

Sammenheng mellom rekombinasjonsrate og genetisk avstand.

Jarle Tufto

21. august 2004

Dersom vi betrakter n loci, synes det klart at det må finnes en sammenheng mellom de $n(n-1)/2$ rekombinasjonsratene mellom lociene. La oss i første omgang betrakte 3 loci, $i = 1, 2, 3$ og definer hendelsen

$$R_{ij} : \text{Rekombinasjon mellom loci } i \text{ og } j, \quad (1)$$

altså hendelsen at det dannes en kjønnsgamet hvor allelene på lociene i og j stammer fra ulike kromosom i kromosomparret.

Da vil hendelsen at det inntreffer rekombinasjon mellom 1 og 3 være

$$R_{13} = R_{12} \cap \bar{R}_{23} \cup \bar{R}_{12} \cap R_{23} \quad (2)$$

altså rekombinasjon mellom 1 og 2 og ikke mellom 2 og 3, eller ikke mellom 1 og 2 men mellom 2 og 3.

Antar vi at R_{12} og R_{23} er uavhengige og definerer rekombinasjonsratene $r_{ij} = P(R_{ij})$ får vi at

$$\begin{aligned} r_{13} &= P(\{R_{12} \cap \bar{R}_{23}\} \cup \{\bar{R}_{12} \cap R_{23}\}) \\ &= P(R_{12} \cap \bar{R}_{23}) + P(\bar{R}_{12} \cap R_{23}) \\ &= P(R_{12})P(\bar{R}_{23}) + P(\bar{R}_{12})P(R_{23}) \\ &= r_{12}(1 - r_{23}) + (1 - r_{12})r_{23} \\ &= r_{12} + r_{23} - 2r_{12}r_{23}. \end{aligned} \quad (3)$$

Det er altså generelt ikke slik at vi kan legge sammen rekombinasjonsrater; den totale rekombinasjonsraten blir mindre en summen.

Generelt innser vi at rekombinasjon mellom to loci er det samme som at et odde antall overkryssninger inntreffer mellom lociene under meiosen. Lar vi den stokastiske variabelen X betegne antallet slike overkryssninger er altså

$$r_{ij} = P(R_{ij}) = P(X \text{ odde}) = P(X = 1 \cup X = 3 \cup \dots) \quad (4)$$

Hvis vi antar at sannsynligheten for at en overkrysning skjer på et lite interval $z, z+\delta z$ av de to kromosomene under meiosen er lik $\lambda(z)\delta z$ og at overkrysning på forskjellige slike små interval inntreffer uavhengig blir det totale antall overkrysninger X på hele intervallet mellom lociene Poissonfordelt. Haldane (1919) var den første til å gjøre bruk av denne modellen.

La parameteren d betegne forventningen $E(X)$ i denne Poissonfordelingen. Rekombinasjonsraten kan da skrives som

$$\begin{aligned} r &= P(X = 1 \cup X = 3 \cup \dots) \\ &= \sum_{i=1,3,5,\dots}^{\infty} \frac{d^i e^{-d}}{i!} \\ &= e^{-d} \sum_{i=1,3,5,\dots}^{\infty} \frac{d^i}{i!}. \end{aligned} \tag{5}$$

Vi kan finne et litt enklere uttrykk for denne litt spesielle uendelige summen (en rekke). Legg først merke til at dersom alle leddene i summen over hadde vært med, ville summen vært lik Taylor-utviklingen av eksponentialfunksjonen

$$e^d = \sum_{i=0}^{\infty} \frac{d^i}{i!} = 1 + d + \frac{d^2}{2!} + \frac{d^3}{3!} + \dots \tag{6}$$

Hadde alle ledd vært med ville vi dermed fått at summen av alle Poisson-sannsynlighetene ble 1.

For å finne summen hvor bare odde ledd inngår ser vi på Taylor-utviklingen av e^{-d} som blir den alternerende rekken

$$e^{-d} = \sum_{i=0}^{\infty} \frac{(-d)^i}{i!} = 1 - d + \frac{d^2}{2!} - \frac{d^3}{3!} + \dots \tag{7}$$

Trekker vi (7) fra (6) faller de like leddene bort og vi står igjen med

$$\begin{aligned} e^d - e^{-d} &= 2\left(d + \frac{d^3}{3!} + \frac{d^5}{5!} + \dots\right) \\ &= 2 \sum_{i=1,3,5,\dots}^{\infty} \frac{d^i}{i!}, \end{aligned} \tag{8}$$

altså 2 ganger summen i (5). Setter vi (8) inn i (5) får vi etter en liten omskriving den enkle formelen

$$r = \frac{1}{2}(1 - e^{-2d}). \tag{9}$$

Denne formelen gir oss sammenhengen mellom rekombinasjonssannsynligheten r og forventet antall overkryssinger d mellom to loci, første gang vist av Haldane (1919). Størrelsen d , det forventede antallet overkryssinger, er et mål på *genetisk avstand* mellom de to lociene. Selv om dette egentlig er en dimensjonsløs størrelse oppgis gjerne slike genetiske avstander i måleenheten Morgan eller i centiMorgan (cM) hvor én Morgan svarer til en avstand på én forventet overkryssing. Når d blir stor ser vi at r går mot $1/2$; det blir som forventet like sannsynlig med et odde som et like antall overkryssinger.

Ved hjelp av (9) kan vi regne oss fram og tilbake mellom rekombinasjonsrater og genetiske avstander. Et vesentlig poeng er at de genetiske avstandene som svarer til gitte rekombinasjonsrater blir additive. Dette følger av setningen om at forventningen til en sum alltid er lik summen til forventningen. Kjenner vi de genetiske avstandene d_{12} og d_{23} beregnet fra r_{12} og r_{23} ved hjelp av (9) blir dermed den totale genetiske avstanden mellom loci 1 og 3

$$d_{13} = E(X_{12} + X_{23}) = E(X_{12}) + E(X_{23}) = d_{12} + d_{23}, \quad (10)$$

hvor X_{ij} er antall overkryssinger mellom loci i og j . Sammenhengen gitt ved (9) kalles i litteraturen en “mapping function” og spiller en sentral rolle i statistiske modeller som brukes for å estimere såkalte koblingskart (linkage maps).

Merk at det ikke er noen lineær sammenheng mellom genetisk avstand og fysisk avstand (målt i antall basepar i DNA-sekvensen). Dette skyldes at overkryssningsintensiteten $\lambda(z)$ langs kromosomet ikke er uniform – vi har en inhomogen Poissonprosess. Typisk svarer en avstand på 10^6 basepar (10Mbp) til en genetisk avstand på omkring $0.01\text{M} = 1\text{cM}$.

Referanser

Haldane, J. B. S., 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**:299–309.