

27611 Eksamen Sommer 2007

Dette sæt indeholder 4 opgaver.

En online version af opgavesættet vil være tilgængeligt fra kursets lektionsplan, under selve eksamen (25. Maj 2007 klokken 9:00 – 13:00). DNA/Protein sekvenser kan kopieres direkte herfra – det er ikke meningen at sekvenserne skal tages ind i hånden.

Lektionsplan:

<http://www.cbs.dtu.dk/dtucourse/27611spring2007/lektionsplan.php>

Svar til opgavesættet skal skrives enten i rå tekst (fx i Notepad/Wordpad/Nedit) eller i Microsoft Word (.doc) format.

Dit studienummer skal fremgå af filnavnet (fx. s022717.doc eller s022717.txt) og skal stå i starten af dokumentet (fx: "Studienummer: s022717")

Svaret skal uploades på CampusNet under kursus 27611 (under "Afløseringer -> Eksamen 2007"). Husk at gemme seneste version af dokumentet inden du uploader svaret.

Underskriv desuden denne forside med studienummer og navn og aflever den til eksamensvagten. Lokalenummer og computernummer **skal** udfyldes med henblik på kontrol af netværkstrafikken.

Navn: _____

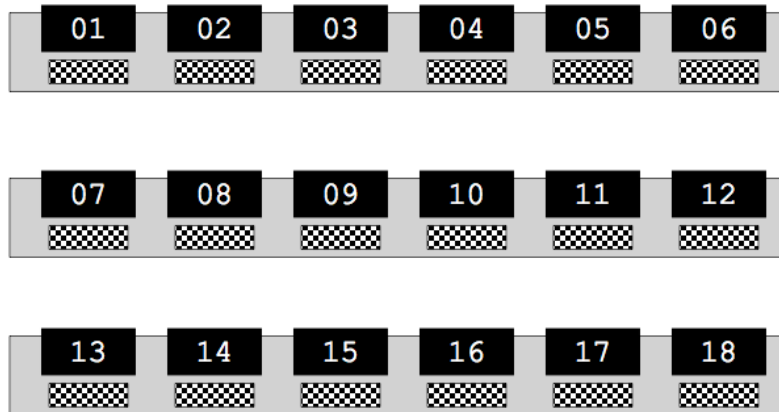
Studienummer: _____

Lokalenummer: _____

Computernummer: _____

(For eksaminander i lokale 062, byg. 208 – skriv nummeret på låget af den bærbare computer. For eksaminander i lokale 052 og 152 i byg. 210, brug oversigten på næste side).

Tavle



Indgang

Oversigt over computernummerering i lokale 052 og 152 i bygning 210.

Hvad gør man hvis en web-server ikke virker:

- 1) Verificer at input-data er i korrekt format. Forkert inputdata er i næsten alle tilfælde årsagen til problemet.
- 2) Rapportér fejlen til eksamensvagten – den kursusansvarlige vil så blive tilkaldt.

Opgave 1: Identifikation af ukendt DNA

Denne opgave tæller 25% af sættet.

Nedenstående sekvens er sekventeret direkte fra DNA som stammer fra en ukendt ikke-kultiverbar mikroorganisme. Det vides ikke hvorfra i genomet sekvensen stammer. Det er nu din opgave at finde ud af så meget som muligt om denne sekvens.

Du skal i dit svar argumentere for valg af værktøjer og databaser, samt dokumentere dine svar med referencer til relevante sekvenser (fx. data i FASTA format, hvis du arbejder videre med sekvensen, eller referencer til GenBank/UniProt entries).

1. Bestem funktionen af sekvensen.
 - a. Er det en sekvens der i forvejen er kendt?
 - b. Er det muligt at finde beslægtede sekvenser med kendt funktion, der gør det muligt at bestemme funktionen?
 - c. Beskriv den sandsynlige funktion.

2.
 - a. Er sekvensen proteinkodende?
 - b. Kan man forvente at sekvensen indeholder en komplet CDS?

3. Er det muligt at afgøre om sekvensen stammer fra en eukaryot eller prokaryot organisme?

4. Sekvensen indeholder enkelte bogstaver, der ikke er A, C, G eller T.
 - a. Hvorfor kan dette forekomme?
 - b. Hvad betyder det når der står "S" eller "K"?

```
>unknown_fragment
AATGGGCACGGGACGCATGTGGCAGGCACCATCGGGSCCGTCGGCAACAACGGTACGGGC
GCAACTGGAATCAATTGGAACGTCCGCATCATGAGCCTGAAGTTCATGAGTTCAGCGGC
AGCGGCTACACCAGCGCCCGGTGCAGGCGATCAACTACGCGGTGCGCATGGGCGCTAAG
GTCATCAATAACAGTTGGGGTGGCGGCAGTTACGATCAGGCGCTGGCATCAACGATCCAG
TTCGCTCAAAGCCGTGGTGTATCGTGGTCAACGCGGCAGGAAACGACGGCGTTAACGTC
GACGCTTCGCCATCGTACCCGGCGAGTCTGAATGGCGCCAACGTGCTGACGGTTGCCGCC
ACCGATCAGAACAACAATCTCGCATCGTTCTCGAACTACGGTGCCGGCACGGTTGACATT
GCCGCTCCGGGTGTGACCATTCTCAGCACTTACACCAGCGKCCGTTATGCATACATGAGC
GGCACATCAATGGCCACTCCGAACGTCGCCGGCGTCGCC
```

Opgave 2:

Denne opgave tæller 30% af sættet.

2A): Psi-Blast

- 1) Hvis du kører en BLAST søgning med en protein sekvens mod NR og finder følgende tre hits, hvilket hit ville du vælge?

- a. 70% id, E værdi = 1.2
- b. 25% id, E værdi = 10
- c. 25%id, Eværdi = 0.001

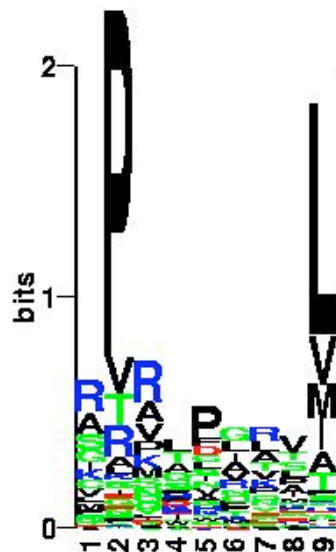
- 2) Hvad er protein sekvens (i FASTA format) for SwissProt entrien P11302?

Brug Psi-Blast til at finde en homolog PDB struktur (med homolog forstås her en sekvens med en signifikant E værdi)

- 3) Hvor mange BLAST iterationer skal du køre for at finde en PDB struktur med en signifikant E værdi?
- 4) Hvad er navnet på den homologe PDB struktur, og hvad er E værdien for hittet?

2B): Spørgsmål Logo'er og vægt matricer

- 1) Logo plottet nedenfor er genereret på baggrund af sekvenser, der vides at have en god binding til MHC. Hvilke er de to mest informative positioner?



- 2) Hvilke aminosyrer på position P2 vil give god binding?
- 3) Nedenfor er angivet en multiple alignment af et sæt peptider, der binder MHC.

KPSEPGGVL
SPALPGLKL
SPKLPVSSL
KPSLPFTSL
SPYQNIKIL

Benyt relationen for udregning af aminosyre frekvenser ud fra de observerede frekvenser og pseudo frekvenser til at udregne vægt matrice (log-odds) værdierne for E og K på position P1. Sæt $\beta=4$, og se bort fra sekvens vægtning.

Opgave 3:

Denne opgave tæller 5% af sættet.

Hvilke af følgende sekvenser er i korrekt FASTA format. (Vælg en eller flere).

(a):

```
<Seq47
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

(b):

```
>(gi|64141:754-928, 1216-1358) Oncorhynchus insulin gene for preproinsulin
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

(c):

```
>(gi|64141:754-928, 1216-1358)
Oncorhynchus insulin gene for preproinsulin
Notice: Introns removed!
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

(d):

```
>seq47
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

(e):

```
>Seq47
  1 ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
  81 CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
 161 TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
 241 GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
 321 CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

(f):

```
>seq47_2
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGT
GTTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGAC
GCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAG
GGTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGT
AAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAAC
TGA
```

(g):

```
LOCUS Seq47
ATGGCCTTCTGGCTCCAAGCTGCATCTCTGCTGGTGTGCTGGCGCTCTCCCCGGGGTAGATGCTGCAG
CTGCCAGCACCTGTGTGGCTCTCACCTGGTGGACGCCCTCTATCTGGTGTGTGGAGAGAAAGGATTCTT
TTACACCCCAAAGAGAGATGTGGATCCCCCTATAGGGTTCTCTCTCCAAAATCAGCAAAGGAGAACGAA
GAGTACCCCTTCAAAGACCAGACGGAGATGATGGTAAAGAGAGGTATTGTAGAGCAGTGTGTACAAGC
CCTGCAACATCTTCGACCTGCAAACTACTGCAACTGA
```

Opgave 4: Sammenligning af insulin fra forskellige organismer

Denne opgave tæller 40% af sættet.

Som du måske husker fra øvelsen "Translation og proteindatabaser", bliver proteinhormonet insulin syntetiseret som et forstadium (*precursor*), hvorefter et signalpeptid og et propeptid spaltes fra før det når sin færdige (*mature*) form, som består af en A-kæde og en B-kæde. (Et **tip** som du får brug for senere: signalpeptider og propeptider findes i UniProt annoteret i feature-tabellen med betegnelsen (FtKey) henholdsvis "signal" og "propep"). Din opgave er nu at sammenligne insulin fra nogle forskellige organismer og finde ud af om signalpeptidet og propeptidet er mere eller mindre konserverede end A- og B-kæderne.

4A: SRS-søgning

Først skal du ved hjælp af SRS fremstille et brugbart datasæt af aminosyresekvenser fra insulin. **Vigtigt:** Brug kun Swiss-Prot delen af UniProt.

1)

Hvor mange entries i Swiss-Prot indeholder *ordet* "insulin" i beskrivelsen (altså *ikke* medregnet sammensætninger som "Insulin-activated" eller "Insulin-like")?

Som du kan se, giver denne søgning en del resultater som ikke *er* insulin, men bare har noget at gøre med insulin. Nu skal du prøve at indsnævre denne søgning på forskellige måder.

2)

Blad ned i resultatlisten til du kommer til hits med navnene "Insulin" og "Insulin precursor". Kig nærmere på nogle af dem. Hvad er helt præcist forskellen mellem dem der hedder "Insulin" og dem der hedder "Insulin precursor"?

Det gælder nu om at begrænse søgningen til insulin-forstadier. Det ville naturligvis være nemmest hvis man kunne søge efter selve sammensætningen "Insulin precursor", men det *virker ikke* i SRS, idet SRS kun indekserer enkeltord, ikke hele sætninger/linier. Besvar i stedet følgende:

3)

Hvor mange entries indeholder begge ordene "insulin" og "precursor" i beskrivelsen?

Som du kan se, er der stadig andre proteiner end insulin med i sættet. Foreslå et eller flere ord, som kan tilføjes til søgningen for at begrænse sættet til insulin-forstadier.

- 4) Hvilke(t) ord valgte du, og hvor mange er der nu tilbage?
- 5) Undgå så de entries der ikke indeholder fuld længde sekvens (fragmenter). Hvor mange er der nu tilbage, og hvordan gennemførte du denne søgning? (**NB:** opgaven *skal* løses i SRS, det er ikke nok at tælle manuelt hvor mange der er!)
- 6) Som sagt skal du analysere signalpeptider og propeptider. Det er derfor nødvendigt at begrænse sættet til de entries der både har et signalpeptid og et propeptid annoteret. Hvor mange entries med begge disse features findes der i *hele* Swiss-Prot (altså ikke kun blandt insulin-forstadier)?
- 7) Kombiner dine to sidste søgninger for at besvare spørgsmålet: Hvor mange insulin-forstadier med annoteret signalpeptid og propeptid er der? (**NB:** hvis du ikke kunne løse spørgsmål 5, så kombiner resultaterne fra 4 og 6 i stedet).

Resultatet af spørgsmål 7 er det ene af de to datasæt du skal bruge i anden halvdel af opgaven. Gem dette datasæt i FASTA format på den computer du arbejder på (**Tip:** i SRS skal du trykke "Save" og derefter sætte "Use view" til "FastaSeqs").

For at få et mere overskueligt datasæt at arbejde videre med, skal du også lave en udgave der er begrænset til primater (se følgende punkter):

- 8) Hvor mange entries fra primater findes der i *hele* Swiss-Prot? (**Tip:** hvis du ikke ved hvad primater hedder på latin, så kig nærmere på det humane entry fra dit tidligere datasæt og check feltet "Taxonomy").
- 9) Kombiner dine to sidste søgninger for at lave det lille datasæt af insulin-forstadier med annoteret signalpeptid og propeptid fra primater. Hvor mange sekvenser indeholder det? Skriv alle entry-navnene (ID'erne) i dit svar!
- 10) Kig nærmere på featuretabellerne i primat-datasættet. Angiv
 - a. sidste position i signalpeptidet,
 - b. første position i propeptidet, og
 - c. sidste position i propeptidet.

Hvis positionen varierer mellem de forskellige entries, så angiv et interval!
Gem også primat-datasættet fra spørgsmål 9 i FASTA format.

4B: Multiple alignment

(**Hjælp** til dem der ikke klarede 4A: Hvis du ikke har fået to datasæt i FASTA format ud af spørgsmål 7 og 9, kan du alligevel godt besvare 4B. Vi har lagt to erstatnings-datasæt på kursus-hjemmesiden, som du kan downloade:

<http://www.cbs.dtu.dk/dtucourse/27611spring2007/eksamen/>

Erstatningen for datasættet fra spørgsmål 9 hedder "insulin-primater-udennavn.fasta", og vi har ændret navnene på sekvenserne, så du kan ikke bruge det til at besvare spørgsmål 9. Erstatningen for datasættet fra spørgsmål 7 hedder "insulin-25-udennavn.fasta", og her har vi både ændret navne og fjernet et antal sekvenser, så du kan ikke bruge det til at besvare spørgsmål 7. Du får også brug for at se på annoteringen af Swiss-Prot entry INS_HUMAN for at besvare 4B, hvis du ikke har besvaret spørgsmål 10).

Brug ClustalW til at lave et multiple alignment af det lille primat-datasæt fra spørgsmål 9. Det er OK at lade alle parametre være default værdier. Du vil observere at insulin fra forskellige primater er temmelig ens. Besvar nu følgende spørgsmål:

11)

Hvor mange positioner i dette alignment er *ikke* 100% konserverede?

12)

Der er et enkelt gap i en af sekvenserne. Forekommer dette i signalpeptidet, A-kæden, propeptidet eller B-kæden?

Lav et tilsvarende alignment af det større insulin-datasæt fra spørgsmål 7. Du skulle nu meget gerne se en større variation i sekvenserne. For at få et overblik over graden af konservering på hver position skal du bruge alignment-editoren **Jalview**: tryk på knappen "Start Jalview" på resultatsiden.

Bemærk at det samlede alignment på grund af gaps er længere end de enkelte sekvenser der indgår i det. Prøv nu at finde hvor grænserne går mellem signalpeptidet, A-kæden, propeptidet og B-kæden i dette alignment. **Tip**: positionen i *alignmentet* fremgår af akse øverst i vinduet, mens positionen i *den enkelte sekvens* vises nederst, når du peger på en aminosyre med musen.

13)

Find en af primatsekvenserne, hvor du jo kender positionerne fra spørgsmål 10, og brug den til at finde, i forhold til det samlede alignment :

- sidste position i signalpeptidet,
- første position i propeptidet, og
- sidste position i propeptidet.

(op til to positioners unøjagtighed bliver regnet som korrekt svar)

Nederst i Jalview-vinduet ser du blokdiagrammer over tre mål for konservering: "Conservation", "Quality" og "Consensus" (% identitet). (**Tip**: Hvis du ikke kan se dem, skal du gå til menuen "View" og sætte markering ved "Show Annotations").

Tænk ikke på forskellen mellem disse tre mål, du skal bare kvalitativt bedømme graden af konservering i de forskellige regioner.

14)

Sammenlign nu signalpeptidet, A-kæden, propeptidet og B-kæden.

a. Er signalpeptidet mere eller mindre konserveret end A- og B-kæderne, eller er der ingen forskel?

b. Er propeptidet mere eller mindre konserveret end A- og B-kæderne, eller er der ingen forskel?