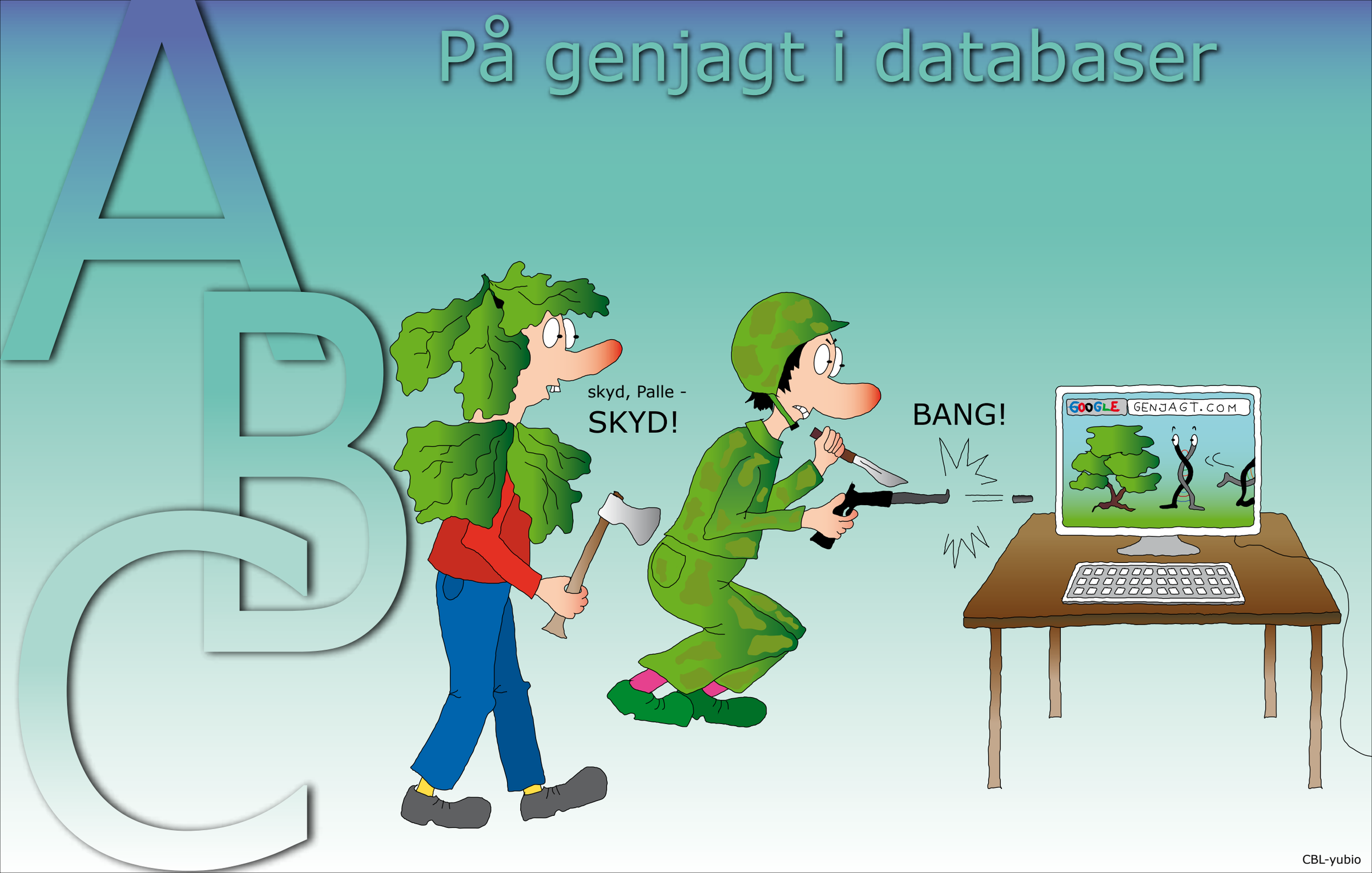
På genjagt i databaser



CBL-yubio

Formål
I dette teoretiske eksperiment skal I gå på jagt i databaser på nettet. I får udleveret forskellige DNA- og aminosyre-sekvenser, som I skal forsøge at finde ud af, hvor stammer fra. Altså hvilke organismer kommer de fra, og hvilket protein kodes der for?

Teori
Der findes mange hjemmesider med informationer om proteiner. En af disse hjemmesider er Protein Data Bank (PDB), der findes her:
http://www.pdb.org/pdb/home/home.do
På hjemmesiden kan man finde både DNA- og aminosyre-sekvenser, ligesom man kan finde 3D-modeller af alverdens proteiner.

På bilag 1 findes DNA-sekvenser opskrevet i et bestemt format, der kaldes FASTA. Dette format kan bruges i specielle programmer til at lave alignments med allerede kendte DNA-sekvenser. Derved kan man måske finde et 100 % match og dermed identificere ens ukendte DNA-sekvens.

Samme FASTA-format kan benyttes ved aminosyre-sekvenser, hvor man med de samme programmer laver alignments med kendte aminosyre-sekvenser. Også her kan man måske finde et match og derved identificere, hvilket protein der er tale om. Disse data findes på bilag 2.

Til at lave DNA-alignments bruges et program, der hedder BLAST, der står for Basic Local Alignment Search Tool. Til at lave aminosyre-alignments bruges et tilsvarende program, der hedder BLASTP. Forkortelsen står for det samme, og det sidste P står for Protein.
Begge programmer findes på denne hjemmeside:
http://blast.ncbi.nlm.nih.gov/Blast.cgi

Når man har en DNA-sekvens, kan man translatere den helt automatisk, og dermed finde ud af hvilke aminosyrer sekvensen koder for. Det kræver dog, at man kender læserammen. Det er ikke ligegyldigt, hvor den ligger. Hjemmesiden Virtual Ribosome kan netop foretage sådanne translationer, hvor det undersøger alle mulige læserammer. Her er vist et eksempel på en sekvens: CATGCGA

Der er umiddelbart tre mulige læserammer:
CAT GCG A
C ATG CGA
CA TGC GA

Men faktisk kan der også læses fra den anden side, så der kommer yderligere tre muligheder:
AGC GTA C
A GCG TAC
AG CGT AC

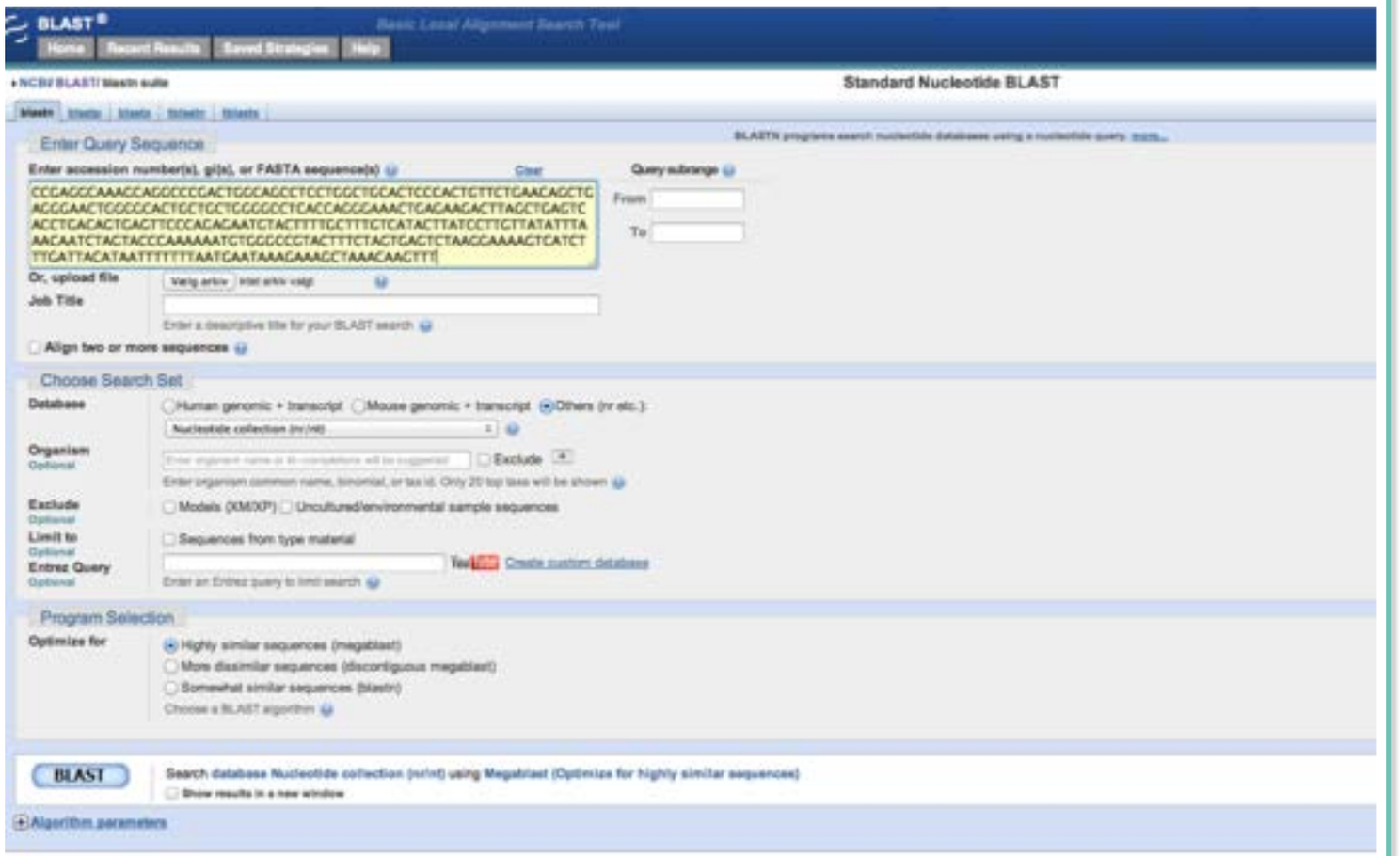
Der findes altså teoretisk 6 mulige læserammer på en hvilken som helst DNA-sekvens. "Virtual Ribosome 2.0" kan finde alle seks og angive startkoder, stopkoder og hvilke aminosyrer, de enkelte kodons/tripletter koder for. Herved kan man finde den mest sandsynlige læseramme, som er den, der indeholder en startkode, et ukendt antal kodons og en stopkode:
https://services.healthtech.dtu.dk/service.php?VirtualRibosome-2.0
Yderligere information om bioinformatik og metoder kan findes på biotech-academys hjemmeside:
https://www.biotechacademy.dk/undervisning/gymnasiale-projekter/bioinformatik-intro/

Materialer
Hjemmesiderne nævnt i teoriafsnittet samt data på de to bilag.

Fremgangsmåde

Undersøgelse af DNA-sekvenser

Åbn hjemmesiden hvor du kan udføre en BLAST-analyse.
Vælg ”nucleotide blast” ovre til venstre på skærmen.
Kopier den første DNA-sekvens fra bilag 1 – sørg for kun at kopiere baserne.
Indsæt baserne som vist nedenfor (gul farve).
Tryk BLAST (blå knap nederst på skærmen).



Nu foretages et BLAST af sekvensen, hvor den sammenlignes med alle kendte DNA-sekvenser. Efter noget tid (der kan gå op til flere minutter) fås et resultat.

(se næste side)

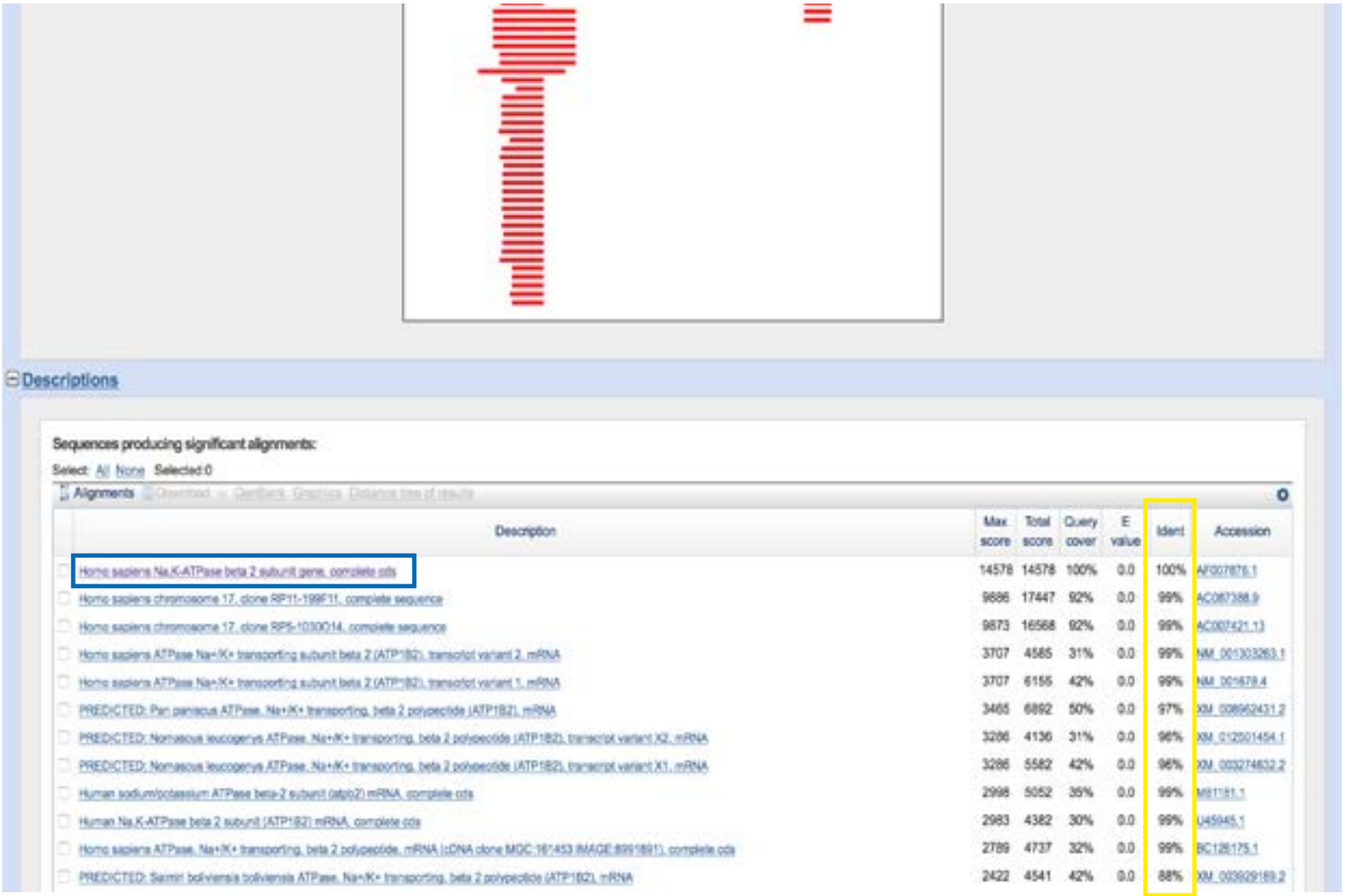
Hvis du scroller ned på skærmen fås en highscoreliste over de bedste alignments ud fra DNA-sekvensen. Længst ude til højre på listen ses en %-sats (se den gule markering). 100 % betyder, at der er fuld alignment med den pågældende sekvens.
Hvis du trykker på den pågældende organisme kommer alignmentet frem (se blå markering).

Nu kommer en ny skærm frem, og helt ovre til højre står der "Related information" - tryk på "Gene".

Nu kommer en lang række oplysninger om genet.
Noter hvilken organisme genet kommer fra.
Hvad koder genet for?

Under overskriften "Genomic context" kan du se, hvor mange exons, genet indeholder. Her ses det også, hvilket kromosom, genet er placeret på.
Noter resultaterne i resultatskemaet.

Udfør en BLAST-analyse af DNA-sekvens nr. 2 fra bilag 1
(kopier den ind ad 2 omgange).
Noter hvilken organisme genet kommer fra.
Noter også, hvilket protein genet koder for.
Hvor mange exons består genet af?
Hvilket kromosom er det placeret på?

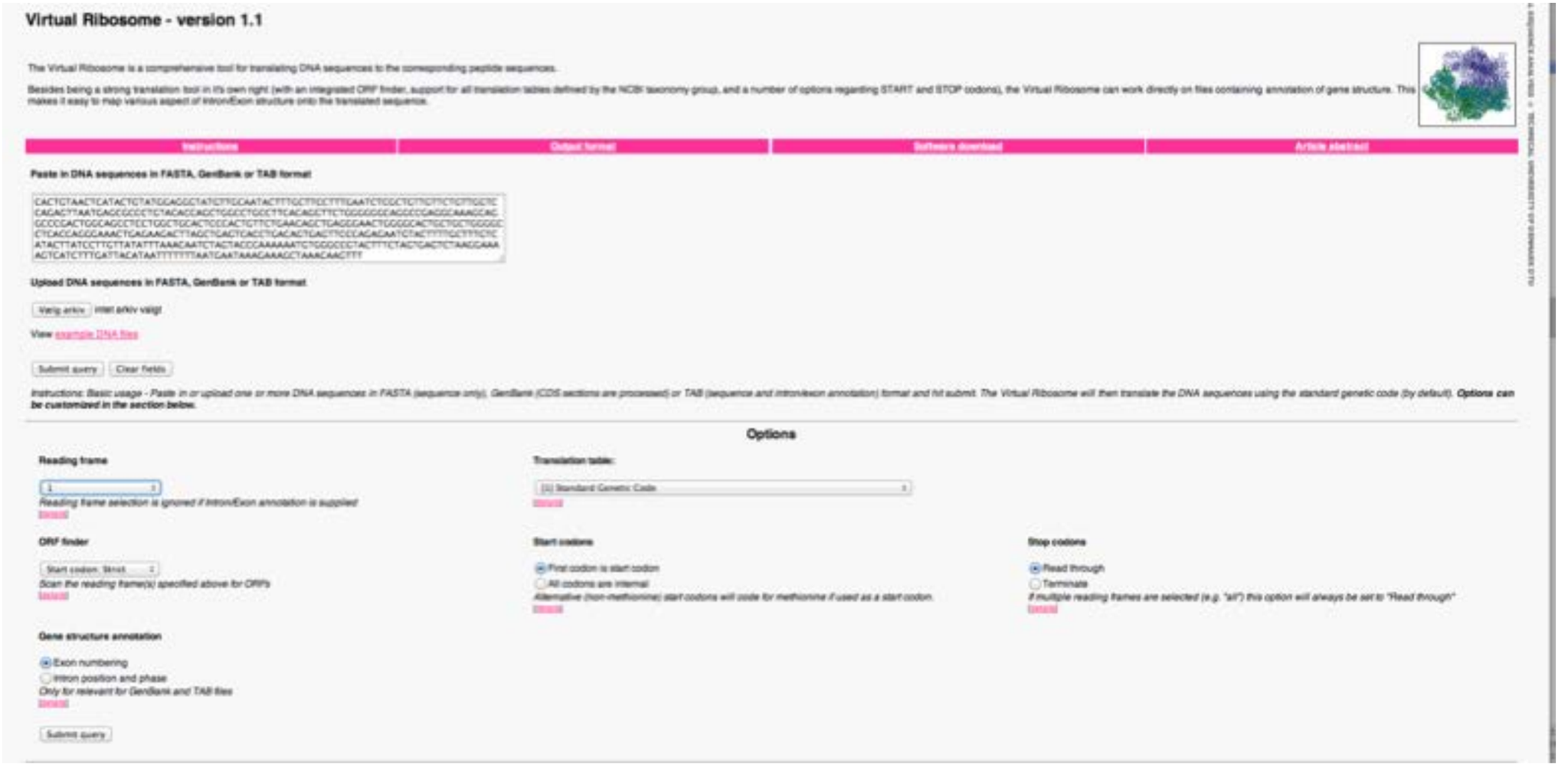


Undersøgelse af aminosyre-sekvenser
Fremgangsmåden er her den samme som ovenfor, men nu skal der vælges en BLASTP-analyse i stedet for. Det gøres ved at vælge "protein blast" på forsiden. Kopier nu den første aminosyresekvens (bilag 2) ind i feltet som før og tryk BLAST. Ved hver af de fire sekvenser besvares følgende:

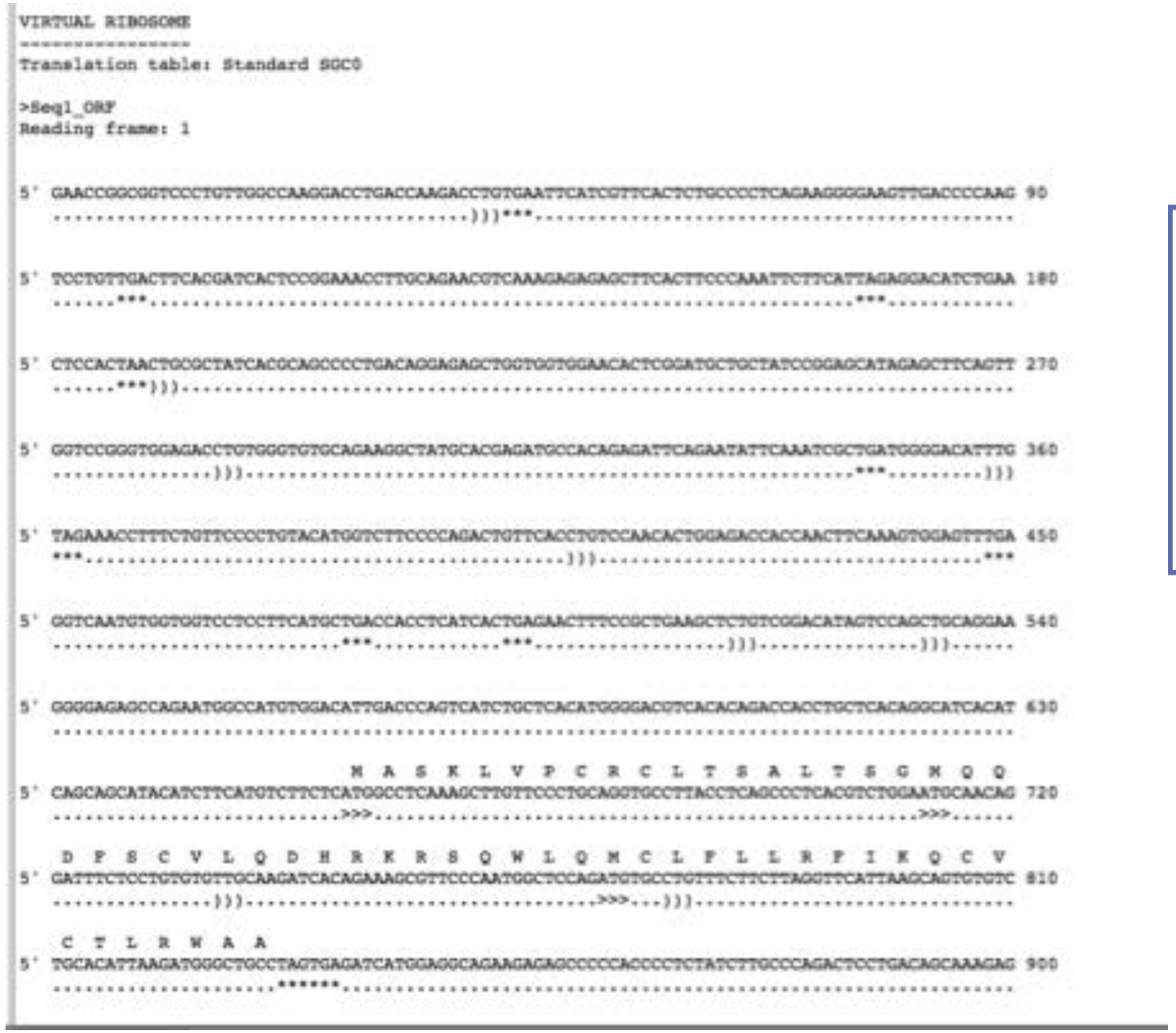
- hvilken organisme kommer proteinet fra
- hvilket protein er det?

Virtual Ribosome

Nu skal vi analysere den første ukendte DNA-sekvens nærmere (nu er den ikke mere ukendt). Kopier sekvensen og åbn Virtual Ribosome. Indsæt den i det øverste felt som vist nedenfor. Midt på skærmen skal I vælge reading frame – altså fra hvilket nukleotid, læserammen begynder. Det er her, der er 6 muligheder. Vælg læseramme ”1”. Ved ”ORF finder” vælges ”start codon: Strict”. Det betyder, at programmet nu leder efter den normale startkode, som er ATG. Vælg ”Submit query” nederst.



Resultaterne ser ud som vist nedenfor:
(ude til højre kan ses signaturforklaringer)



>>>	Startkode (den ene vej)
<<<	Startkode (den anden vej)
******	Stopkode
))) og (((	Alt. Startkoder (irrelevante)

Som det ses, er en stor del af nukleotiderne ikke kodende, mens kun en mindre del er kodende.
De kodende nukleotider har vist aminosyrekoden ovenover sig.

Vælg FASTA øverst på skærmen. Herved får du opsummeret, hvilke aminosyrer genet koder for.
Hvor mange kodes for?

Lav nu endnu en analyse af samme sekvens, men denne gang ændrer du læserammen til 2.
Optæl antallet af aminosyrer, genet koder for, hvis læserammen er sat til 2.

Gør det samme med både læseramme 3, -1, -2 og -3.

Hvilken læseramme giver det største protein?
Antagelsen er, at denne læseramme er den korrekte.

Foretag til sidst en analyse, hvor du vælger ”Alle (6 reading frames)”.
Husk ”ORF finder” skal stadig være ”start codon: Strict”.

Nu fremkommer den bedste løsning – den længste aminosyresekvens.
Var det den samme reading frame, som du fandt (det står noteret øverst på skærmen)?

Hvis der er tid, så udfør lignende analyser på DNA-sekvens 2.

RESULTATER

	Organisme	Protein	Antal Exons	Kromosom-nummer	Bedste læseramme
DNA-sekvens 1					
DNA-sekvens 2					
Aminosyresekvens 1					
Aminosyresekvens 1					
Aminosyresekvens 1					
Aminosyresekvens 1					

Fejlkilder

Diskussion

1. Forklar hvad der forstås ved begrebet læseramme, og forklar hvorfor der er 6 mulige læserammer ved ethvert stykke DNA.

2. Hvilke forskelle er der mellem prokaryoters og eukaryoters DNA?

3. Forklar kort, hvordan DNA bliver til protein hos prokaryoter og eukaryoter.

4. Hvorfor er nogle substitutionsmutationer mere alvorlige end andre?
   Giv konkrete eksempler.

5. Hvorfor har det stor betydning for proteinets struktur, hvis fx cystein udskiftes med en anden aminosyre som følge af en mutation?

6. Hvordan kan BLAST-analyser bruges til slægtskabsanalyser?

BILAG 1: Ukendte DNA-sekvenser

Ukendt sekvens nr. 1 (FASTA-format)

GAACCGGGTCACGAGTGGCCGGAAGTTGCGGCCTTTGCGGATAATCTGAGATGGGGACCACTCTGGACATCAAGATTAAAAGAGCGAATAAAGTGTATCACGCCGGGGAAAT-
GCTCTCTGGTGTGGTGGTCATCTCTAGCAAGGACTCAGTCCAGCACCAGGGAGTCTCTTTGACAATGGAAGGGACTGTAAACCTCCAGCTCAGTGCCAAAAGTGTGGGTGT-
GTTCGAAGCCTTTTACAACTCTGTGAAGCCGATCCAGATTATCAACAGCACCATAGACGTGCTGAAGCCAGGAAAGATTCCCAGTGGCAAGACTGAGGTCCCCTTTGAGTTC-
CCACTGCTTGTGAAAGGCAGCAAGGTCCTGTATGAGACTTACCACGGCGTGTTTGTCAACATTCAGTACACACTGCGCTGTGACATGCGGCGGTCCCTGTTGGCCAAGGAC-
CTGACCAAGACCTGTGAATTCATCGTTCACTCTGCCCCTCAGAAGGGGAAGTTGACCCCAAGTCCTGTTGACTTCACGATCACTCCGGAAACCTTGCAGAACGTCAAAGAGA-
GAGCTTCACTTCCCAAATTCTTCATTAGAGGACATCTGAACTCCACTAACTGCGCTATCACGCAGCCCCTGACAGGAGAGCTGGTGGTGGAACACTCGGATGCTGCTATCCG-
GAGCATAGAGCTTCAGTTGGTCCGGGTGGAGACCTGTGGGTGTGCAGAAGGCTATGCACGAGATGCCACAGAGATTCAGAATATTCAAATCGCTGATGGGGACATTTGTA-
GAAACCTTTCTGTTCCCCTGTACATGGTCTTCCCCAGACTGTTCACCTGTCCAACACTGGAGACCACCAACTTCAAAGTGGAGTTTGAGGTCAATGTGGTGGTCCTCCTTCAT-
GCTGACCACCTCATCACTGAGAACTTTCCGCTGAAGCTCTGTCGGACATAGTCCAGCTGCAGGAAGGGGAGAGCCAGAATGGCCATGTGGACATTGACCCAGTCATCTGC-
TCACATGGGGACGTCACACAGACCACCTGCTCACAGGCATCACATCAGCAGCATACATCTTCATGTCTTCTCATGGCCTCAAAGCTTGTTCCCTGCAGGTGCCTTACCTCAGC-
CCTCACGTCTGGAATGCAACAGGATTTCTCCTGTGTGTTGCAAGATCACAGAAAGCGTTCCCAATGGCTCCAGATGTGCCTGTTTCTTCTTAGGTTCATTAAGCAGTGTGTCTG-
CACATTAAGATGGGCTGCCTAGTGAGATCATGGAGGCAGAAGAGAGCCCCCACCCCTCTATCTTGCCCAGACTCCTGACAGCAAAGAGTCAGCAACCAGCAAGCTCTCTGAT-
TGAGTCTGTAATCCCCTCCAAGGAGTGACAGTGACTTCATCTGGGAATACGGAGTTTTCCCTTCTTGACATCTTGATAGTGTTTCACTAAAGCCTCTAAATGTTTTAGGGCACTG-
GAAATAAAAACCAACAAATAGTCCTGGTTCTCACAGAGGGAGAATGGTGAACAGTGGCTGTTGCTGGAGTGTTTGTGTGTGTGGTACTGGCACACAGTTAATGCTGCCATAT-
TCCCAGAATGTATTGAGAGACAAGGTGCTTGCATAAGGGGAAGAATGCACATGGACTACTCGGTAAGTACTAGGTTCTGAGTGACAGGACACCGTGCTGAGGAAGGGCAATG-
GCCTTGGCCTGTCCCTGACCCCCTGGGTCACATGGCTTTAGCCTCACACTGTAACTCATACTGTATGGAGGGTATGTTGCAATACTTTGCTTCCTTTGAATCTCGCTGTTGTTCT-
GTTGGTCCAGAGTTAATGAGCGCCCTGTACACCAGCTGGCCTGCCTTCACAGGTTCTGGGGGGCAGGCCGAGGCAAAGCAGGCCCGACTGGCAGCCTCCTGGCTGCACTCC-
CACTGTTCTGAACAGCTGAGGGAACTGGGGCACTGCTGCTGGGGCCTCACCAGGGAAACTGAGAAGACTTAGCTGAGTCACCTGACAGTGAGTTCCCAGAGAATGTACTTTT-
GCTTTGTCATACTTATCCTTGTTATATTTAAACAATCTAGTACCCAAAAAATGTGGGCCGTACTTTCTAGTGAGTCTAAGGAAAAGTCATCTTTGATTACATAATTTTTTTAATGAATA-
AAGAAAGCTAAACAAGTTT

Ukendt sekvens nr. 2 (FASTA-format)

TCCCGACCCTAGAGTCCCGTCACGACCCCTGACCCTTACACCACAACTCTCCCGAAGTCCCTCTGCACTACCCTTCTACCCCTTCGGAGACATCCACCTGTCTCGGTCCAC-
CACACCTGTCCCCGACACTCTAACCTCTTCCTCCTCAGACCCCTGACCATCGCCGAGGTTGTCCCCTTGCAGGAGACCCCCACGGGAAATAGGTAGAGACGCCCCGGACAG-
GATCGAAGGGAGGACCGAGGCCGCTCAGGGCCGACCAGAGAAGCGAGAGAAAGAACTAGAGATCGATCTCTAGCTCTTTATTCCTCTGACATTCTGCCCCATCTGCTCCCG-
GACTCTTCCGCCGCATCTGTATCTTAGTCTCTTTCTAACTCCCTGCCTGGGGCCCCCACCTTTGAGCATATATCGGCTTCTCTCTCTGCCGTTGTGTTTCTCTGGGGTTATCTTTCC-
CTCTACTCCGCCCGGACACGCCCTCTATCTCTTCCTCCCACCCTGTCCCAGTCTAGTACCTGGGGTGGGGGTAATGGAGAGACAGCTAGGTCCTAAGAGAAGTCAGGGGG-
GCTGGGCCAACCTCTCCATATTTACATATGTATGAGGGTCGCCTGGGCCAGTGGCGAGGAGGCGGACGTTCTGGGGGTGGGAAGGGGGCGGGCACCCCCAGAGCCGCA-
GAGTATAAAGACCGCGCTCGGCGACCGCGGGCCCCGACTGCTGAGGAGCGGACGCTCCGCCTGGGGGGCCCCCCATCCCTGGCTGTCCCCCAGCTGCGCGTCCCCGCC-
CCACCCCCGCGGCTGAGCCACCACCGGTGCAGTGGTCTCCGCTTGGCGGAGCGAGCCTTGAGCTTCGTTCCACAGCTTCTTTGCATCTTGGATTTCGGGGCGGCCCCCTCC-
CCCACCTCTCTCTGCCTTTTTGTACCCCGCTTTTTTTCTGCGTTCTGCTCGGTTTTTGTAGCCGTCTGTTTTTGCACCCCATTTCGTTTTGTTTCTAGACGGTTTGGCGGGGGGT-
GAAGCTGCATTCATACCCCTTCCTCTTGTTATTCTCCCCTGCTCTGACAGCACCCCTTTTCATCGCAGTTGGGGGGCCTAGGATCGGTGCATCTTCCGCCGCGCTGCCAGCAC-
CCCGCAGCGCGTGGTCGTGCACCCCGGAATCTGCAGCAGCTGCATATCTGAGGGGGGTCTCCTTTGCCCGCGCCGCCTTCGCTCCCCGTGCTTTTGGGTGTGTGGAGGGCT-
TCAGTCGCGGCGCCCCCGCTTCTCCGCAACCCCCCGCCCCGCGCCCGGACTCGCCCCGCGCCACCAAGATGGTCATCCAGAAAGAGAAGAAGAGCTGCGGGCAGGTGGTT-
GAGGAGTGGAAGGAGTTCGTGTGGAACCCGAGGACGCACCAGTTTATGGGCCGCACCGGGACCAGCTGGGGTACGCAGGGCCGGCACGCAAGGGGCGGGGGAAAGCCG-
CGGGGCGACGCCTCGGGGGCGCAGGGTCCCGCCGACGCGCCCCAGCTCCCCTCCCGGGTCCCGGCGTCCAGCTCCCTGCCGGGCTCTGGGCTGGGAGGGGGCCGAATC-
GCCAGTCTAACTCCCCGGCTGGCCGTGCGGAGGCGGAGAAAGTAGGTCACAGCCGCCTTCCGCCCCCCGCGGAGCCCCCTCGGGCGGCGCGGGTCGCCAGCTCCGCCT-
GCGTGCCGCGCCGCGCGCTCACACTCCCCTCTCGGGGCCTGTCGCTCCACACGGGCGTCCCCCACCTCCAAAGAGCGCCCCTTCCCTCCCTCCGGCTCTCACTAGCTCCG-
CAGCCCCGTCTATTTTTAGCTCGTGCCCACCCCCTGGACCCTGGGAACGTTCATGAGGGGGCGGGTCTTGGGGGGTGTGTTAGGGGGGTTCTTCACGGCGGAAGTTGTCT-
GTATCCCACCGCCTGGCCTTGGGAGCCTTCTGGGACTGCTTTGTGGGTTGGGGGGCTGCTGATAGTATGAGTTTTACCGAGGCTGCAGGTTTTAGCCTCCCATGTCGGTGACG-
GAGGGAGGAGTGGTCGCTGTGGTGATTTGTGTGCATCAGCCAGCCAGGTGTCTGTGACAGTCGGATGACTTGGAAGCCTCCCCAGGCTGACCATGGCAGGACTCAGGGAG-
CTGTAGTGGTCGGGGGTTGGGGGGGTGGAGGGGGTCTGGTGACCGGCACAGGTGCAGGTGAGGGGTGGAATTCACTTTACATTTTCCCCATAGAAACAAAGTTATAAATA-
GTGACTGCATACTGCACCTAAAATGCCACGTATCTACAACGAAATTATTACAATGTATTTATAATATATTATTTATTCTAAACATGTATGCATTTGAAATTATCCAATGAAATCTAAC-
CTATTGATGTACCTATTTCTACATTAATACATTATATCATATATTAATATGCTAATAATTATATCACATTAATATATTAATAATATGTTAATTAACCAAACGAAAAGTATATACTGATCA-
CAACACTTGAATTCTTCCAAGAGGGATGGTCAAGCTGGGACGTTGAGACACAGGGGACAGAGGACACTGTGTGACACGATTTACAATCTTTCCACACTGGGCACCGTCCCCA-
TCAGTCCACCCATTCGGGGCCTACACGAAGTGGGTCCCATGCAATCCATTCCCTCAGGGAACTCAAACTCCAGCCCCTGGGATGAGAAGAATCCAGCAATGCTTGGGAGAG-
CCAGAGGACTTCATGGAAGAAGTGTCCTCTGAGATGGAAGGATTGGGAGTCCAGGGTGGTGGGAACAGCCGGCCCTTGGGTCCTTACTTCAGGCGGGGGAGCCATGGAGA-
GATCCCACCAAGGGAAGGCTGTGGAGATTCTGCCTTTCCTCCCTGCCTCTGCCCAGGGTGCTGGGTGTGAACTGAGGGTGGGGTGACTGTTGAAGGTTCTAACAAGCCGTC-
TCTGAGAGATTTGTAGCTAGGCTAGTGTTAGGTCTTTCATTTCAGGAACTGTGTTCAAAGTTTGGCTTCTGAAGGGCACCAGGAGAGAGATGTTGCTATTCAAATCTGAGGGT-
CCAGTCTCTGCGGGGTGGTATGAGGGTTTGCTTGTGAATGGTGGCCAGTACCCGCTTTAAAAGGCACCATGCTAGCACAGCTTTAAGCATGAGTACGAATGCAGAGGTAACA-
GATGTGTGCCTTGTCAGGACTATGCATGGTTGAGAAGTTGGAAATGTAATTGGAGGCAAAATAACAGACCTCCACAAGGTCGGGCTTCACTGTGCCCTAGGACCAGGAGGGG-
GCTGGGAGTCATGGCTAGAAGCCAGACACAACTGCCTGTTTCCAGTTTGTCTCATTTTGCCTCCAGAGGAAGGCTCTAAGACATCCCTGTGGCTCTGTGATCAGTCCCAGTGCA-
GAACTTCAGAGTGGGTAGAGGGGTGTGTGGGGATAGTTGAGGTTATGGTGGGAACCTTGGGCCCTGCTGACCCTGTTTCCTCCTCCCTAGCCTTTATCCTCCTCTTCTACCTC-
GTTTTTTATGGGTTCCTCACCGCCATGTTCACCCTCACCATGTGGGTGATGCTGCAGACTGTCTCCGACCATACCCCCAAGTACCAGGACCGACTGGCCACACCGGGTGAGT-
GTGGAGGCTCCCCCTGCCAGCTACTCTAACTGCTCTTGTGCCCCCAAACCTCCAGAAGGAACTCATAGTTCCTTCCAGGAGTTTGATTTTGATGACCCAATCCCCACGTGCT-
TGGAAGTTCTTGAAATCTGTCCACCTTCCCATTTACTGCAGTTGGGAGCTGTGTGATTTGGGCATGTGGCAGATAGCCACAGGAGATCACCCTCCCATGAAGACGATCTCAGA-

TATTCACCGCTGTCCCCACTCTGGTGTTCCCTGTGTCCATTTTCTTCCCTCTGTGTGCAGTCCCTCATCTTATAGATCCCCCAACTTCTGCCTTTGTTGGCTGTAGGCTTGATGAT-
TCGCCCCAAGACTGAGAACCTTGATGTCATTGTCAATGTCAGTGACACTGAAAGCTGGGACCAGCATGTTCAGAAGCTCAACAAGTTCTTGGAGCGTGAGTGTGGGCCTGGT-
TATGTGTCAGTTCAAGACTTCGGGCAGGGGACTGGGGACCTTGGAAGTGGAACATCTGGCCCCTGAGTCTCTCCCTCCCACCTCTTTAGCTTACAACGACTCTATCCAAGCC-
CAAAAGAATGATGTCTGCCGCCCTGGGCGCTATTACGAACAGCCAGATAATGGAGTCCTCAACTACCCCAAACGTGCCTGCCAATTCAACCGGACCCAGCTGGGCAACTGCTC-
CGGCATTGGGGACTCCACCCACTATGGTTACAGCACTGGGCAGCCCTGTGTCTTCATCAAGATGAACCGGGTATCTATGACCTTGGTCCCCAGGGTGAATGGAGGAAGGATC-
TGGGGACACCACCTGCAGACAATTGCATCCTTTCACTGGGGCTAATGGGCATGAGAAAGACTTGGATGTTTGTGTAGCTGAGAGAAAAAGAGAGGTGGGACTCTTGGAGGC-
TAAGCTCTGCAGTTAGAAGGCTGGATGCCTTTTGTTTTTTTTTTAGACAGGGTCTTGCTATGTTGCCCAGGCTGGCCTTGAACTCCTGGAATTAAGCTATTCCTCCTGCCTCAAC-
CTCCCTAGTAGTTGACTACAGTCCCCGGCTTAGCTTGGGTCTGGATGCCCATCTTCGACAACTTCTTCCTCTGACTCTCTTCACCTTCCACCCTCAACTCCAGGTCATCAACTTC-
TATGCAGGAGCAAACCAGAGCATGAATGTTACCTGTGCTGGGAAGGTGAGTTCGTTGGGCCTTGTCTGCCTGCTCACCTGAGTGGCTCACCTGAGTGTCCTTCATGGTTTCT-
GTGTAATTCACCTGTCTCTCCCTATCTTCTTTGCTCCTAGAGGCCCCATCACCATAGAAACAAGGGGGTAAGAGTGGGCTTTTGGGCTCCACTGTAGCTTGAACTCCGAGGG-
CCCCGCACTCCTCTTGCTTCTCTCTGGGATGCAGAGGCCTGCTCTCCTAGGGGCCAGACACACGCCCTCCTCCACCAACGCCCTGGCCTCTGGCTTCTCTCCCTAAGCTTC-
CACCTTCTCCTTCATTCCCGAGTTGTCCGTATCGTTCGCTCTCCCTCCCATATGGCCCCCACCCGTCAGTTCCTTCTAGGTGTCTGGTAGCTGCTGATCTGTTAGCACCATCTGC-
CACCAGTTCGTTGTCCTTGGGACCCTGTTCCCCGCTTCCCCCCCGGTCTGTCCTTTCTAGAAACTGGCTGCTCCCTCCACATCCCCTTCCTTGCTTCCTATTCAACCCTTAATCA-
TGTATCTCTTCTTTCTTGGCTCTGCTCCAGAAACTGATTCCTGAGGATGGTGTAAGAACTTGGGGTAGGAGTGGAGTAGGAGGCTTTCGTGCCATTAGCAGCCTTCAGGAGTTC-
CTAGAATGATGACAGGGACAGACCATCCCCTCATCTACACACGCACATGCAGACACACACACACAGACTCACAGCTCCAGGGAGGTAGACTTGAGACAGGAATAATTGGGGC-
GAAGGAAGAAGAAAAGAACAAATGGAAGTCTGGTGAGCTCCTGGGTGCCTGCCATCCCTAACTGGCTAACCCCCTATCTTCCTGCACCCCCACAGCGAGATGAAGATGCTGA-
GAATCTCGGCAACTTCGTCATGTTCCCCGCCAACGGCAACATCGACCTCATGTACTTCCCCTACTATGGCAAAAAGTTCCACGTAAGTCCCAGGGGAGGCCCAGGCTGGATGG-
CGGGTGCGGGTGGTGAGCTAGGGAAGGAGGCCGCGTTGCCCCAGGCCTAGACCCTGCACTGCTCTCCGGCCCAGGTGAACTACACACAGCCCCTGGTGGCTGTGAAGTTC-
CTGAATGTGACCCCCAACGTGGAGGTGAATGTAGAATGTCGCATCAACGCCGCCAACATCGCCACAGACGATGAGCGAGACAAGTTCGCCGGCCGCGTGGCCTTCAAACTC-
CGCATCAACAAAACCTGAGGCCCCTTCCTCCCACCCCATCTCTCTCCTGTGGATGCTCCTGGAATGTCCCTGACCCTGCCTGATCCCTCCCTCACCCACCCCAAAGGTATTTT-
TGATAACAGAGCTATGACTTGTCTGAGCCTCACATCCTTTTCCTTGACTTCTCAACCCAGCCTGAAGTCCATTGCGGTTCCGTCACTCGCCTTTCCCACCAACTTCTCCCAACC-
TCAGATCAGTCAGACAGGGAGCTGGGCTAAGATGGCCACAGAGGAGTTAGGAGCCTTTCTAGTTCTGGTTTAGCTGTGAGAGCTATCCACTCTCCTGCCTGCATATCCCCTGA-
GAGTTATAGGAAGTGCCCACTGACCCACCCACCCACCTACACCCCCCGCCACACACACACACAAACGTGCACACGCGTCTCATTTGACCCCTTTGCTTCCAGAGATGAATGTG-
GCACTCCCTCCTTCCATTCCTAAGCTCTAGCCACCGTCCCTTGATCTCTCATACTTTCTCCCTGTCTACACAGTCGCCATCTTGGTGACTTTGAATTTATCTGGCTCCTGGGAG-
GTCTTCTCCTCCTCTCCATCCCTATTCCCTCCTCTGAAAGCACCCCTTGTAATTGAGGACAAGGTGGTTCTGTGGCCTTTTCCCTCTTTGCTGGCACGTTCTGCTTCTCACCCTC-
TGGTGACTCTGTGAGCTGGGAAATGAGGGACTGGAAGTGAGGCCTGTGTTGACCCTTCCTGAAAATCCTCTAGCAGCCCCCGACTTCAGCAGTTTCTTTCTTTGTTTTTTTGA-
GATGGAGTTTCGCTCTTGTTGCCCAGGCTGGAGTGCAATGGTGCAATCTCAGCTCACTGCAACTTCCGCATCCCAGGTTCAAGCGATTCTCCCGCCTCAGGTTCCCGAGTAGC-
TGGGACTACAGGCATGTGCCACCATGCCCGGCTTAATTTCTTTCTTTTTTTTTTTTTTTTTTGCATTTTTTAGTAGAGATGGGGGTTTCTCCTTGTTGGTCAGGCTGGTCTCGAAC-
TCCCGACCTCAGGTGATCCACCTGCCTCGGCCTCCCAAAGTGTTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCTTCAGTTTCTTCCTAGGCCGTTCTGTCACCCAAATA-
GCTGCTACCCAGAGGGGCGGGGTTGACCTAGGCTGAATATCCACTTTGTTTTTATGGATGGCTCCCTTCCCCCATTCGCCTTCCCAGAATATCCTTCAAGTTCCACTTCCCAGG-
GAGCTCTGGGGGAGGGGCGGCCATTCTGGCTCCGTCCCCAGTGGCCACCTTGGAAACATCGGCTGGCTTTGGGACTATTCCACCTCCTTCCCCTGAGCCCAGATCTGCCCC-
CACCATCCTTTCTCTGGCTTCTTTTAGCAAGTTATCAACTAATCACTAACTCCTTCCTTTTCCTCTGCATGCCAGCCTGAAAATTCCAAATCTAGCCTCTGAATGTCTTGGCTCCATC-
TCTTCAGACCCCTTTGCCTTTAAAAAAAAAACAAAAACAAAAACAAAAAAACCCATAATGCCCATAGAATGTCAAATGAGGGGCCTCCTGCCTCCTGCTCTGAATATTCTGTAG-
CTGTAGAGGCATTTTAACCCTTTGTCCTCCAGCATCCCTTCACGTCCTCATCCTCTCTAACCTCCTTTTTCTTTTTTTAATGCTGCAGCCTCCACACTCCACCCACAGGTGACC-
CTTACCTTTTTCTCTAGCTGGATCTGTGTTTCTTCCCTTCGGGCCCCCATGTTTTCCTGCACCCGCCCTACCATGGTCTCTCTCTGCAGTTATTTAATGCCTGTGTCAGATCTACT-
GTAAAAAGAGGATTAAGTAAAATAAAATGAGAGCAATTATATATATAAATATATATCATACACAGAGCCCTGTGTGTGGGTTGTTCCTTTCTGAGCAGTTGGAAGAGCCAGGAAG-
GAGTGGAGGGTGGATCT

BILAG 2: Ukendte aminosyre-sekvenser

Ukendt sekvens nr. 1 (FASTA-format)

SLTKAERTIIGSMWTKISSQADTIGTETLERLFASYPQAKTYFPHFDLNPGSDQLRAHGSKVLAAVGEAVKSIDNVSAALAKLSELHAYVLRVDPVNFKFLSHCLLVTLASHFPADLTA-
EAHAAWDKFLTIVSGVLTEKYR

Ukendt sekvens nr. 2 (FASTA-format)

GIVEQCCTSICSLYQLENYCN

Ukendt sekvens nr. 3 (FASTA-format)

MFALRAASKADKNLLPFLGQLSRSHAAKAAKAAAAANGKIVAVIGAVVDVQFDDNLPPILNALEVDNRSPRLVLEVAQHLGENTVRTIAMDGTEGLVRGQKVLDTGYPIRIPVGAET-
LGRIINVIGEPIDERGPIDTDKTAAIHAEAPEFVQMSVEQEILVTGIKVVDLLAPYAKGGKIGLFGGAGVGKTVLIMELINNVAKAHGGYSVFAGVGERTREGNDLYNEMIEGGVIS-
LKDKTSKVALVYGQMNEPPGARARVALTGLTVAEYFRDQEGQDVLLFIDNIFRFTQAGSEVSALLGRIPSAVGYQPTLATDMGSMQERITTTKKGSITSVQAIYVPADDLTDPAPATT-
FAHLDATTVLSRAIAELGIYPAVDPLDSTSRIMDPNIIGQEHYNVARGVQKILQDYKSLQDIIAILGMDELSEEDKLTVARARKIQRFLSQPFQVAEVFTGHAGKLVPLEQTIKGFSAI-
LAGDYDHLPEVAFYMVGPIEEVVEKADRLAKEAA

Ukendt sekvens nr. 4 (FASTA-format)

MLSAPCCDDRRMCVCPGPRRIGIPVRSSSLPLFSDAMPAPTQLFFPLIRNCELSRIYGTACYCHHKHLCCSSSYIPQSRLRYTPHPAYATFCRPKENWWQYTQGRRYAST-
PQKFYLTPPQVNSILKANEYSFKVPEFDGKMSVLSLDLTAIKLPANAPIEDRRSAATCLQTRGMLLGVFDGHAGCAWSQAVSERLFYYIAGSLVPHETLLEIENAVESGRALL-
PILQWHKHPNDYFSKEASKLYFNSLRTYWQELIDLNTGESTDIDVKEALINAFKRLDNDISLEAQVGDPNSFLNYLVLRVAFSGATACVAHVDGVDLHVANTGDSRAMLG-
VQEEDGSWSAVTLSNDHNAQNERELERLKLEHPKSEAKSVVKQDRLLGLLMPFRAFGDVKFKWSIDLQKRVIESGPDQLNDNEYTKFIPPNYHTPPYLTAEPEVTYHRLRPQDKFLV-
LATDGLWETMHRQDVVRIVGEYLTGMHHQQPIAVGGYKVTLGQMHGLLTERRTKMSSVFEDQNAATHLIRHAVGNNEFGTVDHERLSKMLSLPEELARMYRDDITIIVVQFNSHV-
VGAYQNQEK