

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i: MBV3070

Eksamensdag: 3. juni 2004

Tid for eksamen: 1430 - 1730

Oppgavesettet er på 5 sider oppgaver på norsk/4 pages with exercises in English

Vedlegg: ingen

Tillatte hjelpemidler: ingen

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

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

The English version of the exercises will be found behind the Norwegian version!

Oppgave 1: BLAST

Du bruker BLAST til å søke i en proteinsekvensdatabase. Søkesequensen inneholder ordet TAK. En av sekvensene i proteinsekvensdatabaseen inneholder delsekvensen ATLSGRWARQKN.

- a) Sett opp trebokstavordene i databasesekvensen som BLAST vil sammenligne med TAK.
- b) BLAST bruker substitusjonstabellen BLOSUM62 for å sette poeng på ordene i databasesekvensen. BLOSUM62-tabellen finnes på neste side. Hvilke av ordene i delsekvensen ovenfor vil BLAST analysere videre dersom bare ord med 4 poeng eller mer blir analysert?
- c) Du har sekvensert et stykke DNA og ønsker å finne ut om dette er en kodende sekvens og hva slags protein den eventuelt koder for. Hvilken variant av BLAST vil du benytte for analysen?
- d) Hvorfor vil PSI-BLAST kunne finne fjernere slektninger enn BLASTP kan av en søkesequens

BLOSUM62 Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Oppgave 2. Parvise sekvenssammenstillinger

b) I en sammenligningsmatrise for to sekvenser ser en (ikke nødvendigvis optimal) sti slik ut:

		G	A	R	M	A	R
H							
A							
R							
K							
E							
T							

Sett opp den tilsvarende parvise sekvenssammenstillingen.

b) Hva er affin gapstraff? Gi formelen for affin gapstraff. Hvorfor anser man ofte affin gapstraff for å være mer biologisk ”korrekt” enn konstant gapstraff?

Oppgave 3. Flersekvenssammenstillinger

- ClustalW er et mye brukt program for sammenstilling av flere sekvenser. Hvorfor?
- Gi en kort beskrivelse av de forskjellige trinnene ClustalW benytter ved sammenstilling av sekvenser, med spesiell vekt på hvordan programmet justerer gapstraffen i forskjellige områder av den voksende sammenstillingen.
- Her er en del av en flersekvenssammenstilling av 4 proteinsekvenser:

```

F A P K D L
F R P K D L
F A P K E L
F R G K E L

```

Ekstraher et mønster (pattern) fra sammenstillingen.

Oppgave 4

Kari og Per sitter i kantina og snakker om bioinformatikk. Du nyter en kopp kaffe og Dagbladet på bordet ved siden av og hører bare på dem med et halvt øre. Etter hvert blir utsagnene mer og mer

kontroversielle, så du begynner å følge bedre med.

Kari: ”Utrolig hvor lite veilederen vår kan av bioinformatikk altså. Han ville ikke greie seg en dag uten oss!”

Per: ”Ja bevars! Her om dagen måtte jeg til og med hjelpe ham med å bruke SRS til å få ut alle intronsekvenser på mer enn 500 basepar fra nukleotidsekvensdatabasen!”

Kari: ”Til meg kom han med en proteinsekvens og ante ikke hva han skulle gjøre for å finne ut noe om den. Gi den til meg! sa jeg, og på no time hadde jeg søkt i FASTA og BLAST og funnet haugevis av sekvenser som lignet.”

Per: ”Ja....”

Kari: ”Men noen av dem hadde innmari lav P-verdi, så dem kastet jeg.”

Per: ”Jeg synes det er lettere å se på E-verdiene jeg. Er bare de høye nok vet jeg at jeg er på sporet!”

Kari: ”Jo da, men egentlig burde du se både på E og P, og på Z-verdien også. Det er jo innmari betryggende å ha tre uavhengige mål på statistisk signifikans!”

Per: ”Ja, men det er jo bare E-verdien som er konstant. Det er veldig trygt, synes jeg, i denne stadig foranderlige verden, å ha et mål på likhet mellom to sekvenser som vil være det samme om to, om ti, om hundre år!”

Kari: ”Du skulle vært dikter du, Per.”

Per: ”Å Kari.....”

Kari: ”Men så ville han jo vite mer da. Om sekvensene altså.”

Per: (tar en slurk kaffe og ser spørrende på Kari)

Kari: ”Så jeg heiv dem inn i ClustalW og fikk ut det fine fylogenetiske treet!”

Per: ”Jeg trodde ClustalW lagde flersekvenssammenstillinger jeg?”

Kari: ”Joa, men trær også! Stilig.”

Per: ”Du Kari...”

Kari: ”Men så så han på treet da veit du, og så sa han at han ikke forsto noe og ikke kunne se opp og ned på det. Så etter hvert skjønnte jeg at han gjerne ville ha rot på det.”

Per: ”Ja, det er klart det...”

Kari: ”Så nå for litt siden tok jeg alle hundre sekvensene og foret inn i et program som lager trær basert på Maximum likelihood. Ti minutter etter hadde jeg et nytt tre.”

Per: ”Da ble han vel fornøyd da?”

Kari: ”Nei, for det var ikke rot på treet nå heller! Så derfor har jeg bestemt meg for å bootstrappe det, så kommer nok rota på plass!”

Per: ”Men du, til meg sa han at fylogenetiske trær er vel og bra, men at det egentlig var 3D-strukturen han var mest interessert i.”

Kari: ”Ikke lite frekk altså, lar oss gjøre alt arbeidet!”

Per: ”Så jeg sendte sekvensen hans inn i Uniprot og fikk ut en 3D-modell. Da smilte han fra øre til øre!”

Kari: Gir ikke Swiss-Model enda bedre modeller da?”

Per: ”Jo da, kanskje litt, men har du ikke latt merke til at Swiss-Model setter så gyselige farger på modellene sine da?”

Du orker ikke høre på skvalderet lenger, men reiser deg og sier: ”Dette er det verste jeg har hørt! I løpet av de siste to minuttene har dere sikkert sagt minst 10 gale ting!”

Kari og Per (med fornærmet mine): ”Hah! Hvilke gale ting da??”

Ja, hva er galt i utsagnene til Kari og Per? Gi også en kort forklaring på hvorfor du mener de forskjellige utsagnene var feilaktige.

Oppgave 5

- a) Hva er Pfam databasen og hva brukes den til?

De to viktigste bioinformatiske metodene for å predikere proteiners tre-dimensjonelle struktur er ”model-building by homology” og ”fold recognition / threading”

- b) Gi en kort beskrivelse av prinsippene bak disse to metodene.
- c) La oss si at du har tre homologe proteiner med kjent struktur og at du ønsker en flersekvenssammenstilling (multiple sequence alignment) av disse tre proteinene. En måte å gjøre dette på er å bruke et alignment-program som ClustalW. I dette tilfelle finnes det imidlertid en bedre metode; hvilken? (forklar)
- d) Hvis vi antar at et proteins struktur representerer den ene av alle mulige strukturer med lavest energi, da er det i prinsipp mulig å forutsi et proteins struktur ut fra sekvensen ved hjelp av energiberegninger. Det eneste man da i prinsipp må gjøre er å ta sekvensen og regne gjennom alle strukturer slik at man finner den med lavest (mest gunstig) energi. Beskriv hvorfor dette er svært vanskelig og pr. i dag urealistisk (dvs beskriv de største problemene/utfordringene med dette).

English version

Exercise 1: BLAST

You are making searches in a protein sequence database with BLAST. Your query sequence includes the word TAK. One of the sequences in the database includes the partial sequence ATLSGRWARQKN.

- List the three letter words in the database sequence that BLAST will compare with TAK.
- BLAST uses the substitution matrix BLOSUM62 to score the words in the database sequence. You will find the BLOSUM62 matrix further down on this page. What words in the partial sequence above will BLAST analyse further if only words with a score of 4 and above will be included in the analysis?
- You have sequenced a piece of DNA and wishes to know whether this is a coding sequence and if so, what kind of protein it codes for. What variant of BLAST will you use for your analysis?
- Why is PSI-BLAST able to find more distant relatives of a query sequence than BLASTP?

BLOSUM62 Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Exercise 2. Pairwise alignments

- a) In a comparison matrix for two sequences a (not necessarily optimal) path looks like this:

		G	A	R	M	A	R
H							
A							
R							
K							
E							
T							

Set up the corresponding pairwise alignment.

- b) What is affine gap penalty? Give the formula for affine gap penalty. Why is an affine gap penalty often considered to be more “biologically correct” than a constant gap penalty?

Exercise 3. Multiple alignments

- a) ClustalW is an often used program for constructing multiple alignments. Why?
- b) Give a short description of the various steps used by ClustalW while aligning sequences, with particular emphasis on how the program adjusts the gap penalty in various regions of the growing alignment.
- c) Here is a part of a multiple alignment of 4 protein sequences:

```

F A P K D L
F R P K D L
F A P K E L
F R G K E L

```

Extract a pattern from the alignment.

Exercise 4

Mary and John are talking about bioinformatics in the canteen. You are enjoying a cup of coffee and The Sun at the table next to them and only listen to the conversation with half an ear. As time passes, their statements are getting more and more controversial, so you start to pay more attention:

Mary: "Incredible how little our tutor knows about bioinformatics! He wouldn't manage one day without us!"

John: "Yes sure! The other day I even had to help him use SRS to get all intron sequences longer than 500 base pairs extracted from the nucleotide sequence database!"

Mary: "He came to me with a protein sequence without any idea about what to do with it. Give it to me! I said, and in no time I had made searches in FASTA and BLAST and found lots of similar sequences."

John: "Yes....."

Mary: "But some of them had an incredibly low P value, so those I threw away."

John: "I find it easier to look at the E values. As long as they are sufficiently high I know that I'm on the right track!"

Mary: "Yes sure, but you should actually look at both E and P, and even the Z value. I find it really reassuring to have three independent measures of statistical significance!"

John: "Yes, but it's only the E value that is a real constant. I find it comforting, in this everchanging world, to have a measure of similarity between two sequences that will be the same in two, in ten and in a hundred years!"

Mary: "Oh John, you should have been a poet."

John: "Oh Mary...."

Mary: "But then he wanted to know more, you know. About the sequences, that is."

John: (sips his coffee and sends an enquiring glance towards Mary)

Mary: "So I threw them into ClustalW and got out such a beautiful phylogenetic tree!"

John: "I believed that ClustalW was for making multiple alignments?"

Mary: "Yeah sure, but trees as well, Really cool."

John: "Listen Mary...."

Mary: "But then he looked at the tree, you know, and claimed that he didn't understand anything and could not tell what was up and what was down. So after a while I realized that he wanted a root on the tree."

John: "Yes of course....."

Mary: "So just now I took all hundred sequences and fed them into a program for making trees based on Maximum likelihood. Ten minutes later I had a new tree."

John: “So then he was pleased, I imagine?”

Mary: “No, because there still was no root on the tree! So now I’ve decided to bootstrap it, that will put the root in the right spot!”

John: “But you know, to me he said that phylogenetic trees were OK, but that the 3D structure was what he really was interested in.”

Mary: “Such a nerve, leaving us to do all the work!”

John: “So I sent his sequence into Uniprot and got out a 3D model. That really brought a smile to his face!”

Mary: “Will not Swiss-Model give even better models?”

John: “Sure, maybe a little, but haven’t you noticed the horrible colors Swiss-Model puts on its models?”

You are unable to listen to this nonsense any longer, but get up and say: “I’ve never heard such rubbish! During the last 10 minutes I’m sure you have made at least 10 wrong statements!”

Mary and John (with an insulted attitude): “Hah! What wrong statements?”

So what wrong statements did Mary and John make? Explain briefly why you consider them wrong.

Exercise 5

- a) The Pfam database: what does it contain and what can it be used for?

The two most important bioinformatical methods for prediction of the three-dimensional structure of proteins are ”model-building by homology” and ”fold recognition / threading”

- b) Describe briefly the basic principles behind these two methods
- c) Let’s say you have three homologous proteins whose three-dimensional structures are known and you wish to make a multiple sequence alignment of these. One way to do that is to use an alignment program such as ClustalW. In this case there is a better method for making the alignment; which method, and why is it better?
- d) If we assume that the structure of a protein represents the one conformation with the lowest possible energy, then it would in principle be possible to predict protein structure from sequence by energy calculations. The only thing one would need to do is to calculate the energies for each possible structure and select the structure with the lowest energy. Describe why this is very difficult and, at the moment, rather unrealistic (that is: describe the biggest problems and challenges).