Udledning af Jukes-Cantor-modellen

Når man ser på baserne i DNA, er det muligt at den samme base optræder på samme plads efter en transition, men det kan også være at basen er skiftet ud med en af de 3 øvrige baser. Man kan opskrive en tabel der beskriver sandsynlighederne, som vist i figur 1.

Lodret har man baserne i generation t, og vandret har man baserne i generation t+1.

	A	C	T	G
A	p_{AA}	p_{AC}	p_{AT}	p_{AG}
C	p_{CA}	p_{CC}	p_{CT}	p_{CG}
T	p_{TA}	p_{TC}	p_{TT}	p_{TG}
G	p_{GA}	p_{GC}	p_{GT}	p_{GG}

Figur 1. Tabel over baser og sandsynligheden for en udskiftning/transition mellem baserne.

Der er en sandsynlighed p_{ii} for at basen er den samme, men der er en anden sandsynlighed p_{ij} for at basen skiftes ud i forbindelse med transitionen. Her er sandsynligheden ikke afhængig af hvilken base der skiftes til, men blot af at der sker et skift. 'i' er indekset for den base der var i generation t, og 'j' angiver en ny base efter transitionen.

Man definerer sandsynligheden for at der sker et skift, til at være 'α'. Så har man at summen af sandsynlighederne opfylder

$$p_{AA} + p_{AC} + p_{AT} + p_{AG} = 1,$$

og at sandsynligheden for et skift er uafhængig af hvilken base der skiftes til.

Dvs. $p_{AC} = p_{AT} = p_{AG} = \frac{\alpha}{3}$.

Sandsynligheden for at der ikke skiftes en base, er $p_{AA} = 1-\alpha$.

Som vist i figur 2, kan man ud fra ovenstående opstille en tabel med de pågældende sandsynligheder.

	A	C	T	G
A	$1-\alpha$	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$
C	$\frac{\alpha}{3}$	$1-\alpha$	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$
T	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$	$1-\alpha$	$\frac{\alpha}{3}$
G	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$	$\frac{\alpha}{3}$	$1-\alpha$

Figur 2. Sandsynligheder for skift af baser ved mutation.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Sandsynligheden for at der ikke kommer et skift i base i 1. transition, er:

$$p_{ii}(1) = p_{AA}(1) = 1 - \alpha$$

Bemærk at man beskriver sandsynlighederne som funktion af transition (=generation) som *t*.

Sandsynligheden for skift i base i 1. transition er:

$$p_{ij}(1) = p_{AC}(1) = \frac{\alpha}{3}.$$

Sandsynligheden for at der ikke kommer et skift i samme base i 2. transition, beregnes ved at man benytter hhv. sandsynligheden for ingen skift og sandsynlighederne for to skift, hvor man kommer tilbage til udgangspunktet.

Det svarer til at man tager sandsynlighederne i rækken med A (gul) i figur 3, og man ganger det med sandsynlighederne i søjlen for A (grøn) i figur 4.

	A	C	T	G
A	p_{AA}	p_{AC}	p_{AT}	p_{AG}
C	p_{CA}	p_{CC}	p_{CT}	p_{CG}
T	p_{TA}	p_{TC}	p_{TT}	p_{TG}
G	p_{GA}	p_{GC}	p_{GT}	p_{GG}

Figur 3. Sandsynligheder for udskiftning af baser. Med gult er markeret de teoretiske sandsynligheder for ændringen af A til én af de andre baser eller tilbage til A.

	A	C	T	G
A	p_{AA}	p_{AC}	p_{AT}	p_{AG}
C	p_{CA}	p_{CC}	p_{CT}	p_{CG}
T	p_{TA}	p_{TC}	p_{TT}	p_{TG}
G	p_{GA}	p_{GC}	p_{GT}	p_{GG}

Figur 4. Sandsynligheder for et skift fra én af de andre baser tilbage til A og fra A tilbage til A.

Sandsynligheden for et skift fra basen A og tilbage til A på to transitioner og sandsynligheden for at der ikke sker en transition for A over to transitioner, kan udregnes ved at addere produkterne af hver begivenhed.

$$p_{AA}(2) = p_{AA} \cdot p_{AA} + p_{AC} \cdot p_{CA} + p_{AT} + p_{TA} + p_{AG} \cdot p_{GA}$$

Herefter indsættes værdierne vist i figur 2 for hver sandsynlighed:

$$p_{AA}(2) = (1 - \alpha) \cdot (1 - \alpha) + \left(\frac{\alpha}{3}\right)^2 + \left(\frac{\alpha}{3}\right)^2 + \left(\frac{\alpha}{3}\right)^2$$

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Udtrykkes reduceres herefter:

$$p_{AA}(2) = (1-\alpha)^2 + 3 \cdot \left(\frac{\alpha}{3}\right)^2$$

Hvis man skal finde sandsynligheden for at A skifter til C, tager man sandsynlighederne i rækken med A (gul) vist i figur 3, og ganger dem med sandsynlighederne i søjlen for C (rød) vist i figur 5.

	A	C	T	G
A	p_{AA}	p_{AC}	p_{AT}	p_{AG}
C	p_{CA}	p_{CC}	p_{CT}	p_{CG}
T	p_{TA}	p_{TC}	p_{TT}	p_{TG}
G	p_{GA}	p_{GC}	p_{GT}	p_{GG}

Figur 5. Med rødt er markeret sandsynligheden for at en base skifter til C, og at C forbliver den samme efter en transition.

Som vist ovenfor, beregnes de mulige skift:

$$p_{AC}(2) = p_{AA} \cdot p_{AC} + p_{AC} \cdot p_{CC} + p_{AT} \cdot p_{TC} + p_{AG} \cdot p_{GC}$$

Herefter indsættes værdierne vist i figur 2 for hver sandsynlighed:

$$p_{AC}(2) = (1-\alpha) \cdot \frac{\alpha}{3} + \frac{\alpha}{3} \cdot (1-\alpha) + \left(\frac{\alpha}{3}\right)^2 + \left(\frac{\alpha}{3}\right)^2$$

Udtrykket reduceres:

$$p_{AC}(2) = 2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$$

Bemærk at sandsynlighederne for at der ikke skiftes en base, er ens for de fire baser, og sandsynligheden for skift er ens for alle skiftene; de kan derfor beregnes på tilsvarende vis.

Sandsynlighederne efter to transitioner eller to generationer, beregnet på tilsvarende vis ved skift mellem alle baser og ved bevarelse af baser, er vist i figur 6.

$t = 2$	A	C	T	G
A	$(1-\alpha)^2 + 3 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$
C	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$(1-\alpha)^2 + 3 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$
T	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$(1-\alpha)^2 + 3 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$
G	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$2 \cdot (1-\alpha) \cdot \frac{\alpha}{3} + 2 \cdot \left(\frac{\alpha}{3}\right)^2$	$(1-\alpha)^2 + 3 \cdot \left(\frac{\alpha}{3}\right)^2$

Figur 6. Udregninger for skift og bevarelse af baser over to transitioner.

REGN MED BIOLOGI · ISBN 978-87-90363-98-7

Hvis man betragter sandsynligheden for ændring af baser eller bevarelse på samme måde som vist ovenfor i *t* generationer, så opnår man følgende sandsynligheder:

Sandsynligheden for at en base ikke er skiftet efter *t* transitioner/generationer, er:

(1) $p_{ii}(t) = \frac{1}{4} + \frac{3}{4} \cdot e^{-4\alpha t}$

Sandsynligheden for at en base er skiftet efter *t* transitioner/generationer, er:

(2) $p_{ij}(t) = \frac{1}{4} - \frac{1}{4} \cdot e^{-4\alpha t}$

Sandsynligheden for at starte i en tilfældig base A, C, T eller G er $\frac{1}{4}$. Af figur 2 kan man se at der forekommer 12 skift af baser med en sandsynlighed på $\frac{\alpha}{3}$. For hver base er sandsynligheden for at skifte til en anden base derfor $3 \cdot \frac{\alpha}{3} = \alpha$. Sandsynligheden for at have en transition et tilfældigt sted, er så:

$\frac{1}{4} \cdot 12 \cdot \alpha = 1$

$3\alpha = 1$

$\alpha = \frac{1}{3}$

Leddet der beskriver forlængelsen ud i *t* generationer ($e^{-4\alpha t}$), kan derfor omskrives til:

(3) $e^{-\frac{4}{3}t}$

Definition og estimat af *t*

Forskellen mellem to sekvenser relaterer direkte til den tid (*t*) der er gået siden de to sekvenser repræsenterer samme stamform. Den tid man måler, måles i forskelle mellem to sekvenser eller sagt med andre ord: *t* repræsenterer de faktiske antal ændringer mellem to eller flere sekvenser. Skal man finde *t*, skal man altså finde ud af hvordan en sekvens der er *n* baser lang (x_1, x_2, x_3, x_4, ... x_n), ændres til y_1, y_2, y_3, y_4, ... y_n. Det svarer til at man skal multiplicere sandsynlighederne for, at x_1 vedbliver med at være x_1, med sandsynligheden for at x_1 bliver til y_1 i løbet af tiden *t*. Det skal multipliceres med sandsynligheden for at x_2 vedbliver at være x_2, og multipliceres med sandsynligheden for at x_2 bliver til y_2 i løbet af tiden *t*. Dette bliver man ved med indtil x_n. Denne beregnede sandsynlighed kaldes *L* og skrives således:

$L = p(x_1) \cdot p(x_1 \rightarrow y_1|t) \cdot p(x_2) \cdot p(x_2 \rightarrow y_2|t) \cdot \ldots \cdot p(x_n) \cdot p(x_n \rightarrow y_n|t)$

Herefter tages den naturlige logaritme på begge sider af lighedstegnet:

$\ln(L) = \ln(p(x_1)) + \ln(p(x_1 \rightarrow y_1|t)) + \ln(p(x_2)) + \ln(p(x_2 \rightarrow y_2|t)) + \ldots + \ln(p(x_n)) + \ln(p(x_n \rightarrow y_n|t))$

Idet sandsynligheden for at bevare en base er ens ($\ln(p(x_1)$,+ $\ln(p(x_2)$, +...+ $\ln(p(x_n))$, kan de udtryk samles til en konstant. Det betyder at hele udtrykket kan skrives som:

$\ln(L) = \text{konstant} + \ln(p(x_1 \rightarrow y_1|t)) + \ln(p(x_2 \rightarrow y_2|t)) + \ldots + \ln(p(x_n \rightarrow y_n|t))$

Betragtes figur 6, kan man se at der er 12 sandsynligheder med samme værdi for et skift i baser – som man kan definere som m_1, og 4 sandsynligheder for at baserne bliver de samme. Den værdi kan defineres som m_2. Derfor fås:

$$\ln(L) = \text{konstant} + m_1 \cdot \ln\left(p_{ij}(t)\right) + m_2 \cdot \ln\left(p_{ii}(t)\right)$$

For at finde den maksimale sandsynlighed for disse ændringer differentieres udtrykket med hensyn til *t,* vha. reglen for differentiation af sammensat funktion, og det sættes så lig med 0.

$$\frac{d\left(\ln(L)\right)}{dt} = \frac{m_1}{p_{ij}(t)}\, p_{ij}'(t) + \frac{m_2}{p_{ii}(t)}\, p_{ii}'(t) = 0$$

Vi kender $p_{ii}(t)$ og $p_{ij}(t)$ fra henholdsvis ligning 1 og ligning 2. Tidsfaktoren t er udregnet i udtryk 3. Ligningerne og udtryk 3 erstattes i ovenstående ligning:

$$\frac{m_1}{\frac{1}{4} - \frac{1}{4}\, e^{-\frac{4}{3}t}} \cdot \frac{1}{3}\, e^{-\frac{4}{3}t} + \frac{m_2}{\frac{1}{4} - \frac{3}{4}\, e^{-\frac{4}{3}t}} \cdot -e^{-\frac{4}{3}t} = 0$$

For at gøre udtrykket mere overskueligt sættes $x = e^{-\frac{4}{3}t}$:

$$\frac{4m_1}{1 - x} \cdot \frac{x}{3} - \frac{4m_2}{1 + 3x} \cdot x = 0$$

$$4m_1x + 12m_1x^2 - 12m_2x + 12m_2x^2 = 0$$

$$(12m_1 + 12m_2)x^2 + (4m_1 - 12m_2)x = 0$$

$$x = \frac{12m_2 - 4m_1}{12m_1 + 12m_2} = \frac{3m_2 - m_1}{3m_1 + 3m_2}$$

x erstattes igen af $e^{-\frac{4}{3}t}$:

$$e^{-\frac{4}{3}t} = \frac{3m_2 - m_1}{3m_1 + 3m_2}$$

Den naturlige logaritme tages på begge sider af lighedstegnet:

$$-\frac{4}{3}t = \ln\left(\frac{3m_2 - m_1}{3m_1 + 3m_2}\right)$$

$$t = -\frac{3}{4}\ln\left(\frac{3m_2 - m_1}{3m_1 + 3m_2}\right)$$

Der lægges $3m_1$ til og trækkes $3m_1$ fra i tælleren:

$$t = -\frac{3}{4}\ln\left(\frac{3m_2 + 3m_1 - 3m_1 - m_1}{3m_1 + 3m_2}\right)$$

$$t = -\frac{3}{4}\ln\left(1 - \frac{4m_1}{3(m_1 + 3m_2)}\right)$$

$m_1 + m_2$ udgør dels den del af sekvensen hvor der er sket substitutioner, og dels den del hvor der ikke er sket substitutioner. Det svarer samlet til hele længden af sekvenserne. Derfor kan $m_1 + m_2$ erstattes med n som betegner længden af sekvenserne i et alignment.

$$t = -\frac{3}{4}\ln\left(1 - \frac{4m_1}{3n}\right)$$

Antal ændringer delt med sekvenslængden er p-afstanden. Det kan med de variabler der er anvendt i udtrykket, skrives som p-afstanden = $\frac{m_1}{n}$. Derved bliver Jukes-Cantor-afstanden (t)

$$t = -\frac{3}{4}\ln\left(1 - \frac{4}{3} \cdot p\right)$$

Eksempel: Er p-afstanden fundet i en afstandsmatrix, på 0,1000, er Jukes-Cantor-værdien (t) der korrigerer for tilbagemutationer.

$$t = -\frac{3}{4}\ln\left(1 - \frac{4}{3} \cdot 0{,}1000\right)$$

$t = 0{,}1073$

REGN MED BIOLOGI · ISBN 978-87-90363-98-7