

Navn: _____

Parti: _____

Journalen leveres senest tirsdag 4. oktober 2005 i kassen utenfor labben.

BIO 1000

LAB-ØVELSE 3

Fylogenetisk analyse – 27. september 2005

Faglig ansvarlig: Øystein Flagstad

Hovedansvarlig for lab-øvelsen: Øystein Flagstad

Gruppe	Gruppeansvarlig	Hjelpelærer
1	Peder Haugen	Nemanja Jevremovic
2	Øystein Flagstad	Kjetill Voje
3	Øystein Flagstad	Kjetill Voje
4	Pål Trosvik	Beatriz Decenciere
5	Peder Haugen	Nemanja Jevremovic
6	Pål Trosvik	Beatriz Decenciere

Kontakt-adresser

Øystein Flagstad

oflagsta@ulrik.uio.no

Peder Haugen

p.m.haugen@bio.uio.no

Pål Trosvik

Pal.Trosvik@matforsk.no, paltr@student.matnat.uio.no

Beatriz Decenciere

beatriz.decenciere@bio.uio.no

Kjetill Voje

kjetillv@student.matnat.uio.no

Nemanja Jevremovic

njevremovic@gmail.com

NB! HUSK KALKULATOR

Bakgrunn

Fylogenetisk analyse søker å rekonstruere den evolusjonære historien til en art eller en gruppe av arter. Selve analysen leder fram til et fylogenetisk tre som representerer en spesifikk hypotese om evolusjonshistorien til gruppen av arter som studeres. Vi skal i denne laboratorieøvelsen bruke både morfologiske og molekylære karakterer til å illustrere to sentrale fylogenetiske metoder: (1) Parsimoni, (2) Maximum Likelihood.



Kort om metodene

(1) Parsimoni (MP) søker å minimere det totale antall evolusjonære endringer i en fylogeni. Blant alle mulige fylogenetiske trær, vil treet med færrest evolusjonære endringer reflektere den korrekte hypotesen. Et fylogenetisk tre basert på denne metoden kalles for et kladogram.

(2) I henhold til maximum likelihood (ML)-prinsippet finnes det et tre blant alle mulige fylogenetiske trær som reflekterer det mest sannsynlige scenariet av evolusjonære hendelser. For å finne dette mest sannsynlige scenariet må vi bruke en såkalt substitusjonsmatrise, der visse mutasjoner kan være mer sannsynlige enn andre.

Mer detaljerte metodebeskrivelser er gitt i Appendix A og B. Her finner du beskrivelser av hvordan man manuelt kan konstruere fylogenetiske trær i henhold til de to metodene.

Praktisk

Alle partier får utdelt representanter for seks insektordener [Odonata (øyenstikkere), Hemiptera (teger), Diptera (fluer og mygg), Hymenoptera (veps etc), Coleoptera (biller) og Lepidoptera (sommerfugler)]. Edderkoppen (Araenae) skal brukes som utgruppe.

Oppgave 1

Vi skal undersøke de åtte morfologiske karakterene som er gitt i tabellen på side 3.

- (a) Fullfør karaktermatrisen over primitive (opprinnelige) og avledete karakterer (se Fig 25.11 i boka). Anta at utgruppen alltid har den opprinnelige karakteren
- (b) Bruk metoden som er beskrevet i **Appendix A** for å lage et kladogram.
- (c) Hvor mange monofyletiske grupper finner du totalt i treet (se Fig 25.10 i boka), og hvilke karakterer skiller hovedgruppene blant insektene?
- (d) Kladogrammet dere kom fram til inneholder en såkalt polytomi, dvs at det ikke er fullstendig løst opp. Er polytomier alltid et uønsket resultat av for lite informasjon? Hvis ikke, hva kan slike ufullstendig oppløste trær indikere? Diskuter hvordan denne formen for evolusjon kan oppstå (**hint**: tenk på det dere lærte i kapittel 24).

Oppgave 2

Kladogrammet i oppgave 1 representerer flere ulike trær siden innbyrdes slektskap mellom alle taxa i polytomien kan permuteres på flere ulike måter. I denne oppgaven skal vi bruke den molekylære datamatriksen gitt nedenfor for å teste to av hypotesene som er representert i kladogrammet.

```
Karakternummer      0000000001111111111
                    123456789012345678
Araneae              ACTCAACTTGAGAAATTG
Odonata              CCCGGACTTACCAGGCTA
Hemiptera            CCCAGAATTAACCAGCGA
Coleoptera           CTAATTATTACCGTCAAC
Hymenoptera          CCAAGTAAGACCCACATC
Lepidoptera          CCAATTATCACCGGCACT
Diptera              CCAAATACCACCGTCAAC
```

- (a) Permuter polytomien på ulike måter, og velg ut to fullstendig oppløste trær som du synes virker som sannsynlige hypoteser for insektylogenien. Tegn inn de evolusjonære hendelsene (mutasjonene) på begge trærne.
- (b) Avgjør hvilken av de to hypotesene som er mest sannsynlig ved å bruke ML-prinsippet (Appendix B). Baser utregningene dine på substitusjonsmatriksen nedenfor.

Substitusjonsmatrise til bruk i ML-analysen

fra / til	A	C	G	T
A	0,8	0,05	0,1	0,05
C	0,05	0,8	0,05	0,1
G	0,1	0,05	0,8	0,05
T	0,05	0,1	0,05	0,8

Oppgave 1

(a) Karaktermatrisen

Karakter	Araneae	Odonata	Hemiptera	Hymenoptera	Lepidoptera	Diptera	Coleoptera
1) Tre beinpar							
2) Fasettøyne							
3) Larve	0	0	0	1	1	1	1
4) Puppe	0	0	0	1	1	1	1
5) Vinger							
6) Dekkvinger							
7) Andre vingepar tilbakedannet							
8) Skjellklede vinger							

(b) Matrise over ant. felles avledete karakterer (kun inngruppe-taxa; se Appendix A)

Kladogram med de evolusjonære endringene inntegnet

(c) Monofyletiske grupper (se Fig 25.10 i boka)

(d) Polytomier

Oppgave 2

- (a) Tegn to av trærne som er representert i kladogrammet fra oppgave 1. Hvis du har noen formening, velg gjerne de to trærne som du mener best reflekterer innsekt-fylogenen. Kartlegg alle evolusjonære hendelser på trærne.
- (b) ML-estimat av dine to hypoteser. Du behøver ikke vise alle utregningene i detalj. Eksemplifiser imidlertid prosedyren gjennom å vise evolusjonære scenarier [trær og transisjons-sannsynligheter (p_{cd})] for én karakter fra hver hypotese. For begge hypotesene skal du også vise det endelige produktet av sannsynligheter for alle karakterer, samt de totale ML-estimatene. Sett hypotesene opp mot hverandre ved å vurdere størrelsesforholdet mellom de to estimatene.

Appendix A - Parsimoni

Parsimoni (MP) er en såkalt søkemetode, der et optimalitetskriterium ligger til grunn for å finne det beste treet blant alle mulige fylogenetiske trær. Optimalitetskriteriet i denne sammenhengen søker å minimere det totale antall evolusjonære endringer i en fylogeni. Blant alle mulige fylogenetiske trær, vil treet med færrest evolusjonære endringer reflektere den korrekte hypotesen. Et fylogenetisk tre basert på denne metoden kalles for et kladogram.

Manuell utarbeidelse av et kladogram

Utarbeidelsen av kladogrammet baserer seg på felles avledete trekk fra en utgruppe, i vårt tilfelle felles avledete trekk fra edderkoppen. Et felles avledet trekk er et karaktertrekk som deles av minst to taxa i inngruppen og som samtidig skiller seg fra utgruppen. I datasettet nedenfor er alle felles avledete trekk uthevet. For karakter 9 ser vi at taxon 1 og 2 deler en variant av et avledet trekk, mens taxon 3 og 4 deler en annen variant.

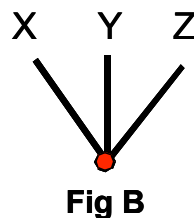
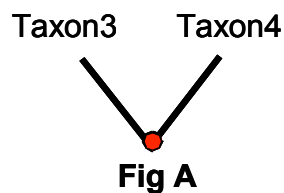
```

Utgruppe  GAGCTATTGA
Taxon1    GAACTAATAA
Taxon2    GAACCAATAA
Taxon3    CAACCAATCG
Taxon4    CAACCGATCA
    
```

- 1) Begynn med å konstruere en matrise over felles avledete trekk for alle par av taxa i datasettet deres. For datasettet ovenfor, vil denne matrisen se slik ut.

	Taxon2	Taxon3	Taxon4
Taxon1	3	2	2
Taxon2		3	2
Taxon3			5

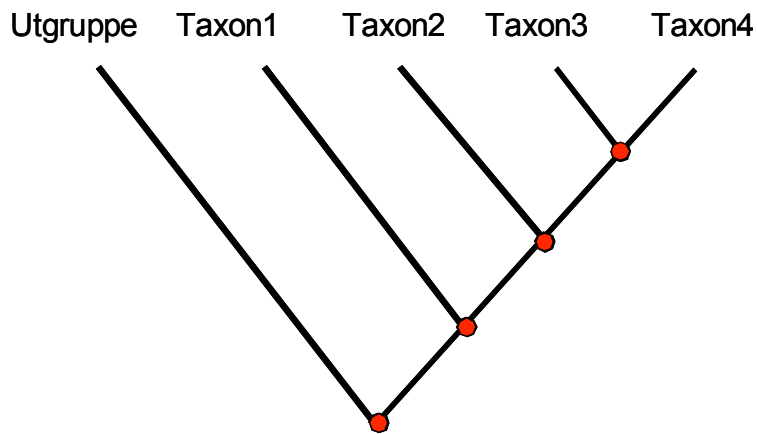
- 2) Finn det taxon-paret som deler flest avledete karakterer; i eksempelet vårt er dette Taxon 3 og 4. Bind disse sammen i form av to grener på et tre (Fig A). Dersom mer enn to taxa deler det maksimale antall avledete karakterer, knytt sammen alle disse taxa i en såkalt polytomi (Fig B).



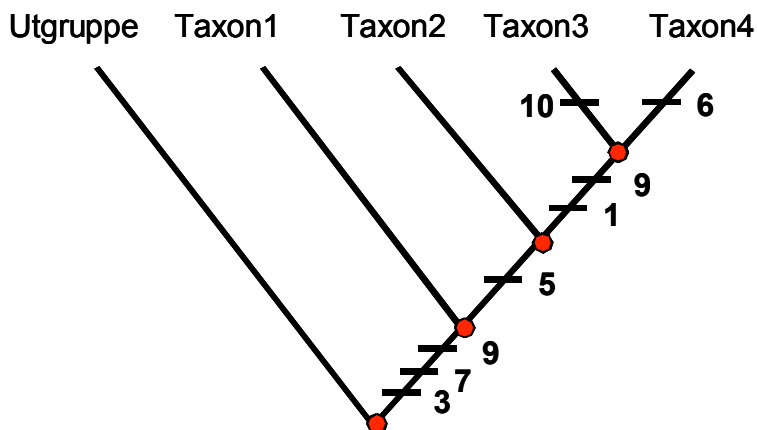
- 3) De såkalte terminale taxa (Taxon 3 og 4 ovenfor) har en felles forfar, symbolisert med sirkelen. Denne felles forfaren har også de samme fem felles avledete trekk. Forfaren har imidlertid ikke avledete trekk som Taxon 3 og 4 har alene (f.eks. karakter 6 for Taxon4 og karakter 10 Taxon3). Konstruer et nytt taxon (Taxon A nedenfor) som representerer denne felles forfaren. Lag en ny karaktermatrise der de første taxa som ble gruppert erstattes av deres felles forfar.

Utgruppe	GAGCTATTGA
Taxon 1	GAAC T AATAA
Taxon 2	GAAC C AATAA
Taxon A	CAAC C AAT C A

- 4) Fortsett prosedyren beskrevet under punkt 2 og 3 til alle taxa har fått sin plass i fylogenieen.



- 5) Sett på de evolusjonære endringene i treet, og indiker hvilke karakterer som endrer seg på de ulike plassene i treet.



Appendix B - Maximum Likelihood (ML)

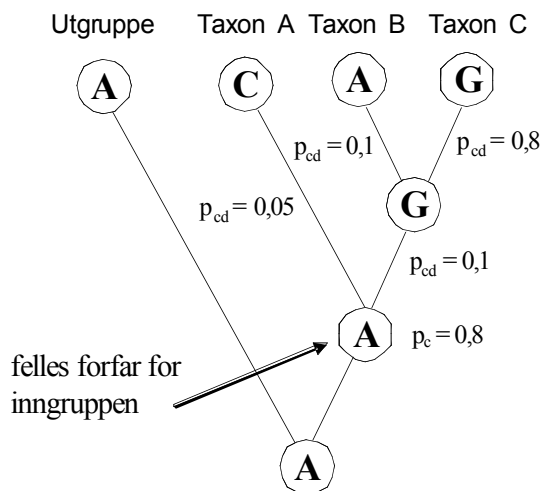
Maximum Likelihood (ML) er i likhet med parsimoni en søkemetode, der et optimalitetskriterium ligger til grunn for å finne det beste treet blant alle mulige fylogenetiske trær. Optimalitetskriteriet for denne metoden søker å finne det treet som reflekterer det mest sannsynlige scenariet av evolusjonære hendelser. Et ML-estimat av en fylogeni forutsetter tre elementer:

- (1) En modell som reflekterer sekvens-evolusjon (altså selve mutasjonsprosessen). Denne modellen kan representeres i en såkalt substitusjonsmatrise (se nedenfor).
- (2) Et fylogenetisk tre (hypotesen)
- (3) Et datasett

For å forstå ML-prinsippet, bør vi se for oss evolusjonshistorien som en tilfeldig prosess der,

- (1) Inngruppens felles forfar har karakteren c med sannsynlighet p_c
- (2) Alle evolusjonære endringer er uavhengige av hverandre. Karakter c endres til d med sannsynlighet p_{cd} i henhold til substitusjonsmatrisen. Vi kan kalle p_{cd} for transisjons-sannsynligheten.
- (3) Hver karakter evolverer uavhengig av hverandre

Gitt hypotesen (det fylogenetiske treet), kan vi således tegne inn et evolusjonært scenario for hver karakter i datamatriksen, som eksemplifisert for én karakter nedenfor.



Substitusjonsmatrise

fra / til	A	C	G	T
A	0,8	0,05	0,1	0,05
C	0,05	0,8	0,05	0,1
G	0,1	0,05	0,8	0,05
T	0,05	0,1	0,05	0,8

Gitt det fylogenetiske treet (hypotesen), er altså sannsynligheten for at Taxon A, B og C skal ha henholdsvis C, A og G for denne karakteren $0,8 \times 0,05 \times 0,1 \times 0,1 \times 0,8 = 0,00032$ eller $3,2 \times 10^{-4}$

Vi kan så fortsette å tegne evolusjonære scenarier for hver av karakterene i datamatriksen vår, og regne ut en sannsynlighet for hvert av disse scenariene. Siden vi forutsetter at alle karakterer evolverer uavhengig av hverandre, kan vi til slutt bruke produktregelen, og multiplisere alle sannsynlighetene med hverandre. Sluttproduktet blir det endelige ML-estimatet av den spesifikke hypotesen vi har lagt fram for den aktuelle fylogien.

Maximum likelihood er regneteknisk uhyre krevende siden metoden i prinsippet skal søke seg gjennom alle mulige trær, og generere et ML-estimat for hvert av disse trærne. I en fylogeni som inneholder 50 taxa, finnes det for eksempel 3×10^{78} mulige trær. Det sier seg selv at en slik formidabel oppgave blir tung selv for en uhyre kraftig datamaskin. Men dersom formålet er å sette et begrenset antall hypoteser opp mot hverandre, blir oppgaven langt mer overkommelig.

I sammenligningen av to eller flere hypoteser, kan det være hensiktmessig å vurdere ML-estimatene i størrelsesordener på 10. Man kan således stille seg spørsmålet om en bestemt hypotese er tilnærmet like sannsynlig som en alternativ hypotese, om den er ca 10 ganger mer sannsynlig, 100 ganger, 1000 ganger eller enda mer sannsynlig i forhold til den alternative hypotesen. Dette, kjære studenter, blir en av mange spennende oppgaver i denne krevende, men forhåpentligvis lærerike laboratorieøvelsen.

Lykke til....