



Databasesøgning med BLAST

Denne vejledning giver en introduktion til databasesøgning med forskellige programmer i BLAST-familien. Vejledningen indeholder først en grundig introduktion og gennemgang af de forskellige BLAST-programmer og dernæst en grundig kagebogsgennemgang af en af BLAST-programmernes mange facetter. Til sidst følger øvelser og opgaver som kan løses ved hjælp af BLAST. Vejledningen er skrevet til brug i MEGA, men kan uden problemer generaliseres til brug i en almindelig internet-browser. Se link 3 i tema 12 i Bioteknologi 6 for en introduktion til MEGA.

Introduktion

BLAST er en forkortelse for Basic Local Alignment Search Tool. Navnet angiver meget præcist hvad det er BLAST kan, nemlig at finde små stykker sekvens i en database der svarer til en søgesekvens man allerede har. Helt grundlæggende har man en søgesekvens (query sequence) og en database denne søgesekvens skal sammenlignes med. Valg af database skal overvejes nøje da forskellige valg har forskellige fordele og ulemper.

Der findes forskellige BLAST-metoder, men fælles for dem alle er at de kun søger efter lokale alignments. En lokal alignment er en alignment mellem en mindre del af en søgesekvens og en mindre del af en databasesekvens der er meget lig hinanden. De mange forskellige BLAST-metoder kan overordnet inddeles i to hovedtyper, de proteinbaserede metoder og de nucleotidbaserede metoder. Herunder følger en grundig introduktion til de forskellige BLAST-typer.

Protein BLAST

Protein BLAST-metoderne kan ofte finde relevante sekvenser fra meget fjernt beslægtede organismer idet proteinsekvenser udvikler sig meget langsommere end DNA-sekvenser. Der findes i øjeblikket flere forskellige protein BLAST-algoritmer. BLASTP er den letteste at starte med. For de særligt interesserede er der herunder også en gennemgang af flere af alternativerne. Nye brugere af BLAST kan med fordel nøjes med at læse om BLASTP.

BLASTP

Søgesekvensen er protein og databasen man sammenligner med, indeholder kun proteinsekvenser. Bemærk at søgesekvensen skal angives i et-bogstavs-aminosyre-koderne, se Bioteknologi 6, side 10. Princippet i metoden er at ens søgesekvens sammenlignes med en database med andre proteinsekvenser.

Analysere man den genetiske kode, ses det at ikke alle aminosyresubstitutioner er lige sandsynlige, og mange aminosyreændringer kræver mere end en mutation i det samme codon. Derfor forventer vi at der er forskel på, hvor ofte vi ser forskellige aminosyresubstitutioner. Denne information udnyttes ved at man bruger såkaldte substitutionsmatricer i søgningen. En substitutionsmatrice er en empirisk bestemt matrice der viser, hvor ofte bestemte aminosyrer ændres til andre aminosyrer når man sammenligner homologe sekvenser. Denne information benyttes når man skal finde homologe sekvenser i en database. Benytter man forskellige substitutionsmatricer kan det give forskellige søgeresultater, men med mindre man er ekspert eller af anden årsag har information der gør det relevant at skifte substitutionsmatrice, kan det anbefales blot at benytte standardmatricen (BLOSUM62) der i langt de fleste tilfælde fungerer fornuftigt.

Alternative protein-BLAST metoder

BLASTX

Søgesekvensen er en nucleotidsekvens, men denne translateres til en proteinsekvens inden søgningen. Databasen indeholder kun proteinsekvenser. BLASTP-algoritmen bruges med den translaterede søgesekvens. BLASTX bruges ofte når en nucleotidbaseret BLAST-metode som BLASTN ikke har fundet relevante resultater i nucleotiddatabaserne. Metoden kan selvfølgelig kun bruges når ens nucleotidsekvens koder for et protein. Dette kunne eksempelvis være en mRNA-sekvens.

TBLASTN

TBLASTN er en særlig udgave af BLASTP hvor søgesekvensen er en proteinsekvens, og databasen man benytter, er en nucleotiddatabase der oversættes til protein ved hjælp af den genetiske kode. Oversættelsen af nucleotidsekvenserne til proteinsekvenser sker i alle tre mulige læserammer. Bortset fra det er det en helt normal BLASTP-algoritme der køres. Denne metode benyttes primært når BLASTP ikke har fundet relevante resultater. Manglen på resultater med BLASTP kan skyldes at der ikke findes så mange proteinsekvenser i proteindatabaserne som nucleotidsekvenser i nucleotiddatabaserne. Forklaringen på dette er at det i dag er lettere, hurtigere og billigere at lave nucleotidsekventering end proteinsekventering.

TBLASTX

I TBLASTX er søgesekvensen en nucleotidsekvens, men denne translateres til protein inden søgningen, og databasen indeholder translaterede nucleotidsekvenser som i TBLASTN-metoden. Denne metode er meget lig TBLASTN, den eneste forskel er at søgesekvensen er en nucleotidsekvens. TBLASTX bruges primært til at lede efter sekvenser fra fjernt beslægtede organismer når BLASTN ikke kan finde relevante resultater i nucleotiddatabaserne, og man ved at ens nucleotidsekvens er proteinkodende.

PSI-BLAST

PSI-BLAST er i dag nok den mest benyttede protein BLAST-metode. Metoden tager udgangspunkt i en standardkørsel med BLASTP og laver så en sekvensprofil baseret på ligheder i BLASTP-resultaterne. Derefter søges databasen igennem igen baseret på denne sekvensprofil. På denne måde findes mere fjernt beslægtede sekvenser der ikke dukker op med en normal BLASTP-søgning. Denne proces gentages så en række gange hvor de nye resultater hver gang bruges til at opdatere sekvensprofilen. Efter et antal søgninger skulle resultaterne gerne pege på det samme sæt proteiner hver gang, og dermed er metoden færdig. PSI-BLAST er en meget følsom algoritme og finder derfor i nogle tilfælde sekvenser der ikke er homologe, men ensartede ved et tilfælde. Derfor er det altid vigtigt at brugeren har en kritisk tilgang til resultaterne af en PSI-BLAST-søgning.

PHI-BLAST og DELTA-BLAST

PHI-BLAST er en udvidelse af PSI-BLAST hvor man kan definere, at man leder efter særlige proteinmotiver. Et motiv er i denne sammenhæng blot et kort stykke af proteinsekvensen man ved er meget bevaret i mange arter. Det kan eksempelvis være et bestemt motiv der findes i en række membranreceptorproteiner. Et eksempel kunne være et ni aminosyrer langt motiv hvor de fem centrale aminosyrer varierer, men de to yderste er bevarede i alle undersøgte organismer. Et sådant motiv kunne fx skrives på denne måde: MW?????NG. Til at definere motivet man leder efter, bruges et simpelt programmeringssprog. Ovenstående motiv skrives eksempelvis sådan: M-W-x(5)-N-G. Med lidt træning kan man hurtigt lære at bruge dette simple programmeringssprog. DELTA-BLAST er en nyudviklet metode meget lig PHI-BLAST. Metoden skulle ifølge udviklerne være markant hurtigere og lidt bedre end PHI-BLAST.

Nucleotid BLAST

BLASTN

I BLASTN er søgesekvensen en nucleotidsekvens, enten RNA eller DNA, og databasen man sammenligner med, indeholder kun nucleotidsekvenser. Nucleotid BLAST-metoderne har den store fordel i forhold til de proteinbaserede metoder, at de kan bruges til at undersøge de store dele af genomet, der ikke koder for proteiner. I menneskets tilfælde drejer det sig om ca. 99 % af genomet. Den mest almindeligt benyttede nucleotid BLAST kaldes BLASTN. Derudover findes der et væld af alternative metoder hvor der her kun nævnes MegaBLAST, og i langt de fleste tilfælde kan man nøjes med at bruge BLASTN.

I modsætning til BLASTP baserer BLASTN sig ikke på en empirisk bestemt substitutionsmatrice, men blot på en simpel lokal alignment algoritme. BLASTN kan bruges til at sammenligne og finde homologe sekvenser i alle kombinationer af RNA- og DNA-sekvenser, ligegyldigt om de er proteinkodende eller ej.

MegaBLAST

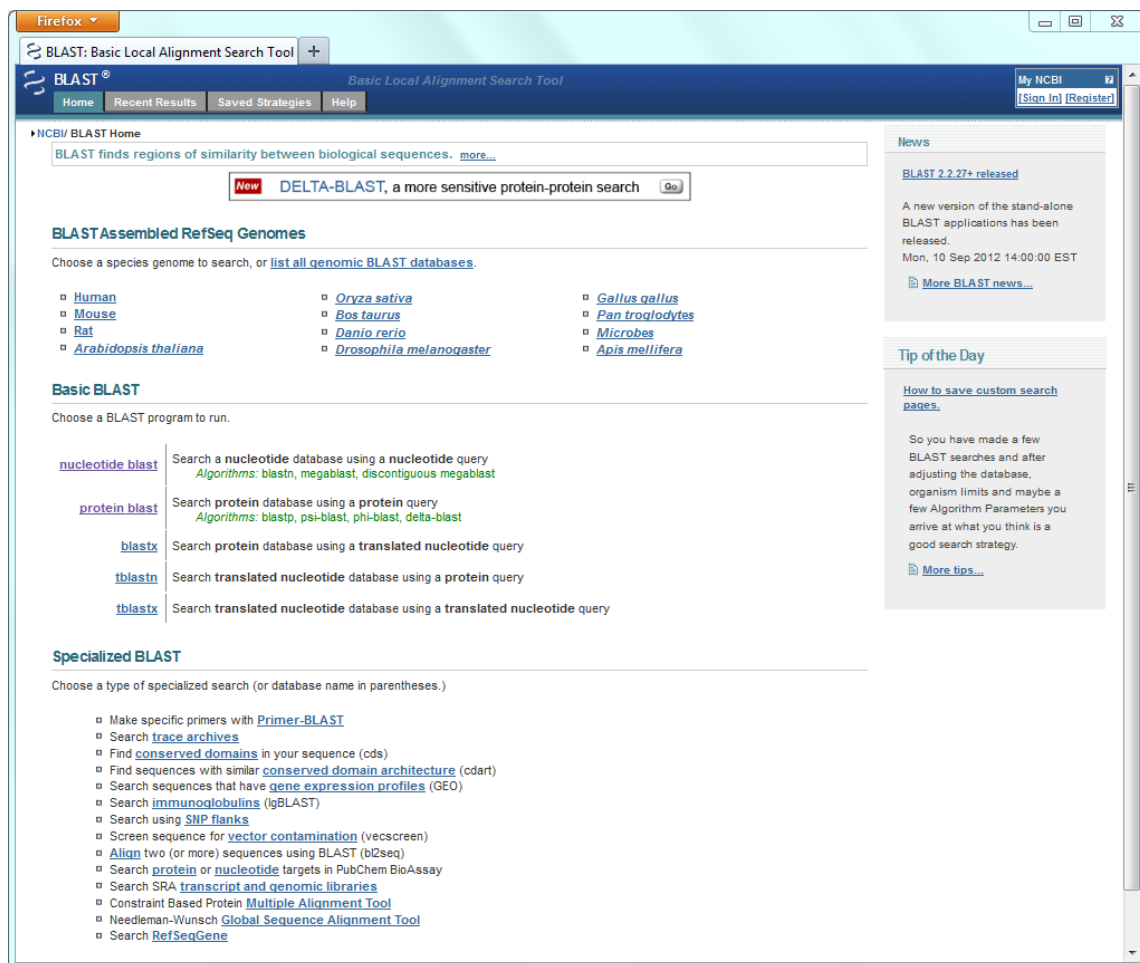
MegaBLAST er en markant hurtigere nucleotidbaseret algoritme end BLASTN. Metoden kan dog kun benyttes til at finde meget ensartede sekvenser. Hvis forskellen mellem søgesekvensen og databasesekvenserne overskrider et par procent, er MegaBLAST ikke følsom nok til at finde de homologe sekvenser, og man bør i stedet benytte den mere generelle, men også markant langsommere BLASTN.

BLAST i praksis

I denne del af vejledningen gennemgås en BLASTP-søgning. Desuden præsenteres forskellige funktioner og muligheder for at tilpasse søgealgoritmen. Der findes en lang række forskellige databaser online som man kan sammenligne sin søgesekvens med. En af de mest benyttede databaser er Genbank der kan findes på adressen:

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

Når man åbner ovenstående link, kommer man til et skærmbillede som vist herunder, hvor man kan vælge hvilken type BLAST, man ønsker at udføre.

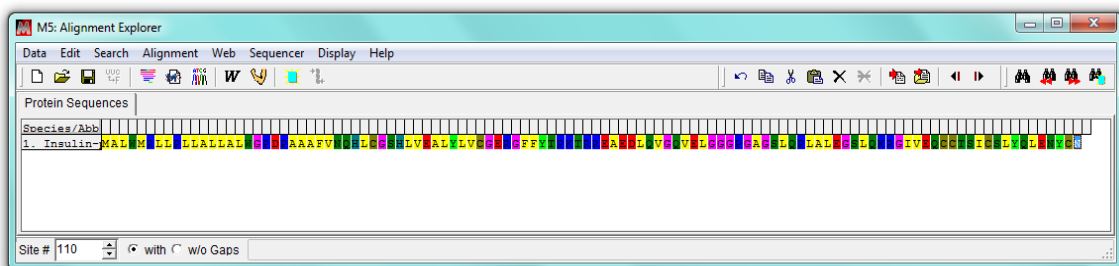


BLAST kan køres direkte fra alignment explorer delen af MEGA. BLAST-funktionen findes under fanebladet 'Web' og funktionen 'Do BLAST search'. Hvis man markerer en sekvens i alignment explorer og vælger 'Do BLAST search', åbnes Genbanks BLAST-side i et specielt MEGA-browser-vindue. Dette har den fordel at sekvenser man måtte finde og ønske at tilføje sin alignment, let kan tilføjes ved at trykke på knappen med det røde plus og teksten 'Add to alignment'. Hvorvidt man ønsker at benytte MEGA eller blot en almindelig browser er principielt ligegyldigt. MEGA giver dog en let og hurtig måde at tilføje nye sekvenser til en alignment på, så i den resterende del af denne vejledning vil denne metode blive benyttet.

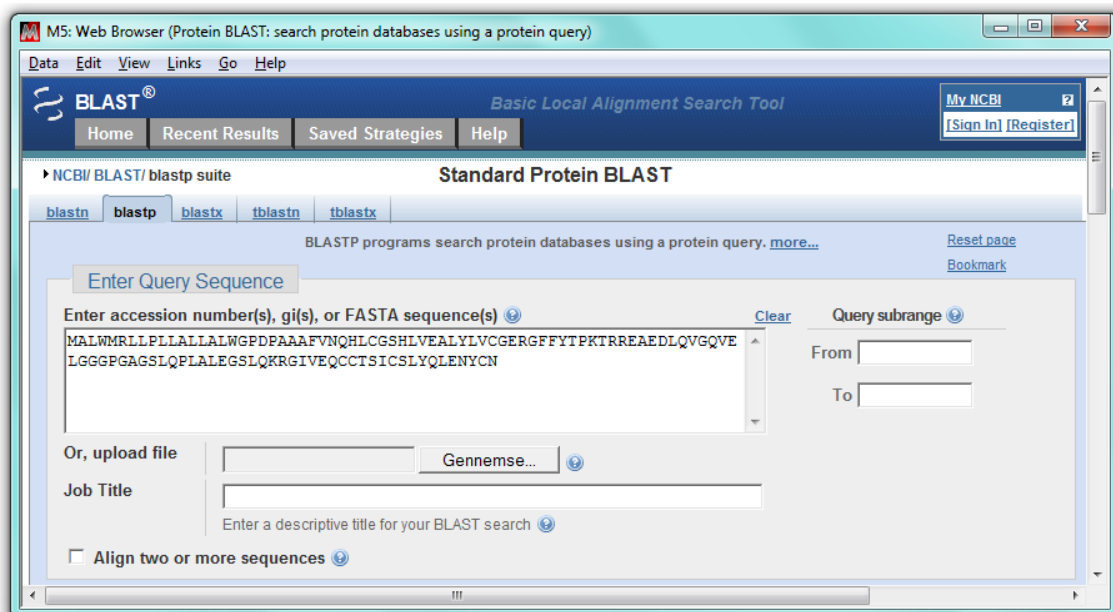
Min første protein-BLAST i MEGA

I denne øvelse bruges BLASTP til at lede efter homologe sekvenser til menneskets præproinsulin-protein og danne en multipel alignment ud fra resultaterne. Denne øvelse er lavet som en grundig kagebogsgennemgang af BLASTP-programmets mange facetter.

1. Åbn MEGA og opret en ny protein-alignment, dette gøres ved at vælge knappen 'Align' og herefter funktionen 'Edit/Build Alignment'. Væg herefter protein, og så skulle der gerne være en blank alignment explorer klar til brug.
2. Download søgesekvensen fra linket:
www.bioteknologibogen.dk/bioteknologi-6/data/Insulin-preproprotein-menneske.txt
3. Indlæs sekvensen ved hjælp af fanebladet 'Edit' og herefter funktionen 'Insert Sequence From File'. Dit skærbillede skulle nu gerne se nogenlunde ud som vist herunder:

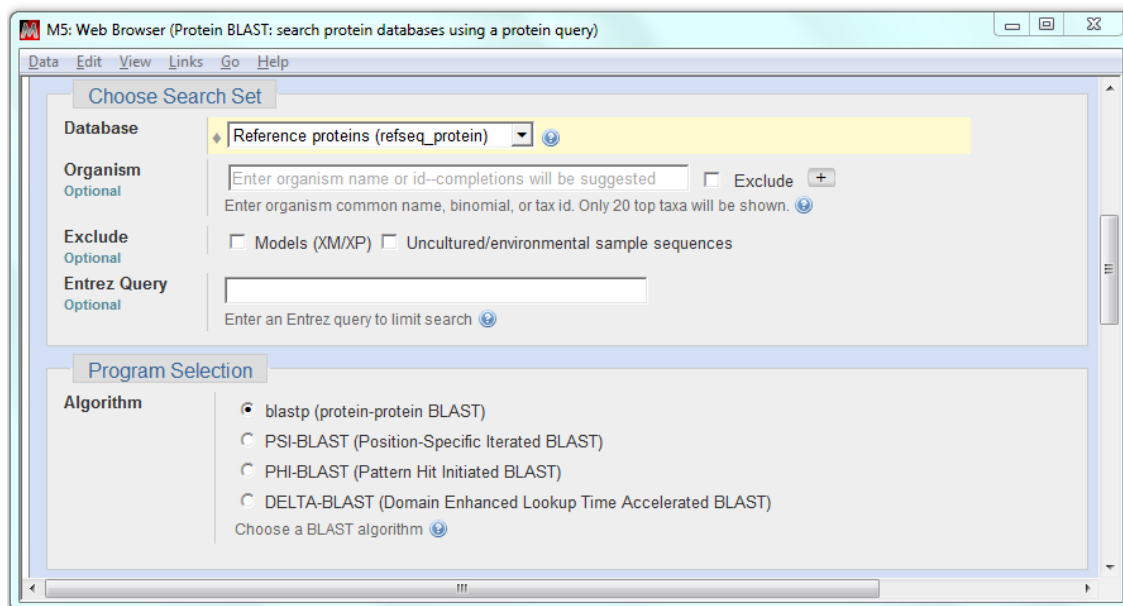


4. Markér sekvensen ved at trykke på navnet og vælg 'Web' og funktionen 'Do BLAST Search'. Nu åbnes et MEGA browser-vindue hvor den øverste del ser ud som vist herunder:



Læg mærke til at søgesekvensfeltet 'Enter Query Sequence' allerede er udfyldt med den ønskede sekvens. Under feltet 'Job Title' kan man give søgningen et navn. Hvis ikke man navngiver søgningen, får den et standardnavn. Læg også mærke til at man let kan skifte mellem de forskellige BLAST-programmer.

5. Brug nu scroll-knappen til at gå længere ned på siden og find de to underafsnit der hedder 'Choose Search Set' og 'Program Selection'. Skærbilledet herunder viser dette.



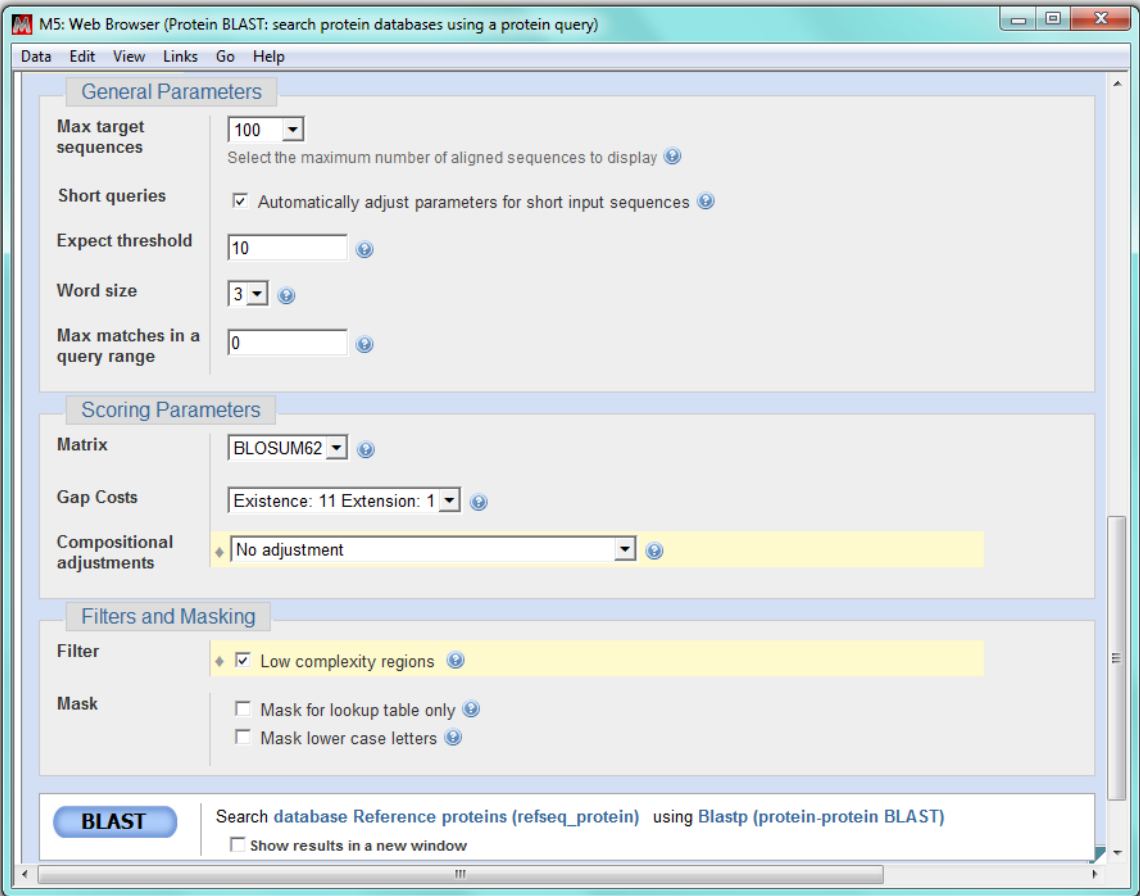
Her skal man træffe en række valg som for eksempel hvilken database man vil sammenligne med søgesekvensen. I feltet 'Organism' kan man tilvælge eller fravælge arter i databasen. Dette er særligt nyttigt hvis man har prøvet at køre en søgning og fået hundredvis af resultater fra eksempelvis mennesket, og man ikke er interesseret i disse. Ved at skrive human og trykke på exclude-knappen undgår man sekvenser fra mennesket i næste søgning. Sidst, men ikke mindst skal man vælge hvilken algoritme man ønsker at benytte blandt BLASTP, PSI-BLAST, PHI-BLAST eller DELTA-BLAST.

De forskellige databaser opdateres løbende, og der vil derfor forekomme ændringer med tiden, men da denne vejledning blev skrevet, var der følgende database-muligheder:

- Non-redundant protein sequences (nr) – Standarddatabasen med alle sekvenser i Genbank
 - Reference proteins (refseq_protein) – Kun sekvenser hvis funktion kendes
 - UniprotKB/Swissprot (swissprot) – Kun sekvenser hvis funktion er eksperimentelt bestemt
 - Patented protein sequences (pat) – En database over sekvenser der er søgt patent på
 - Protein Data Bank proteins (pdb) – En database over sekvenser med kendte proteinstrukturer
 - Metagenomic proteins (env_nr) – Sekvenser fundet i økologiske undersøgelser, art ofte ukendt
6. Vælg databasen 'Reference proteins (refseq_protein)' og algoritmen BLASTP til at lave søgningen i databasen. Hvis du har valgt dette, bør det se ud som vist på skærbilledet ovenfor. Det er værd at afprøve hvad der sker, hvis du søger i forskellige databaser eller bruger forskellige algoritmer.
7. Brug nu scroll-knappen til at gå længere ned på siden og find de tre underafsnit der hedder 'General parameters', 'Scoring parameters' og 'Filters and Masking'. Skærbilledet på næste side viser dette.

Bioteknologi 6, Tema 12 – Øvelser og opgaver

Linkadresserne fungerer pr. 1.11.2012. Forlaget tager forbehold for evt. ændringer i adresserne



The screenshot shows the NCBI BLAST web interface. The title bar reads 'M5: Web Browser (Protein BLAST: search protein databases using a protein query)'. The interface is divided into three main sections: General Parameters, Scoring Parameters, and Filters and Masking. In the General Parameters section, 'Max target sequences' is set to 100, 'Short queries' is checked, 'Expect threshold' is 10, 'Word size' is 3, and 'Max matches in a query range' is 0. In the Scoring Parameters section, the 'Matrix' is BLOSUM62, 'Gap Costs' are set to Existence: 11 and Extension: 1, and 'Compositional adjustments' are set to 'No adjustment'. In the Filters and Masking section, the 'Filter' is set to 'Low complexity regions' (checked), and 'Mask' options are unchecked. At the bottom, there is a 'BLAST' button and a search bar containing 'Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)'. There is also a checkbox for 'Show results in a new window'.

General Parameters

Her anbefaler jeg at man i langt de fleste tilfælde beholder standardværdierne, men hvis man vil ændre dem er her en kort beskrivelse af nogle af de væsentligste af funktionerne.

'Max target sequences' – Hvis man ønsker flere eller færre resultater end de 100, som er standard, kan det ændres her.

'Word size' – Hvis man ønsker en mere følsom søgning på mindre dele af ens søgesekvens, kan man ændre denne fra 3 til 2, dette resulterer i længere søgetid og højere sensitivitet, men måske også i et mindre overskueligt resultat.

'Expect threshold' – Denne funktion definerer søgningens specificitet ved at sætte en grænse for hvor statistisk signifikant et resultat skal være, før det medtages i resultaterne. Her betyder værdien 10 at vi kan forvente, at der i gennemsnit vil dukke 10 ikke-homologe sekvenser op der ved et rent tilfælde er ens nok med søgesekvenser, til at de kommer frem. Hvis man ønsker større sikkerhed for, at de fundne sekvenser er homologe med ens søgesekvens, skal værdien sættes ned.

Scoring parameters

Også her foreslår jeg at standardværdierne benyttes, men bemærk at MEGA desværre ikke benytter BLAST-standardværdierne. Ændringer fra standardværdierne er markeret med gult.

1. I feltet 'Compositional adjustments' vælges standardværdien som er 'Composition-based statistics'
2. I feltet 'Matrix' kan man vælge mellem flere forskellige empirisk baserede substitutionsmatricer, BLOSUM62 er standardvalget og bør bibeholdes med mindre man har god grund til andet.
3. Den sidste funktion 'Gap Costs' kontrollerer hvor mange og lange gaps '-' der kan indsættes i sekvensalignments, før en alignment i stedet deles i to. Heller ikke her bør den uøvede bruger ændre på standardværdien.

Filters and Masking

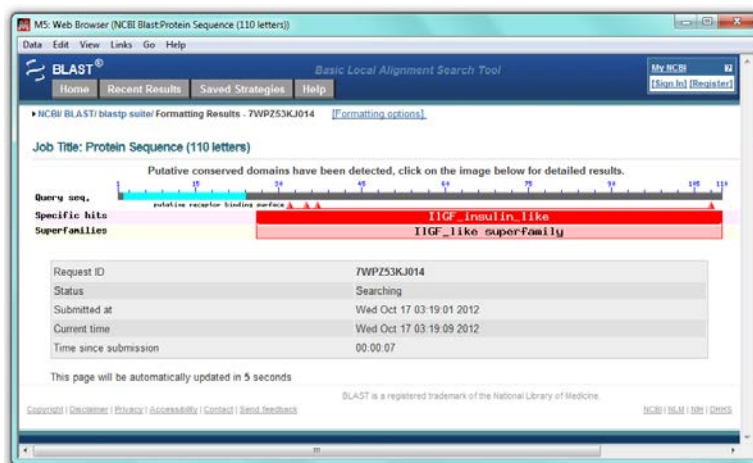
De tre funktioner herunder har alle at gøre med repetitive elementer i søgesekvensen. Repetitive elementer er dele af en sekvens der findes mange forskellige steder i genomet, og bliver de ikke ignoreret (maskeret) i søgningen, vil resultatet blive en lang række ofte ikke homologe sekvenser. Se siderne 54-60 i bioteknologi 6 for en grundigere gennemgang af forskellige typer af repetitive elementer.

Hvis man ved at en del af ens søgesekvens er repetitiv, kan man markere dette ved at skrive denne del af søgesekvensen i søgesekvensfeltet med små bogtaver og herefter markere 'Mask lower case letters'.

Man kan også lade søgeprogrammet definere dele af sekvensen der ser repetitiv ud ved at markere 'Low complexity regions'.

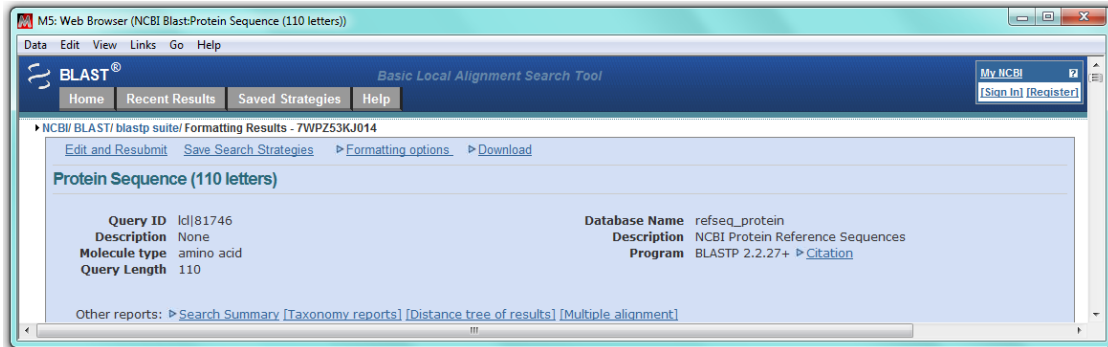
Hvis man vælger at lade programmet lede efter repetitive regioner, anbefaler jeg at man også vælger 'Mask for lookup table only'.

4. Marker alle tre ovenævnte muligheder.
5. Sæt kryds i feltet 'Show results in a new window'.
6. Start BLASTP-søgningen ved at trykke på den blå BLAST-knap. Nu dukker følgende vindue op, vinduet er et såkaldt ventevindue der opdateres løbende, mens søgningen er i gang:



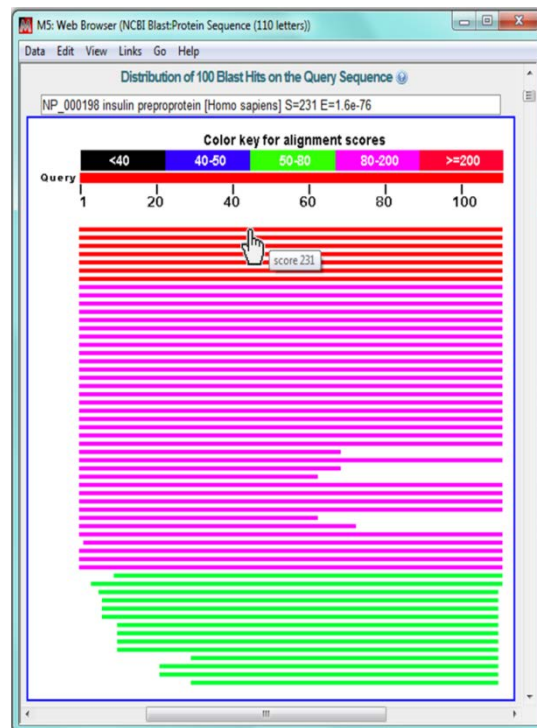
BLAST-Resultater

Efter forhåbentlig kort tid ændres vinduet til at vise resultaterne af din søgning. Øverste del af resultatvinduet er vist herunder:



Det vigtige her er at du kan få forskellige analyser af dine resultater, og du kan ændre på hvordan dine resultater bliver vist. Ændringer til resultatopsætningen gøres i 'Formatting options'. Du kan desuden downloade resultaterne til en fil med 'Download'-funktionen. På linjen i bunden af billedet der starter med 'Other reports', kan du få en række interessante oversigter over dine resultater.

1. Scroll ned til skærbilledet ser ud som vist her til højre. Hver linje repræsenterer et resultat, og jo højere oppe på listen, jo større lighed har resultatet med ens søgesekvens.
2. Placer musen henover den øverste røde linje. Nu vises information omkring denne sekvens i det øverste hvide felt. Her ses at sekvensen har Genbanknavnet NP_000198, er et insulin-præproprotein fra mennesket, alignment scoren er 231 og E-værdien er $1,6 \times 10^{-76}$. E-værdien siger groft sagt hvor mange ikke-homologe sekvenser i databasen vi forventer, vil være lige så godt et match med søgesekvensen ved et tilfælde. Eller sagt på en anden måde, det er et mål for hvor sandsynligt det er, at den fundne sekvens ikke er homolog med søgesekvensen, og denne værdi er ekstrem lav hvilket indikerer, at der er meget stærkt statistisk belæg for at den fundne sekvens er homolog med søgesekvensen.

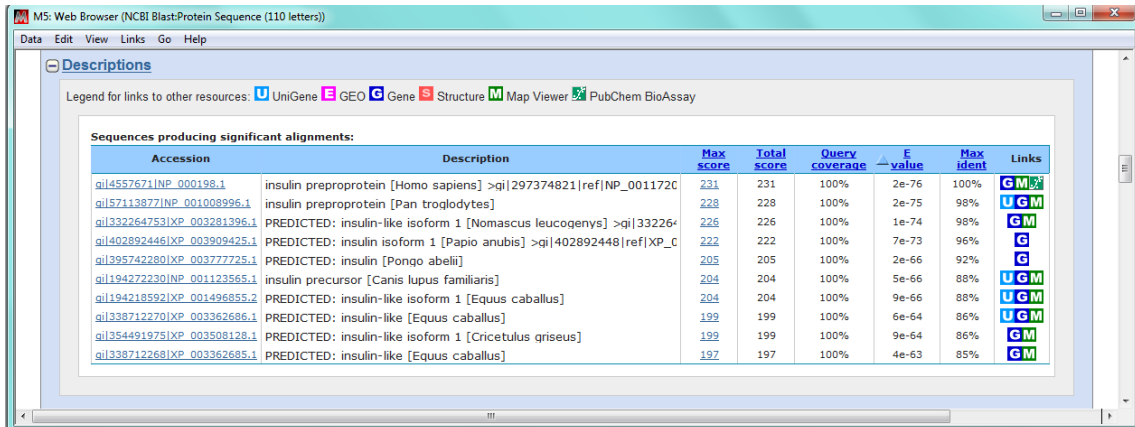


Jo lavere E-værdi, jo mere sandsynligt er det at søgesekvensen og resultatsekvensen er homologe sekvenser.

Bioteknologi 6, Tema 12 – Øvelser og opgaver

Linkadresserne fungerer pr. 1.11.2012. Forlaget tager forbehold for evt. ændringer i adresserne

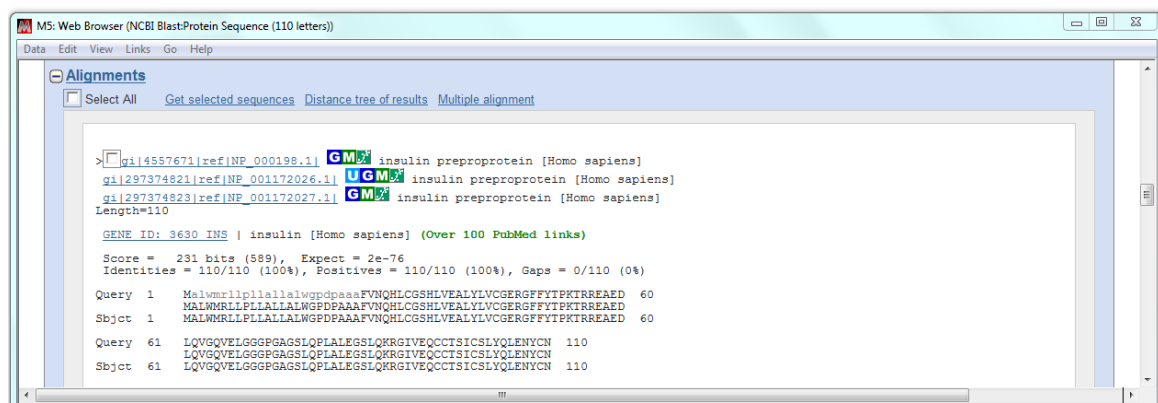
3. Scroll lidt længere ned på resultatsiden til underafsnittet 'Descriptions'. Se billedet herunder. Læg mærke til at jeg her har valgt — øverst på siden under 'Formatting options' — kun at vise 10 resultater.



Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
gi 4557671 NP_000198.1	insulin preproprotein [Homo sapiens] >gi 297374821 ref NP_001172C	231	231	100%	2e-76	100%	G M P
gi 57113877 NP_001008996.1	insulin preproprotein [Pan troglodytes]	228	228	100%	2e-75	98%	G M P
gi 332264753 XP_003281396.1	PREDICTED: insulin-like isoform 1 [Nomascus leucogenys] >gi 332264	226	226	100%	1e-74	98%	G M P
gi 402892446 XP_003909425.1	PREDICTED: insulin isoform 1 [Papio anubis] >gi 402892448 ref XP_C	222	222	100%	7e-73	96%	G M P
gi 395742280 XP_003777725.1	PREDICTED: insulin [Pongo abelii]	205	205	100%	2e-66	92%	G M P
gi 194272230 NP_001123565.1	insulin precursor [Canis lupus familiaris]	204	204	100%	5e-66	88%	G M P
gi 194218592 XP_001496855.2	PREDICTED: insulin-like isoform 1 [Equus caballus]	204	204	100%	9e-66	88%	G M P
gi 338712270 XP_003362686.1	PREDICTED: insulin-like [Equus caballus]	199	199	100%	6e-64	86%	G M P
gi 354491975 XP_003508128.1	PREDICTED: insulin-like isoform 1 [Cricetulus griseus]	199	199	100%	9e-64	86%	G M P
gi 338712268 XP_003362685.1	PREDICTED: insulin-like [Equus caballus]	197	197	100%	4e-63	85%	G M P

Her er de 10 bedste resultater opsummeret. I forhold til før er der tre kolonner mere. Den ene hedder 'Query coverage' og viser hvor stor en procentdel af søgesekvensen der er alignet til denne resultatsekvens. E-værdien (E-value) er ekstremt lille for alle 10 resultater. Kolonnen 'Max ident' viser i hvor mange procent søgesekvensen og resultatsekvensen er ens i den dannede parvise alignment. Nederste match på listen har beskrivelsen 'PREDICTED: insulin-like' og er fra en hest *Equus caballus*. At en sekvens kaldes predicted betyder at den ikke er eksperimentelt undersøgt, men at bioinformatiske undersøgelser tyder på at dette er et insulinlignende protein. Generelt er sekvens-ID-navne der starter med NP eksperimentelt undersøgte, mens ID-navne der starter med XP, kun er undersøgt ved hjælp af bioinformatik.

4. Undersøg hvilke arter de 10 sekvenser kommer fra og hvilken funktion de enkelte proteiner har.
5. Der er flere sekvenser fra heste på listen, hvad kan årsagen til dette være?
6. Scroll længere ned på siden til underafsnittet 'Alignments'. Nu ses følgende som en del af skærbilledet:



```

>gi|4557671|ref|NP_000198.1| G M P insulin preproprotein [Homo sapiens]
gi|297374821|ref|NP_001172C26.1| U G M insulin preproprotein [Homo sapiens]
gi|297374823|ref|NP_001172027.1| G M P insulin preproprotein [Homo sapiens]
Length=110
GENE ID: 3630 INS | insulin [Homo sapiens] (Over 100 PubMed links)
Score = 231 bits (589), Expect = 2e-76
Identities = 110/110 (100%), Positives = 110/110 (100%), Gaps = 0/110 (0%)

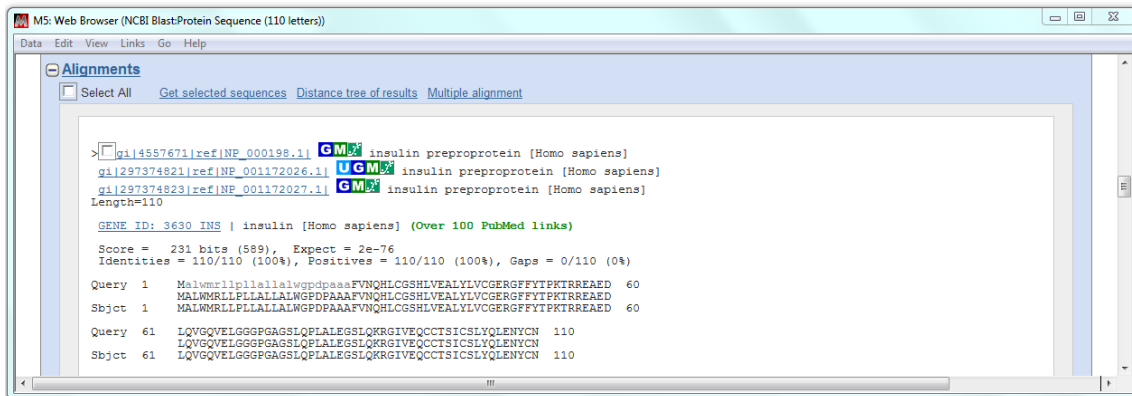
Query 1  MALVMRLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED 60
Sbjct 1  MALVMRLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED 60

Query 61  LQVGVQLGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIQSLYQLENYCN 110
Sbjct 61  LQVGVQLGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIQSLYQLENYCN 110
  
```

Nederst ser vi den dannede parvise alignment mellem søgesekvensen og resultatsekvensen. Scroller man endnu længere ned på siden, vises resten af resultatsekvenserne også. Ser man nærmere på den parvise alignment, ser man at søgesekvensen (query) er øverst og resultatsekvensen (Sbjct) er nederst, mens der mellem dem er en såkaldt konsensussekvens. I dette tilfælde hvor de to sekvenser er ens, er konsensussekvensen blot den samme sekvens.

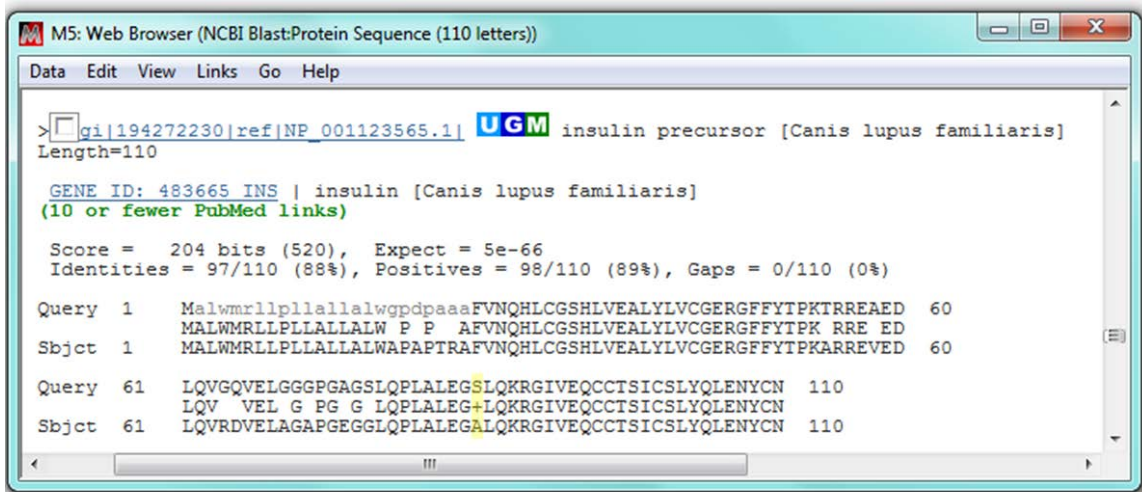
Bioteknologi 6, Tema 12 – Øvelser og opgaver

Linkadresserne fungerer pr. 1.11.2012. Forlaget tager forbehold for evt. ændringer i adresserne



En anden ting der er værd at bemærke, er at første del af søgesekvensen, bortset fra den første methionin-aminosyre, er skrevet med små bogstaver i en grå farve. Dette betyder at BLASTP har behandlet denne del af sekvensen som et repetitivt område, og derfor har denne del af søgesekvensen ikke været brugt til at finde de resulterende sekvenser, men som jeg nævnte tidligere kan man stadig få programmet til at udvide den parvise alignment til at dække hele søgesekvensen, hvis det giver mening. Det er præcis det der er sket i dette tilfælde.

7. Hvor mange aminosyrer er henholdsvis ens og forskellige i de to viste sekvenser?
8. Tryk på det blå 'GENE ID: 3630 INS' link. Du kommer nu til PubMed-artikeldatabasen hvor du kan finde og læse videnskabelige artikler om denne sekvens.
9. Vend tilbage til BLAST-resultaterne og scroll ned til alignment nummer 6. Databasesekvensen er fra en hund (*Canis lupus familiaris*). Den del ser ud som vist på nedenstående skærmbillede:

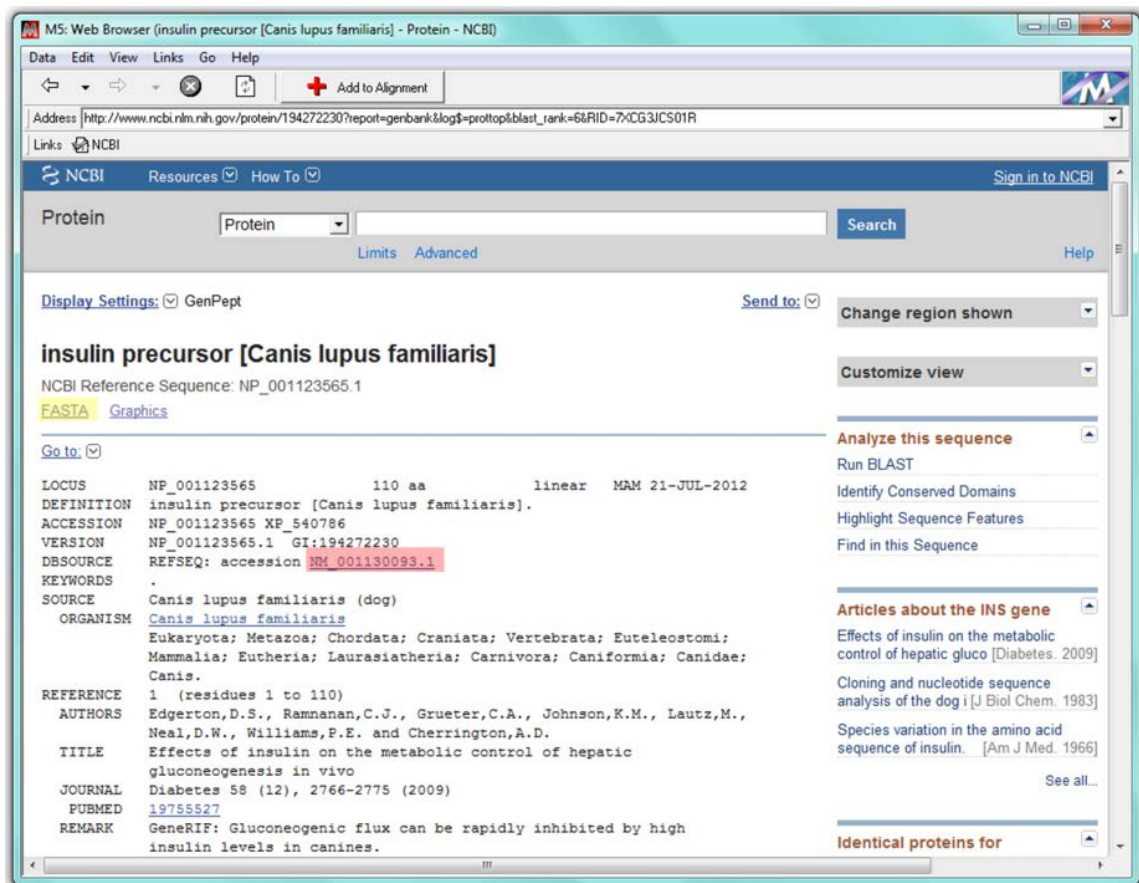


Vi ser med det samme at der i dette tilfælde kun er 97 ud af de 110 aminosyrer, der er i søgesekvensen som er ens i resultatsekvensen, hvilket giver en identitet på ca. 88 %. Hvis vi ser på den viste alignment, kan vi nu se at konsensussekvensen har 12 felter, hvor der ikke står noget bogstav, det skyldes at søgesekvensen og resultatsekvensen i denne alignmentposition er forskellige. Et enkelt sted ser vi et '+' i konsensussekvensen, på billedet har jeg valgt at fremhæve dette med gult. Aminosyrerne i de to sekvenser er henholdsvis S (serin) og A (alanin). Et sådan '+' betegnes som en 'positiv'. Positive er summen af identiske aminosyrer og nogle særlige forskelle i enkelte positioner i alignmenten, disse forklares nærmere i boksen øverst på næste side.

Positive

En positiv i en BLAST-alignment indikerer at denne aminosyre har udviklet sig specielt interessant. Forklaringen findes ved at sammenligne den substitutionsmatrice der bruges i søgningen (BLOSUM62), og den fulde liste af resultater. Hvis en bestemt ændring oftere forekommer mellem søgesekvensen og forskellige sekvenser i databasen end hvad man statistisk ville forvente ud fra substitutionsmatricen, kalder vi den for en positiv. En positiv er en indikation af at netop denne aminosyre tilsyneladende er så funktionel vigtig, at den samme ændring er sket flere gange igennem evolutionen i flere forskellige arter end vi vil forvente sker ved et tilfælde.

10. Undersøg hvorvidt den samme aminosyreændring S/A i samme position i proteinet kan findes i andre sekvensalignments. Hvis du finder nogle, så forklar hvorvidt der er tale om parallel evolution.
11. Tryk nu på linket ud for hundens sekvens 'gi | 194272230 | NP_001123565.1'. Nu kommer du ind på Genbanksiden for denne sekvens. Der er mange informationer i disse databaser, her vil vi kun fokusere på to ting. Skærbilledet bør nu se ud som vist herunder:



MS: Web Browser (insulin precursor [Canis lupus familiaris] - Protein - NCBI)

Address: [http://www.ncbi.nlm.nih.gov/protein/194272230?report=genbank&log\\$=prottop&blast_rank=6&RID=7XCG3JCS01R](http://www.ncbi.nlm.nih.gov/protein/194272230?report=genbank&log$=prottop&blast_rank=6&RID=7XCG3JCS01R)

Links: [NCBI](#)

Protein: [Limits](#) [Advanced](#) [Help](#)

Display Settings: ☒ GenPept [Send to:](#)

insulin precursor [Canis lupus familiaris]
NCBI Reference Sequence: NP_001123565.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NP_001123565 110 aa linear MAM 21-JUL-2012

DEFINITION insulin precursor [Canis lupus familiaris].

ACCESSION NP_001123565 XP_540786

VERSION NP_001123565.1 GI:194272230

DBSOURCE REFSEQ: accession [NM_001130093.1](#)

KEYWORDS .

SOURCE [Canis lupus familiaris](#) (dog)

ORGANISM [Canis lupus familiaris](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Laurasiatheria; Carnivora; Caniformia; Canidae;
Canis.

REFERENCE 1 (residues 1 to 110)

AUTHORS Edgerton,D.S., Ramnathan,C.J., Grueter,C.A., Johnson,K.M., Lautz,M.,
Neal,D.W., Williams,P.E. and Cherrington,A.D.

TITLE Effects of insulin on the metabolic control of hepatic
gluconeogenesis in vivo

JOURNAL Diabetes 58 (12), 2766-2775 (2009)

PUBMED [19755527](#)

REMARK GeneRIF: Gluconeogenic flux can be rapidly inhibited by high
insulin levels in canines.

Analyze this sequence
[Run BLAST](#)
[Identify Conserved Domains](#)
[Highlight Sequence Features](#)
[Find in this Sequence](#)

Articles about the INS gene
[Effects of insulin on the metabolic control of hepatic gluco \[Diabetes. 2009\]](#)
[Cloning and nucleotide sequence analysis of the dog i \[J Biol Chem. 1983\]](#)
[Species variation in the amino acid sequence of insulin. \[Am J Med. 1966\]](#)
[See all...](#)

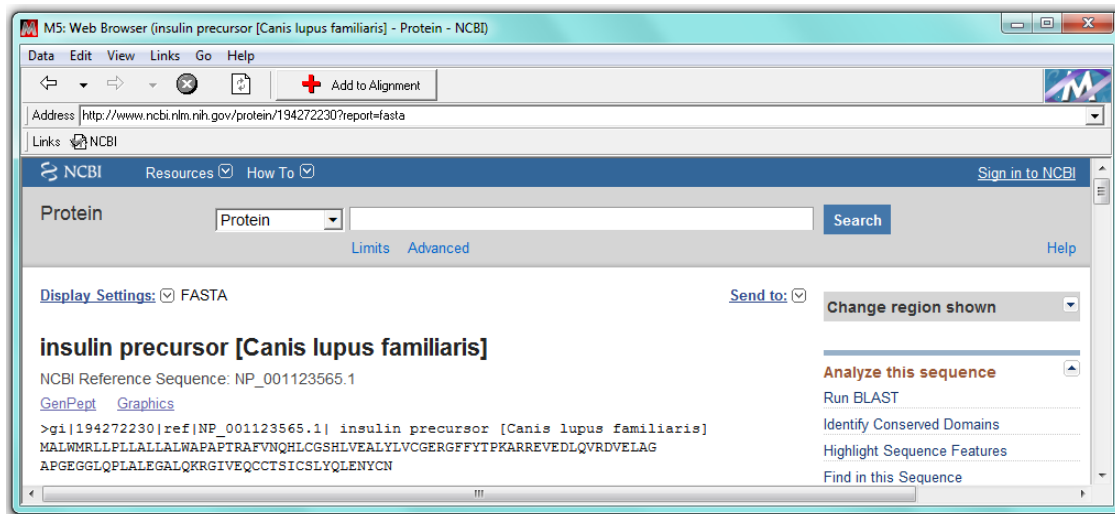
Identical proteins for

Hvis vi starter med det med rødt markerede link, så er det et link til mRNA for dette protein. Hvis man skulle ønske at arbejde med mRNA'et i stedet for proteinsekvensen, kan den findes ved at trykke på dette link. Den anden vigtige ting er 'FASTA'-linket der er markeret med gult på skærbilledet. FASTA er et sekvensformat og trykker man på linket, får man proteinsekvensen frem i et let tilgængeligt format der nemt kan tilføjes til ens alignment.

Bioteknologi 6, Tema 12 – Øvelser og opgaver

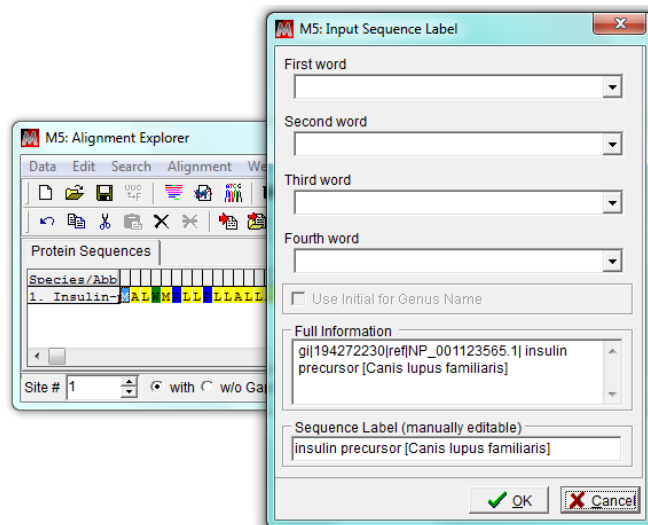
Linkadresserne fungerer pr. 1.11.2012. Forlaget tager forbehold for evt. ændringer i adresserne

12. Tryk på 'FASTA' linket, nu bør skærmbilledet se ud som vist herunder:



Nu ser du den 110 aminosyrer lange sekvens.

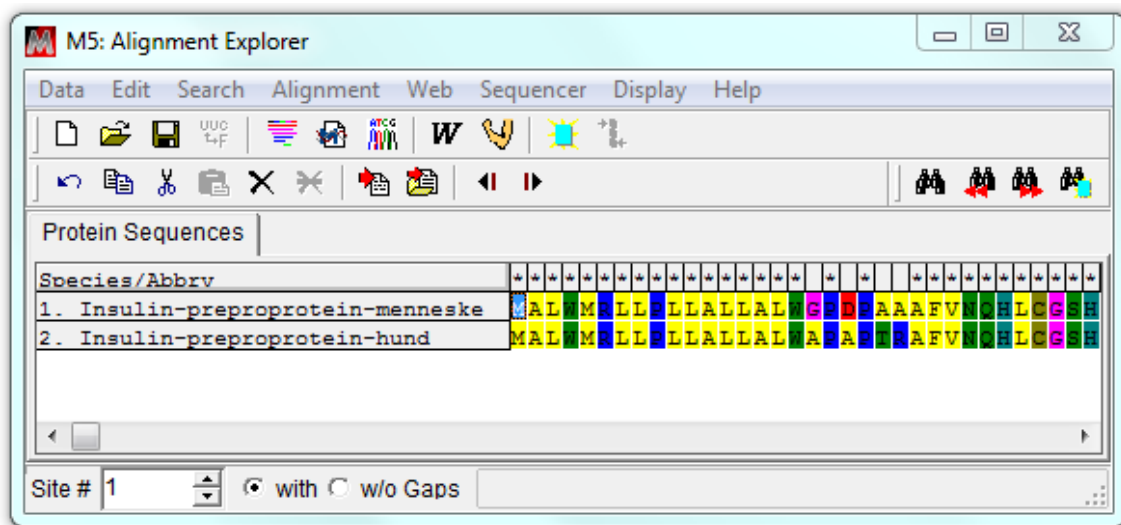
13. Tryk på knappen øverst med det røde plus og titlen 'Add to alignment'. Nu kommer alignment explorer-vinduet frem samt et vindue hvor du kan indtaste information om den tilføjede sekvens.



14. Det eneste der skal ændres, er den nederste tekstboks hvor der står 'Sequence Label'. Erstat det skrevne med 'Insulin-preproprotein-hund' og tryk på OK-knappen. Dit alignment explorer-vindue skulle nu gerne have to sekvenser, den oprindelige fra mennesket og den nye fra hunden. Skærmbilledet bør se ud i stil med det viste øverst på næste side.

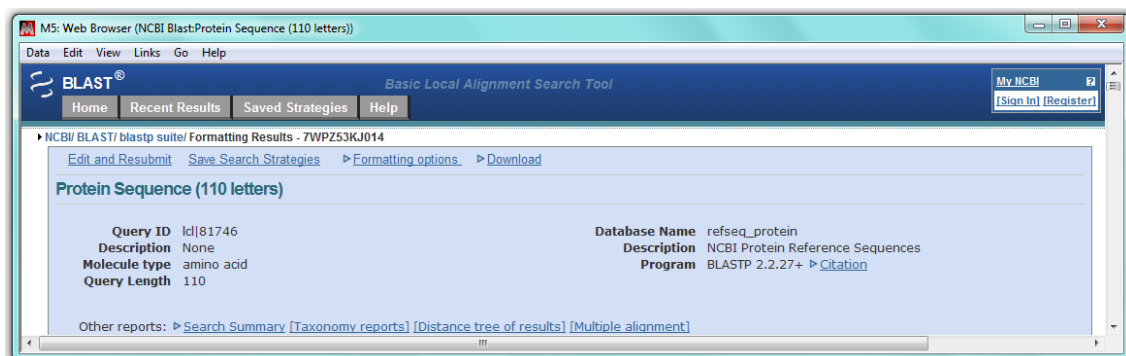
Bioteknologi 6, Tema 12 – Øvelser og opgaver

Linkadresserne fungerer pr. 1.11.2012. Forlaget tager forbehold for evt. ændringer i adresserne



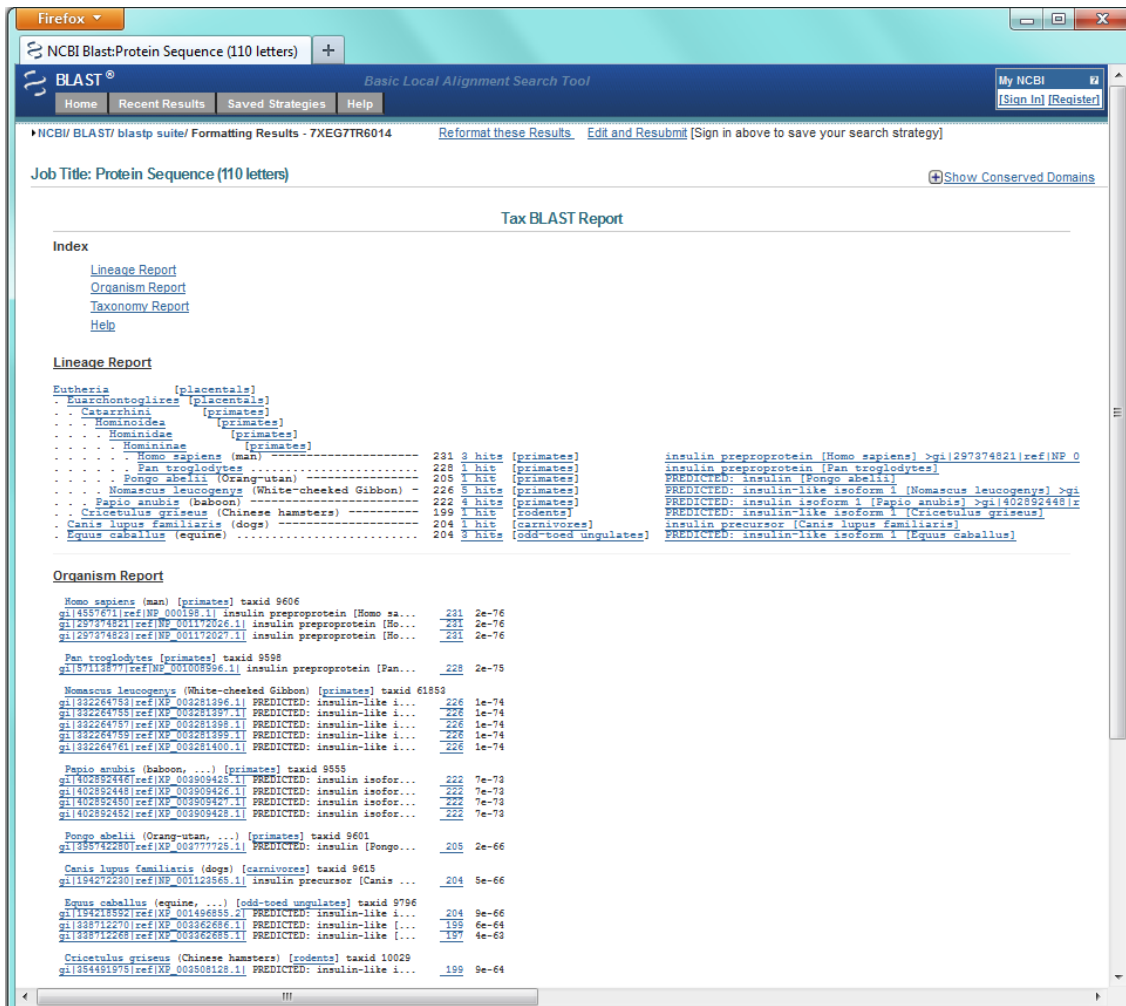
Da de to sekvenser var præcis lige lange, er de allerede alignet ganske fornuftigt. Læg mærke til at der flere steder er en forskellig aminosyre i de to proteinsekvenser. I alt kan vi tælle 13 forskelle ud af de 110 aminosyrer, altså ca. 88 % identitet. Dette passer meget godt med det vi ser med andre gener fra mennesket og hunden, og er resultatet af to udviklingslinjer der sidst havde en fælles forfar for ca. 100 mio. år siden.

15. Vend tilbage til BLASTP-resultatsiden. Vinduet ser ud som vist herunder. Hvis ikke det kan genfindes, kan det altid genskabes ved at køre en ny BLASTP med den oprindelige søgesekvens, resultatet skulle meget gerne blive det samme som før! Vi vil nu se nærmere på resten af resultatinformationerne.



Yderligere resultatsider

1. Vælg det link der hedder 'Taxonomy reports'. Et nyt vindue åbner med en oversigt over hvilke arter resultatsekvenserne kommer fra og disse arters taksonomiske placering i livets træ. Hvis man kun vælger at se på 10 sekvenser, ser det ud som vist herunder.



The screenshot shows the NCBI BLAST Tax BLAST Report interface. The top section displays the job title 'Protein Sequence (110 letters)' and a 'Show Conserved Domains' button. Below this, the 'Tax BLAST Report' is shown, which includes a 'Lineage Report' and an 'Organism Report'.

Lineage Report

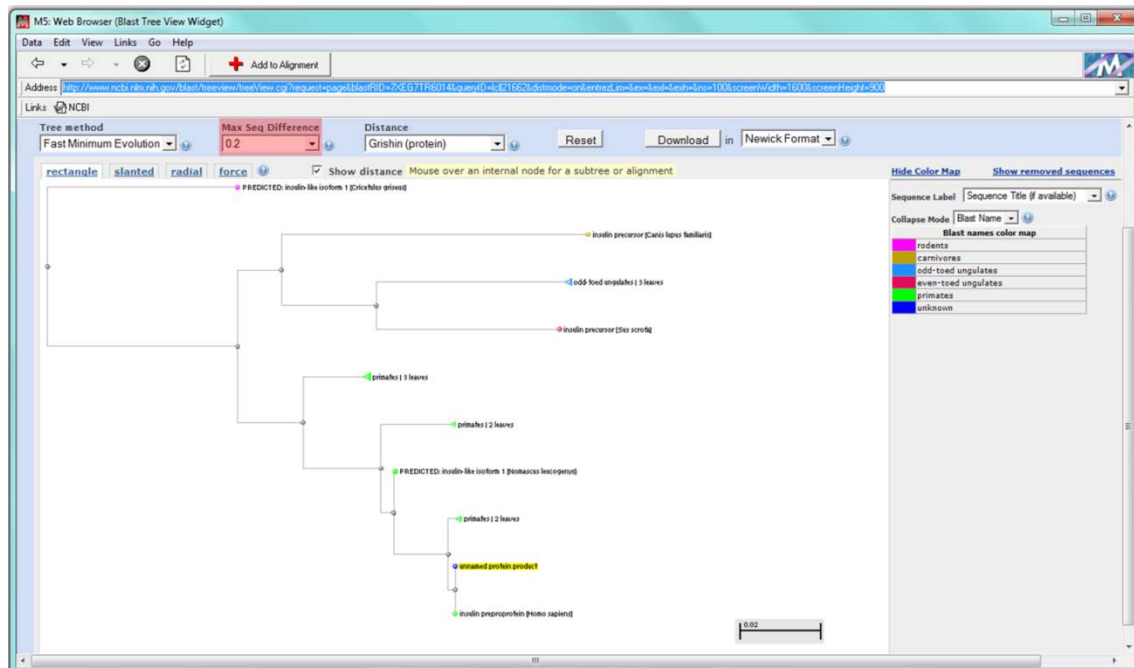
Lineage	Hit Count	Organism
Eutheria [placentalia]	231	3 hits [primates]
Marschallia [placentalia]	228	1 hit [primates]
Catarrhini [primates]	208	1 hit [primates]
Hominidae [primates]	226	3 hits [primates]
Homininae [primates]	222	4 hits [primates]
Homo sapiens (man)	199	1 hit [rodents]
Pan troglodytes	204	1 hit [carnivores]
Pongo abelii (Orang-utan)	204	3 hits [odd-toed ungulates]
Nomascus leucogenys (White-cheeked Gibbon)		
Papio anubis (baboon)		
Cricetulus griseus (Chinese hamsters)		
Canis lupus familiaris (dogs)		
Equus caballus (equine)		

Organism Report

Organism	Hit Count	Accession	Score
Homo sapiens (man) [primates] taxid 9606	231	gi 4357671 ref NP_000198.1	2e-76
Pan troglodytes [primates] taxid 9598	228	gi 15113377 ref NP_001008896.1	2e-75
Nomascus leucogenys (White-cheeked Gibbon) [primates] taxid 61858	226	gi 332264758 ref XP_002281396.1	1e-74
Papio anubis (baboon, ...) [primates] taxid 9555	222	gi 402892448 ref XP_002809426.1	7e-73
Pongo abelii (Orang-utan, ...) [primates] taxid 9601	208	gi 129737482 ref NP_00172027.1	2e-76
Canis lupus familiaris (dogs) [carnivores] taxid 9615	204	gi 139472230 ref NP_001038868.1	5e-66
Equus caballus (equine, ...) [odd-toed ungulates] taxid 9796	204	gi 129737482 ref NP_00172027.1	2e-76
Cricetulus griseus (Chinese hamsters) [rodents] taxid 10029	199	gi 254491975 ref XP_002808129.1	9e-64

Denne rapport viser at der er fundet sekvenser fra otte forskellige organismer. Man kan også se i den øverste hierarkiske opbygning at alle otte organismer er placental pattedyr. Derudover kan man se at det bedste match er mennesket, efterfulgt af chimpanse, gibbon, bavian, orangutang, hund, hest og kinesisk hamster. Har man flere end 10 sekvenser med, vil der også være ikke-pattedyr på listen.

2. Gå tilbage til resultatsiden og tryk på linket 'Distance tree of results'. Nedenstående skærbillede viser resultatet:



Det er muligt du får et noget mere rodet resultat frem på skærmen med mange flere sekvenser. Dette kan ændres under menuen 'Max. Seq Difference', markeret med rødt på billedet, hvor jeg har valgt 0.2 svarende til at jeg kun vil se resultater fra de sekvenser, der er mindre end 20 % forskellige fra min søgesekvens. Søgesekvensen er markeret med gult. Hvordan skal dette fylogenetiske træ så analyseres? Umiddelbart er det let at se at søgesekvensen er identisk med en fra mennesket. Vandrette afstande i træet viser genetisk afstand mellem sekvenserne. Grønne trekanten betyder at der er flere end en sekvens, og dem kan man få at se ved at trykke på en grøn trekant og vælge 'Expand/Collapse'. Man kan også manipulere på træets form og udseende på andre måder, men ikke dets biologiske indhold! Indholdet er bestemt af den biologiske information gemt i sekvenslighederne. For mere information omkring fylogenetiske træer, se Bioteknologi 6, link 5-7 i tema 12, eller siderne 27-41.

3. Hvilke to sekvenser, udover menneskesekvensen, er mest ens med søgesekvensen?
4. Fra hvilken art stammer den sekvens der er mindst ens med søgesekvensen?

Opgaver til databasesøgning

Herunder er der tre opgaver der alle kan løses ved hjælp af databasesøgninger med BLASTP.

Opgave 1 – Fusk med fiskesælget

Et supermarked sælger en fisk der ifølge reklamen er frit fanget atlantehavslaks (*Salmo salar*), men er den nu også det? Du er udsendt af Fødevarestyrelsen for at undersøge et rygte om at dette supermarked nogle gange snyder og sælger billigere dambrugsopdrættede fisk under påstanden at det er frit fanget laks. Ud fra en vævsprøve fra en af fiskene i køledisken har du udvundet og sekventeret følgende korte proteinsekvens på 40 aminosyrer:

Proteinsekvens fra fisk i køledisken:

LEMYCAPVKSGKAARSVRAQRHTDMPRTPKVSTAVQSVDR

1. Har supermarkedet rent mel i posen? – Og hvis ikke kødet er fra Atlanterhavslaksen, hvilken art er det så fra?

Opgave 2 – Undersøgelse af resistente bakterier

I en gymnasieklasse har man dyrket to stammer af *E. coli*-bakterier på petriskåle hvoraf den ene viste resistens overfor penicillin, mens den anden ikke gjorde. Ved hjælp af protein-gelelektroforese undersøgte man hvilke proteiner de to bakteriestammer hver i sær producerede. Eleverne identificerede og oprensede et unikt proteinbånd der kun var til stede hos den resistente bakteriestamme. Eleverne indsendte det oprensede protein til sekventering på Århus Universitet og fik følgende proteinsekvens tilbage:

MFKTTLCALLITASCSTFAAPQKINDIVHRTITPLIEQQKIPGMAVAVIYQGKPYFTWGYADI
AKKQPV

1. Hvilket protein er der tale om, og kan det have noget at gøre med den resistente *E. coli* stammes evne til at overleve i et miljø hvor der er tilsat penicillin?

Opgave 3 – Er odderen derude?

En forsker arbejder med at undersøge odderens (*Lutra lutra*) udbredelse i Danmark. Ifølge lokale naturinteresserede har de set oddere i to mindre søer på Fyn. Disse folk har fundet ekskrementer ved begge søbredder der måske er fra oddere. Fra ekskrementerne er det lykkedes forskeren at udvinde følgende proteinsekvenser:

Sø-1 sekvens:

MVLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPPTTKTYFPHFDLSPGSAQVKAHGK
KVADALTTA

Sø-2 sekvens:

MVLSPADKTNIKTAWEKIGSHGGEYGAEAVERMFLGFPTTKTYFPHFDFTHGSESIKAHGKK
VSEALTKA

1. Hvilket protein er sekvenserne fra, og er det bevis på at der er oddere i de to søer som de lokale påstår?