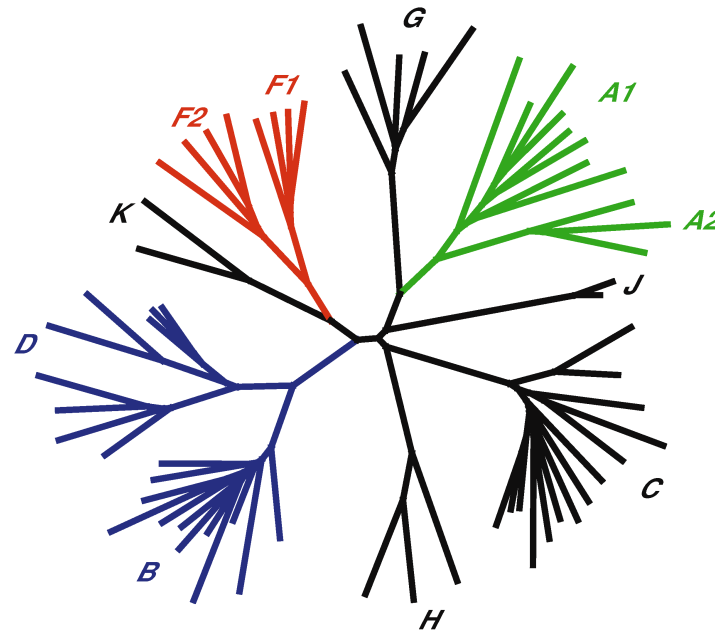


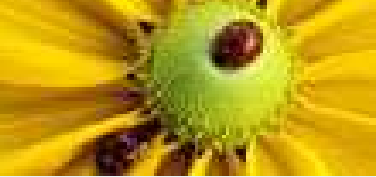
## Gruppe M:

- hauptverantwortlich für die globale Pandemie
- > 96% aller Infektionen

## Gruppe M:

Subtypen A-D, F-H, J, K  
Subsubtypen A1, A2, F1, F2





# Geographische Verteilung von Subtypen und CRFs

Verallg. HMMs

Profil-HMMs

Rekombination bei HIV

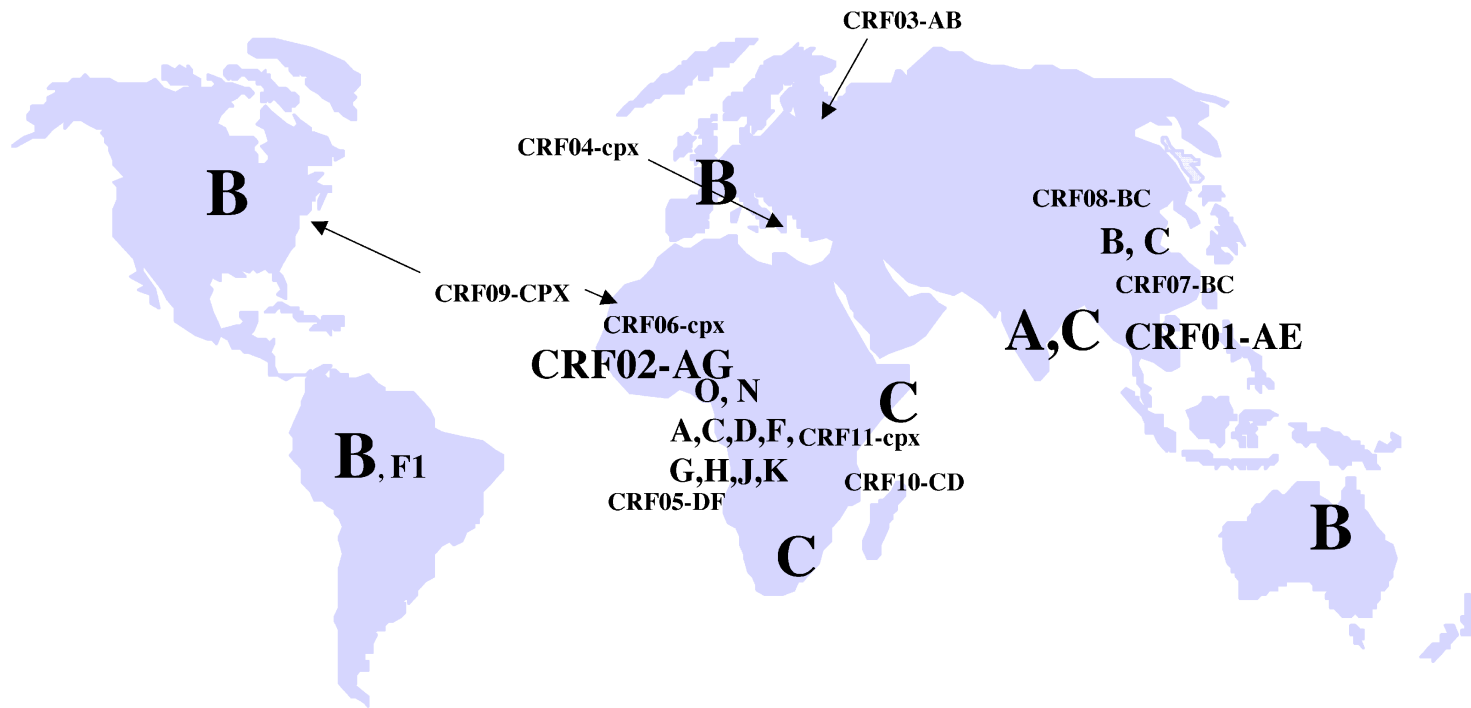
● HIV

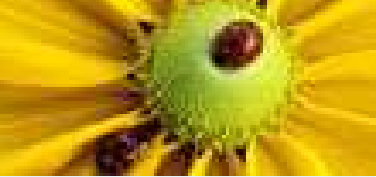
● geographische Verteilung

● CRFs

● Rekombination

jpHMM





# Circulating Recombinant Forms (CRFs)

Verallg. HMMs

---

Profil-HMMs

---

Rekombination bei HIV

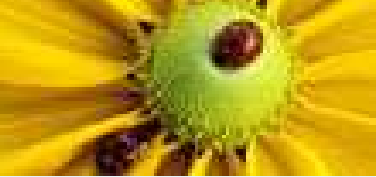
---

- HIV
- geographische Verteilung
- CRFs
- Rekombination

jpHMM

---

- Formen des HIV, die sich aus mehreren Stämmen rekombiniert und ausgebreitet haben
- entstehen durch Mehrfachinfektion eines Individuums
- zur Zeit 34 solcher CRFs bekannt



# Circulating Recombinant Forms (CRFs)

Verallg. HMMs

Profil-HMMs

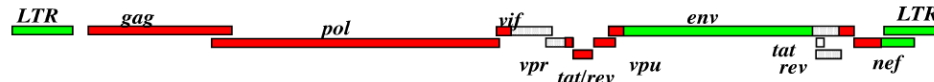
Rekombination bei HIV

- HIV
- geographische Verteilung
- CRFs
- Rekombination

jpHMM

## Beispiele:

**CRF01 A/E**



**CRF02 A/G**



**CRF03 A/B**



**CRF04 cpx**



**CRF05 D/F**



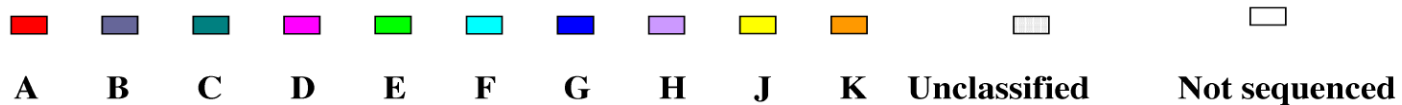
**CRF06 cpx**



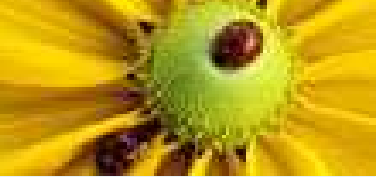
**CRF07-BC**



**CRF08-BC**



M. Peeters, Review, 2000



# Erkennung von Rekombination wichtig

Verallg. HMMs

---

Profil-HMMs

---

Rekombination bei HIV

---

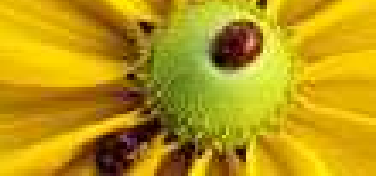
- HIV
- geographische Verteilung
- CRFs
- **Rekombination**

jpHMM

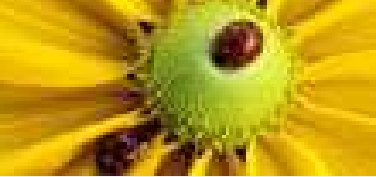
---

- Subtypen unterschiedliche Ausbreitungsraten, Fitness, Krankheitsverlauf
- CRFs teilweise fitter als Elternstämme
- Entwicklung von Impfstoffen und Anti-Retroviralen Medikamenten
- Verstehen der Dynamik der Ausbreitung





# jpHMM - ein “springendes” Profil-HMM



# Springendes Alignment

## Springendes Alignment (Spang et al., 2002)

Verallg. HMMs

---

Profil-HMMs

---

Rekombination bei HIV

---

jpHMM

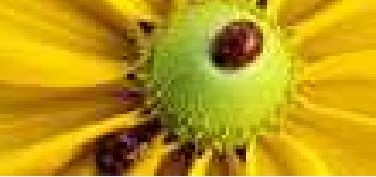
- Springendes Alignment
- jpHMM
- Decoding
- Beam-Search-Algorithmus
- Beam Search
- Beispielausgabe von jpHMM

Subtyp 1      **AATTG**  
                  **AA-TG**  
                  **AC-TT**

Subtyp 2      **ACATG**  
                  **CCACG**

---

Eingabesequenz      **ATTAAG**



# Springendes Alignment

## Springendes Alignment (Spang et al., 2002)

Subtyp 1     **A--ATTG**  
                 **A--A-TG**  
                 **A--C-TT**

Subtyp 2     **A--CATG**  
                 **C--CACG**

---

Eingabesequenz     **ATTAA-G**

- Eingabesequenz wird nur zu einem **Referenz**-Subtyp aligniert
- Der Referenz-Subtyp kann wechseln (“Sprung”)

Verallg. HMMs

Profil-HMMs

Rekombination bei HIV

jpHMM

● Springendes Alignment

● jpHMM

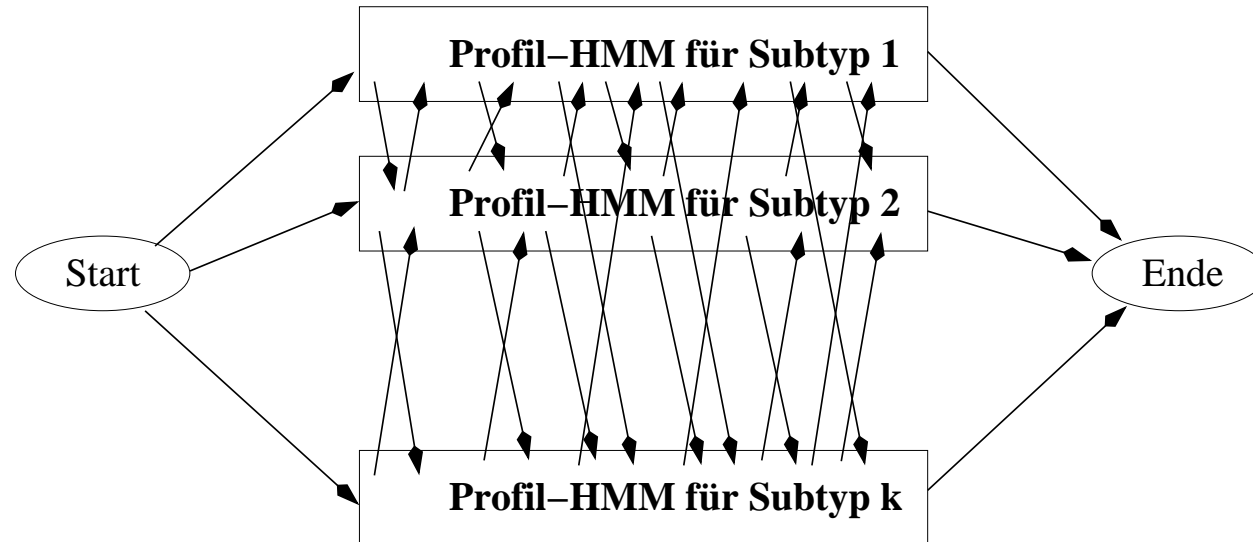
● Decoding

● Beam-Search-Algorithmus

● Beam Search

● Beispielausgabe von jpHMM

# Springendes Profil-HMM (jpHMM)



- jeweils ein Profil-HMM modelliert reine HIV-Sequenzen eines Subtyps
- Sprünge sind von jedem Subtypen zu jedem anderen an fast jeder Stelle erlaubt

Verallg. HMMs

Profil-HMMs

Rekombination bei HIV

jpHMM

● Springendes Alignment

● jpHMM

● Decoding

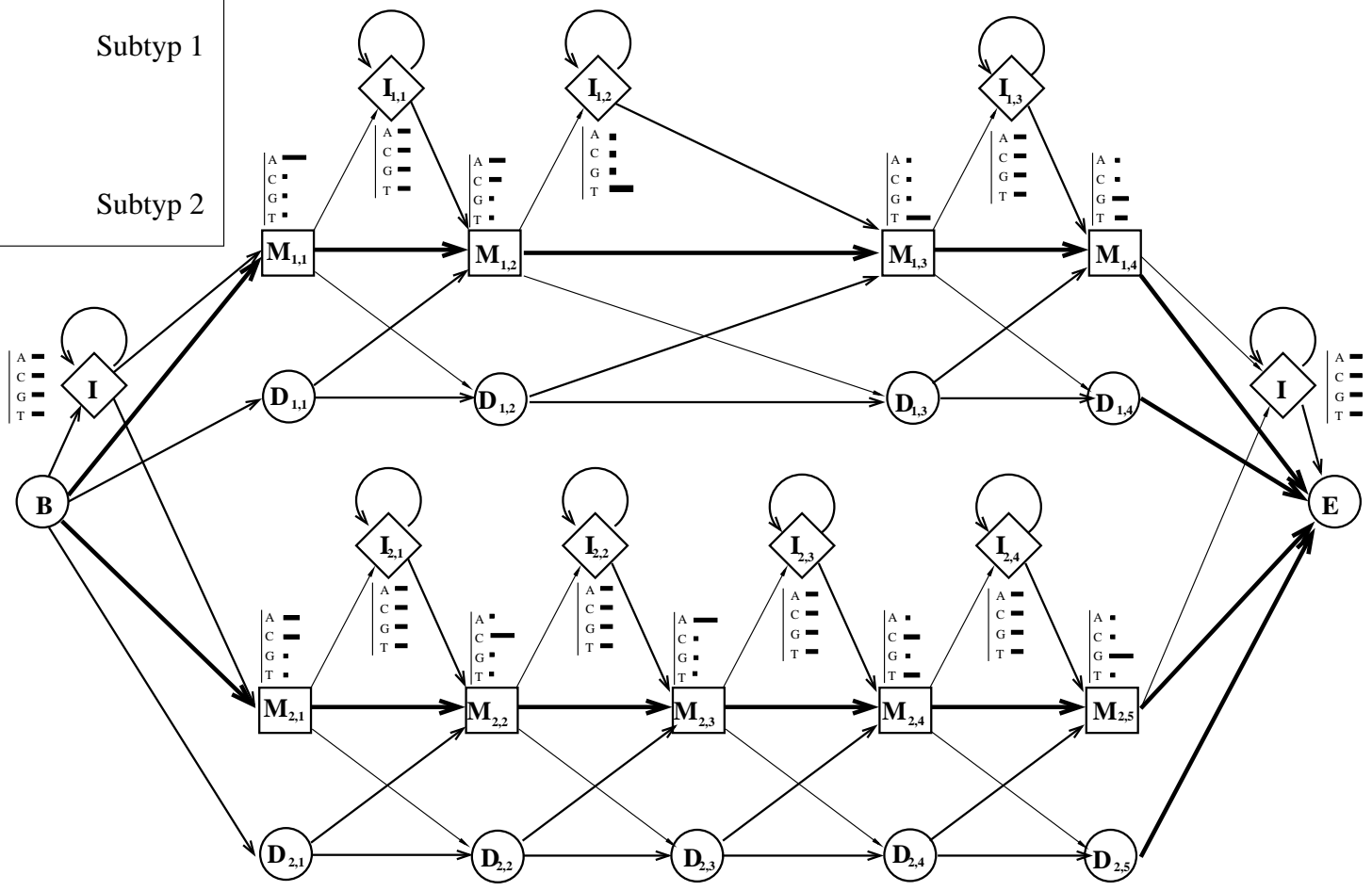
● Beam-Search-Algorithmus

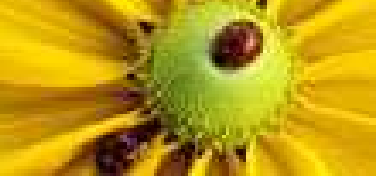
● Beam Search

● Beispielausgabe von jpHMM

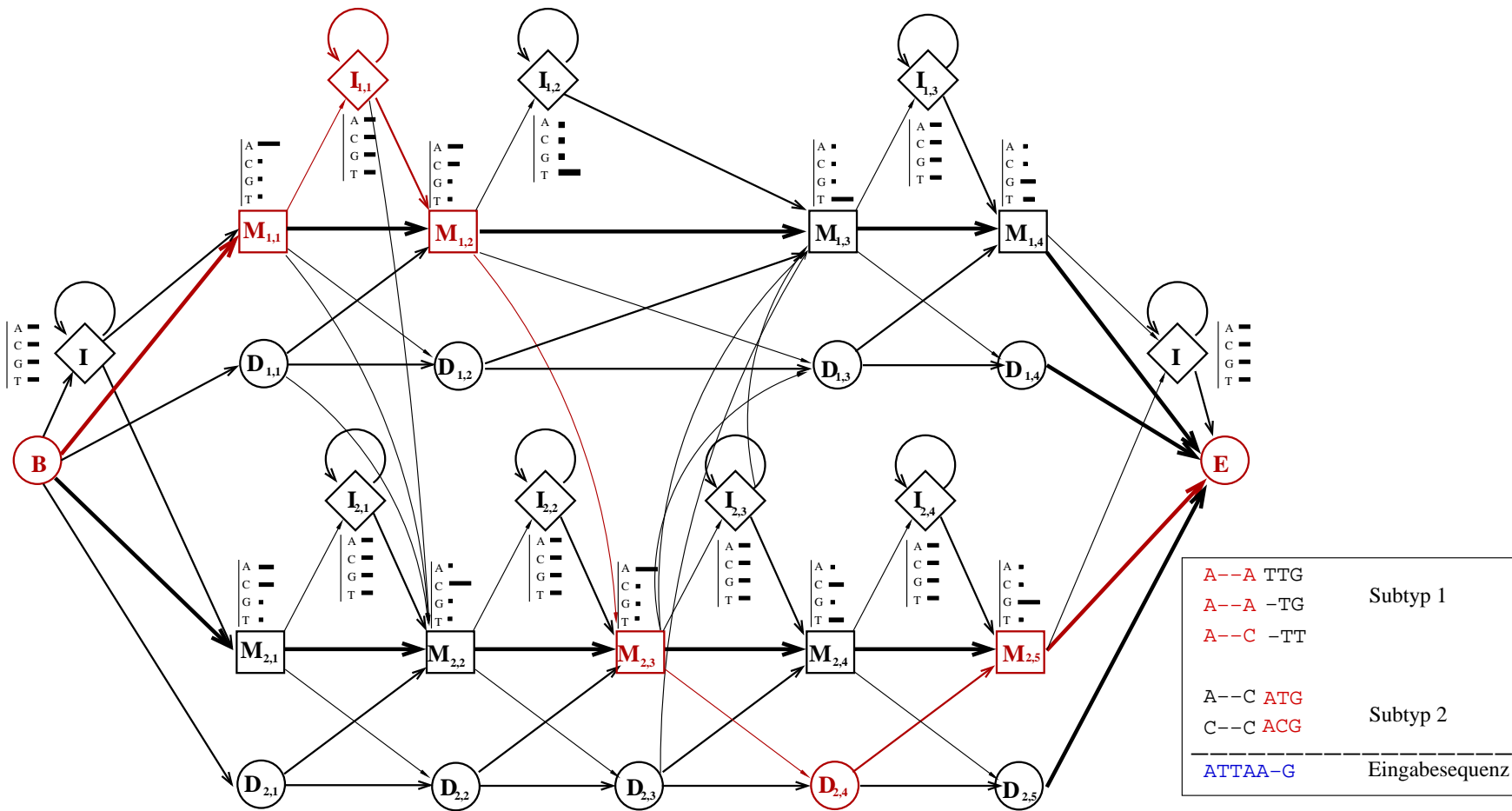
# Springendes Profil-HMM

|       |          |
|-------|----------|
| AATTG | Subtyp 1 |
| AA-TG |          |
| AC-TT |          |
| ACATG | Subtyp 2 |
| CCACG |          |

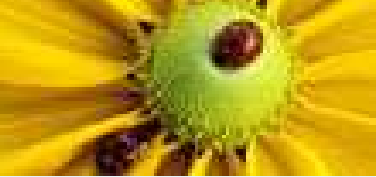




# Springendes Profil-HMM



Jeder **Pfad** entspricht einem springenden Alignment und definiert die Rekombinations-Bruchstellen und Eltern-Subtypen.



# Decoding the Jumping Profile HMM

Wollen den wahrscheinlichsten – oder wenigstens einen wahrscheinlichen Pfad –  $\hat{g}$  durch das HMM finden, gegeben die unklassifizierte HIV-Eingabesequenz  $\sigma$ .

Größe eines HIV Genoms:  $n \approx 10000$  Zeichen

Anzahl der Subtypen: 14

Anzahl der Zustände im HMM:  $|Z| \approx 10000 \cdot 14 \cdot 3 = 420000$

↪ Zu groß für den Viterbi-Algorithmus.

Verwenden heuristische Beschleunigung um Speicherplatz und Zeit und sparen:

## Beam-Search-Algorithmus

Idee: Berechne die Viterbi-Rekursion *nicht* auf einer (dynamisch ermittelten) Teilmenge von unwahrscheinlichen Zustand/Position-Kombinationen.

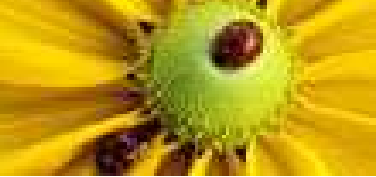
Verallg. HMMs

Profil-HMMs

Rekombination bei HIV

jpHMM

- Springendes Alignment
- jpHMM
- Decoding
- Beam-Search-Algorithmus
- Beam Search
- Beispielausgabe von jpHMM



# Beam-Search-Algorithmus

modifizierte Viterbi-Variable  $\gamma'_{q,t} \leq \gamma_{q,t}$

wird in jedem Schritt  $t$  nur für Teilmenge  $\mathcal{A}_t$  von Zuständen gespeichert.

$$\mathcal{A}_t = \{q \mid \gamma'_{q,t} \geq B\gamma_t^*\} \quad (8)$$

Dabei ist

$$\gamma_t^* = \max_q \gamma'_{q,t}$$

und  $0 < B \ll 1$  der “Beam”.

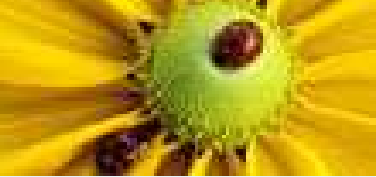
$\mathcal{A}_t$ : Menge der **aktiven** Zustände.

$\gamma'_{q,t} := 0$ , falls  $q \notin \mathcal{A}_t$ ,

d.h. die modifizierten Viterbi-Variablen der inaktiven Zustände werden auf 0 gesetzt und brauchen nicht gespeichert zu werden.

Im nächsten Schritt  $t + 1$  der Rekursion brauchen die modifizierten Viterbi-Variablen  $\gamma'_{q,t+1}$  nur von Zuständen berechnet zu werden, die von Zuständen in  $\mathcal{A}_t$  durch einen Pfad mit nur einer Emission erreichbar sind.





# Beam search algorithm

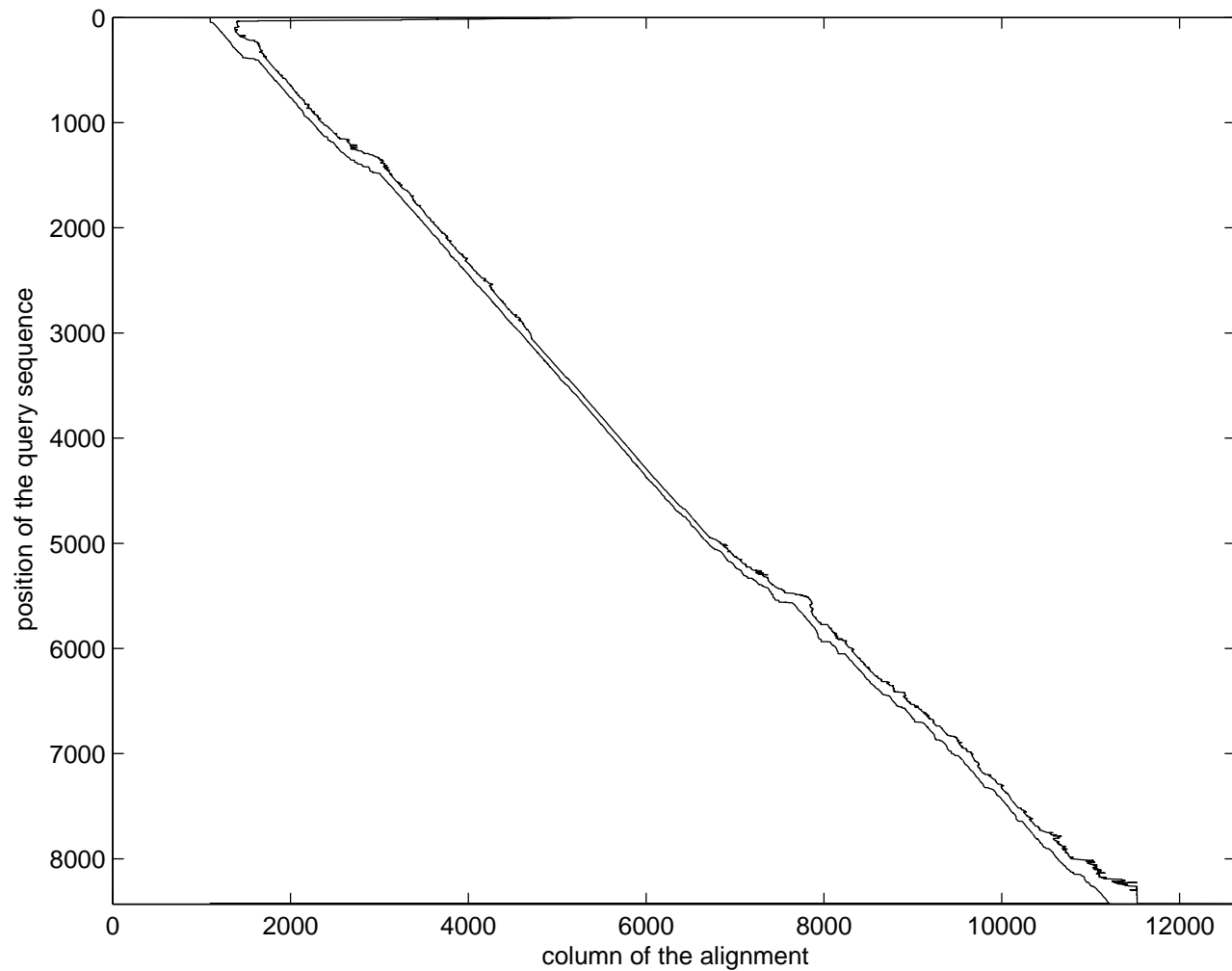
Verallg. HMMs

Profil-HMMs

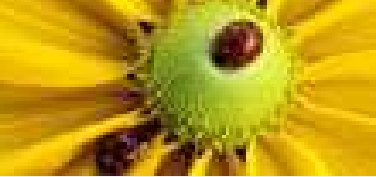
Rekombination bei HIV

jpHMM

- Springendes Alignment
- jpHMM
- Decoding
- Beam-Search-Algorithmus
- Beam Search
- Beispielausgabe von jpHMM



Spalten, die aktive Zustände enthalten bei  $B = 10^{-20}$ .  
Durchschnittliche Anzahl aktiver Zustände: ca. 2000



# Beam search algorithm

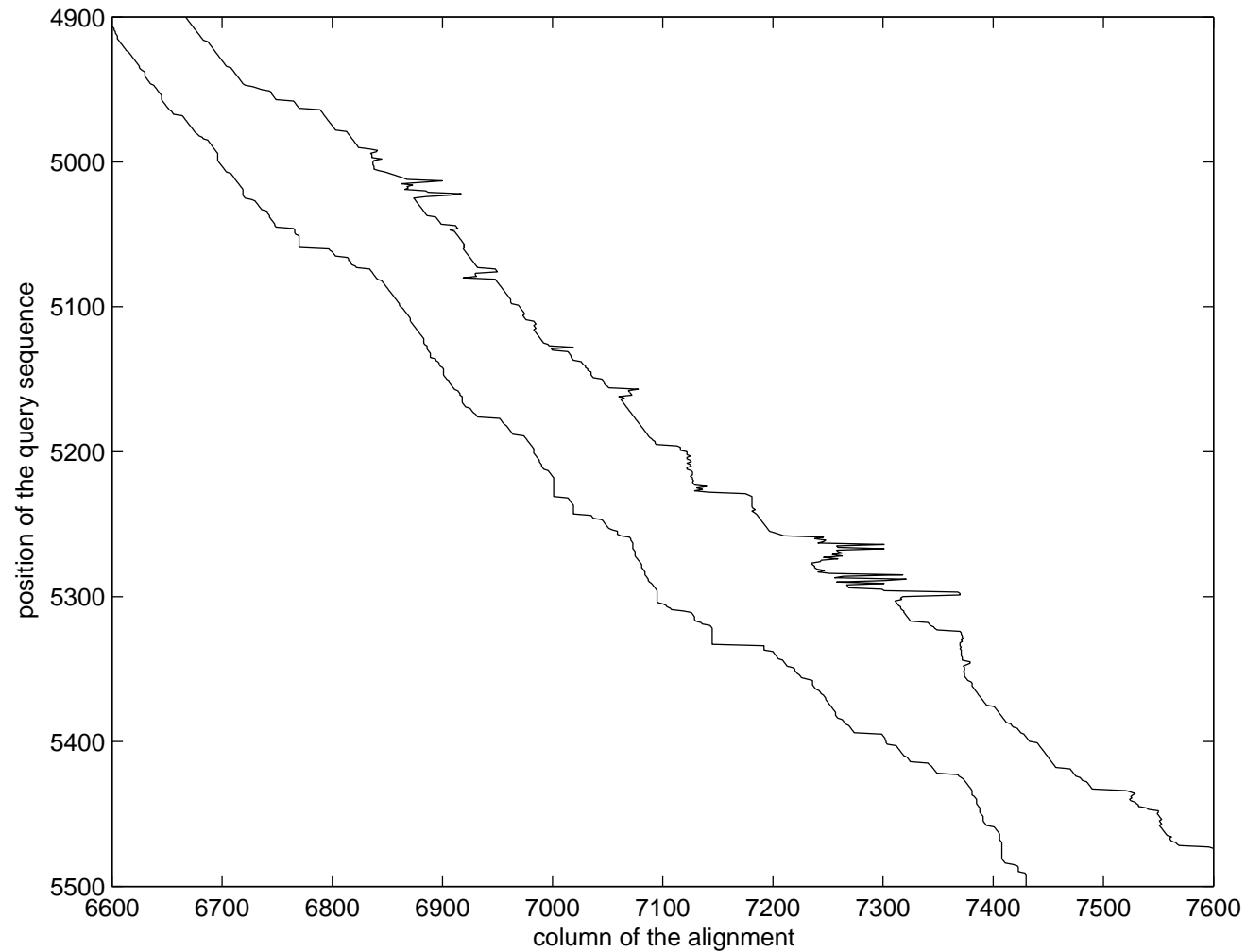
Verallg. HMMs

Profil-HMMs

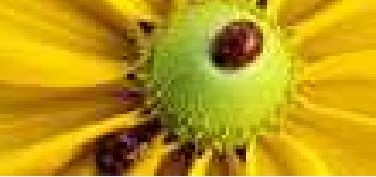
Rekombination bei HIV

jpHMM

- Springendes Alignment
- jpHMM
- Decoding
- Beam-Search-Algorithmus
- Beam Search
- Beispielausgabe von jpHMM



Spalten, die aktive Zustände enthalten bei  $B = 10^{-20}$ .  
Durchschnittliche Anzahl aktiver Zustände: ca. 2000



# Beispielausgabe von jpHMM

Verallg. HMMs

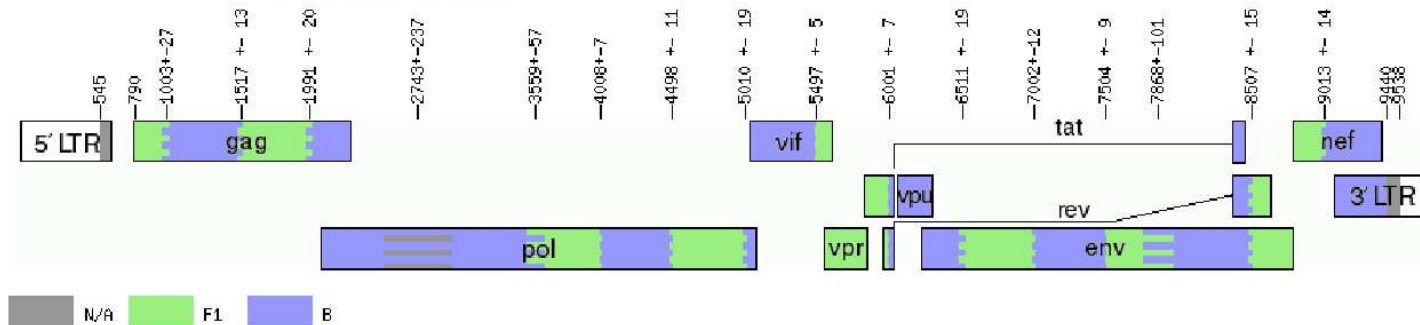
Profil-HMMs

Rekombination bei HIV

jpHMM

- Springendes Alignment
- jpHMM
- Decoding
- Beam-Search-Algorithmus
- Beam Search
- Beispielausgabe von jpHMM

Genome map (based on [HXB2 numbering](#))



Note:

- Numbers in the above figure denote intervals for recombination breakpoints based on HXB2 numbering.
- The uncolored regions denote missing information due to input fragment sequence.
- The gray regions denote missing information due to uninformative subtype models (subtype: N/A).
- The sequence regions of less than 10 nucleotides long are too short to be mapped onto the genome map.

Posterior probabilities of the subtypes (based on [HXB2 numbering](#))

