UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Skriftlig eksamen/Written Examination

Emne/Course: **MOL204** **Anvendt bioinformatikk I /** **Applied bioinformatics I**	Semester: **H2012**
Dato/Date: **17. desember/17 December**	Kl. (fra- til)/Time (from-to): **9:00-13:00**
Tillatte hjelpemidler (i samsvar med emnebeskrivelsen)/Permitted examination support material(according to the course description): **kalkulator/calculator**	Antall sider/Number of pages: **7**

Annen informasjon:
Alle spørsmål skal besvares. Dersom ikke annet er angitt, forventes korte og konsise svar. Les oppgavetekstene nøye. Sammenlign gjerne med den engelske teksten. I noen av spørsmålene er det brukt engelske ord (markert med 'en:') slik de forekommer i læreboken. Bruk ikke for lang tid på noe enkeltspørsmål. Hopp heller over spørsmål som synes å ta for lang tid, og ta heller disse til slutt. Dersom et spørsmål er uklart, gjør da oppmerksom på hvordan du har bestemt deg for å forstå det. Ta med alle relevante momenter i hver oppgave eller henvis til andre svar dersom noen av oppgavene synes å overlappe. Dersom du illustrerer dine svar med figurer, husk at disse skal vurderes både som original og kopi - unngå derfor bruk av farger. Bruk heller andre metoder for å fremheve ulike deler av figurene. Tentative poeng er angitt for hver oppgave. Totalt utgjør de **85 poeng**. Bruk poengene til å vurdere hvor mye arbeid det lønner seg å legge i hvert svar. **Merk: ingen spørsmål krever lange utredninger.**

Additional information:
Answer all questions. If not otherwise stated, brief and concise answers are expected. Read the questions carefully. Do not spend too much time on each question. It is better to proceed to the next questions and return to time consuming questions at the end. If a question appears unclear or ambiguous, then explain how you have interpreted the question. Include all relevant aspects in each answer. If one answer appears to overlap with another, you may cross-reference them. If you illustrate any of your answers with figures, remember that these will be evaluated both from the original and from the copy - avoid therefore the use of colours. Use other methods to highlight different parts of the figures. For each question is given a tentative number of points to indicate how the question contributes to the total of **85 points**. Use these points to judge how much time it is worth spending on each question. **Note: none of the questions require long answers.** English text on pages 5-7.

Oppgåve 1 – Databasar og databasesøk (totalt 12p)

A (4p) EMBL databasen er ein sentral bioinformatisk ressurs. Kva slag data finn ein her? Nemn to andre databasar med same type data, og forklar korleis EMBL databasen er relatert til dei. Kva tyder det at desse databasane er primære databaser?

B (4p) EMBL databasen er òg relatert til databasen UniProtKB. Forklar kva UniProtKB er, og korleis data flyt frå EMBL databasen til UniProtKB. UniProtKB databasen er samansett av to deler – forklar kva dei er, og korleis dei er relaterte til kvarandre.

C (4p) Ein *oppføring* (en:entry/record) i databasar som EMBL og UniProtKB består av mange *felt* (en:fields). Eit av desse felta inneheld *tilgangsnummer* (en:accession number) – kva er det?. Nemn eit verktøy som kan nyttast for å søkje i databasar som EMBL og UniProtKB, og korleis du kan søke spesifikt i felt.

Oppgave 2 – Parvise sammenstillinger (en: alignments) (totalt 18p)

Vi har gitt to proteinsekvenser *q*:LDIR og *d*:IRR og følgende skåringsmatrise (utdrag fra PAM250):

	R	D	I	L
R	6	-1	-2	-3
D		4	-2	-4
I			5	2
L				6

To ulike matriser *H1* og *H2* brukt for å finne de beste globale sammenstillingene ved dynamisk programmering, ser delvis utfylt slik ut:

H1

	d	I	R	R
q	0	-2	-4	-6
L	-2	2	0	-2
D	-4	0	1	-1
I	-6	1		
R	-8	-1		

H2

	d	I	R	R
q	0	-3	-6	-9
L	-3	2	-1	-4
D	-6	-1	1	-2
I	-9	-1		
R	-12	-4		

A (2p) I tillegg til skåringsmatrisen trenger vi å vite gapstraffen for å fylle ut matrisene *H1* og *H2*. Hva slags gapstraff er brukt her, og hva er kostnaden *g* for hvert gap?

B (6p) Fyll ut de manglende verdiene i matrisene *H1* og *H2*. Hva er skårene *S1* og *S2* for de beste sammenstillingene?

C (6p) Finn de(n) beste sammenstillingen(e), og forklar kort prosedyren. Illustrer ved å tegne en eller flere stier (en:paths) gjennom matrisen.

D (4p) Forklar kort prosedyren for lokal sammenstilling. Hvorfor brukes lokale sammenstillinger oftere enn globale til søk i proteinsekvensdatabaser?

Oppgåve 3 – Sekvensbaserte søk i databasar (totalt 15p)

A (3p) Definer omgrepa *homolog*, *ortolog*, og *paralog*.

B (2p) Smith-Waterman er ein algoritme som kan nyttast til å gjere sekvenssøk. Kvifor nyttast denne metoden vanlegvis ikkje til dette føremonet?

C (4p) Nemn ein heuristisk metode som er meir vanleg brukt til sekvenssøk, og forklar kort prinsippet for og hovudstega i denne metoden. Kva er veikskapen til denne metoden samanlikna med Smith-Waterman?

D (3p) Kva fortel E-verdien om resultatet til eit sekvenssøk?

E (3p) Dersom du skal søkje etter fjerne slektningar av eit protein er det betre å nytta BLOSUM45 som skårematrise enn BLOSUM62. Kvifor?

Oppgave 4 – Multiple sekvenssammenstillinger (totalt 11p)

A (3p) Forklar hva progressiv multippel sammenstilling er og hvordan denne metoden brukes i Clustal.

B (2p) Iterativ sammenstilling kan brukes til å forbedre en mulippel sammenstilling. Forklar kort hvorfor.

C (3p) Flere programmer for multippel sekvenssammenstilling kan gjøre bruk av informasjon om proteinenes struktur (f.eks. Clustal og 3Dcoffee). Forklar kort hvorfor man i noen situasjoner kan få sammenstillinger med bedre kvalitet når strukturinformasjon inkluderes.

D (3p) Når man skal lage et fylogenetisk tre for et sett av homologe proteinsekvenser, er det bedre å benytte en multippel sekvenssammenstilling enn parvise sammenstillinger av alle mulige par av sekvensene. Forklar kort hva årsaken til dette er.

Oppgåve 5 – Proteindomener og protein struktur (totalt 15p)

A (4p) Kva er eit konservert globulært proteindomene? Kva for eigenskaper er typiske for slike domener?

B (3p) Kvifor er SMART eit godt verktøy til å identifisere kjente konserverte globulære domener i proteiner?

C (4p) PDB (Protein Data Bank), CATH og SCOP er alle databasar som inneheld informasjon om proteinstrukutur. Kva for data kan ein finne i dei tre databasane, og forklar kort korleis dei er relaterte til kvarandre.

D (4p) Den mest brukte metoden for å modellere proteiners tredimensjonale struktur er homologimodellering. Skisser kort dei ulike stega som nyttast i homologimodellering.

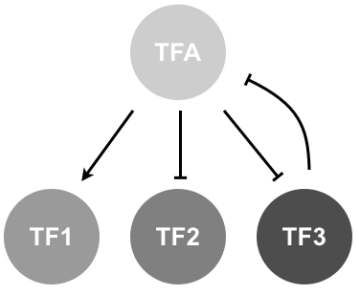
Oppgave 6 – Systembiologi (totalt 14p)

A (5p) Forklart kort hvordan man innen systembiologi betrakter *komponenter*, *interaksjoner* og *reaksjoner*, og *dynamiske endringer* til å bygge modeller av regulatoriske og metabolske nettverk.

B (3p) Høyereordens biologiske systemer sies å vise emergente egenskaper (en: emergent properties). Forklar, gjerne med et eksempel, hva en systembiolog mener med dette begrepet.

C (4p) Anta at du kjenner de spesifikke bindingssetene til 50 transkripsjonsfaktorer i gjærgenomet (*Saccharomyces cereviciae*). Du har også tilgang til målinger av endringer i genuttrykk for alle de ca. 6000 genene i gjær, som følge av at hver av de 50 transkripsjonsfaktorene muteres (tap av funksjon). Forklar hvordan du kan bruke denne informasjonen for å lage et genregulatorisk nettverk for gjær.

D (2p) En liten del av et genregulatorisk nettverk er vist i figuren under.



TFA er en transkripsjonsfaktor som kan aktivere transkripsjonsfaktoren TF1, og hemme faktorene TF2 og TF3. I tillegg er TF3 er en TFA-hemmer. Hvilken effekt vil en TF3 mutasjon ha på genuttrykket til TFA, TF1, og TF2?

End of Norwegian text – English text on next pages

Question 1 – Databases and database searches (total 12p)

A (4p) The EMBL database is a central bioinformatical resource. What kind of data does it contain? Mention two other databases that contain the same type of data, and explain how the EMBL database is related to them. What does it mean that these databases are primary databases?

B (4p) The EMBL database is also related to the UniProtKB database. Explain what UniProtKB is, and how data flows from the EMBL database to UniProtKB. The UniProtKB database is composed of two parts – explain what they are and how the relate to each other.

C (4p) An *entry* or *record* in databases such as EMBL and UniProtKB consist of many *fields*. One of these fields contain the *accession number* – what is it? Mention a tool that can be used to search in databases such as EMBL and UniProtKB, and how you can search specifically in fields.

Question 2 – Pairwise alignments (total 18p)

Two protein sequences *q*:LDIR og *d*:IRR and the following scoring matrix (excerpt from PAM250) are given:

	R	D	I	L
R	6	-1	-2	-3
D		4	-2	-4
I			5	2
L				6

Two different matrices *H1* and *H2* for finding the best global alignments by dynamic programming, are shown below, partially completed:

H1

		d	I	R	R
	q	0	-2	-4	-6
	L	-2	2	0	-2
	D	-4	0	1	-1
	I	-6	1		
	R	-8	-1		

H2

		d	I	R	R
	q	0	-3	-6	-9
	L	-3	2	-1	-4
	D	-6	-1	1	-2
	I	-9	-1		
	R	-12	-4		

A (2p) In addition to the scoring matrix we need to know the gap penalty to be able to complete the matrices *H1* and *H2*. What kind of gap penalty is used here, and what is the cost *g* for each gap?

B (6p) Fill in the missing values in the matrices *H1* and *H2*. What are the scores *S1* and *S2* for the best alignments?

C (6p) Find the best possible alignment(s) and explain briefly the procedure. Illustrate by drawing one or more paths through the matrix.

D (4p) Explain briefly the procedure for local alignment. Why are local alignments more frequently used than global alignments for searches in protein sequence databases?

Question 3 – Sequence based searches in databases (total 15p)

A (3p) Define the concepts *homolog*, *ortholog*, and *paralog*.

B (2p) Smith-Waterman is an algorithm that can be used to perform sequence searches. Why is this method usually not used for this purpose?

C (4p) Mention a heuristic method that is more commonly used for sequence searches, and explain briefly principle for and the main steps in this method. What is the weakness of this method compared to Smith-Waterman?

D (3p) What does the E-value tell you about the result of a sequence search?

E (3p) When looking for distant relatives of a protein it is better to use the BLOSUM45 scoring matrix than BLOSUM62. Why?

Question 4 – Multiple sequence alignments (total 11p)

A (3p) Explain what progressive multiple sequence alignment is and how this method is used in Clustal.

B (2p) Iterative alignment can be used to improve a multiple sequence alignment. Explain briefly why.

C (3p) Several multiple sequence alignment programs can make use of information on protein structure (e.g. Clustal and 3Dcoffee). Explain briefly why one in certain situations can obtain alignments of higher quality when structural infromation is incorporated.

D (3p) When making a phylogenetic tree for a set of homologous protein sequences, it is better to use a multiple sequence alignment as basis rather than all possible pairwise alignments. Explain briefly why.

Question 5 – Protein domains and protein structure (total 15p)

A (4p) What is a conserved globular domain? What kind of properties are typical of such domains?

B (3p) Why is SMART a great tool for identifying known conserved globular domains in proteins?

C (4p) PDB (Protein Data Bank), CATH and SCOP are all databases containing information on protein structures. What kind of data is available in the three databases, and explain briefly how they relate to each other.

D (4p) The most used method for modelling the three dimensional structure of proteins is homology modelling. Describe briefly the different steps used in homology modelling.

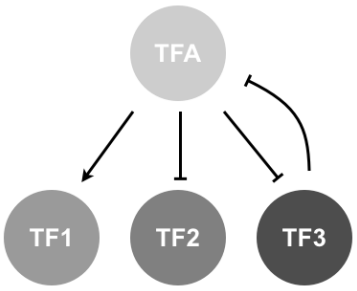
Question 6 – Systems biology (total 14p)

A (5p) Explain briefly how one in systems biology consider *components*, *interactions* and *reactions*, and *dynamic changes* when building models of regulatory and metabolic networks.

B (3p) Higher order biological systems are said to show emergent properties. Explain what a systems biologist mean by this concept. You may use an example to illustrate.

C (4p) Assume that you know the spesific binding sites for 50 transcritption factors in the yeast genome (*Saccharomyces cerevisiae*). You also have access to measurements of changes in gene expression resulting from mutations (loss of function) of each of the 50 transcription factors. Explain how this information can be used to make a gene regulatory network for yeast.

D (2p) A small part of the gene regulatory network is shown in the figure below.



TFA is a transcription factor that can activate the transcription factor TF1, and inhibit factors TF2 and TF3. In addition, TF3 is a TFA-inhibitor. What effect would a TF3 mutation have on the gene expression of TFA, TF1, and TF2?