5

Genetiske afstande og afstandsmatricer

Denne vejledning indeholder en række små øvelser og opgaver der illustrerer, hvordan man ud fra genetiske sekvenser kan udregne en gennemsnitlig evolutionær afstand mellem to sekvenser og benytte disse såkaldte p-afstande til at opbygge afstandsmatricer ud fra en multipel alignment. Vejledningen vil først give en kort introduktion til hvordan disse beregninger kan laves i hånden vha. små eksempler, hvorefter det vises hvordan dette hurtigt og nemt kan laves i programmet MEGA. Se link 3 i tema 12 i Bioteknologi 6 for en introduktion til MEGA.

Beregning af evolutionær afstand mellem to sekvenser

Den evolutionære afstand mellem to genetiske sekvenser kan defineres på mange måder. I denne vejledning vil vi udelukkende beskæftige os med to måder. De to metoder er det totale antal forskelle og den såkaldte p-afstand.

Beregning af det totale antal forskelle mellem to sekvenser

Givet en parvis alignment af to sekvenser kan man let optælle antallet af forskelle mellem de to sekvenser. Herunder er vist et simpelt eksempel med et lille udsnit af hæmoglobin A-proteinet fra henholdsvis mennesket (*Homo sapiens*) og koen (*Bos taurus*). Med rødt er markeret de to kolonner i den parvise alignment hvor de to sekvenser har forskellige aminosyrer.

Menneske	MVLSPADKTNVKAAW
Ko	MVLSAADKGNVKAAW

Bemærk at der er anvendt et-bogstavskoden for aminosyrer, se Bioteknologi 6, side 10. Det totale antal forskelle mellem de to sekvenser i denne lille alignment er altså to. Ofte ignoreres kolonner hvor der er indsat et 'gap' i den ene af de to sekvenser, i beregningen af det totale antal forskelle, men man kan selvfølgelig vælge at inddrage disse kolonner også i beregningen. Grunden til at man normalt udelader kolonner med gaps, er at det ikke er let at argumentere biologisk for hvordan de skal tælles med. Man kan selvfølgelig vælge den simpleste løsning at ethvert gap tæller som en forskel på lige fod med en aminosyreændring, men det har jeg valgt ikke at gøre i denne vejledning. Det betyder at hvis man har en alignment med mange gaps, så får man måske ikke et retvisende billede af den evolutionære afstand mellem de to sekvenser. Hvis vi tager ovenstående eksempel igen, men laver en anden alignment, fås følgende:

Menneske	MVLSP-ADK-TNVKAAW
Ko	MVLS-AADKG-NVKAAW

Det totale antal forskelle mellem de to sekvenser i denne alignment er nu 0. Det kan jo se mærkeligt ud at den totale afstand her er 0, og det er det også. Problemet er dog ikke i måden at udregne det totale antal forskelle, men mere i at den viste alignment sandsynligvis ikke viser den rigtige evolutionære historie af de to sekvenser, og derfor giver det ikke så meget mening at beregne en evolutionær afstand mellem sekvenserne. Dette illustrerer vigtigheden i at have en god alignment når man skal lave biologisk sekvensanalyse. Ovenstående metode kan udføres på præcis samme måde hvad enten sekvenserne er nucleotider eller protein.

Beregning af p-afstand mellem to sekvenser

Det totale antal forskelle mellem to sekvenser er ikke et særligt informativt mål for den evolutionære afstand mellem to sekvenser da det jo afhænger af sekvensernes længde. Hvis vi antager at de evolutionære ændringer sker med ca. samme hastighed alle steder i genomet, vil vi forvente at det totale antal forskelle vil stige lineært som funktion af sekvensalignmentens længde. Derfor har vi brug for et andet mål, den såkaldte p-afstand, der lettere kan bruges til sammenligning mellem forskellige sekvenser, arter osv.

Princippet i udregning af p-afstanden er blot at man ud fra en parvis alignment af to sekvenser og det totale antal forskelle, dividerer antal forskelle med længden af den parvise alignment fraregnet eventuelle gaps. Se eventuelt side 29 i Bioteknologi 6 for en nærmere forklaring af begrebet p-afstand. Lad os tage et eksempel. Herunder er vist en alignment af et udvalgt stykke af insulingenet og den tilsvarende proteinsekvens fra henholdsvis mennesket og koen.

Nucleotidalignment

```
Menneske    GGG GGC CCT GGT GCA GGC AGC CTG CAG CCC TTG GCC CTG GAG GGG
Ko          GGA GGC CCG GGC GCG GGC --- --- --- --- --- GGC CTG GAG GGG
```

Proteinalignment

```
Menneske    GGPGAGSLQPLALEG
Ko          GGPGAG-----GLEG
```

Forskellene er markeret med rødt i de to alignments. I nucleotidalignmenten optælles det totale antal forskelle til fem, mens det i proteinalignmenten optælles til en. Forskellen på antal forskelle i nucleotid- og proteinalignmenten skyldes at fire ud af de fem ændringer i nucleotidalignmenten er synonyme, så de ændrer ikke den aminosyre det pågældende codon koder for. Der er ti codons af tre nucleotider i den øverste alignment når man fraregner codons med gaps. Tilsvarende er der ti aminosyrer i den nederste alignment når man fraregner gaps. P-afstandene bliver altså:

Nucleotidalignment
P-afstand = totale antal forskelle / antal nucleotider = 5 / 30 = 0,17

Proteinalignment
P-afstand = totale antal forskelle / antal aminosyrer = 1 / 10 = 0,10

Manuel opbygning af afstandsmatrice

En afstandsmatrice er blot en tabel der viser samtlige parvise evolutionære afstande baseret på en multipel alignment. I en multipel alignment med N sekvenser vil der være:

$$\text{Antal parvise afstande} = \sum_{i=1}^{N-1} i$$

Hvilket blot betyder at hvis der eksempelvis er N = 5 sekvenser, så er der 1+2+3+4 = 10 parvise afstande.

Afstandsmatricer fås i mange former, her vil vi fokusere på hvordan man laver en, der viser det totale antal forskelle, men det er trivielt at gøre det tilsvarende med p-afstande i stedet. Her tages udgangspunkt i et lille udsnit af en alignment af DNA-polymerase gamma-I fra fire arter, mus (*Mus musculus*), rotte (*Rattus norvegicus*), menneske (*Homo sapiens*) og gær (*Saccharomyces cerevisiae*). Udsnittet er vist herunder:

Mus	SWAWAEGWTRYGP
Rotte	KWVWAEGWTRYGP
Menneske	AWAWAEGWTRYGP
Gær	EWLRKPGWVKYVP

Da der er N = 4 sekvenser, skal vi lave 1+2+3 = 6 parvise afstande. Disse indsættes i en simpel tabel for overskuelighedens skyld. Tabellen kan se ud som vist herunder. Bemærk at der er indsat et 0 alle steder hvor de to sekvenser er ens og den ene halvdel af tabellen er hvid. Grunden til dette er at det ikke gør nogen forskel, om man sammenligner mus med rotte eller omvendt rotte med mus, resultatet bliver det samme. I Bioteknologi 6 kan du på side 30 se et eksempel hvor de blanke felter i stedet viser p-afstanden.

	Mus	Rotte	Menneske	Gær
Mus	0			
Rotte		0		
Menneske			0	
Gær				0

Når man skal udfylde en afstandsmatrice, tager man to sekvenser ad gangen og beregner. Lad os starte med at sammenligne mus og rotte. Forskellene er markeret med rødt.

Mus	SWAWAEGWTRYGP
Rotte	KWVWAEGWTRYGP

Alt i alt er det totale antal forskelle mellem de to sekvenser to. Dette indføres nu i afstandsmatricen.

	Mus	Rotte	Menneske	Gær
Mus	0	2		
Rotte		0		
Menneske			0	
Gær				0

På tilsvarende vis kan man nu lave en sammenligning mellem sekvenserne fra mus og menneske.

Mus SWAWAEGWTRYGP
Menneske AWAWAEGWTRYGP

I dette tilfælde findes kun en forskel. Dette indsættes i afstandsmatricen, og denne ser nu ud som vist herunder. Med rødt er markeret den afstand vi lige har beregnet.

	Mus	Rotte	Menneske	Gær
Mus	0	2	1	
Rotte		0		
Menneske			0	
Gær				0

På fuldstændig tilsvarende vis kan de resterende fire parvise afstande udregnes, og resultatet bliver følgende:

	Mus	Rotte	Menneske	Gær
Mus	0	2	1	8
Rotte		0	2	8
Menneske			0	8
Gær				0

Det er selvfølgelig et meget lille eksempel og skulle man finde på at lave en evolutionær analyse af dette resultat, vil man komme frem til det overraskende resultat at mennesket og musens sekvenser er dem der ligner hinanden mest, og derfor er de måske tættest beslægtet af de fire arter. Dette er selvfølgelig ikke rigtigt da mus og rotte er meget tættere beslægtet med hinanden, end nogen af dem er med mennesket. Årsagen til dette er at datasættet er meget lille, og tilfældigheder spiller derfor en meget stor rolle. Havde man analyseret hele proteinet, får man nedenstående matrice:

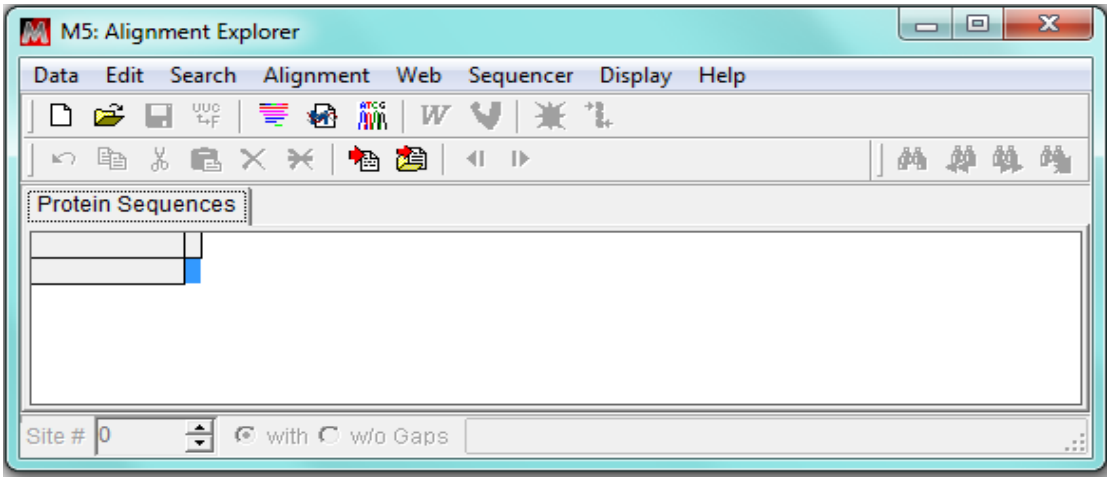
	Mus	Rotte	Menneske	Gær
Mus	0	41	84	559
Rotte		0	99	555
Menneske			0	557
Gær				0

Som man kan se, passer resultaterne nu med hvad vi umiddelbart ville forvente. Musen og rottens sekvenser er klart mere ens end de er med nogle andre sekvenser, og de tre pattedyr er markant mere ens end nogle af dem er med gær. Det kan umiddelbart virke som en større opgave at tælle 559 forskelle op i en alignment, og det er det også hvis man gør det i hånden, og risikoen for at lave tællefejl er meget stor. Derfor bruger vi i stedet computeren til at lave disse beregninger for os og som vi vil se i næste afsnit, så er det meget hurtigt og let at lave disse beregninger i MEGA.

Udregning af afstandsmatricer i MEGA

Efter at have lært og forstået hvordan disse matricer kan laves i hånden, er der ingen grund til at gøre det i hånden fremover. Ved hjælp af MEGA er det let, hurtigt og garanteret uden fejl at lave en afstandsmatrice hvis ellers ens alignment er indlæst/indtastet korrekt!

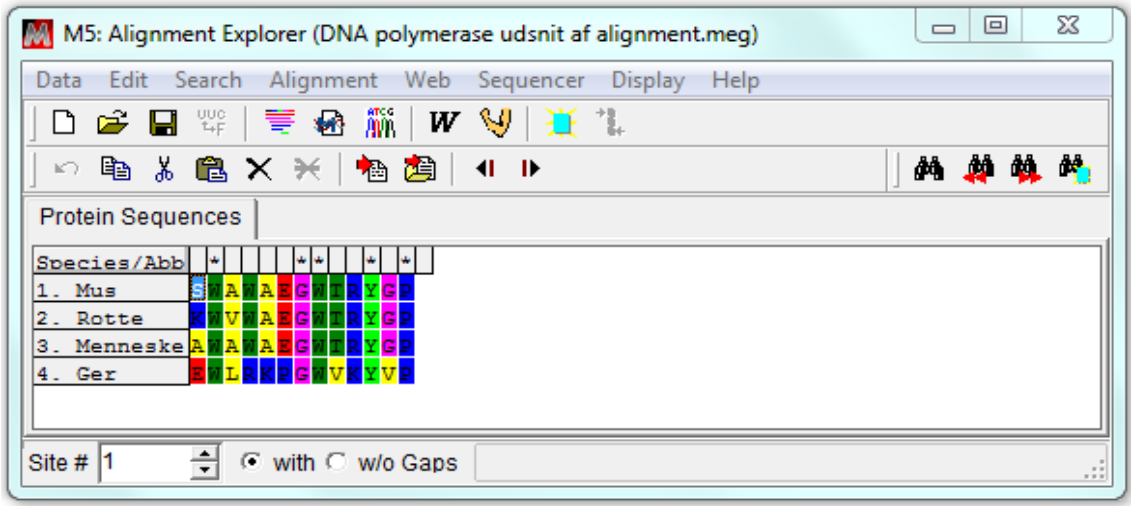
1. Åben MEGA og start en ny alignment explorer ved at vælge 'Align' og 'Edit/Build Alignment'. Vælg 'Create a new alignment' og herefter 'Protein'. Skærmbilledet skulle gerne se ud som vist herunder:



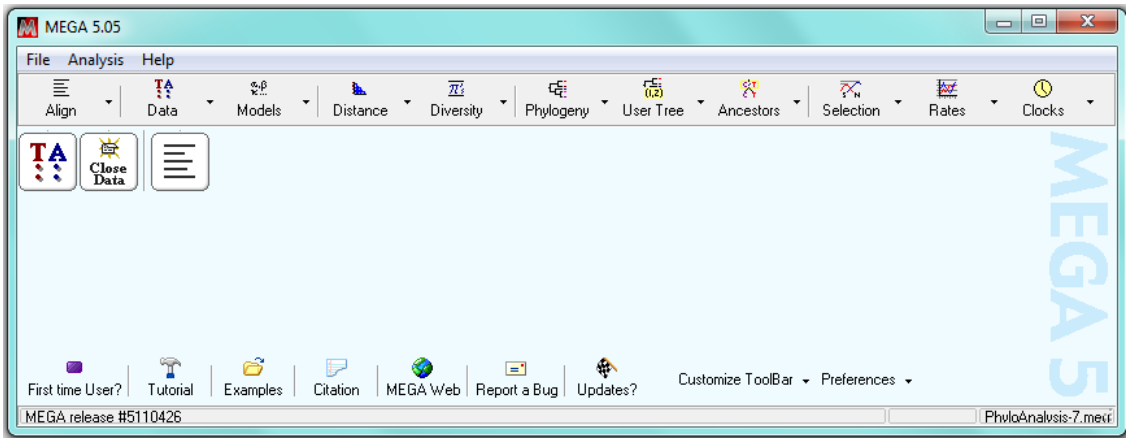
2. Indtast de fire korte sekvenser der blev brugt tidligere ved at vælge 'Edit' og 'Insert Blank Sequence' fire gange. Herefter navngives de fire sekvenser Mus, Rotte, Menneske og Ger (MEGA kan ikke acceptere danske ø, æ og å) ved at højreklikke på sekvensnavnene og vælge 'Edit Sequence Name'.

Mus	SWAWAEGWTRYGP
Rotte	KWVWAEGWTRYGP
Menneske	AWAWAEGWTRYGP
Gær	EWLRKPGWVKYVP

3. Indtast ovenstående fire sekvenser. Nu ses skærmbilledet vist herunder:

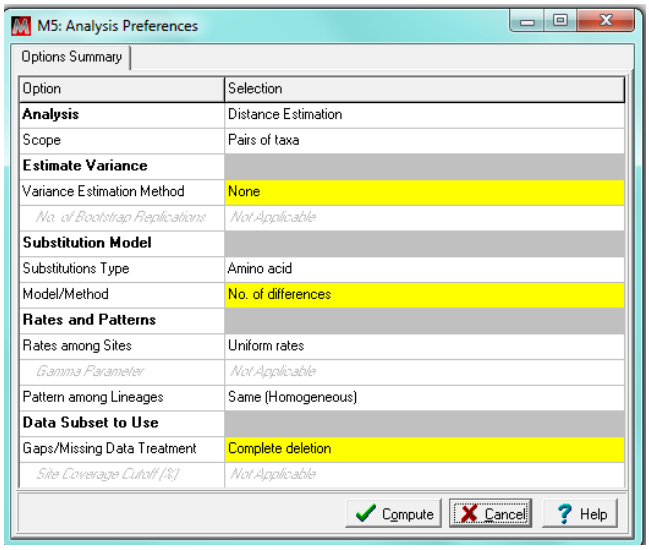


4. Vælg 'Data' og 'Phylogenetic Analysis' og svar nej til at det er proteinkodende nucleotidsekvenser. Find MEGAs hovedvindue frem, det bør ligne nedenstående skærmbillede:



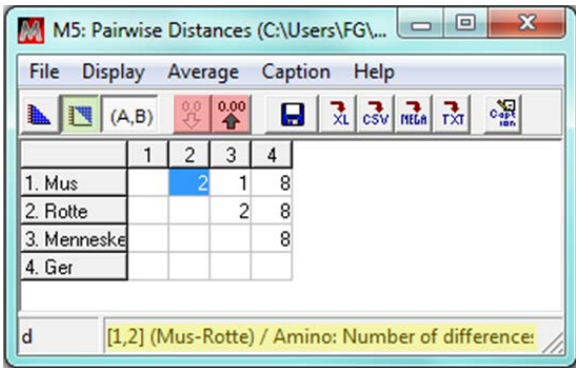
Der er nu tre ikoner på dit MEGA-skrivebord. Det første ikon hvorpå der blandt andet er tegnet et rødt T og et blåt A, giver adgang til at se dine data i vinduet 'Sequence Data Explorer'. Det næste ikon lukker dine data, og sidste ikon er dit 'Alignment Explorer'-vindue. MEGA kan altid kun have et datasæt åbnet på skrivebordet ad gangen så hvis du åbner et andet, vil det nuværende forsvinde!

5. Klik på knappen 'Distance' og vælg 'Compute Pairwise Distances' hvorefter følgende vindue bør dukke op:



Der er nu forskellige muligheder, det er muligt at din opsætning er lidt anderledes end skærmbilledet viser. Det vigtige her er at der under 'Substitution Model' og ud for 'Model/Method' i det gule felt står 'No. of differences', da det betyder at programmet giver en afstandsmatrice med de totale antal forskelle.

6. Indstil mulighederne i de gule felter hvis ikke det ser ud som vist herover og tryk derefter på 'Compute'-knappen. Nu dukker der et nyt vindue op med den beregnede afstandsmatrice som vist øverst på næste side:



Man kan trykke på de forskellige indgange i tabellen. Her er vist hvad der sker når man trykker på indgangen der sammenligner sekvens 1 (mus) med sekvens 2 (rotte). Det eneste der ændrer sig, er den med gult markerede tekst i bunden af vinduet. Antal decimaler kan ændres med knapperne markeret med rødt på skærmbilledet. Tabellen kan vendes ved at trykke på den med grønt markerede knap.

7. Sammenlign den beregnede afstandsmatrice med nedenstående der blev beregnet i hånden. Er de to identiske?

	Mus	Rotte	Menneske	Gær
Mus	0	2	1	8
Rotte		0	2	8
Menneske			0	8
Gær				0

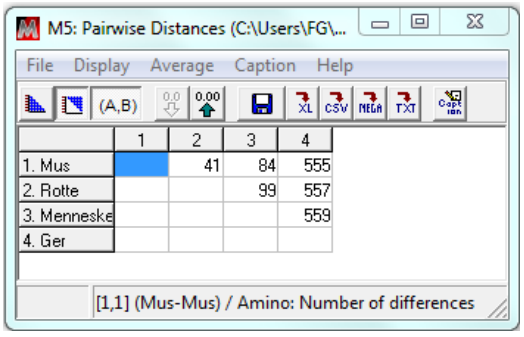
Hvorfor bruge MEGA?
Med så små datasæt som ovenstående er der måske ikke den store tid sparet ved at bruge MEGA, men der er sikkerhed for at optællingerne og beregningerne er rigtige og når datasæt bliver bare en smule større, er det langt hurtigere at indtaste data i MEGA i stedet for at lave beregningerne i hånden.

8. Download den alignment der blev benyttet tidligere i udregningen af afstandsmatricen for hele polymerase gamma-I-proteinet, den kan hentes på nedenstående link:

   www.bioteknologibogen.dk/bioteknologi-6/data/DNA_pol_alignment.meg

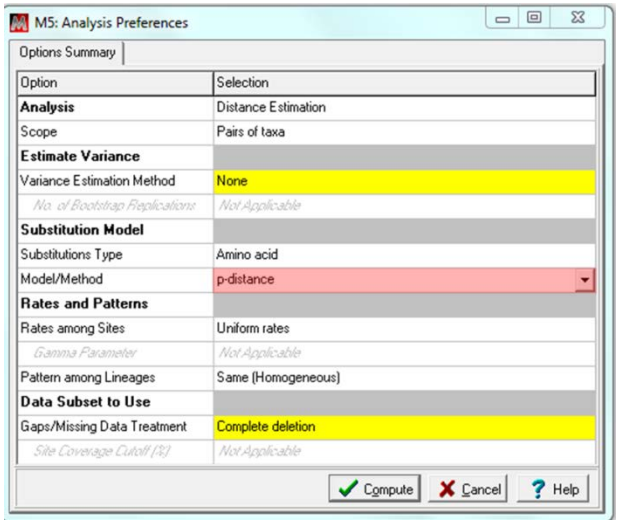
9. Indlæs filen i MEGA, den letteste måde er at dobbeltklikke på filen så åbner den automatisk i MEGA. Før dette gøres, er det dog smart lige at lukke alle andre MEGA-vinduer så man ikke forvirrer sig selv senere. Virker dobbeltklik-metoden ikke, kan filen hentes ind i MEGA ved 'File' og 'Open a File/Session' i stedet.

10. Brug MEGA til at beregne en afstandsmatrice for denne markant større alignment. Resultatet bør se ud som vist herunder:

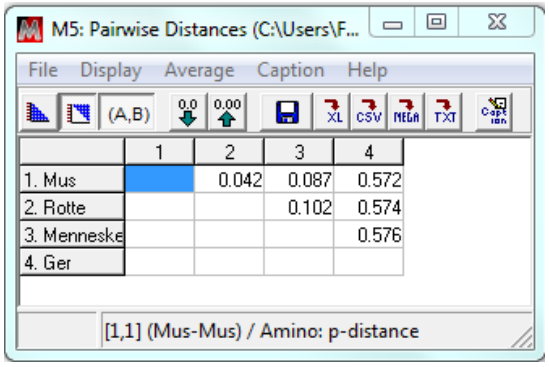


Hvis ikke billedet ser ud som vist, kan du ændre antal decimaler og vende tabellen som beskrevet ovenfor. Du kan også ændre rækkefølgen af arterne i matricen ved at markere og trække et artsnavn. Hvis dette ikke løser problemet, er det nok ikke den rigtige sekvensfil der er indlæst.

11. Luk vinduet med afstandsmatricen og prøv i stedet at få MEGA til at lave en afstandsmatrice baseret på de parvise p-afstande. Eneste ændring der er nødvendig, er under 'Substitution Model' og 'Model/Method' hvor 'No. of differences' skal ændres til 'p-distance'. Det er markeret med rødt på nedenstående skærmbillede.



12. Klik på 'Compute'-knappen. Et vindue i stil med nedenstående dukker nu op:



Husk at du kan ændre antallet af decimaler. Resultatet viser at i dette protein er der forskel på menneske og mus i 8,7 % af aminosyrerne.

Opgaver til afstandsmatricer

Herunder er der tre opgaver der alle omhandler afstandsmatricer.

Opgave 1. Slægtskabsanalyser

Til Biologi A august-eksamen 2008 handlede opgave 3 om slægtskabsanalyser og arters tilpasning til vand. I opgaven blev der angivet en multipel alignment som er vist herunder:

Kænguru	CACCACCACCAATACA
Blåhval	ACCGATTCCCCACCCA
Flodhest	ACCGGCATCCCGCCCA
Zebra	ACTCACACCTCATTCA
Næsehorn	ACTCACCCCTTTCTCA
Ringsæl	ACCAACCATTTATACA

Desuden var der en delvist udfyldt afstandsmatrice som lettere redigeret er vist herunder:

	Kænguru	Blåhval	Flodhest	Zebra	Næsehorn	Ringsæl
Kænguru	0					
Blåhval		0				
Flodhest	10		0			
Zebra	8		8	0		
Næsehorn	9		9	4	0	
Ringsæl	6		10	7	7	0

1. Beregn de manglende afstande i hånden og udfyld matricen.

2. Indtast den multiple alignment i MEGA og beregn den fulde afstandsmatrice.

Opgave 2. Menneskeaberne

Mitokondrie-DNA bruges ofte til at fastslå slægtskab mellem arter. Download nedenstående fil fra hjemmesiden.

www.bioteknologibogen.dk/bioteknologi-6/data/Primater_mitokondrie_udsnit.meg

1. Indlæs filen i MEGA og lav en afstandsmatrice ud fra alignment i filen. Bemærk at man i MEGA kan ændre rækkefølgen i matricen ved at trække i de forskellige sekvensnavne. Udfyld afstandsmatricen herunder.

	Chimpanse	Gorilla	Bavian	Bonobo	Makak	Orangutang	Menneske
Chimpanse	0						
Gorilla		0					
Bavian			0				
Bonobo				0			
Makak					0		
Orangutang						0	
Menneske							0

2. Forklar ud fra din afstandsmatrice hvor tæt mennesket er beslægtet med de forskellige andre primater.

Opgave 3. Truede tigre

En af de vejledende eksamensopgavesæt der udkom i 2007 indeholdt en opgave om truede tigre. I opgaven fik man en multipel nucleotidalignment af seks tigre, som vist herunder.

Sibirisk tiger	GCACCGTACCCCCCTCACTTTGTGGCACCTCTATATAATGCTACTAGGCTGCCG
Sydkinesisk tiger	ACGCCGCACTCCCTCCGCTTTGTGGCATCTCTACATGATGCCATCAAGCCACTG
Indokinesisk tiger I	GTACCGCACCCCCCTCGCTTTATAGCACTTCTATATAATGCTACTAGGCTGCTG
Indokinesisk tiger II	GCGCCGCACCCCCCTCGCTTTGTGATATCTTTACGTAATGCTACTAGGCTGCCG
Sumatra tiger	ACGCCGCACCCCCTTCGCTTTGCGGCGTCTCTACATAACGCCATTAGGTTGCTG
Bengalsk tiger	GCGCCGGACCCCCCTTGCTCTGTGGCATCTCTACATAACGTCATTAGACTGCTG

Derudover var der i opgaven en delvist udfyldt afstandsmatrice, denne er vist herunder i et lidt anderledes format.

	Bengalsk tiger	Sumatra tiger	Indokinesisk tiger II	Indokinesisk tiger I	Sydkinesisk tiger
Siberisk tiger					
Sydkinesisk tiger	15	11	16	18	
Indokinesisk tiger I	15	15	12		
Indokinesisk tiger II	13	13			
Sumatratiger	10				

1. Beregn de manglende afstande i hånden og udfyld matricen.

2. Indtast den multiple alignment i MEGA og beregn den fulde afstandsmatrice. Vær opmærksom på at afstandsmatricen i opgaven er opbygget på en lidt anderledes måde end MEGA viser den.