

## Bayes Teoria, ratkaisut-2, (28.01.2011)

Oletetaan että diploidi yksilö on heterotsygottinen lokuksissa  $l = 1, 2, 3$ , eli niissä geenikohdissa hänen perimä koostu pareista

$$l = 1 \quad G_1 = \{A, a\}$$

$$l = 2 \quad G_2 = \{B, b\}$$

$$l = 3 \quad G_3 = \{C, c\}$$

jossa alleli-tyypit  $A \neq a$ ,  $B \neq b$  ja  $C \neq c$ . Oletetaan että lokuksen järjestyks on  $1 \leftrightarrow 2 \leftrightarrow 3$ .

Lokusten väliset rekombinatio todennäköisyydet ovat  $\rho(1, 2) = 0.1$  ja  $\rho(2, 3) = 0.2$ .

## Kysymys 1

Olettamalla rekombinaatioiden riippumattomuus eri geenien välissä, laske todennäköisyys että yhdessä gameetissa ei tapahdu rekombinaatiota lokusten 1,2,3: välissä

$$\mathbf{R.} \quad (1 - \rho_{1,2})(1 - \rho_{2,3}) = 0.9 \times 0.8 = 0.72$$

## Kysymys 1

Olettamalla rekombinaatioiden riippumattomuus eri geenien välissä, laske todennäköisyys että yhdessä gameetissa ei tapahdu rekombinaatiota lokusten 1,2,3: välissä

$$\mathbf{R.} \quad (1 - \rho_{1,2})(1 - \rho_{2,3}) = 0.9 \times 0.8 = 0.72$$

## Kysymys 2

Laske todennäköisyys että yhdessä gameetissa tapahtuu kahta rekombinaatiota (rekombinaatio välissä [1, 2] ja rekombinaatio välissä [2, 3]).

$$\mathbf{R.} \quad \rho_{1,2}\rho_{2,3} = 0.02.$$

Näistä genotyypeistä voi muodostaa  $2^3 = 8$  haplotyyppiä. Kyseisestä yksilöltä on kerätty  $N = 100$  gametteja, eli siemeniä jotka mitattiin laboratoriossa. Siemen sisältää yhden haplotyyppi joka on koottu yksilön haplotyyppi-parista rekombinaatio-todennäköisyyksien mukaisesti.

Eri rekombinantteja esintyi siemenien datassa kappaleita

$$\#(a, b, c) = 7$$

$$\#(a, b, C) = 3$$

$$\#(a, B, c) = 8$$

$$\#(a, B, C) = 37$$

$$\#(A, b, c) = 34$$

$$\#(A, b, C) = 7$$

$$\#(A, B, c) = 0$$

$$\#(A, B, C) = 4$$

**Kysymys 4** Datan perusteella mikä on yksilön todennäköisimmin haplotyyppaus ?

**Kysymys 5** Laske posteriori todennäköisyys datan ehdolla että yksilön haplotyyppit ovat  $(a, B, C)$  ja  $(A, b, c)$ .

**Kysymys 6** Oletetaan nyt että rekombinaatio todennäköisyydet  $\rho(1, 2), \rho(2, 3)$  ovat tuntemattomia. Silloin datan todennäköisyys on  $\rho(1, 2), \rho(2, 3)$  funktio. Miten lähtisit laskemaan parametrin  $(\rho(1, 2), \rho(2, 3))$  suurimman uskottavuuden estimaattori ?

**Vihjeet** Yksilön haplotyytit ovat komplementaariset, siksi on 4 eri vaihtoehtoa, jotka ovat a priori yhtä todennäköisiä.

I:  $(a, B, C)$  ja  $(A, b, c)$ , vai

II:  $(a, b, c)$  ja  $(A, B, C)$ , vai

III:  $(a, b, C)$  ja  $(A, B, c)$ , vai

IV:  $(a, B, c)$  ja  $(A, b, C)$ .



Jokaiselle vaihtoehdoille laske ensin datan ehdollinen todennäköisyys rekombinaatio todennäköisyyden perusteella, ja sitten laske posterioria Bayesin kaavalla.  
Taksulaskin tai tietokone tule tarpeen.

## Ratkaisut

Olkoon  $H \in \{(I), (II), (III), (IV)\}$  yksilön oikea haplotyyppi ja  $X = (X_1, X_2, X_3)$  gametti.

Laskemme ensin mallin avulla  $P(X|H)$ .

Esimerkiksi

$$\begin{aligned} P(X = (a, B, c)|H = (I)) &= \\ P(X_1 = a|H = (I))P(X_2 = B|X_1 = a, H = (I)) &\times \\ \times P(X_3 = c|X_1 = a, X_2 = B, H = (I)) & \\ = P(X_1 = a|H = (I))P(X_2 = B|X_1 = a, H = (I)) &\times \\ \times P(X_3 = c|X_1 = a, H = (I)) & \\ \frac{1}{2}(1 - \rho_{12})\rho_{23} & \end{aligned}$$

siis todennäköisyys puolella on aloitettu kopiomaan lokuksessa 1 haplotyyppiä ( $aBC$ ) (toinen vaihtoehto silloin kun  $H = (I)$  olisi ollut ( $Abc$ ), lokusten 1 ja 2 välissä ei tapahtunut rekombinaatiota todennäköisyydellä  $(1 - \rho_{12})$  ja lokusten 2 ja 3 välissä on tapahtunut rekombinaatiota todennäköisyydellä  $\rho_{23}$ .

Yksinkertaisuuden vuoksi poikeataan alkuperäisestä tehtävästä ja oletetaan  $\rho_{12} = \rho_{23} = \rho$ .  
Jokaiselle gameetille joka esiintyy aineistossa lasketaan montako rekombinaatioita olisi pitänyt tapahtua gameetti-kopioinnin prosessissa jokaisessa tilanteissa  $H = \{(I), (II), (III), (IV)\}$ .

	(I)	(II)	(III)	(IV)
(a ,b ,c)	1	0	1	2
(A ,B ,C)	1	0	1	2
(a, B ,c )	1	2	1	0
(A ,b ,C)	1	2	1	0
( a, B ,C)	0	1	2	1
(A ,b ,c)	0	1	2	1
( a, b ,C)	2	1	0	1

Olkoon  $0 < \rho < 1/2$ , (silloin  $(1 - \rho) > \rho$ ), esimerkiksi  $\rho = 0.1, 1 - \rho = 0.9$ .

$$\begin{aligned} P(\text{data}|I) &= \\ &2^{-N} (\rho(1 - \rho))^{\#(abc)+\#(ABC)+\#(aBc)+\#(AbC)} \\ &\times (1 - \rho)^{2(\#(aBC)+\#(Abc))} \rho^{2\#(abC)} \\ &= 2^{-100} (\rho(1 - \rho))^{26} (1 - \rho)^{142} \rho^6 \\ &= 2^{-100} \rho^{32} (1 - \rho)^{168} = \exp(-160.69) \end{aligned}$$

$$\begin{aligned}
P(\text{data}|\Pi) &= \\
&2^{-N} (\rho(1-\rho))^{\#(aBC)+\#(Abc)+\#(abC)} \times \\
&\times (1-\rho)^{2(\#(abc)+\#(ABC))} \rho^{2(\#(aBc)+\#(AbC))} \\
&= 2^{-100} (\rho(1-\rho))^{74} (1-\rho)^{22} \rho^{30} \\
&= 2^{-100} \rho^{104} (1-\rho)^{96} = \exp(-318.89)
\end{aligned}$$

$$\begin{aligned}
P(\text{data} | (III)) &= \\
&2^{-N} (\rho(1 - \rho))^{\#(abc) + \#(ABC) + \#(aBc) + \#(AbC)} \times \\
&\times (1 - \rho)^{2(\#(abC))} \rho^{2(\#(aBC) + \#(Abc))} \\
&= 2^{-100} (\rho(1 - \rho))^{26} (1 - \rho)^6 \rho^{142} \\
&2^{-100} \rho^{168} (1 - \rho)^{32} = \exp(-459.52)
\end{aligned}$$



$$\begin{aligned}
P(\text{data} | (IV)) &= \\
&2^{-N} (\rho(1 - \rho))^{\#(aBC) + \#(Abc) + \#(abC)} \times \\
&\times (1 - \rho)^{2(\#(aBc) + \#(AbC))} \rho^{2(\#(abc) + \#(ABC))} \\
&= 2^{-100} (\rho(1 - \rho))^{74} (1 - \rho)^{30} \rho^{22} \\
&= 2^{-100} \rho^{96} (1 - \rho)^{104} = \exp(-301.32)
\end{aligned}$$

Koska apriori  $P(I) = P(II) = P(III) = P(IV) = 1/4$  ja  $P(data|I)$  on paljon suurempi kuin muiden vaihtoehtojen uskottavuuksien, voidaan heti sanoa että  $P(I|data) \sim 1$ .  
Lasketaan:

$$\begin{aligned} P(data) &= \\ &P(I)P(data|I) + P(II)P(data|II) + \\ &+ P(III)P(data|III) + P(IV)P(data|IV) \\ &\sim P(I)P(data|I) \sim P(I)P(data|I) + 0 \\ &= 0.4 \times \exp(-160.69) + 0 \end{aligned}$$

$$P(I|data) \sim \frac{P(I)P(data|I)}{P(I)P(data|I) + 0} = 1$$

Käytännössä on varma että vaihtoehto (I) on tosi, kun  $\rho = 0.1$  (oletetusti tunnettu).

Kun  $\rho$  on tuntematon, datan todennäköisyys on polynomi

$$P(\text{data}|\rho) = 2^{-102} \times \left( \rho^{32}(1 - \rho)^{168} + \rho^{104}(1 - \rho)^{96} + \rho^{168}(1 - \rho)^{32} + \rho^{96}(1 - \rho)^{104} \right)$$

Voitaisiin laskea numeerisesti polynomiaalisen derivaatan nolla-kohtia ja vertailla  $P(\text{data}|\rho)$  arvoja.

Käytännössä kun  $0 < (0.5 - \rho)$  on tarpeeksi suuri  
 $P(\text{data}|\rho) \sim P(I)P(\text{data}|I, \rho)$ , jossa  
 $P(\text{data}|I, \rho) \propto \rho^{32}(1 - \rho)^{168}$ ,  
Kun tämä maksimoidaan seuraa että suurimman  
uskottavuuden estimaattori on

$$\hat{\rho} \sim \frac{32}{(168 + 32)} = 0.16$$