NTNU

**Det medisinske fakultet**
**Institutt for kreftforskning og molekylær medisin**

# EKSAMENSOPPGAVE I MTEK3001
# Anvendt bioinformatikk og systembiologi

**Onsdag 2. juni 2010, kl. 0900-1300**

Antall studiepoeng: 7.5
Tillatte hjelpemidler: Godkjent kalkulator
Antall sider (inkludert forside): 4

**Språk:**
**Bokmål / Engelsk**

Kontaktperson under eksamen:
Professor Finn Drabløs, 72 57 33 33 / 915 76 023

**Sensurfrist: 23. juni**
Sensuren kunngjøres på internett http://studweb.ntnu.no/

**Oppgaver er gitt *både* på norsk ("Oppgave") *og* på engelsk ("Problem").**
**Problems are given *both* in Norwegian ("Oppgave") *and* in English ("Problem").**

---

**Oppgave 1** – Identifisere proteiner i en prøve (25 p)

Du har mottatt en kompleks prøve av proteiner, og du ønsker å separere denne prøven eksperimentelt for å identifisere og karakterisere utvalgte proteiner.

a) Du separerer prøven din i to trinn og ender opp med en todimensjonal (2D) gel. Beskriv kort hvordan du gjør denne separasjonen basert på to viktige egenskaper for proteiner. Hvilke egenskaper er dette? (5 p)

b) Din prøve er del av et komparativt eksperiment hvor du sammenligner proteinnivå i ulike prøver. Beskriv hvordan du kan sammenligne uttrykksnivå for proteiner med 2D geler. Forklar kort en eksperimentell teknikk du kan bruke for å bestemme identiteten til utvalgte proteiner fra gelen. (10 p)

c) Innledende sekvensanalyse viser at et av proteinene du har identifisert har et "coiled coil" område og ett enkelt transmembrant domene. Forklar hvordan disse to egenskapene kan identifiseres ved å bruke bare proteinsekvensen og eksperimentelt bestemte aminosyre-egenskaper. (10 p)

**Problem 1** – Identifying proteins in a sample (25 p)

You have been given a complex sample of proteins, and you want to separate this sample experimentally in order to identify and characterise selected proteins.

a) You do the separation of your sample in two steps and end up with a two-dimensional (2D) gel. Describe briefly how you do this separation based on two important properties of proteins. What are these properties? (5 p)

b) Your sample is part of a comparative experiment, where you compare protein levels in different samples. Describe how you can compare protein expression levels by 2D gels. Explain briefly an experimental technique you can use to determine the identity of selected proteins from the gel. (10 p)

c) Initial sequence analysis shows that one of the proteins you have identified has a coiled coil region and a single transmembrane domain. Explain how these two features may be identified, using only the protein sequence and experimentally determined amino acid properties. (10 p)

---

**Oppgave 2** – Finne lignende proteiner (25 p)

Du har nå klart å identifisere et av proteinene i den komplekse proteinprøven som enzymet incredulase, og du har funnet proteinsekvensen for enzymet. Nå ønsker du å finne beslektede proteiner.

a) Forklar termene homologer, ortologer og paraloger, og forskjellen mellom similaritet og homologi. (5 p)

b) Vi kan sammenligne sekvenser ved optimal parvis sekvenssammenstilling. Forklar forskjellen på lokal og global sammenstilling. Formelen som brukes til å fylle matrisen for dynamisk programmering for den enkleste formen for parvis sammenstilling består av tre ledd. Forklar hva disse leddene representerer og hvordan de brukes for å beregne et element i matrisen. Forklar forskjellen mellom en lineær og en forbedret ("affine") modell for mellomrom i sekvenssammenstilling. (10 p)

c) For sekvenssammenstilling trenger vi en substitusjonsmatrise. Forklar hvordan PAM-matrisen har blitt generert. Det er ikke nødvendig å oppgi formler. Beskriv forholdet

mellom PAM-avstand og prosent identitet mellom sekvenser i sekvenssammenstilling. (10 p)

**Problem 2** – Finding similar proteins (25 p)

You have now been able to identify one of the proteins in your complex protein sample as the enzyme incredulase, and you have found the protein sequence of the enzyme. Now you want to find related proteins.

a) Explain the terms homologs, orthologs and paralogs, and the difference between similarity and homology. (5 p)

b) We can compare sequences by optimal pairwise alignment. Explain the difference between local and global alignment. The formula used to fill the dynamic programming matrix for the simplest form of pairwise alignments consists of three terms. Explain what these terms represent and how they are used to compute an element of the matrix. Explain the difference between a linear and an affine gap penalty model for sequence alignment. (10 p)

c) For sequence alignment we need a substitution matrix. Explain how the PAM matrix has been generated. No formulas are needed. Describe the relationship between PAM distance and percentage identity between sequences in sequence alignments. (10 p)

---

**Oppgave 3** – Analysere protein evolusjon (25 p)

Du har funnet et sett proteiner som ut fra sekvenslikhet synes å være beslektet med ditt incredulase enzym. Nå ønsker du å bruke disse sekvensene til å gjøre en fylogenetisk analyse av incredulasefamilien av proteiner.

a) Forklar hva som menes med en utgruppe og hvordan vi kan bruke en utgruppe til å finne roten av et fylogenetisk tre. (5 p)

b) Forklar forskjellen mellom et cladogram, et additivt tre og et ultrametrisk tre. Forklar forskjellen mellom synonyme og ikke-synonyme mutasjoner, hvordan dette påvirker mutasjonshastighet på kodonnivå og mulig påvirkning på fylogenetisk analyse. (10 p)

c) Figur 1 viser en avstandsmatrise for incredulase og beslektede proteiner, basert på en multippel sekvenssammenstilling. Bruk UPGMA til å konstruerer et fylogenetisk tre for sekvenssettet, basert på avstandsmatrisen. (10 p)

**Problem 3** – Analysing protein evolution (25 p)

You have found a set of proteins that by sequence similarity seem to be related to your incredulase enzyme. Now you want to use these sequences to do a phylogenetic analysis of the incredulase family of proteins.

a) Explain what we mean by an outgroup and how we can use an outgroup to root a phylogenetic tree. (5 p)

b) Explain the difference between a cladogram, an additive tree and an ultrametric tree. Explain the difference between synonymous and nonsynonymous mutations, how this affects mutation rate at the codon level and possible consequences for phylogenetic analysis. (10 p)

c) Figure 1 shows a distance matrix for a set of sequences, based on a multiple alignment. Use UPGMA to construct a phylogenetic tree for the sequence set, based on the distance matrix. (10 p)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| A | - | 6 | 8 | 1 | 2 | 6 |
| B | | - | 8 | 6 | 6 | 4 |
| C | | | - | 8 | 8 | 8 |
| D | | | | - | 2 | 6 |
| E | | | | | - | 6 |

*Figur 1 / Figure 1*

---

**Oppgave 4** – Forutsigelse av proteinstruktur (25 p)

Du tror at reduksjon av aktiviteten til incredulase kan være en effektiv behandling for en alvorlig sykdom, og du ønsker å utvikle en inhibitor for enzymet. For å gjøre det trenger du tertiærstrukturen for proteinet, men forsøk på eksperimentell strukturbestemmelse har ikke lyktes. Derfor prøver du beregningsbasert strukturforutsigelse, og starter med sekundærstrukturen.

a) Prediksjon av sekundærstruktur kan gjøres med neurale nett, ofte ved at det brukes en dobbel arkitektur med to neurale nett for forbedret ytelse. Beskriv typiske inndata til hvert av disse to nettverkene. (5 p)

b) Kvaliteten på prediksjon av sekundærstruktur estimeres med $Q_3$ målet. Beskriv dette målet, og forklar hvorfor det noen ganger kan gi et for optimistisk estimat av prediksjonskvalitet, sammenlignet med for eksempel Sov målet. Andre typer prediksjoner vurderes ofte basert på sensitivitet (Sn) og spesifisitet (Sp). Forklar hvordan disse kvalitetsmålene blir estimert. (10 p)

c) Forklar kort hovedtrinnene i homologibasert modellering av tertiærstruktur av et protein, fra du starter med en enkelt proteinsekvens til du ender opp med en optimalisert tertiær struktur. (10 p)

**Problem 4** – Predicting protein structure (25 p)

You believe that reducing the activity of incredulase may be an efficient treatment of a serious disease, and you want to design an inhibitor of the enzyme. In order to do so you need the tertiary structure of the protein, but attempts on experimental structure determination have not been successful. You therefore try computational structure prediction, starting with secondary structure.

a) Secondary structure prediction can be done with neural networks, often using a double architecture with two neural networks for improved performance. Describe typical input to each of these two networks. (5 p)

b) The quality of secondary structure prediction is often estimated with the $Q_3$ measure. Describe this measure, and explain why it sometimes may give a too optimistic estimate of prediction quality, compared to for example the Sov measure. Other types of predictions are often assessed using sensitivity (Sn) and specificity (Sp). Explain how these quality measures are estimated. (10 p)

c) Explain briefly the main steps in homology-based modelling of tertiary structure of a protein, starting with a single protein sequence and ending with an optimised tertiary structure. (10 p)