

Et genom - mange historier

Genomsekvenser er guldminer af information om artsdannelse og naturlig selektion.

Men man skal stille de rigtige spørgsmål for at få de interessante svar.

Af Thomas Mailund,
Kasper Munch, Asger Hobolth,
Mikkel H. Schierup og
Peter F. Gammelby

■ I Douglas Adams' romanse-
rie *The Hitchhiker's Guide to the
Galaxy* blev en gigantisk super-
computer sat til at finde svaret
på det ultimative spørgsmål om
Livet, Universet og Alting. Efter
7,5 mio. års arbejde kom com-
puterens svar: »42«.

Når man ikke stiller spørgs-
målene præcist nok, risikerer
man at få svar, som man ikke
bliver klogere af.

På Aarhus Universitet har vi
også en supercomputer. I de
kommende år skal den blandt
meget andet bearbejde store
mængder data, som vi på Bioin-
formatics Research Centre
(BiRC) for nylig fik tilsendt på
en harddisk fra universitetet i
Barcelona.

Harddisken indeholder data
fra i alt 106 genomer (fra 96
individer af fem forskellige arter
af menneskeaber og 10 men-
nesker), som et internationalt
hold af forskere i Great Ape
Genome Diversity Project har



Kamilah, en vestlig lavlandsgorilla fra San Diego Zoo, lagde celler til den første sekventering af artens genom. Efter en tur i BiRC's Coal-HMM-model afslørede dataene, at vores fælles herkomst er et genetisk kludetæppe. 15 pct. af menneskets genom er mere beslægtet med gorillaen end med vores nærmeste slægtning, chimpansen.

sekventeret. Projektet har som
mål til fulde at forstå, hvordan
menneskeaberne, og mennesket

selv, har udviklet sig over de sid-
ste 20 millioner år, og specielt
at forstå hvordan de forskellige

arter er opstået.

Diskene blev ikke sendt til
BiRC fordi vi har adgang til en
supercomputer; vi er med, fordi
vi ved at kombinere statistiske,
matematiske og datalogiske
redskaber har udviklet nogle
unikke og effektive metoder til
både at stille de rigtige spørgs-
mål, programmere computeren
til at lede de rette steder efter
svarene, og til at fortolke sva-
rene med.

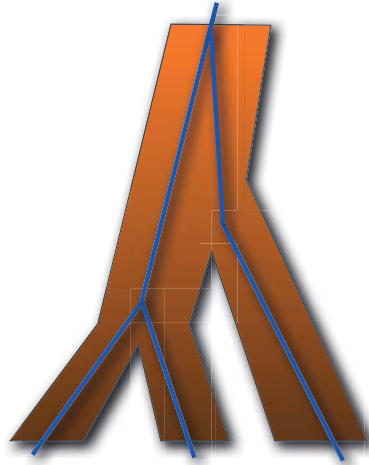
Rygraden i vores metode er
Coalescent Hidden Markov
Models (CoalHMM), som ud
af genom-dataene kan trække
information om artsdannelses-
tidspunkter for millioner af år
siden, artsdannelsesprocesser og
betydningen af naturlig selek-
tion – endda fra DNA fra ét
enkelt individ.

Med den har vi siden 2005
været med til at vise, at vist, at
mennesket og chimpansen split-
tede op for 4,5 mio. år siden, at
mennesket kun er ca. 1,2 pct.

Boks 1 - Topologien for slægtskab

Et eksempel på tre befolkninger, hvor to er tættere beslægtet med hinanden end med den tredje. Det brune træ illustrerer befolkningernes forhold og viser hvilke to befolkninger, der er tættest beslægtet (topologien i træet) samt hvor længe befolkningerne har været adskilt (grenlængder i træet). Det blå træendi er et eksempel på hvor beslægtede tre individer er. Topologien er den samme som for befolkningstræet: de to individer fra de tættest beslægtede befolkninger er også tættest beslægtede. Grenlængderne er derimod længere, specielt er der langt fra det tredje individ til de to andre i forhold til afstanden mellem befolkningerne.

Forskellen skyldes, at ikke alle individer inden for en befolkning er lige tæt beslægtede. Søkende er tættere beslægtede end grandfætre og så videre. Dette gør sig også gældende i forfærende befolkninger, så hvis vi kikker på, hvor langt tilbage i

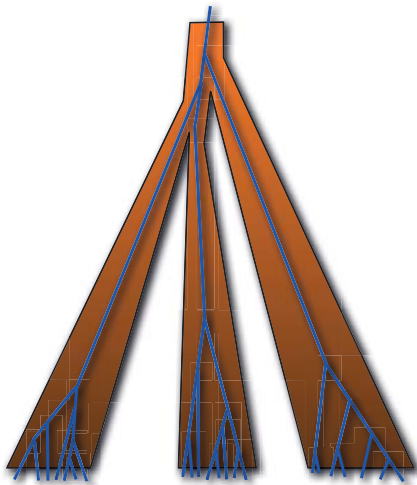
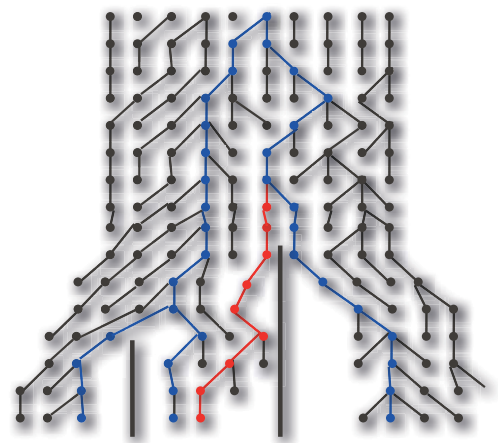


tiden vi skal gå for at finde den nærmeste forfader til to individer, skal vi først gå tilbage til det tidspunkt, da de to befolkninger delte sig fra en stambefolkning, og derefter yderligere et stykke tid tilbage, bestemt af, hvor tæt beslægtede de to individers forfædre var på det tidspunkt. Dette er illustreret i figuren øverst til

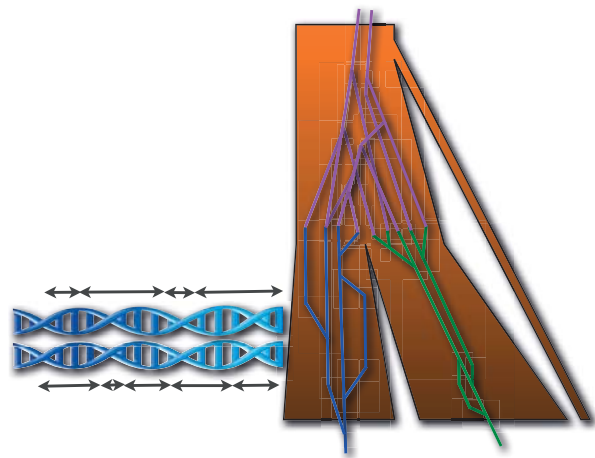
højre, hvor hver prik repræsenterer et individ – levende eller død – og hvor slægtskab er illustreret med kanter. De blå individer svarer til det blå træ til venstre.

Topologien for slægtskabet mellem individer vil ikke altid passe med deres befolkninger. Hvis befolkningerne er tæt nok beslægtede, vil den største

variation i graden af slægtskab ikke findes imellem befolkninger – det vil i stedet være den variation der var i stambefolkningen. Havde vi f.eks. valgt den røde linje i den midterste befolkning ville vi se at individ 2 er tættere beslægtet med individ 3, selv om befolkning 1 og 2 er tættest beslægtet.



Hvis alle individer inden for den samme art har en fælles forfader tættere på i dag end opdelingen af stamarterne, så vil vi ikke kunne se variationen i tiden i stamarten, for lige meget hvilke individer vi kikker på vil vi se den samme forfader i stamarten.



Rekombination er imidlertid udtryk for, at forskellige dele af vores DNA har oplevet forskellige forfædre. Gener langt fra hinanden kan stort set betragtes som uafhængige, så i stedet for at vælge forskellige par af individer for at se variationen i tiden til den fælles forfader, kan man vælge forskellige gener.

forskellig fra chimpansen og bonobo, 1,7 pct. forskellig fra gorillaen og 2,5 pct. forskellig fra orangutangen.

Vi har vist, at der er dele af vores genom hvor vi ligner gorillaen mere end chimpansen, og at vores gener i disse områder også mest udtrykkes som gorillaens. Vi kan se, at naturlig selektion har spillet en overraskende stor rolle i udviklingen

af vores genom, ikke mindst på vores X kromosom og omkring vores gener, og måske især de gener der har med immunsystemet at gøre. Vi kan se, at bonobo og chimpansen først gik fra hinanden for ca. 1 million år siden, men at der alligevel er områder, hvor chimpansen ligner mennesket mere end den ligner bonobo (ca. 2 pct. af arvmassen).

Desuden kan vi se, at der i menneskets linje helt op til artsdannelsen med orangutang ikke har været perioder med meget få individer; tværtimod var stamfaderen til mennesket og chimpansen talrig.

En arts oprindelse

Når en art udvikler sig er der i virkeligheden tale om, at sammensætningen af indivi-

der i arten, eller populationen, udvikler sig. Individer fødes, får unger og dør. Ikke alle individer får lige mange unger, og over tid vil få individer være blevet forfædre til alle levende individer i populationen. Det er således kun enkelte individer, hvis arvelige træk breder sig til hele arten, på bekostning af andre. Det er denne proces der giver anledning til at alle individer i

en art er beslægtede gennem et stamtræ.

Når en art deler sig i to, vil det i det simpleste tilfælde ske ved, at denne population deles op i to nye populationer, som derefter udvikler sig uafhængigt af hinanden. Når en population deles i to, vil individer i den nye population A være lige så tæt beslægtet med individer i population B som de er med de andre i population A. Man kan sige, at halvdelen af stamtræets grene fortsætter i hver population. Som tiden går efter opsplitningen vil individer i de to nye populationer blive mere og mere beslægtede indbyrdes, indtil alle individer i hver population er efterkommere af samme individ.

Kort efter opdelingen vil man ud fra variationen kunne se, hvor langt tilbage i tiden man skal gå for at finde den sidste fælles forfader mellem to individer, afhængigt af hvilke to individer man vælger. (se boks 1).

Ud fra denne variation kan man lære meget om stambefolkningen. For eksempel kan vi se, om stambefolkningen var stor eller lille. Stor variation i tiden tilbage til den fælles forfader betyder, alt andet lige, en stor befolkning frem for en lille. Vi kan se, om stambefolkningen gik igennem en flaskehals på noget tidspunkt. I så fald vil der være usædvanlig mange, der finder deres fælles forfader inde i flaskehalsen.

Statistik fra en enkelt observation

Hvis man statistisk skal afgøre, om én model passer bedre til data end en anden, eller bestemme værdien af parametrene i en model, kræver det normalt, at man har flere observationer. Jo flere jo bedre.

Når vi skal lære noget om, hvornår en stambefolkning delte sig i to, vil de relevante observationer f.eks. være dem, der viser variationen i tiden tilbage til den sidste fælles forfader. Når man ser på tæt beslægtede befolkninger kan denne variation findes ved at vælge mange forskellige individer,

men for meget fjernt beslægtede befolkninger – ultimativt forskellige arter – virker denne tilgang ikke.

Hvis alle individer inden for den første art har en sidste fælles forfader inden for deres egen art, og det samme gør sig gældende i den anden art, vil vi nemlig se præcis den samme afstand mellem et par individer, lige meget hvilket par vi vælger. Dermed har vi reelt kun én enkel observation fra slægtskabet i stamarten, og ud fra den kan vi intet sige om variationen i stamarten (se boks 1 nederst).

Her kommer rekombination til hjælp. Rekombination sker, når vores DNA-streng gennem generationerne bliver ”klippet” i stykker og sat sammen på nye måder. Det medfører, at forskellige stykker af arvemassen kan betragtes som uafhængige observationer af processen. I stedet for at lave

statistik på flere individer – hvilket ikke fører til flere observationer – kan vi lave statistik på flere forskellige stykker af vores genom.

DNA stykker er dog kun uafhængige i den tid, hvor en rekombination har afkoblet dem fra. Mens de er koblet sammen har de samme forhistorie og er altså ikke uafhængige, og hvis man inddrager dem i analysen kan man komme til at undervurdere variationens størrelse i tiden tilbage til forfædrene.

Det problem kan løses på to måder:

Man kan kikke på DNA stykker der ligger langt fra hinanden i arvemassen. Sådanne stykker er adskilt af mange rekombinationer og er derfor uafhængige i langt den største del af deres historie, og kan derfor betragtes som stort set uafhængige observationer. Mange statistiske metoder til at

det er ved at maksimere $P(A,B)$ mht. disse parametre at vi kan estimere parametrene og derved lære om arternes opdeling og udvikling.

Problemet er dog ikke helt løst ved at kunne beregne $P(A,B)$. A og B er forhistorien bag to DNA stykker, men blot ud fra to DNA stykker kan vi ikke direkte afgøre hvad deres forhistorie er. Vi bliver derfor nødt til at integrere over alle mulige forhistorier for den observerede DNA, vægtes med sandsynligheden for at disse historier ligger til grund for den observerede DNA.

Dette klarer vi ved at bruge en statistisk teknologi der hedder skjulte Markov modeller, og vores analysemetode kalder vi CoalHMM'er (Coalescence Hidden Markov Models; coalescence henviser til den populationsgenetiske model af forhistorier og hidden Markov models henviser til skjulte Markov modeller).

Det er ved kombinationen af skjulte Markov modeller og metoder til at beregne den kombinerede sandsynlighed for forhistorier, $P(A,B)$, at vi over de sidste fem til ti år har udviklet statistiske kraftfulde metoder til genetisk analyse af stamarter ud fra DNA fra nulevende arter.

undersøge stamarter er baseret på denne idé. Fælles for dem er at man så kun kan kikke på en lille del af genomet; man skal vælge stykker langt fra hinanden og kan ikke samtidig analysere den DNA der ligger imellem dem.

Alternativt kan man eksplicit modellere rekombinationsprocessen og hvordan forhistorien for ét DNA stykke afhænger af nabo-DNA. På denne måde kan man teoretisk udnytte den fulde arvemasse. Rekombinationsprocessen er i midlertid meget kompleks, og selv med supercomputere er det ikke muligt at modellere den fulde proces for et helt genom uden at ty til forsimplende antagelser.

”Coalescent skjult Markov model”

Vore analyser af stamarter bygger på, at vi kan finde segmenter af genomet med den samme historie ved at ”glide”

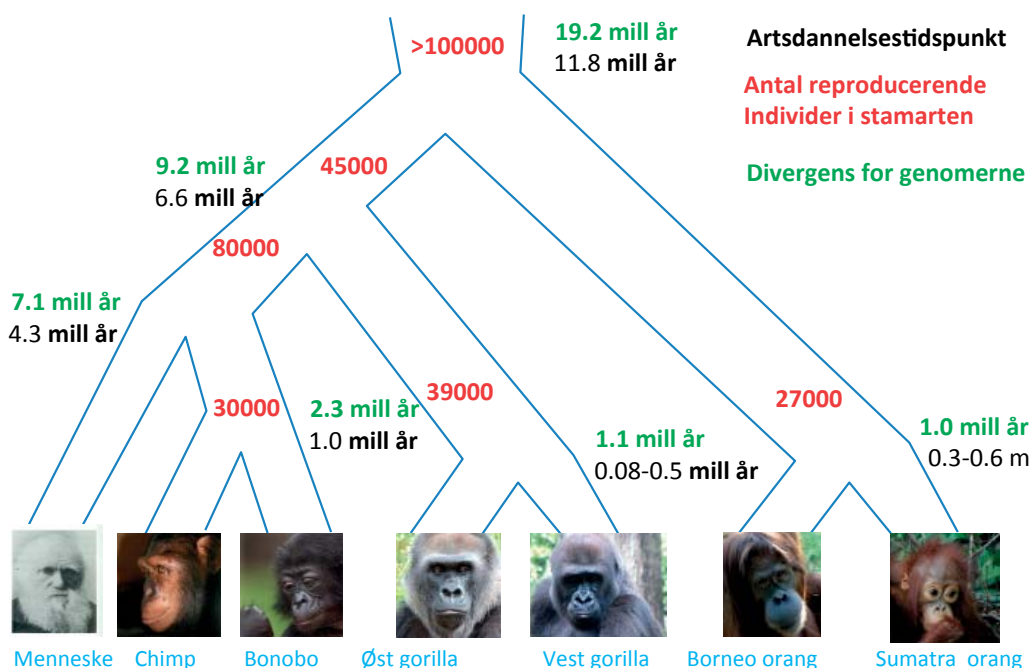
Hvordan man udregner sandsynligheder i en Markov kæde

Boks 2: Problemet bliver i sin enkelhed et spørgsmål om at kunne opskrive den kombinerede sandsynlighed for forhistorien af to nabo-DNA-stykker. Kalder vi forhistorien for stykkerne for A og B, så skal vi kunne beregne $P(A,B)$ for alle mulige historier af DNA-stykkerne. Derfra kan vi beregne den betingede sandsynlighed, $P(B | A)$, hvorefter vi med approksimationen kan bruge denne betingede sandsynlighed til at beregne den samlede sandsynlighed for ikke blot A og B, men A, B, C og D (eller hvor mange stykker vi har brug for), da $P(A, B, C, D)$ approksimeres som $P(A) P(B | A) P(C | B) P(D | C)$.

Uden approksimationen ville vi skulle beregne $P(A, B, C, D) = P(A) P(B | A) P(C | A, B) P(D | A, B, C)$, der vokser i kompleksitet med hvert nyt stykke, og derfor i praksis kun kan beregnes for få stykker.

Det at beregne den kombinerede sandsynlighed $P(A,B)$, og den afledte sandsynlighed $P(B|A)$, er dog heller ikke lige til, og det meste af vores teoretiske arbejde over de sidste mange år har fokuseret på effektive måder at beregne denne sandsynlighed på.

Sandsynligheden $P(A,B)$ afhænger af flere modelparametre, inklusiv tiden hvor arterne delte sig, og



På menneskeabernes artstræ kan man se, at selv om f.eks. mennesket og chimpansen gik fra hinanden for ca. 4,3 mio år siden, skal man i gennemsnit 7,1 mio år tilbage for at finde vore genomers stamfædre - og at vi skal hele 19,2 mio. år tilbage for at finde stamfædrene til alle menneskeabernes genomer.

henover genomet og estimere, om det næste basepar har den samme historie som det foregående.

Her benytter vi os af antagelsen om Markov egenskaben, hvilket her betyder, at sandsynligheden for, at det næste basepar har samme historie som det foregående, kun afhænger af det foregående basepar og ikke resten af sekvensen. Dette er kun en approksimation til den mere komplicerede rekombinationsproces, men det har vist sig at være en meget god approksimation, der betyder at vi kan nøjes med at modellere afhængigheden mellem nabo-stykker og ikke afhængigheden mellem alle stykker i hele genomet.

Dette reducerer kompleksiteten betydeligt og gør det muligt at modellere hele genomer hurtigt og effektivt. Vi kan således opstille alternative hypoteser for artsdannelse ved at udregne sandsynlighederne for at observere vores data under de forskellige modeller (se boks 2 for et eksempel)

Vores CoalHMM'er er

unikke i dels at kunne analysere hele genomer og i at kunne analysere stamarter der levede millioner af år før nu. Der eksisterer andre metoder der kan analysere stamarter millioner af år gamle ved at bruge isolerede DNA-stykker, der antages uafhængige, og andre metoder, der via Markov-model-tilgangen kan analysere hele genomer, men alle andre metoder antager enten uafhængige DNA stykker eller har en approksimation af $P(A,B)$ der fejler, når man går millioner af år tilbage.

Hvordan skilles arter?

Samtidig med at vi forbereder arbejdet med de 106 genomer er vi netop nu i gang med at undersøge, på hvilken måde forskellige tæt beslægtede arter er blevet adskilt. Er det sket som en pludselig opstået barriere for genudveksling (dette kaldes allopatrisk artsdannelse), eller har det været en gradvis proces, hvor genudvekslingen mellem forskellige populationer er blevet mindre over tid indtil den helt er

ophørt (dette benævnes oftest sympatrisk artsdannelse)?

Vi har undersøgt genomerne for tre par af forskellige beslægtede aber:

1. Chimpansen og bonobo
2. Sumatra og Borneo orangutan
3. Den østlige og vestlige gorilla.

Resultaterne viser at chimpansen og bonobo pludseligt gik fra hinanden for ca. 1 million år siden, mens at de to gorillaarter og de to orangutangarter langsomt gik fra hinanden og først havde den sidste genudveksling for hhv. ca. 50.000 og 300.000 år siden.

Nøglen til denne forskel ligger måske i, at bonoboer og chimpanser i dag er adskilt af Congo-floden som muligvis fik sit nuværende leje for ca. 1 million år siden, og da chimpanser hader vand, vil den have været en pludselig barriere for genudveksling.

Vi kan således opsummere vores nuværende viden om artsdannelse i menneskeabernes i deres artstræ (se ovenstående figur) ■

Om forfatterne:



Thomas Mailund, datalog, er lektor på BiRC (Bioinformatics Research Centre), Aarhus Universitet
mailund@birc.au.dk



Kasper Munch, biolog, er postdoc på BiRC
kaspermunch@birc.au.dk



Asger Hobolth, statistiker, er lektor på BiRC
asger@birc.au.dk



Mikkel H. Schierup, biolog, er professor på BiRC
mheide@birc.au.dk



Peter F. Gammelby er journalist på Aarhus Universitet, Science and Technology
gammelby@science.au.dk