

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

**Eksamen i: MBV3070**  
**Eksamensdag: 10. juni 2005**  
**Tid for eksamen: 9:00 – 12:00**  
**Oppgavesettet er på 3 side(r)**  
**Vedlegg: Ingen**  
**Tillatte hjelpemidler: Ingen**

*Kontroller at oppgavesettet er komplett  
før du begynner å besvare spørsmålene.*

### Oppgave 1

Din veileder har i en årrekke vært interessert i enzymet kamikase fra mollusken *Arena globalis*. Din oppgave er å bringe nye teknikker inn i prosjektet, og du har allerede klonet og sekvensert cDNA som koder for enzymet.

- a) Hvordan vil du finne aminosyresekvensen til proteinet som dette cDNA koder for? Oppgi om mulig et eller flere datamaskinprogrammer som kan hjelpe deg med dette.

Nukleotidsekvensen viser seg å inneholde en åpen leseramme som koder for et protein på 270 aminosyrer som du antar er kamikase-enzymet.

- b) Din veileder vil gjerne vite om det finnes lignende enzymer i andre organismer. Hvordan vil du hjelpe ham med å finne svar på dette?

Det viser deg at ditt protein viser sekvenslikhet med en rekke enzymer fra forskjellige arter, som alle er serin proteaser. Dette kan tyde på at også kamikase er en serin protease, en opplysning du vet vil revolusjonere kamikase-feltet. Men du vet at slik sekvenslikhet ikke alltid vil bety at proteinene har identisk funksjon og bestemmer deg for en videre bioinformatisk analyse av sekvensen.

- c) Hva slags bioinformatiske analyser kan du tenke deg å benytte for å bli sikrere på at kamikase virkelig er en serin protease?

Analysen din styrker hypotesen om at kamikase virkelig er en serin protease. Ved hjelp av ClustalW lager du nå en flersekvenssammenstilling av forskjellige serin proteaser og kamikase. Et område i den ser slik ut (kamikase er nederst):

LT FATVR--GAGHEVPLFEP  
IA FLTIK--GAGHMVPTDKP  
IT FLTIK--GAGHMVPTDKP  
LA FTLSNSVGH--MAPSKDP  
LQEV LIRNAGH--MVPRDQP  
FT FLRIYDAGH--MVPYDQP  
ID LLTVKGAGH--MVPYDRA

- d) Du ser et behov for manuell redigering av flersekvenssammenstillingen. Hvorfor? Nevn et eller flere datamaskinprogrammer som kan være nyttige for slik manuell redigering av flersekvenssammenstillinger.
- e) Din veileder vil gjerne innføre setestyrte mutasjoner i kamikase for å prøve å endre substratspesifisiteten uten å miste den enzymatiske aktiviteten. Basert på den manuelt redigerte sammenstillingen, er det noen aminosyrer i området i sammenstillingen over du ville anbefalt ham ikke å prøve å endre? Hvorfor?

I nabolaboratoriet har de nettopp fullført sekvenseringen av genomet til *Cafeteria vulgaris*, en svært fjern slektning av *Arena globalis*. Det ville være interessant å vite om denne organismen inneholder proteiner som er homologer til kamikase.

- f) Hva vil du gjøre for å finne mulige homologe sekvenser? Hva vil det si at sekvenser er homologe? Vil det at du søker i en genomsekvens kunne medføre problemer?

Du finner virkelig et gen i *C. vulgaris* –genomet som koder for et protein med en sekvens som ligner så mye på kamikase-sekvensen at dere kan konkludere med at dette proteinet også må ha kamikase-aktivitet. Dere er klare for å publisere. Til publikasjonen trenger du et fylogenetisk tre og bestemmer deg for å bruke flersekvenssammenstillingen ovenfor med kamikasesekvensen fra *C. vulgaris* innført og en distansebasert metode for konstruksjonen av treet. For området som tilsvarer flersekvenssammenstillingen over ser *vulgaris*-sekvensen slik ut:

VDFLTVRGS GHFVPEDKP

- g) Vis avstandene mellom de forskjellige sekvensene i denne delen av den nye flersekvenssammenstillingen, i tabellform.

Det endelige treet lager du ved UPGMA-metoden. Dere sender inn artikkelen til et tidsskrift som bruker refereer, vitenskapelige eksperter som skal vurdere kvaliteten av arbeidet. Refereene har diverse kommentarer til manuskriptet deres. Blant annet er de misfornøyd med valg av metode for konstruksjon av treet.

- h) Dersom du hadde vært referee, hvilke kommentarer ville du ha gitt til valget av metode?

## Oppgave 2

Man har i dag tilgang til fullstendige genomsekvenser fra mange forskjellige organismer.

- a) Gjør rede for de viktigste forskjellene man har funnet mellom genomsekvenser fra prokaryote og eukaryote organismer. Eksempler verdsettes!

Som følge av genomsekvensering har man utviklet metoder som gjør det mulig å studere alle genene fra en organisme i ett og samme forsøk. Én slik metode er mikromatrise ("microarray")-teknologi.

- b) Beskriv prinsippene for mikromatriseteknologien og gi eksempler på hvilke opplysninger mikromatrise-eksperimenter kan gi.

## Oppgave 3

- a) Man sammenligner sekvensene til to proteiner som begge består av rundt 300 aminosyrer. Et av disse proteinene har kjent tredimensjonal struktur. Hvor stor sekvenslikhet må disse proteinene ha for at man kan bygge en tredimensjonal modell av det ene proteinet med ukjent struktur? (Forklar; det holder med en kort forklaring)
- b) Hva er en Ramachandran plott og hvordan kan en slik plott brukes til å sjekke for mulige feil i proteinstrukturer eller strukturmodeller?
- c) Når man har en ukjent protein sekvens kan man lete etter mulige homologer med hjelp av vanlig sekvenssøk (f eks BLAST), men også med hjelp av metoder som kalles "threading" eller "Environmental template method" eller "structural profile method". Forklar hvorfor den siste metoden kan være mer sensitiv (dvs bedre til å finne fjerne slektskaper, dvs "remote homologues") enn vanlige sekvenssøk med f eks BLAST.
- d) Du har en ukjent proteinsekvens, du søker mot Pfam, og du finner en treff som er statistisk klart signifikant. Diskuter i hvilken grad treffet gir deg strukturell informasjon og i hvilken grad treffet gir deg funksjonell informasjon.

# UNIVERSITETET I OSLO

## Det matematisk-naturvitenskapelige fakultet

**Exam in MBV3070**

**Day of exam: June 10<sup>th</sup> 2005**

**Exam hours: 9:00 – 12:00**

**This examination paper consists of 3 pages.**

**Appendices: None**

**Permitted materials: None**

*Make sure that your copy of this examination paper is complete before answering.*

### Exercise 1

Your tutor has for a long time been interested in the enzyme kamikase from the mollusc *Arena globalis*. Your job is to bring new and modern techniques into the project, and you already have cloned and sequenced cDNA coding for the enzyme.

- a) What will you do to find the amino acid sequence of the protein coded by this cDNA? If possible, name one or more computer programs that may help you in this.

The nucleotide sequence turns out to contain an open reading frame coding for a 270 amino acid protein that you assume to be the kamikase enzyme.

- b) Your tutor would like to know if similar enzymes exist in other organisms. What will you do to help him answer this question?

It turns out that your protein shows sequence similarity to a number of enzymes from various species, all of them serine proteases. This suggests that even kamikase might be a serine protease, something that would revolutionize the whole kamikase field. You know, however, that sequence similarity will not always mean that the corresponding proteins have identical function, and you decide to perform further bioinformatical analyses of the kamikase sequence.

- c) What types of bioinformatical analyses will you consider to use to become more convinced that kamikase really is a serine protease?

Your analysis strengthen the hypothesis that kamikase really is a serine protease. Using ClustalW, you now produce a multiple sequence alignment containing several serine proteases and kamikase. A part of the alignment looks like this (kamikase at the bottom):

```

LTFATVR--GAGHEVPLFEP
IAFLTIK--GAGHMVPTDKP
ITFLTIK--GAGHMVPTDKP
LAFTLSNSVGH--MAPSKDP
LQEVLI RNAGH--MVPRDQP
FTFLRIYDAGH--MVPYDQP
IDLLTVKGAGH--MVPYDRA

```

- d) You observe a need for manual editing of the sequence alignment. Why? Name one or more computer programs that can be useful for such manual editing of multiple sequence alignments.
- e) Your tutor would like to introduce site directed mutations in kamikase to try to alter the substrate specificity without losing the enzymatic activity. Based upon the manually edited sequence alignment, are there any amino acid residues in the area of the alignment shown above that you would recommend him not to alter? Why?

In the neighbouring laboratory, sequencing of the genome of *Cafeteria vulgaris*, a distant relative of *Arena globalis*, has just been completed. It would be interesting to know if this organism contains proteins that are homologues of kamikase.

- f) What would you do to find possible homologues? What does it mean that sequences are homologous? Could the fact that you are searching in a genome sequence give you any problems?

You really find a gene in the *C. vulgaris* genome that codes for a protein with a sequence that is so similar to the kamikase sequence that you conclude that even this protein must have kamikase activity. You are ready to publish. For the publication you need a phylogenetic tree and decide to use the multiple sequence alignment above, after having included the *C. vulgaris* sequence, and a distance-based method for the tree construction. For the region corresponding to the multiple sequence alignment above, the *vulgaris* sequence is like this:

```
VDFLTVRGS GHFVPEDKP
```

- g) In a table, show the distances between the various sequences in this part of the new multiple sequence alignment.

You produce the final tree using the UPGMA method. You submit the paper to a journal using referees, scientific experts that will evaluate the quality of your work. The referees have several comments to your manuscript. Among others, they are not too pleased about your choice of tree construction method.

- h) If you had been a referee, what comments would you have had to the choice of method?

## Exercise 2

Today, complete genome sequences from many different organisms are available.

- a) Describe the main differences one has observed between genome sequences from prokaryotic and eukaryotic organisms. Examples will be rewarded.

As a consequence of genome sequencing, methods have been developed to allow the study of all genes from an organism in a single experiment. One such method is the microarray technology.

- b) Describe the principles behind microarray technology and give examples of what types of information microarray experiments can provide.

### **Exercise 3**

- a) The sequences of two proteins, both containing about 300 amino acid residues, are compared. One of these proteins has a known three-dimensional structure. How high should the sequence identity between the proteins be to allow the building of a three-dimensional model of the protein with unknown structure? (Explain, a short explanation will be sufficient).
- b) What is a Ramachandran plot and how can such a plot be used to look for possible errors in protein structures or structural models?
- c) With a new protein sequence one can look for possible homologues using ordinary sequence searches (e.g. BLAST), but also by using methods called “threading” or “environmental template method” or “structural profile method”. Explain how this last method may be more sensitive (i.e. better at finding distant relatives, i.e. “remote homologues” than ordinary sequence queries, for instance using BLAST).
- d) You have a previously unknown protein sequence, you query Pfam, and you get a hit that statistically is clearly significant. Discuss to what extent this hit gives you structural information and to what extent it gives you functional information.