Avdeling for allmenne fag

EKSAMEN

4315 Bioinformatikk

27.05.2004

Tid:	6 timer (09:00 – 15:00)
Målform:	Bokmål
Sideantall:	3 sider oppgave, 1 side vedlegg
Hjelpemidler:	PC m/ internett
Merknader:	Ingen bøker eller skriftlige notater kan brukes. Kandidatene får ikke bruke e-post/chat/msn e.l.
Vedlegg:	1 side ("Vedlegg til eksamen i 4315 Bioinformatikk")

Eksamensresultatene blir offentliggjort på lister med kandidatnummer, lagt i hyllene ved eksamensoppslagstavla, eller du kan gå inn på Arena høgskole – da trenger du heller ikke kandidatnummeret.

Eksamen i 4315 Bioinformatikk, 27.05.2004.

I oppgavesettet foreslås/oppgis en del nettadresser. Du kan velge å bruke disse eller andre du liker bedre etter eget valg. Svar klart og konsist på oppgavene. Hvis du ikke finner noe greit svar på en oppgave forklar kort hva du har gjort, slik at det kan vurderes hvor nær en løsning du var.

- Lykke til!

Oppgave 1

I denne oppgaven skal du hovedsaklig studere et virus som heter Simian Virus 40 (SV40).
SV40 binder seg til et protein som kalles p53. I siste del av oppgaven skal du også se på dette proteinet.
I PDBs "molecule of the month" er både SV40 og p53 beskrevet. Det er ikke direkte nødvendig for oppgaven at du ser på disse beskrivelsene, men det kan være med på å gi deg et overblikk. Hvis du ønsker kan du derfor sjekke:
p53: http://www.rcsb.org/pdb/molecules/pdb31_1.html
SV40: http://www.rcsb.org/pdb/molecules/pdb47_1.html

I) Ta utgangspunkt i følgende SwissProt post (primary accesion no.): **P03070**
Finn denne i SwissProt (http://us.expasy.org/sprot/) og svar på følgende fra NiceProt View-fila:

a) Hva heter proteinet?
b) Hvilken virusfamilie hører SV40 til?
c) Hvor mange aminosyrer langt er P03070 og hva er molekylvekten?
d) Denne posten er i Swiss-Prot, Hva er sammenhengen/forskjellen mellom Swiss-Prot og TrEMBL?

II) Du skal så studere egenskaper og karakteristiske trekk til P03070. Hent sekvensen som FASTA-format eller bruk accesion-nummeret direkte. Gjør og svar på følgende:

a) Hva er et FASTA-format? (Hvordan er det bygd opp / hva karakteriserer det)
b) bruk ProtParam (http://us.expasy.org/tools/protparam.html) til å beregne proteinets molekyl vekt. Stemmer svaret med oppgave Ic? Hvilken aminosyre er det flest av i proteinet og hvor mange prosent? Metoden protParam bruker til å beregne molekylvekt tar ikke hensyn til såkalte "post-translational modifications". Nevn tre eksempler på slike modifikasjoner.
c) Bruk Peptide Cutter (http://us.expasy.org/tools/peptidecutter/) til å bestemme hvor mange kuttinger enzymet "Proline-endopeptidase" gjør i proteinet og i hvilke posisjoner.
d) Bruk ett av verktøyene Pfscan (http://hits.isb-sib.ch/cgi-bin/PFSCAN),
CD-search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) eller InterProScan (http://www.ebi.ac.uk/InterProScan/) til å identifisere domenene i proteinet. Hvor mange domener har en betydelig treff? Hva er navnene deres? Generelt sett, hva er poenget med å identifisere domener i et protein?
e) Til denne oppgaven anbefales resurser administrert av NCBI (http://www.ncbi.nlm.nih.gov/).
De samme domenene du fant i oppgave IId finnes som 3D-strukturer i PDB. Gjør et proteinBLAST-søk mot PDB som database og finn PDB-kode for dem. Oppgi også E-verdien i BLAST-søket. Hva er en E-verdi, hva forteller den om?
Hvilke aminosyrer (nummer X til Y) i P03070 er involvert i binding til p53 proteinet?

III) Som nevnt binder P03070 seg til p53. I denne oppgaven skal du se på dette proteinet.

a) Genet som p53-proteinet "kommer fra" heter TP53 (som også er HUGO-kode). Bruk f.eks. Ensembl (http://www.ensembl.org/) eller ulike resurser under NCBI (http://www.ncbi.nlm.nih.gov/) til å finne ut hvilket humant kromosom og hvilket kromosombånd TP53 genet ligger på. Mutasjoner i TP53 genet er assosiert med cancer (kreft). Blant annet et familierelatert syndrom oppkalt etter de to som oppdaget det. Hva kalles dette syndromet?

b) Proteinstrukturen MMDB:3973 / PDB:1TUP viser den delen av p53 som binder DNA.
Det er for hvert av domenene A, B og C et helix-område som ligger mot DNA. Dette området har sekvens PGRDRRTEEE. I rom nærmest dette området (også ganske nær DNA) ligger et område på 9 aminosyrer uten fast struktur. Finn sekvensen til dette området. (Hint: Bruk 3D-verktøyet Cn3D. Hvis dette ikke synes å være installert på PC'en din trykk 'START - alle programmer - diverse - Cn3D', eller spør om hjelp). Forøvrig er adressene til MMDB (NCBI structure) og PDB henholdsvis http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml og http://www.rcsb.org/pdb/index.html evt. http://pdb.protein.osaka-u.ac.jp/pdb/).

c) I oppgave II fant du den DNA bindende delen av SV40 proteinet P03070. Sammenlikn denne med den DNA bindende delen av p53 (bruk 1TUP:A). Bruk verktøyet Lalign (http://www.ch.embnet.org/software/LALIGN_form.html) til dette. Gjør en global sammenlikning, med BLOSUM35 som " scoring matrix", og finn ut hvor mange prosent likhet det er mellom de to sekvensene. Er det sannsynlig at de har strukturell likhet? Begrunn svaret.

d) I kreftforskning brukes ofte smågnagere som mus og rotter i laboratorieforsøk. Det er ikke trivielt å avgjøre om disse er egnet som modeller for forskning på menneskets p53, men du skal gjøre en forenklet vurdering basert på sekvensdata. Gjør en multisekvenssammenlikning vha ClustalW eller Tcoffee av p53 fra menneske (P04637), mus(P02340), rotte (P10361) og kanin(Q95330). Avgjør fra dette resultatet om slike forsøksdyr synes egnet til formålet. Begrunn svaret. Hvor godt konservert, mellom disse 4 artene, er de to korte sekvensene du så på i oppgave IIIb?
(Adressen til Tcoffee er f.eks http://www.ch.embnet.org/software/TCoffee.html eller http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi . Tilsvarende kan ClustalW finnes flere steder: http://pir.georgetown.edu/pirwww/search/multaln.html, http://www.ebi.ac.uk/clustalw/index.html eller http://www.ch.embnet.org/software/ClustalW.html)

Oppgave. 2

a) Forklar kortfattet hva man mener med ortologe, paraloge og xenologe gener.

b) Figur 1 (vedlegg) viser fylogrammer for to gener, 16s ribosomal RNA og beta-lactamase (som koder for resistens mot penicillin) fra syv bakteriearter: *Proteus mirabilis, Proteus vulgaris, Klebsiella pneumoniae, Klebsiella oxytoca, Pseudomonas aeruginosa, Pseudomonas fluorescens* og *Synechococcus*. *Proteus* og *Klebsiella* er gram-negative gamma-proteobakterier i gruppen *enterobacteriaceae*, *Pseudomonas* er gram-negative gamma-proteobakterie i gruppen *pseudomonadaceae* mens Synechococcus er photosyntetisk cyanobakterie.
Sammenlikn de to trærne og foreslå hvorfor de er forskjellige.

c) Figurarket/vedlegget skal leveres tilbake som en del av besvarelsen. Angi på figurene med bokstaver (N eller G) de følgende delene: Nodene (N) og Interne grener (G)

d) Hvilken betydning har grenenes lengde i et fylogram?

e) Hva representerer nodene i et fylogenetisk tre?

f) Hva er en utgruppe i en fylogenetisk analyse?

Oppgave 3

I denne oppgaven tenker vi oss at du er teknisk sakkyndig for politiet.
Det har vært et alvorlig utbrudd av infeksjon med den dødelig matbårne *Escherichia coli* stammen, O157:H7. To mulige kilder har blitt identifisert: En landsomspennende hamburgerkjede og en lokal kebab-restaurant. *E. coli* O157:H7 har blitt påvist i kjøttdeig fra begge steder. Tre personer har mistet livet i utbruddet og flere er syke. Det er snakk om store erstatningsbeløp og muligens kriminelle handlinger. Politiet frykter at noen av de mistenkte kan rømme til utlandet, og det haster med å få en sikker identifikasjon av kilden slik at de kan gå til eventuell pågripelse.

Du blir bedt om å identifisere kilden. Du har følgende O157:H7 isolater:

- Kjøttdeig fra hamburgerkjeden. 3 isolater.
- Kjøttdeig fra kebabrestauranten. 2 isolater.
- Isolater fra syke personer. 6 isolater.

Du bestemmer deg for å sammenligne isolatene vha genomisk restriksjonsanalyse (DNA fingerprinting) og gelelektroforese. Du har tilgang til utstyr til:

- Agarose gel-elektroforese som separerer DNA fragmenter i området 1000 - 10 000 bp
- Pulsfelt gel-ektroforese som separerer fragmenter i størrelsesområdet 50 000 - 700 000.

Du finner at du har tilstrekkelig mengder av kun tre restriksjonsenzymer: SimI, SacI og Sse232I.

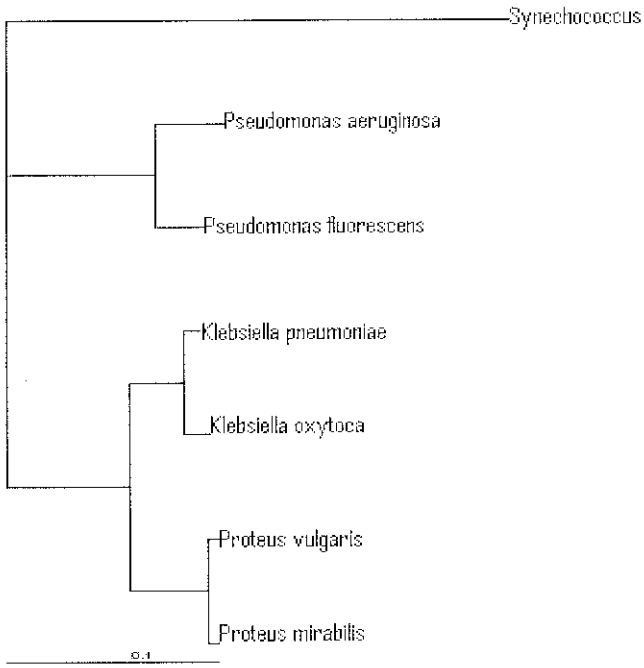
I) For å finne en kombinasjon av enzym og elektroforesesystem som skiller mellom *E. coli* 0157:H7 stammer, skal du tester dem med verktøyet "Restriction Digest Tool" under TIGR-CMR (http://www.tigr.org/tigr-scripts/CMR2/restrict_display.pl) og bruke genomisk sekvens til E.Coli O157:H7 isolat Sakai.

a) Beskriv kort de 6 mønstrene du ser.
b) Hvilke(t) av de tre restriksjonsenzymene virker best egnet til oppdraget, for hver av elektroforese metodene? Begrunn svaret.

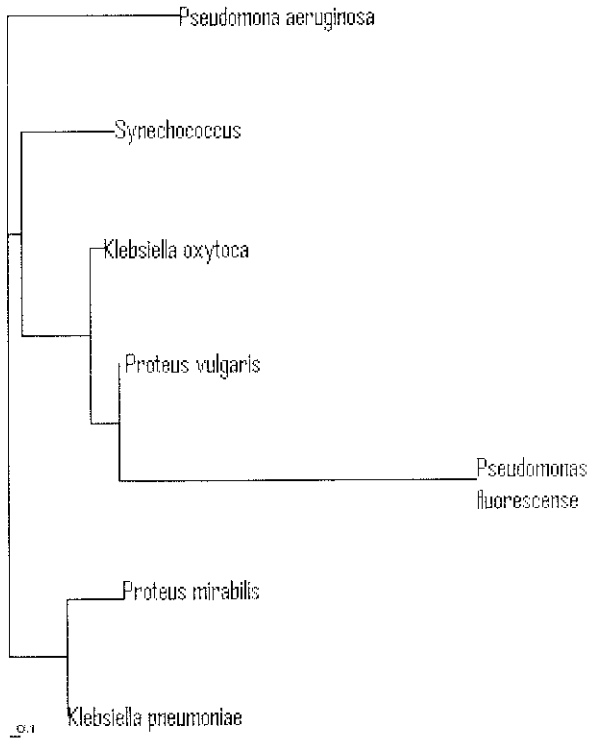
II) Enzymene kutter ved følgende sekvenser:
SimI: GGGTC.
SacI: GAGCTC.
Sse232I: CGCCGGCG

a) Forklar kort hvorfor de forskjellige enzymene gir forskjellige mønster.
b) Beskriv (kortfattet) hvordan du ville gå fram for å identifisere kilden for utbruddet.

Vedlegg til eksamen i 4315 Bioinformatikk. **Studentnr:** _____



Figur 1A: Fylogenetisk tre - 16s ribosomal RNA



Figur 1B: Fylogenetisk tre - Beta-lactamase