Substitučné modely

Askar Gafurov

14.11.2019

Substitučné modely, označenie

Nech $P(b|a,t)$ je pravdepodobnosť, že ak začneme s bázou a, tak po čase t budeme mať bázu b.

Matica pravdepodobností prechodu:

$$S(t) = \begin{pmatrix} P(A|A,t) & P(C|A,t) & P(G|A,t) & P(T|A,t) \\ P(A|C,t) & P(C|C,t) & P(G|C,t) & P(T|C,t) \\ P(A|G,t) & P(C|G,t) & P(G|G,t) & P(T|G,t) \\ P(A|T,t) & P(C|T,t) & P(G|T,t) & P(T|T,t) \end{pmatrix}$$

Substitučné modely, požiadavky

- $S(0) = I$
- $\lim_{t\to\infty} S(t) = \begin{pmatrix} \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \\ \pi_A & \pi_C & \pi_G & \pi_T \end{pmatrix}$

  Rozdelenie π nazývame limitné (equilibrium)
- $S(t_1 + t_2) = S(t_1)S(t_2)$ (multiplikatívnosť)
- Pre Jukes-Cantorov model by navyše malo platiť

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$S(t) = \begin{pmatrix} 1-3s(t) & s(t) & s(t) & s(t) \\ s(t) & 1-3s(t) & s(t) & s(t) \\ s(t) & s(t) & 1-3s(t) & s(t) \\ s(t) & s(t) & s(t) & 1-3s(t) \end{pmatrix}$$

$$S(2t) = S(t)^2 =$$

$$= \begin{pmatrix} 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 & 2s(t)-4s(t)^2 \\ 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 2s(t)-4s(t)^2 & 1-6s(t)+12s(t)^2 \end{pmatrix}$$

$$\approx \begin{pmatrix} 1-6s(t) & 2s(t) & 2s(t) & 2s(t) \\ 2s(t) & 1-6s(t) & 2s(t) & 2s(t) \\ 2s(t) & 2s(t) & 1-6s(t) & 2s(t) \\ 2s(t) & 2s(t) & 2s(t) & 1-6s(t) \end{pmatrix}$$

pre $t \to 0$

Matica rýchlostí (matica intenzít, substitution rate matrix)

- Matica rýchlostí pre Jukes-Cantorov model:
$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$
- Pre veľmi malý čas t je $S(t) \approx I + Rt$
- Rýchlosť α je pravdepodobnosť zmeny za jednotku času pre veľmi krátke t, resp. derivácia $s(t)$ vzhľadom na t v bode 0
- Riešením diferenciálnych rovníc pre Jukes-Cantorov model dostávame $s(t) = (1 - e^{-4\alpha t})/4$

Riešenie pre Jukes-Cantorov model

$$S(t) = \begin{pmatrix} (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 \end{pmatrix}$$

Matica rýchlostí sa zvykne normalizovať tak, aby na jednotku času pripadla v priemere jedna substitúcia, čo dosiahneme ak $\alpha = 1/3$

Substitučné matice, zhrnutie

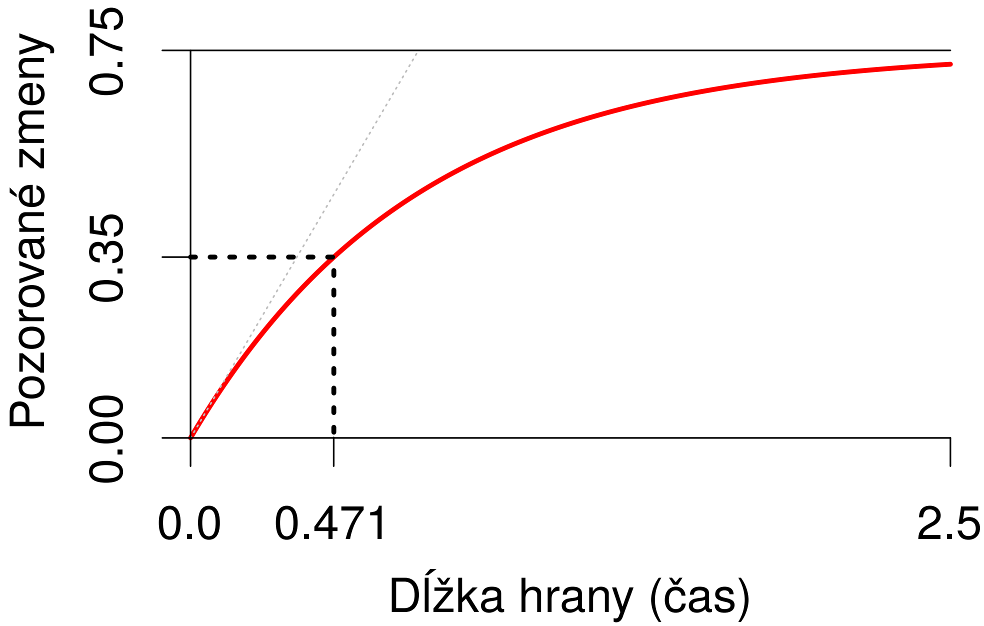
- $S(t)$: matica 4×4, kde políčko $S(t)_{a,b} = P(b|a,t)$ je pravdepodobnosť, že ak začneme s bázou a, tak po čase t budeme mať bázu b.
- Jukes-Cantorov model predpokladá, že $P(b|a,t)$ rovnaká pre každé $a \neq b$
- Pre daný čas t máme mimo diagonály $s(t)$, na diagonále $1 - 3s(t)$
- Matica rýchlostí R: pre J-C model mimo diagonály α, na diagonále -3α
- Pre veľmi malé t máme $S(t) \approx I - Rt$
- Rýchlosť α je pravdepodobnosť zmeny za jednotku času pre veľmi malé t, resp. derivácia $s(t)$ vzhľadom na t v bode $t = 0$
- Riešením diferenciálnych rovníc pre J-C model dostávame $s(t) = (1 - e^{-4\alpha t})/4$
- Matica rýchlostí sa zvykne normalizovať tak, aby na jednotku času pripadla v priemere jedna substitúcia, čo dosiahneme ak $\alpha = 1/3$

Korekcia evolučných vzdialeností

$\Pr(X_{t_0+t} = C \mid X_{t_0} = A) = \frac{1}{4}(1 - e^{-\frac{4}{3}t})$

Očakávaná frekvencia pozorovaných zmien na bázu za čas t:

$D(t) = \Pr(X_{t_0+t} \neq X_{t_0}) = \frac{3}{4}(1 - e^{-\frac{4}{3}t})$



Korekcia pozorovaných vzdialeností

$$D = \frac{3}{4}(1 - e^{-\frac{4}{3}t}) \qquad \Rightarrow \qquad t = -\frac{3}{4}\ln(1 - \frac{4}{3}D)$$

Zložitejšie modely

- Všeobecná matica rýchlostí R

$$R = \begin{pmatrix} \cdot & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ \mu_{CA} & \cdot & \mu_{CG} & \mu_{CT} \\ \mu_{GA} & \mu_{GC} & \cdot & \mu_{GT} \\ \mu_{TA} & \mu_{TC} & \mu_{TG} & \cdot \end{pmatrix}$$

- μ_{xy} je rýchlosť, akou sa báza x mení na inú bázu y
- Presnejšie $\mu_{xy} = \lim_{t\to 0} \frac{\Pr(y \mid x,t)}{t}$
- Diagonálu dopočítame tak, aby súčet každého riadku bol 0
- Existujú modely s menším počtom parametrov (kompromis medzi J-C a ľubovoľnou maticou)

Kimurov model

- Zachytáva, že puríny sa častejšie menia na iné puríny (A a G) a pyrimidíny na ine pyrimidíny (C a T)
- Má dva parametre: rýchlosť tranzícií α, transverzií β
- $R = \begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \beta \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix}$

HKY model (Hasegawa, Kishino, Yano)

- Rozšírenie Kimurovho modelu, umožňuje aj rôzne pravdepodobnosti A, C, G a T v limite (v ekvilibriu)
- Ak nastavíme čas v evolučnom modeli na nekonečno, nezáleží na tom, z ktorej bázy sme začali, frekvencia výskytu jednotlivých báz sa ustáli v tzv. ekvilibriu.
- V Jukes-Cantorovom modeli je pravdepodobnosť každej bázy v ekvilibriu 1/4.
- V HKY si zvolíme aj frekvencie jednotlivých nukleotidov v ekvilibriu $\pi_A, \pi_C, \pi_G, \pi_T$ so súčtom 1
- Parameter κ: pomer tranzícií a transverzií (α/β)
- Matica rýchlostí:

$$\mu_{x,y} = \begin{cases} \kappa\pi_y & \text{ak mutácia z } x \text{ na } y \text{ je tranzícia} \\ \pi_y & \text{ak mutácia z } x \text{ na } y \text{ je transverzia} \end{cases}$$

Od R k $S(t)$

- Pre zložité modely nevieme odvodiť explicitný vzorec na výpočet $S(t)$, ako sme mali pri Jukes-Cantorovom modeli
- Vo všeobecnosti $S(t) = e^{Rt}$
- Exponenciálna funkcia matice A sa definuje ako $e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$
- Ak R diagonalizujeme $R = UDU^{-1}$, kde D je diagonálna matica, tak $e^{Rt} = Ue^{Dt}U^{-1}$ a exponenciálnu funkciu uplatníme iba na prvky na uhlopriečke D
- Diagonalizácia vždy existuje pre symetrické R (na diagonále vlastné hodnoty)