

UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i / Exam in: MBV3070

Eksamensdag / Day of exam: 13. august 2004

Tid for eksamen / Exam hours: 14³⁰ - 17³⁰ (3 timer / hours)

Oppgavesettet er på 2 sider norsk oppgavetekst / 2 pages with exercises in English

Vedlegg / Appendices: Ingen / None

Tillatte hjelpemidler / Permitted materials: Ingen / None

*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 exam follows behind the Norwegian version!

Oppgave 1.

Vi har to sekvenser SATLTA og STVRLS. Med bruk av dynamisk programmering finner vi følgende globale sammenstilling mellom dem:

SAT--LTA

S-TVRLS-

- a) Tegn opp matrisen (tabellen) som er brukt ved den dynamiske programmeringen, og tegn inn hvordan pilene går mellom cellene i matrisen for at en skal få denne sammenstillingen ved "tilbakespolingen" (back propagation). Du skal selvsagt ikke fylle inn verdier i matrisen.

		S	A	T	L	T	A
S							
T							
V							
R							
L							
S							

- b) Skriv de generelle formlene for straff av gap med lengde k for affin og lineær gapstraff, og forklar hva de betyr. Diskuter forskjellen på lineær og affin gapstraff når det gjelder virkningen de har på hvor mange og hvor lange gap som blir innført ved sammenstilling av sekvenser.

Affine gap cost: $w(k) = h + gk$ or $w(k) = i + j(k-1)$ for $k = 1$ or higher, $w(k) = 0$ for $k = 0$ where $w(k)$ is gap cost, k is gap length, $h + g$ and i are gap opening costs and g and j gap extension costs

Linear gap cost: $w(k) = m \cdot k$ where m is the gap opening and gap extension cost.

Linear gap cost implies that a long, continuous gap will give the same penalty as several short gaps, whereas one from a biological point of view would like less penalty since fewer and longer insertions/deletions are more likely than more/shorter. Affine gap costs will give less penalty for fewer/shorter and can be seen as giving more biologically relevant alignments

Oppgave 2.

- a) Et av trinnene i FASTA-programmet er oppsetting av en "hash table" eller "lookup table" med utgangspunkt i søkesekvensen. Lag en slik tabell for delsekvensen ATLSGRWARQNK.

A	1, 8
C	-
D	-
E	-
F	-
G	5
H	-
I	-
K	12
L	3
M	-

N	11
P	
Q	10
R	6, 9
S	4
T	2
V	-
W	7
Y	-

- b) I et senere trinn lager FASTA en "offset vector" med utgangspunkt i "hash table" og en databasesekvens. Hvordan blir denne vektoren for databasesekvensen FARTAKMERALT dersom du bruker "hash table" fremstilt ovenfor?

-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	
					3			2	1		1	1	1	1	2					1	1	1	

- c) Med utgangspunkt i vektoren, hvilken foreløpige parvise sammenstilling av de to sekvensene vil FASTA analysere videre?

A T L S G R W A R Q K N - - - - -
 - - - - - F A R T A K M E R A L T

Oppgave 3

- a) For poengberegning for flersekvenssammenstillinger benyttes ofte "sum of pairs" (SP-poeng). Sett opp et uttrykk for beregning av poengverdien for kolonnen merket med pil i flersekvenssammenstillingen nedenfor.

```

M Q P I L L L
M L R - L L -
M K - I L L L
M P P V L I L

```



$SP(Q,L,K,P) = p(Q,L) + p(Q,K) + p(Q,P) + p(L,K) + p(L,P) + p(K,P)$
 where $p(a,b)$ is the score for the amino acid pair a and b from a normal substitution matrix (PAM, BLOSUM...)

- b) Hva er en profil? Skisser kort hvordan en profil kan konstrueres med utgangspunkt i en flersekvenssammenstilling.

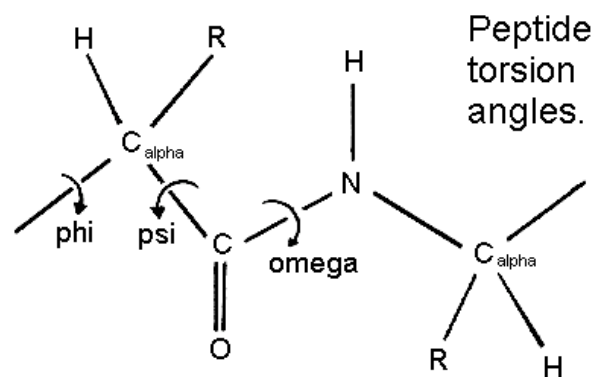
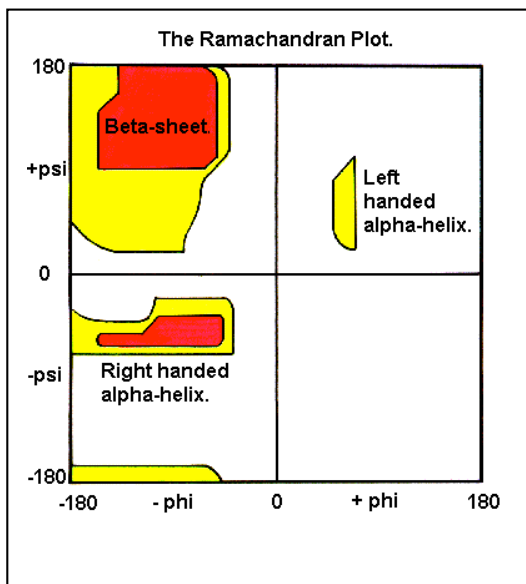
A profile is constructed from a multiple alignment by giving scores for all 20 amino acids for each position in the alignment, by taking into account a) how frequently the amino acid is found in the relevant position in the alignment and b) how likely it is to replace one of the other amino acids in the same position (from PAM, BLOSUM...). A profile can be regarded as a position specific substitution matrix and will allow detection of more distant relatives than most alternatives.

- c) Skisser hvordan programmet PSI-BLAST skiller seg fra andre BLAST-varianter ved søk i databaser og hvorfor dette programmet ofte vil finne fjerne slektninger som ikke kan påvises med andre søkeprogrammer.

BLAST works in an iterative way. In the first round, a normal BLAST search is performed, and hits with a relevant score are kept. A multiple alignment and a corresponding profile is constructed by the program on the fly, and a new search is made, using the profile rather than the original search string. New hits will be made, and the program will construct new alignments and new profiles that are again used for searching the database. The process is repeated as long as one feels safe or until no new hits are made.

Oppgave 4

- a) Hva viser et Ramachandran plott



No significant rotation around omega. No significant variation in bond angles and lengths. Thus, phi and psi are the main structural determinants. A Ramachandran plot shows combinations of phi and psi that are found in protein structures, and represents energetically favourable combinations of phi and psi. In such plots proline and glycine depart from the other amino acids.

psi are the main structural determinants. A Ramachandran plot shows combinations of phi and psi that are found in protein structures, and represents energetically favourable combinations of phi and psi. In such plots proline and glycine depart from the other amino acids.

- b) Hva gjenspeiler B-faktor når man fremviser proteinstruktur?

B-factor: says something about how well the position was defined by the electron density (-> an indicator of mobility and accuracy). B factors normally vary and will be high in some regions (e.g. loop regions on the surface), but high B factor values over longer regions may imply that the structure is not correct

- c) Hvordan kan man analysere likhet mellom to fastlagte proteinstrukturer?

The structures are classified according to content of secondary structure elements (SSE), architecture (relative arrangement of SSE), topology (how the SSEs are connected), fold (topologically defined arrangement of SSEs), and compared based on this. An important trick is to initially regard the structures as collections of ordered SSEs. Two overlapping

SSEs constitute a “unit of structural similarity”.

Programs:

VAST (SSE-based Vector-Alignment Search Tool); database: MMDB – Entrez

DALI (based på contact networks; Distance Matrix Alignment Program); database: FSSP (Families of structurally similar proteins). OBS! Commonly used

English version

Exercise 1

Consider the two sequences SATLTA and STVRLS. Using dynamic programming, we find a global alignment between the two as:

```
SAT--LTA
S-TVRLS-
```

- Write the matrix (table) that is used for the dynamic programming and indicate with arrows on the matrix how you can get this alignment with back propagation. You do not have to fill in the values in the matrix.
- Write the general formulae for cost of gaps of length k for affine and linear gap costs and explain what they mean. Discuss the difference between linear and affine gap costs with respect to the effects they have on how many and how long gaps are generated in sequence alignments.

Exercise 2

- One step in the FASTA program involves construction of a “hash table” or “lookup table” based on the query sequence. Make such a table for the partial sequence ATLSGRWARQNK.
- In a later step FASTA makes an “offset vector” based on the “hash table” and a database sequence. What will this vector look like for the database sequence FARTAKMERALT ?
- Based on the offset vector, FASTA will make a further analysis of one pairwise alignment of the query sequence and the database sequence. Which alignment?

Exercise 3

- For the scoring of multiple alignments, sum of pairs (SP-score) is often used. Give an expression for calculating the score for the column marked with an arrow in the multiple alignment below:

```
M Q P I L L L
M L R - L L -
M K - I L L L
M P P V L I L
```



- What is a profile? Describe briefly how a profile can be constructed starting with a multiple alignment.
- Describe how the program PSI-BLAST differs from other BLAST variants for database searches, and why this program will find distant relatives that cannot be identified using other search programs.

Exercise 4

- a) What does a Ramachandran plot show?
- b) What does the B factor reflect when protein structure is shown?
- c) How can similarity between two determined protein structures be analysed?