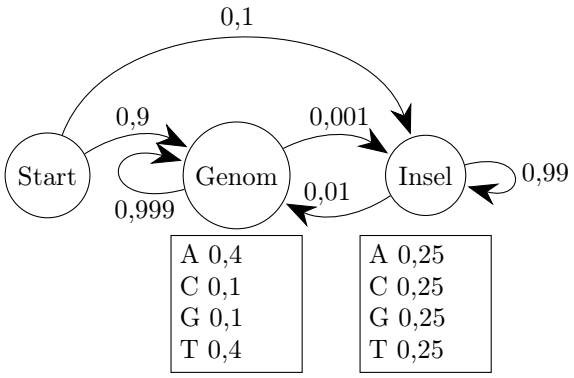
1 HMM für CpG-Inseln



2 HMM-Begriffe

- Zustand / state
  - $S_1, \ldots, S_n$
- Startzustand / start state
  - S_1
- Übergangswahrscheinlichkeiten / Transition probabilities
  - $T(i, j)$ (von S_i zu S_j)
  - z.B. $T(\text{Genom}, \text{Insel}) = 0{,}001$
- Ausgabealphabet / Emission alphabet
  - z.B. {A,C,G,T}
- Ausgabewahrscheinlichkeiten / Emission probabilities
  - $E(i, c)$ (in S_i, Nuk. c)
  - z.B. $E(\text{Genom}, \text{A}) = 0{,}4$
- Schweigezustände / silent states
- Endzustand / end state
- Pfad / path : Sequenz von Zuständen

3 Viterbi-Algorithmus

- Eingabe: HMM, Seq. x
- Ausgabe: Die Wahrscheinlichkeit des optimalen Pfads, d.h. $\max_\pi \Pr(x \text{ und } \pi|\text{HMM})$ (π ist ein Pfad)
  - Mit Backtracking können wir den optimalen Pfad bestimmen, wie beim Needleman-Wunsch-Algorithmus.
- $A(i, k)$: Wahrscheinlichkeit des optimalen Pfads von S_1 nach S_n, der $x_1 x_2 \cdots x_k$ ausgibt

1: **for** $k = 1,2, \ldots, |x|$ **do**
2: $A(1, k) = 0$
3: **end for**
4: $A(1,0) = 1$
5: **for** $i = 2,3, \ldots, n$ **do**
6: $A(i,0) = 0$
7: **end for**
8: **for** $k = 1, \ldots, |x|$ **do**
9: **for** $i = 2, \ldots, n$ **do**
10: $A(i, k) = \max_{j \in \{1,\ldots,n\}} A(j, k-1) \cdot T(j, i) \cdot E(i, x_k)$
11: **end for**
12: **end for**
13: Wkt. des optimalen Pfads $= \max_{i=2}^{n} A(i, |x|)$

3.1 Backtracking

1: $i = \text{argmax}_i \, A(i, |x|)$
2: $k = |x|$
3: **while** $k > 0$ **do**
4: Nuk. x_k hat Zustand S_i
5: $i = \text{argmax}_{j \in \{1,\ldots,n\}} A(j, k-1) \cdot T(j, i) \cdot E(i, x_k)$
6: $k = k - 1$
7: **end while**

3.2 Logarithmen / Scores

- kann beliebige Scores zum Addieren benutzen, z.B. +1 und -1
- $A(i, k) = \log_2$ Wkt. optimalen Pfads

1: **for** $k = 1,2, \ldots, |x|$ **do**
2: $A(1, k) = -\infty$
3: **end for**
4: $A(1,0) = 0$
5: **for** $i = 2,3, \ldots, n$ **do**
6: $A(i,0) = -\infty$
7: **end for**
8: **for** $k = 1, \ldots, |x|$ **do**
9: **for** $i = 2, \ldots, n$ **do**
10: $A(i, k) = \max_{j \in \{1,\ldots,n\}} A(j, k-1) + T(j, i) + E(i, x_k)$
11: **end for**
12: **end for**
13: Wkt. des optimalen Pfads $= \max_{i=2}^{n} A(i, |x|)$

4 Forward-Algorithmus

Der Viterbi-Algorithmus in Abschnitt 3 ändert sich ein bisschen, um der Forward-Algorithmus zu werden: In Zeilen 10 und 13 wird "max" zu "$\sum$":

10: $A(i, k) = \sum_{j \in \{1,\ldots,n\}} A(j, k-1) \cdot T(j, i) \cdot E(i, x_k)$

13: Wkt. aller Pfade $= \sum_{i=2}^{n} A(i, |x|)$

Und das war es schon.

5 Parametereinschätzung mit bekannter Zustandssequenz

- Eingabe:
  - Menge von Zuständen
  - Ausgabealphabet
  - Menge von Trainingssequenzen mit Zustand für jedes Nuk
- Ausgabe:
  - Übergangs- und Ausgabewahrscheinlichkeiten
- Maximum-Likelihood-Methode

6 HMM mit Dinukleotiden

Hier ist das HMM für CpG-Inseln mit Dinukleotiden.

Tabelle der Zustände (alle Ausgabewkt. sind 1 für die aufgelisteten Nukleotide)

Zustand	**Ausgaben**
Start	*(keine)*
gC	C
gG	G
iC	C
iG	G
gCC	C
gCG	G
gGC	C
gGG	G
iCC	C
iCG	G
iGC	C
iGG	G

Visuelle Darstellung des HMMs

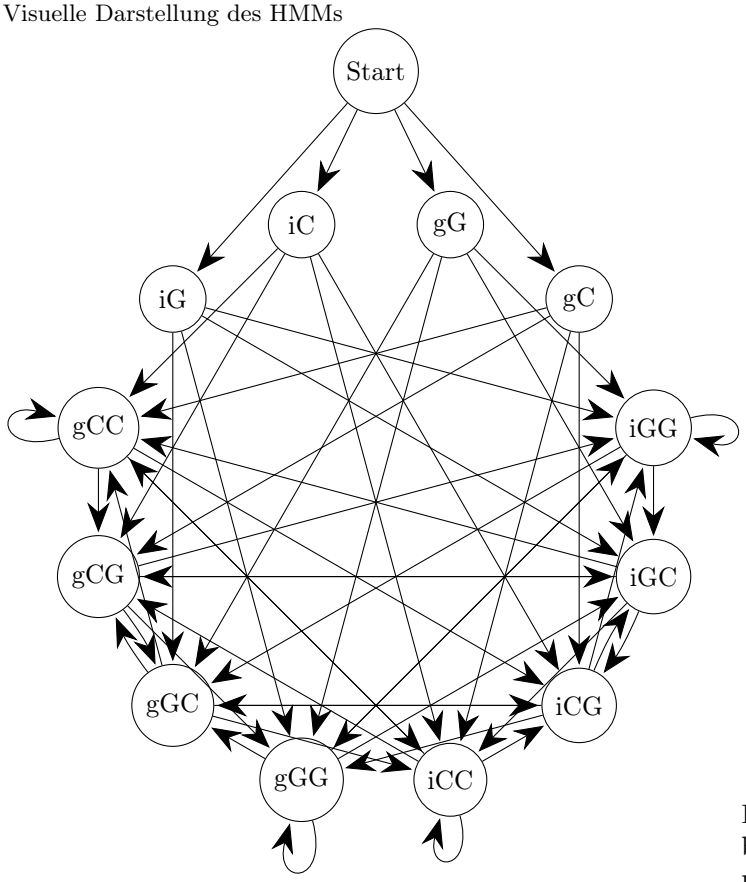


Tabelle der Übergänge $T(i, j)$

Zustände von	**auf**	**Wahrscheinlichkeit (Wkt.)**
Start	gC	$Pr(g) \cdot \Pr(C)$
Start	gG	$Pr(g) \cdot \Pr(G)$
Start	iC	$Pr(i) \cdot \Pr(C)$
Start	iG	$Pr(i) \cdot \Pr(G)$
gC	gCC	$\Pr(g \to g) \cdot \Pr(C\|g, C)$
gC	gCG	$\Pr(g \to g) \cdot \Pr(G\|g, C)$
gC	iCC	$\Pr(g \to i) \cdot \Pr(C\|g, C)$
gC	iCG	$\Pr(g \to i) \cdot \Pr(G\|g, C)$
gG	gGC	$\Pr(g \to g) \cdot \Pr(C\|g, G)$
gG	gGG	$\Pr(g \to g) \cdot \Pr(G\|g, G)$
gG	iGC	$\Pr(g \to i) \cdot \Pr(C\|g, G)$
gG	iGG	$\Pr(g \to i) \cdot \Pr(G\|g, G)$
iC	gCC	$\Pr(i \to g) \cdot \Pr(C\|i, C)$
iC	gCG	$\Pr(i \to g) \cdot \Pr(G\|i, C)$
iC	iCC	$\Pr(i \to i) \cdot \Pr(C\|i, C)$
iC	iCG	$\Pr(i \to i) \cdot \Pr(G\|i, C)$
iG	gGC	$\Pr(i \to g) \cdot \Pr(C\|i, G)$
iG	gGG	$\Pr(i \to g) \cdot \Pr(G\|i, G)$
iG	iGC	$\Pr(i \to i) \cdot \Pr(C\|i, G)$
iG	iGG	$\Pr(i \to i) \cdot \Pr(G\|i, G)$
gCC	gCC	$Pr(g \to g) \cdot \Pr(C\|g, C)$
gCC	gCG	$Pr(g \to g) \cdot \Pr(G\|g, C)$
gCG	gGC	$Pr(g \to g) \cdot \Pr(C\|g, G)$
gCG	gGG	$Pr(g \to g) \cdot \Pr(G\|g, G)$
gCC	iCC	$Pr(g \to i) \cdot \Pr(C\|g, C)$
gCC	iCG	$Pr(g \to i) \cdot \Pr(G\|g, C)$
gCG	iGC	$Pr(g \to i) \cdot \Pr(C\|g, G)$
gCG	iGG	$Pr(g \to i) \cdot \Pr(G\|g, G)$
gGC	gCC	$Pr(g \to g) \cdot \Pr(C\|g, C)$
gGC	gCG	$Pr(g \to g) \cdot \Pr(G\|g, C)$
gGG	gGC	$Pr(g \to g) \cdot \Pr(C\|g, G)$
gGG	gGG	$Pr(g \to g) \cdot \Pr(G\|g, G)$
gGC	iCC	$Pr(g \to i) \cdot \Pr(C\|g, C)$
gGC	iCG	$Pr(g \to i) \cdot \Pr(G\|g, C)$
gGG	iGC	$Pr(g \to i) \cdot \Pr(C\|g, G)$
gGG	iGG	$Pr(g \to i) \cdot \Pr(G\|g, G)$
iCC	gCC	$Pr(i \to g) \cdot \Pr(C\|i, C)$
iCC	gCG	$Pr(i \to g) \cdot \Pr(G\|i, C)$
iCG	gGC	$Pr(i \to g) \cdot \Pr(C\|i, G)$
iCG	gGG	$Pr(i \to g) \cdot \Pr(G\|i, G)$
iCC	iCC	$Pr(i \to i) \cdot \Pr(C\|i, C)$
iCC	iCG	$Pr(i \to i) \cdot \Pr(G\|i, C)$
iCG	iGC	$Pr(i \to i) \cdot \Pr(C\|i, G)$
iCG	iGG	$Pr(i \to i) \cdot \Pr(G\|i, G)$
iGC	gCC	$Pr(i \to g) \cdot \Pr(C\|i, C)$
iGC	gCG	$Pr(i \to g) \cdot \Pr(G\|i, C)$
iGG	gGC	$Pr(i \to g) \cdot \Pr(C\|i, G)$
iGG	gGG	$Pr(i \to g) \cdot \Pr(G\|i, G)$
iGC	iCC	$Pr(i \to i) \cdot \Pr(C\|i, C)$
iGC	iCG	$Pr(i \to i) \cdot \Pr(G\|i, C)$
iGG	iGC	$Pr(i \to i) \cdot \Pr(C\|i, G)$
iGG	iGG	$Pr(i \to i) \cdot \Pr(G\|i, G)$

$\Pr(C|g, G)$ soll heißen: die Wkt. ein C-Nukleotid auszugeben, gegeben, dass der vorherige Zustand g=Genom war, und dass das vorherige Nukleotid G war.