Modeller for molekylær evolution i DNA-sekvenser

Af: Kresten Cæsar Torp, Aalborghus Gymnasium

Forudsætninger:
Du skal kende til DNA's opbygning og punktmutationer. Lav evt. opgaven intro til mutationer. Du skal også kende til de grundlæggende evolutionsmekanismer: Mutation og naturlig selektion.

I 1963 foreslog forskerne Emile Zuckerkandl og Linus Pauling i artiklen "Molecules as Documents of Evolutionary History", at man må kunne følge hvordan evolutionen er forløbet ved at sammenligne nukleotid- og aminosyresekvenser mellem arter og mellem individer indenfor arter. Et evolutionært forløb, mente de, vil kunne ses som en kæde af efterfølgende mutationer.
De mutationer der indtræffer, vil være uafhængige af hvilke mutationer der er gået forud. Derfor vil der opstå en række af sådanne kæder af mutationer, som resulterer i et træ med mange forgreninger, et fylogenetisk træ.

Pauling og Zuckerkandl benyttede proteinet hæmoglobin som eksempel. Pauling havde tidligere bestemt aminosyresekvensen af hæmoglobin og dets 3D-struktur. De foreslog, at hvis mutationer foregår tilfældigt, må man også kunne anvende antallet af forskelle i aminosyrer mellem et protein fra to arter som et molekylært ur til at datere, hvornår arter blev splittet op i evolutionen.

Men hvordan griber man det an? I denne opgave vil du lære om hvordan man i dag anvender forskellige modeller til at undersøge, hvilken molekylær evolution der er foregået. Du vil også lære om de forudsætninger og begrænsninger der knytter sig til modellerne.

Når vi opstiller modeller, laver vi en reduceret udgave af virkeligheden, med fokus på et bestemt aspekt af denne. Dvs. at vi tager et udsnit af det vi observerer, og koncentrerer os om det. Enhver model har derfor sit begrænsede anvendelsesområde, og den bygger på bestemte forudsætninger. Når man er bevidst om disse, kan man til gengæld få udbytte af at modellen til at beregne, fortolke og forudsige.
Her skal vi se på nogle af den mest enkle modeller der anvendes i molekylær evolution.

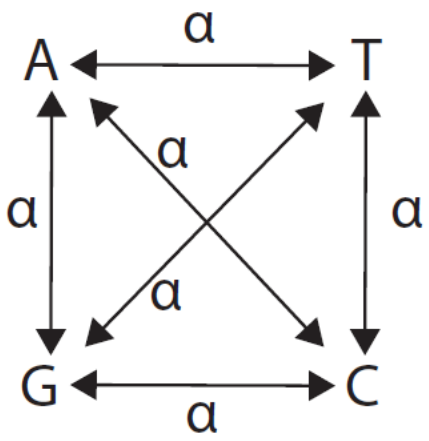
Indhold

novo nordisk fonden

Jukes-Cantor-modellen

Den nok mest enkle model for molekylær evolution er Jukes-Cantormodellen fra 1969. Den bygger på to forudsætninger:

- Den medtager kun substitutionsmutationer.
- Den antager, at alle substitutionsmutationer sker med samme sandsynlighed (angivet med sandsynighedsparameteren α på figuren).
- Den antager, at substitutionsmutationer indtræffer uafhængigt af hinanden, og uafhængigt af hvilke mutationer der er gået forud.
- Den antager, at alle nukleotider har samme sandsynlighed for over tid at mutere tilbage igen, fx: A→G→A.



Figur 1. Jukes-Cantors model

Jukes-Cantor-modellen, vist i figur 1, er enkel at anvende, og erfaringen er, at den i mange tilfælde er tilstrækkelig til at kortlægge den evolutionære forbindelse fx mellem arter eller mellem individer eller underpopulationer indenfor en art. Man kan fx benytte den til at beregne den genetiske afstand mellem DNA-sekvenser man undersøger. Det kan man fx anvende til at skabe overblik over evolutionære sammenhænge mellem arter, eller sammenhængene mellem forskellige stammer af virus indenfor en historisk sygdom eller en igangværende epidemi.

Opgave

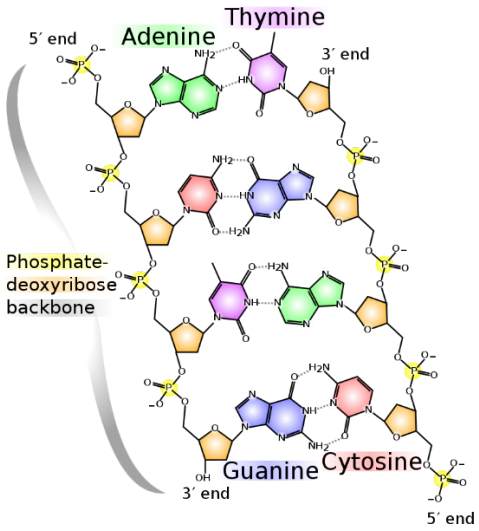
1. Forklar modellen i figur 1.
    a. Hvordan viser den de mulige substitutionsmutationer i DNA?
    b. Hvordan viser den sandsynligheden for at der indtræder substitutionsmutationer?
2. Angiv hvilke begrænsninger modellen har. Begrund dit svar.
    a. Hvilke mutationstyper overser den fx?
    b. Forklar, hvilken betydning tilbagemutationer har for modellens anvendelse.
3. Giv eksempler på spørgsmål, man ikke vil kunne undersøge med modellen.
4. Diskuter, hvorfor modellen, trods sine begrænsninger, kan være anvendelig til at kortlægge evolutionære forbindelser og genetisk afstand mellem arter.

➢ Lav opgaven om den matematiske baggrund for Jukes-Cantor-modellen. Filen findes på www.sysbio.dk under temaet Bioinformatik > Molekylær evolution.

novo nordisk fonden

Kimuras første model (1980)

Kimuras første model udvider Jukes-Cantors model med det forhold, at ikke alle substitutionsmutationer er lige sandsynlige. Det skyldes kemiske forskelle mellem nukleotiderne i DNA.
Kernebaserne adenin (A) og guanin (G) er såkaldte *pyrimidiner*, mens thymin (T) og cytosin (C) er *puriner*. I hvert trin i et DNA-molekyle er en purin parret med en pyrimidin efter *baseparringsprincippet*.
Figur 2 viser den kemiske struktur af et udsnit af et DNA-molekyle.



Figur 2. Udsnit af et DNA-molekyle.
Kilde: Af Madprime (diskussion · bidrag) - Vektorgrafikken blev lavet med Inkscape.., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1848174

Opgave

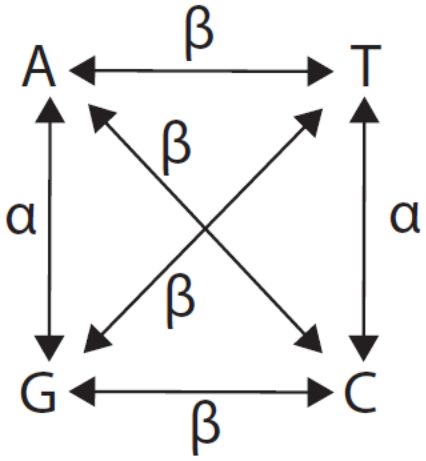
1. Angiv hvilke baser der bindes til hinanden ved baseparringsprincippet.
2. Forklar ud fra figur 2 den kemiske baggrund for baseparringsprincippet.
3. Forklar, hvilken betydning baseparringsprincippet har, når DNA skal replikeres (kopieres til to nye DNA-strenge) før celledelinger.

Substitutionsmutationer mellem puriner (C↔T) eller mellem pyrimidiner (A↔G) kaldes *transitionsmutationer*. Substitutionsmutationer mellem pyrimidiner og puriner (C↔A, C↔G, T↔A og T↔G) kaldes *transvertionsmutationer*.
Hvis mutationer ellers sker tilfældigt, vil transitionsmutationer forekomme ca. to gang oftere end transvertionsmutationer. Indenfor de proteinkodende gener forekommer de ca. tre gange oftere. Dvs.

$$\frac{Antal\ transitionsmutationer}{Antal\ transversionsmutationer} \approx 2$$

I Kimuras model indgår derfor to sandsynlighedsparametre, én for transitionsmutationer og én for transvertionsmutationer. På figur 3 er de angivet med henholdsvis α og β.

novo nordisk fonden



Figur 3. Kimuras første model

Opgave

4. Forklar Kimuras model ud fra figur 3.
5. Diskuter, i hvilke tilfælde, det kunne være en fordel at anvende Kimuras model frem for Jukes-Cantor. Tænk i hvilke spørgsmål man ønsker besvaret fra data.
6. Diskuter, i hvilke tilfælde man vil få tilstrækkelig information ved at anvende Jukes-Cantors model.

Der er flere mulige forklaringer på hvorfor transitionsmutationer er mere hyppige end transversionsmutationer.

- Mutationer sker oftest ved DNA-replikation. De enzymer, der sætter nye nucleotider på DNA-strengen kan lettere ved en fejl sætte en anden pyrimidin på en purin, end de kan sætte en purin på og omvendt.
- Når transitioner er endnu mere hyppige i de proteinkodende gener, skyldes det, at en transitionsmutation i tredje baseposition i en triplet oftere vil være tavs, end hvis der sker en transvertionsmutation. Figur 4 viser den genetiske kode. Med blå baggrundsfarve er markeret fire tripletter, hvor en transitionsmutation i tredje base vil medføre en tavs mutation, mens transvertionsmutationer ikke vil være tavse. De øvrige tilsvarende tripletter er ikke markerede. Sker der aminosyreudskiftninger får det indflydelse på proteinernes funktionalitet, og så vil individet, og dermed den nye genetiske variant, oftere blive fraselekteret i næste generation. Tavse mutationer vil derimod ikke give selektionsmæssige fordele eller ulemper, og vil derfor nedarves som SNP's.

Første nucleotid		Andet nucleotid: T		Andet nucleotid: C		Andet nucleotid: A		Andet nucleotid: G		Tredje nucleotid
T		TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)	T
		TTC		TCC		TAC		TGC		C
		TTA	Leu (L)	TCA		TAA	Stop	TGA	Stop	A
		TTG		TCG		TAG		TGG	Trp (W)	G
C		CTT	Leu (L)	CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)	T
		CTC		CCC		CAC		CGC		C
		CTA		CCA		CAA	Gln (Q)	CGA		A
		CTG		CCG		CAG		CGG		G
A		ATT	Ile (I)	ACT	Thr (T)	AAT	Asn (N)	AGT	Ser (S)	T
		ATC		ACC		AAC		AGC		C
		ATA		ACA		AAA	Lys (K)	AGA	Arg (R)	A
		ATG	Met (M)*	ACG		AAG		AGG		G
G		GTT	Val (V)	GCT	Ala (A)	GAT	Asp (D)	GGT	Gly (G)	T
		GTC		GCC		GAC		GGC		C
		GTA		GCA		GAA	Glu (E)	GGA		A
		GTG		GCG		GAG		GGG		G

Figur 4. Den genetiske kode baseret på DNA-nucleotider på den kodende streng. * angiver startcodon. Den vil svare til koden på mRNA, hvor T blot erstattes med U. Den farvede markeringer angiver fire tripletter, hvor en transitionsmutation i tredje baseposition vil være tavs i modsætning til transversionsmutationer. Aminosyrerne er angivet ved deres tre- og ét-bogstavforkortelser.

Opgave

1. Se på de fire tripletter i figur 4, markerede med farve:
    a. Forklar, hvorfor en transitionsmutation i tredje base vil bevirke en tavs mutation.
    b. Angiv, om transversionsmutationer her vil bevirke missense- eller nonsense-mutationer.
2. Marker på samme måde de øvrige steder ind i tabellen i figur 4, hvor transitionsmutationer tilsvarende vil være tavse, mens transversionsmutationer vil have betydning for hvilken aminosyre der indsættes.
    a. Angiv om transversioner her fører til missense- eller nonsense-mutationer.
3. Forklar hvordan selektion kan medføre, at transitionsmutationer vil optræde mere hyppigt i en population end transversionsmutationer.
4. Giv eksempler på forskellige spørgsmål man kan få belyst ved hjælp af DNA-undersøgelser af forekomsten af forskellige SNP's i en population eller hos forskellige arter.

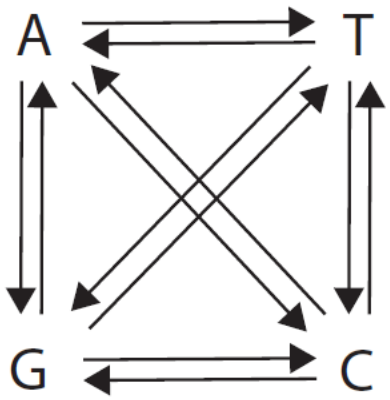
Modeller med flere parametre

Forskellen mellem transitionsmutationer og transversionsmutationer er ikke den eneste årsag til at forskellige mutationer ikke optræder med samme hyppighed:
Kemiske forskelle mellem A og G og mellem C og T kan have indflydelse.
Sekvenser af ens nucleotider kan påvirke sandsynligheden for bestemte mutationer.
I genomet forekommer særlige GC-rige områder. G bindes stærkere til C end A bindes til T. Derfor kan GC-rige områder anvendes til at skabe stabile og kompakte områder i DNA i forbindelse med kromosomernes struktur.
Flere modeller angiver derfor en sandsynlighedsparameter for hver mutation.
Sandsynlighedsparametrene kan fx bestemmes empirisk, men kan variere efter hvilken organisme og hvor i genomet, der er tale om. Ofte defineres endnu en sandsynlighedsparameter, som tager højde for om mutationerne foregår i et GC-rigt område.

Det diskuteres, hvilke modeller der er mest relevante at anvende i hvilke situationer. Ofte viser det sig, at man får besvaret spørgsmålet tilstrækkeligt med de simple modeller, og at de mere komplicerede måske nok kan være nøjagtige i visse situationer, men også kan øge usikkerheden i andre.
Generelt er en model ikke bedre end de data der ligger bag de parametre man er i stand til at putte i den.

Opgave

1. Forklar, ud fra figur 2, hvorfor G bindes stærkere til C end A bindes til T. Du skal fokusere på hydrogenbindingerne mellem baserne.
2. Bestem antallet af sandsynlighedsparametre for en model, hvor hver mutation tildeles sin egen sandsynlighedsparameter. Skriv sandsynlighedsparametrene på figur 5, og tæl op:



Figur 5. Model med mange parametre.

3. Forklar, hvorfor en model ikke er bedre end de data der ligger til grund for dens parametre.