Esempio HMM CpG: Forward, Backward e Posteriori (prime 3 posizioni)

Modello HMM (Isole CpG) visto durante il corso di Genomica Computazionale, Corso di Laurea Specialistica in Genomica Funzionale, 2025-2026.

Fabrizio Mafessoni

Stati nascosti: nonCpG, CpG
Transizioni:
P(nonCpG→nonCpG)=0.9, P(nonCpG→CpG)=0.1
P(CpG→nonCpG)=0.5, P(CpG→CpG)=0.5
Emissioni:
P(AT|nonCpG)=0.98, P(CG|nonCpG)=0.02
P(AT|CpG)=0.20, P(CG|CpG)=0.80
Probabilità iniziali: π(nonCpG)=π(CpG)=0.5
Sequenza osservata (prime 3 posizioni): AT, AT, CG
(La sequenza completa usata per i backward è: AT AT CG CG CG AT AT AT CG AT AT.)

Forward α (prime 3 posizioni)

Definizione: α_t(s) = P(o_1, ..., o_t, z_t = s)
Ricorrenza: α_t(s) = P(o_t | z_t = s) × Σ_{s'} α_{t-1}(s') · P(z_t = s | z_{t-1} = s')

Posizione t = 1 (osservazione AT)

$\alpha_1(\text{nonCpG}) = \pi(\text{nonCpG}) \cdot P(\text{AT}|\text{nonCpG})$
$= 0.5 \cdot 0.98$
$= 0.49$

$\alpha_1(\text{CpG}) = \pi(\text{CpG}) \cdot P(\text{AT}|\text{CpG})$
$= 0.5 \cdot 0.20$
$= 0.10$

Posizione t = 2 (osservazione AT)

$\alpha_2(\text{nonCpG}) = P(\text{AT}|\text{nonCpG}) \times [\alpha_1(\text{nonCpG}) \cdot P(\text{nonCpG}\to\text{nonCpG})$
$+ \alpha_1(\text{CpG}) \cdot P(\text{CpG}\to\text{nonCpG})]$
$= 0.98 \times [0.49 \cdot 0.9 + 0.10 \cdot 0.5]$
$= 0.98 \times [0.441 + 0.05]$
$= 0.98 \times 0.491$
≈ 0.48118

$\alpha_2(\text{CpG}) = P(\text{AT}|\text{CpG}) \times [\alpha_1(\text{nonCpG}) \cdot P(\text{nonCpG}\to\text{CpG})$
$+ \alpha_1(\text{CpG}) \cdot P(\text{CpG}\to\text{CpG})]$

$= 0.20 \times [0.49 \cdot 0.1 + 0.10 \cdot 0.5]$
$= 0.20 \times [0.049 + 0.05]$
$= 0.20 \times 0.099$
$= 0.0198$

Posizione t = 3 (osservazione CG)

$$\begin{aligned}\alpha_3(\text{nonCpG}) &= P(\text{CG}|\text{nonCpG}) \times [\alpha_2(\text{nonCpG}) \cdot P(\text{nonCpG} \to \text{nonCpG}) \\ &\quad + \alpha_2(\text{CpG}) \cdot P(\text{CpG} \to \text{nonCpG})] \\ &= 0.02 \times [0.48118 \cdot 0.9 + 0.0198 \cdot 0.5] \\ &= 0.02 \times [0.433062 + 0.0099] \\ &= 0.02 \times 0.442962 \\ &\approx 0.00885924\end{aligned}$$

$$\begin{aligned}\alpha_3(\text{CpG}) &= P(\text{CG}|\text{CpG}) \times [\alpha_2(\text{nonCpG}) \cdot P(\text{nonCpG} \to \text{CpG}) \\ &\quad + \alpha_2(\text{CpG}) \cdot P(\text{CpG} \to \text{CpG})] \\ &= 0.80 \times [0.48118 \cdot 0.1 + 0.0198 \cdot 0.5] \\ &= 0.80 \times [0.048118 + 0.0099] \\ &= 0.80 \times 0.058018 \\ &\approx 0.0464144\end{aligned}$$

Backward β (posizioni t = 3,2,1)

Definizione: β_t(s) = P(o_{t+1}, ..., o_T | z_t = s)
Ricorrenza: β_t(j) = Σ_k P(z_{t+1}=k | z_t=j) · P(o_{t+1}|k) · β_{t+1}(k)
Qui mostriamo i calcoli per t=3,2,1 (usando la sequenza completa).

Dalla sequenza completa si ottiene (calcolato una volta con la ricorrenza backward):
$\beta_4(\text{nonCpG}) \approx 0.00284896$
$\beta_4(\text{CpG}) \approx 0.01087190$

Posizione t = 3 (β_3, osservazione successiva = CG a t=4)

$$\begin{aligned}\beta_3(\text{nonCpG}) &= P(\text{nonCpG} \to \text{nonCpG}) \cdot P(\text{CG}|\text{nonCpG}) \cdot \beta_4(\text{nonCpG}) \\ &\quad + P(\text{nonCpG} \to \text{CpG}) \cdot P(\text{CG}|\text{CpG}) \cdot \beta_4(\text{CpG}) \\ &= 0.9 \cdot 0.02 \cdot 0.00284896 + 0.1 \cdot 0.80 \cdot 0.01087190 \\ &\approx 0.0000513 + 0.0008698 \\ &\approx 0.0009210\end{aligned}$$

$$\begin{aligned}\beta_3(\text{CpG}) &= P(\text{CpG} \to \text{nonCpG}) \cdot P(\text{CG}|\text{nonCpG}) \cdot \beta_4(\text{nonCpG}) \\ &\quad + P(\text{CpG} \to \text{CpG}) \cdot P(\text{CG}|\text{CpG}) \cdot \beta_4(\text{CpG}) \\ &= 0.5 \cdot 0.02 \cdot 0.00284896 + 0.5 \cdot 0.80 \cdot 0.01087190 \\ &\approx 0.0000285 + 0.0043488\end{aligned}$$

≈ 0.0043773

Posizione t = 2 (β_2, osservazione successiva = CG a t=3)

$$\beta_2(\text{nonCpG}) = P(\text{nonCpG}\rightarrow\text{nonCpG})\cdot P(\text{CG}|\text{nonCpG})\cdot\beta_3(\text{nonCpG}) + P(\text{nonCpG}\rightarrow\text{CpG})\cdot P(\text{CG}|\text{CpG})\cdot\beta_3(\text{CpG})$$
$$= 0.9\cdot 0.02\cdot 0.0009210 + 0.1\cdot 0.80\cdot 0.0043773$$
$$\approx 0.0000166 + 0.0003501$$
$$\approx 0.0003668$$

$$\beta_2(\text{CpG}) = P(\text{CpG}\rightarrow\text{nonCpG})\cdot P(\text{CG}|\text{nonCpG})\cdot\beta_3(\text{nonCpG}) + P(\text{CpG}\rightarrow\text{CpG})\cdot P(\text{CG}|\text{CpG})\cdot\beta_3(\text{CpG})$$
$$= 0.5\cdot 0.02\cdot 0.0009210 + 0.5\cdot 0.80\cdot 0.0043773$$
$$\approx 0.0000092 + 0.0017509$$
$$\approx 0.0017601$$

Posizione t = 1 (β_1, osservazione successiva = AT a t=2)

$$\beta_1(\text{nonCpG}) = P(\text{nonCpG}\rightarrow\text{nonCpG})\cdot P(\text{AT}|\text{nonCpG})\cdot\beta_2(\text{nonCpG}) + P(\text{nonCpG}\rightarrow\text{CpG})\cdot P(\text{AT}|\text{CpG})\cdot\beta_2(\text{CpG})$$
$$= 0.9\cdot 0.98\cdot 0.0003668 + 0.1\cdot 0.20\cdot 0.0017601$$
$$\approx 0.000324 + 0.0000352$$
$$\approx 0.000359$$

$$\beta_1(\text{CpG}) = P(\text{CpG}\rightarrow\text{nonCpG})\cdot P(\text{AT}|\text{nonCpG})\cdot\beta_2(\text{nonCpG}) + P(\text{CpG}\rightarrow\text{CpG})\cdot P(\text{AT}|\text{CpG})\cdot\beta_2(\text{CpG})$$
$$= 0.5\cdot 0.98\cdot 0.0003668 + 0.5\cdot 0.20\cdot 0.0017601$$
$$\approx 0.0001797 + 0.0001760$$
$$\approx 0.0003557$$

Posteriori γ_1, γ_2, γ_3

Definizione generale: $\gamma\_t(s) = P(z\_t = s \mid O) = \alpha\_t(s)\cdot\beta\_t(s) / P(O)$
dove $P(O) = \alpha\_T(\text{nonCpG}) + \alpha\_T(\text{CpG}) \approx 2.11327\times 10^{-4}$.

t = 1

$$\gamma_1(\text{nonCpG}) = \alpha_1(\text{nonCpG})\cdot\beta_1(\text{nonCpG}) / P(O)$$
$$\approx (0.49 \cdot 0.000359) / 2.11327\times 10^{-4}$$
$$\approx 0.8317$$

$$\gamma_1(\text{CpG}) = \alpha_1(\text{CpG})\cdot\beta_1(\text{CpG}) / P(O)$$
$$\approx (0.10 \cdot 0.0003557) / 2.11327\times 10^{-4}$$

≈ 0.1683

t = 2

$\gamma_2(\text{nonCpG}) = \alpha_2(\text{nonCpG}) \cdot \beta_2(\text{nonCpG}) \,/\, P(O)$

$\approx (0.48118 \cdot 0.0003668) \,/\, 2.11327 \times 10^{-4}$

≈ 0.8351

$\gamma_2(\text{CpG}) \quad = \alpha_2(\text{CpG}) \cdot \beta_2(\text{CpG}) \,/\, P(O)$

$\approx (0.0198 \cdot 0.0017601) \,/\, 2.11327 \times 10^{-4}$

≈ 0.1649

t = 3

$\gamma_3(\text{nonCpG}) = \alpha_3(\text{nonCpG}) \cdot \beta_3(\text{nonCpG}) \,/\, P(O)$

$\approx (0.00885924 \cdot 0.0009210) \,/\, 2.11327 \times 10^{-4}$

≈ 0.0386

$\gamma_3(\text{CpG}) \quad = \alpha_3(\text{CpG}) \cdot \beta_3(\text{CpG}) \,/\, P(O)$

$\approx (0.0464144 \cdot 0.0043773) \,/\, 2.11327 \times 10^{-4}$

≈ 0.9614

Si vede che alle posizioni 1 e 2 (AT) la probabilità a posteriori di nonCpG è alta (~83–84%), mentre alla posizione 3 (CG) la probabilità di CpG è molto alta (~96%), in corrispondenza dell'osservazione di una potenziale CpG.