

Bioinformatica tentamen D1 voor 2MNW op woensdag 30 maart 2005 van 9.30-12.30 uur in zaal Q105

Naam:

Studentnummer:

NB: er zijn extra vellen achteraan bijgevoegd die je kunt gebruiken om antwoorden verder uit te werken, mocht je over een pagina heen gaan. Vermeld duidelijk welke vraag je beantwoordt op de extra vellen.

Het tentamen bestaat uit 11 meerkeuzevragen (deel A) en 8 open vragen (deel B)

A. Meerkeuzevragen:

1. In *sequence alignment* methoden wordt rekening gehouden met de volgende evolutionaire operaties geobserveerd in sequenties (meerdere antwoorden zijn mogelijk):
 - a. Mutaties
 - b. Insertions/Deletions
 - c. Repeats
 - d. Inversions
2. Een bioloog wil twee sequenties alignen, gebruik makend van de BLOSUM62 matrix en 11 als gap opening penalty en 1 als gap extension penalty. De bioloog vergist zich en geeft de gap penalties op als -11 en -1, zodat het alignment programma gap “bonuspunten” uitdeelt i.p.v. gap penalties op de gebruikelijke wijze toe te passen. De bioloog maakt een global alignment en een semi-global alignment van de twee sequenties. Het alignment dat de grootste kans heeft het hoogst te scoren is het:
 - a. Global alignment
 - b. Semi-global alignment
 - c. De twee alignments zullen beide altijd dezelfde score opleveren.
 - d. Beide programma's zullen geen score opleveren want ze gaan “crashen” omdat ze proberen oneindig lange gaps te maken.
3. Welke van de volgende 4 antwoorden is NIET juist: een *sequence alignment* wordt gemaakt om
 - a. een mogelijke homologie tussen de sequenties vast te stellen
 - b. ideeën over structureel en/of functioneel belangrijke aminozuren op te doen.
 - c. conserveringspatronen van aminozuren te zien.
 - d. te bewijzen dat er mogelijk geen homologie tussen de sequenties bestaat.
4. Bij het maken van een alignment tussen twee sequenties die homoloog zijn, en dus twee op elkaar lijkende tertiaire structuren hebben, is de corresponderende secundaire structuur die we het meest verwachten in posities van het alignment waar gaps voorkomen:
 - a. de β -strand.
 - b. de α -helix.
 - c. de loop structuur.
 - d. geen van de drie voorgaande structuren want er is geen verschil in het verwachte aantal gaps voor ieder van deze structuren.

5. Welke van de volgende 4 antwoorden is NIET juist: een *sequence alignment* wordt gemaakt om
- een mogelijke homologie tussen de sequenties vast te stellen
 - ideeën over structureel en/of functioneel belangrijke aminozuren op te doen.
 - conserveringspatronen van aminozuren te zien.
 - te bewijzen dat er mogelijk geen homologie tussen de sequenties bestaat.
6. Microarrays worden gebruikt om relatieve gen expressie te bepalen door middel van het meten van de relatieve concentratie van
- ionen in twee typen cellen
 - mRNA in twee typen cellen
 - metabole pathways in twee typen cellen
 - DNA moleculen in twee typen cellen
7. Twee stellingen:
- (a) De meeste prokaryote genen bestaan uit meerdere exonen
- (b) Exonen zijn meestal langer dan intronen
- Stelling (a) en (b) zijn juist.
 - Stelling (a) is juist, (b) is onjuist
 - Stelling (a) is onjuist, (b) is juist
 - Stelling (a) en (b) zijn beide onjuist
8. Het feit dat het DNA van *P. falciparum* (Malaria parasiet) voor 82% uit A en T bestaat, betekent voor het alignen van dit genoom met andere genomen die ook *A/T-rich* zijn dat
- de alignment score hoger zal zijn dan wanneer sequenties zonder een dergelijke scheve compositie aligned worden, maar dit betekent biologisch dan niet veel.
 - De alignment score hoger zal zijn dan wanneer sequenties zonder een dergelijke scheve compositie aligned worden, en dit is biologisch van belang.
 - de alignment score lager zal zijn dan wanneer sequenties zonder een dergelijke scheve compositie aligned worden, maar dit betekent biologisch dan niet veel.
 - de alignment score lager zal zijn dan wanneer sequenties zonder een dergelijke scheve compositie aligned worden, en dit is biologisch dan van belang.
9. Het *framework* model voor proteïen folding is
- een *two-step* proces dat verloopt via het vormen van secundaire structuur
 - een *single-step* proces dat verloopt via het vormen van secundaire structuur
 - een *two-step* proces dat **niet** verloopt via het vormen van secundaire structuur
 - een *single-step* proces dat **niet** verloopt via het vormen van secundaire structuur
10. De *furthest neighbour clustering* (complete linkage) methode wordt een *space dilating* cluster methode genoemd omdat
- de oorspronkelijke datapunten naar elkaar toe schuiven en dus steeds dichter bij elkaar komen te liggen
 - als afstand tussen clusters de grootste afstand geselecteerd wordt uit de afstanden tussen punten in de ene en punten in de andere cluster, waardoor de verwachte afstand tussen groeiende clusters toeneemt.
 - chaining* optreedt en hierdoor de dendrogrammen steeds compacter worden.
 - deze methode pas heel laat en met tegenzin clustert.

11. Voor twee objecten zijn de volgende eigenschappen gemeten:

	Eigenschap 1	Eigenschap 2	Eigenschap 3
Object A:	15	24	27
Object B:	9	27	33

Een wiskundige heeft als afstand tussen deze twee objecten 9 gemeten. Dit betekent dat als afstandmaat is genomen:

- a. City block distance
- b. Euclidean distance
- c. Single linkage clustering
- d. Complete linkage clustering

B. Open vragen:

B.1. *Reverse engineering* -- hieronder staat de DNA codon tabel:

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Gegeven de volgende eiwit sequentie in single-letter code:

SLRD

Opdracht: Reverse-engineer twee DNA sequenties die beide coderen voor deze eiwit sequentie, gebruikmakend van de bovenstaande codon tabel, zodanig dat de twee DNA sequenties **maximaal** verschillen. Bepaal de afstand tussen twee codons door niet-identieke posities te tellen. Geef de twee DNA sequenties en bereken de *Hamming distance* (het aantal niet-identieke posities in de twee DNA sequenties).

B.2. Neem de twee DNA sequenties van de vorige vraag:

a) Geef het biologisch juiste sequence alignment voor deze twee coderende DNA sequenties. Heb je hier een alignment programma voor nodig? Verklaar je antwoord.

b) Wat zou er kunnen gebeuren wanneer je maximaal verschillende *reverse engineered* DNA sequenties maakt en die “align” met bijv. de onderstaande *nucleotide exchange matrix* en gap penalty waarden 0 voor gap-opening en 1 voor gap extensie? Zal dit alignment altijd hetzelfde zijn als onder a)? Verklaar je antwoord.

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

c) Hoe zou je (mede op grond van antwoord a) en b)) gen sequenties die coderen voor hetzelfde eiwit het best kunnen alignen, gebruikmakend van de DNA sequenties of de corresponderende eiwit sequenties? Verklaar je antwoord.

/

B.3. Een moleculair biologe wil twee sequenties alignen en zeker stellen dat de alignment score biologisch significant is. Daartoe neemt de biologe de twee sequenties, aligned ze met een semi-globaal *Dynamic Programming* algoritme, gebruikmakend van de BLOSUM62 amino acid exchange matrix (zie matrix op volgende pagina) en gap penalties 1 (gap-opening) en 1 (gap-extension), en bepaalt zo de alignment score. Hierna doet zij hetzelfde nog eens met dezelfde matrix en zelfde gap penalties, maar nu met één van de sequenties omgedraaid (tegenrichting, d.w.z. van C- naar N-terminaal).

(Herinner je dat bij semi-globaal alignment geen end gap-penalties gegeven worden)

De twee sequenties die de biologe wil alignen zijn SNPAILV en ADTLL.

Opdracht: maak beide alignments en geef de alignment score (BLOSUM62, Gopen=1, Gextend=1). Gebruik de twee zoek matrices en de BLOSUM62 matrix op de volgende pagina:.

Sequenties in oorspronkelijke volgorde:

	S	N	P	A	I	L	V
A							
D							
T							
L							
L							

Alignment score:

Éen sequentie in tegenrichting:

	V	L	I	A	P	N	S
A							
D							
T							
L							
L							

Alignment score:

BLOSUM62 matrix

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

B.4. De biologe van de vorige vraag stelt dat wanneer de alignment score van het oorspronkelijke alignment duidelijk hoger is dan de score van het alignment met één van de sequenties in tegenrichting, dat dan de kans dat het alignment biologisch significant is duidelijk hoger is. Deze stelling van de bioloog is:

- a. onzinnig, omdat sequenties omdraaien niets met biologie te maken heeft.
- b. onjuist, omdat hij voor het test alignment beide sequenties had moeten omdraaien om een zinnige vergelijking tussen het oorspronkelijke en het test alignment te krijgen.
- c. juist, want een alignment met één van de sequenties in tegenrichting kan dienst doen als een zogenaamd *gerandomiseerd* alignment, waarbij beide sequenties nog steeds dezelfde aminozuurcompositie hebben, maar de volgorde van de sequenties voor het alignment programma nu praktisch gesproken *random* is.
- d. onjuist, omdat er wel een systematisch verschil is tussen het oorspronkelijke alignment en het *gerandomiseerde* alignment met één van de sequenties in tegenrichting, alleen is de score van het laatste alignment altijd hoger en niet lager zoals gesteld door de bioloog.

Welk antwoord is juist? Geef je reden.

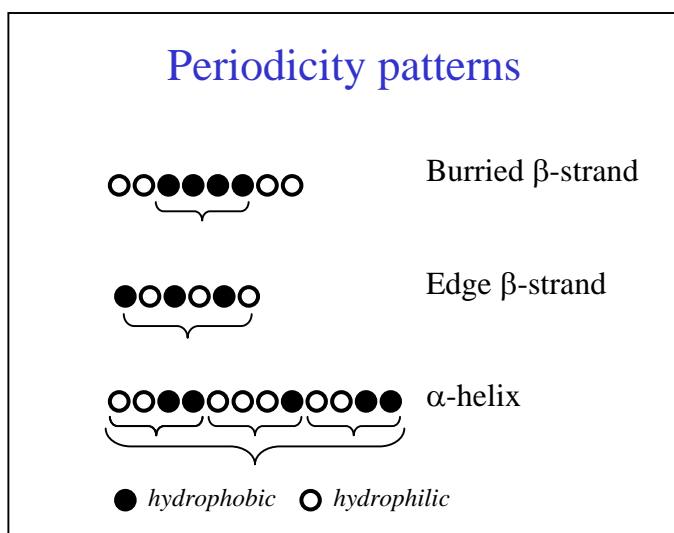
B.5. Een moleculair bioloog heeft een multiple alignment waarvoor hij graag de secundaire structuur zou willen bepalen. Onderstaand zijn twee segmenten van het multiple alignment: voorspel hiervoor de secundaire structuur (d.w.z. één secundaire structuur per alignment).

De fragmenten van het multiple alignment zijn:

DFKWRVCA	en	MVDRLIKEFYTS DNQ
DFRCTLRI		MIDRLLREFYTTDDQ
EYKCDL DL		MIERLLRDSYSTNDQ
EIKLKI KF		VIDKILRDSFGSNNN
DLKLEIDY		PAAKI IDDAFGSDEE

Neem als hydrofiele aminozuren: D, E, G, H, K, N, Q, R, S, en T.
 Neem als hydrofobe aminozuren: A, C, F, I, L, M, P, V, W, en Y.

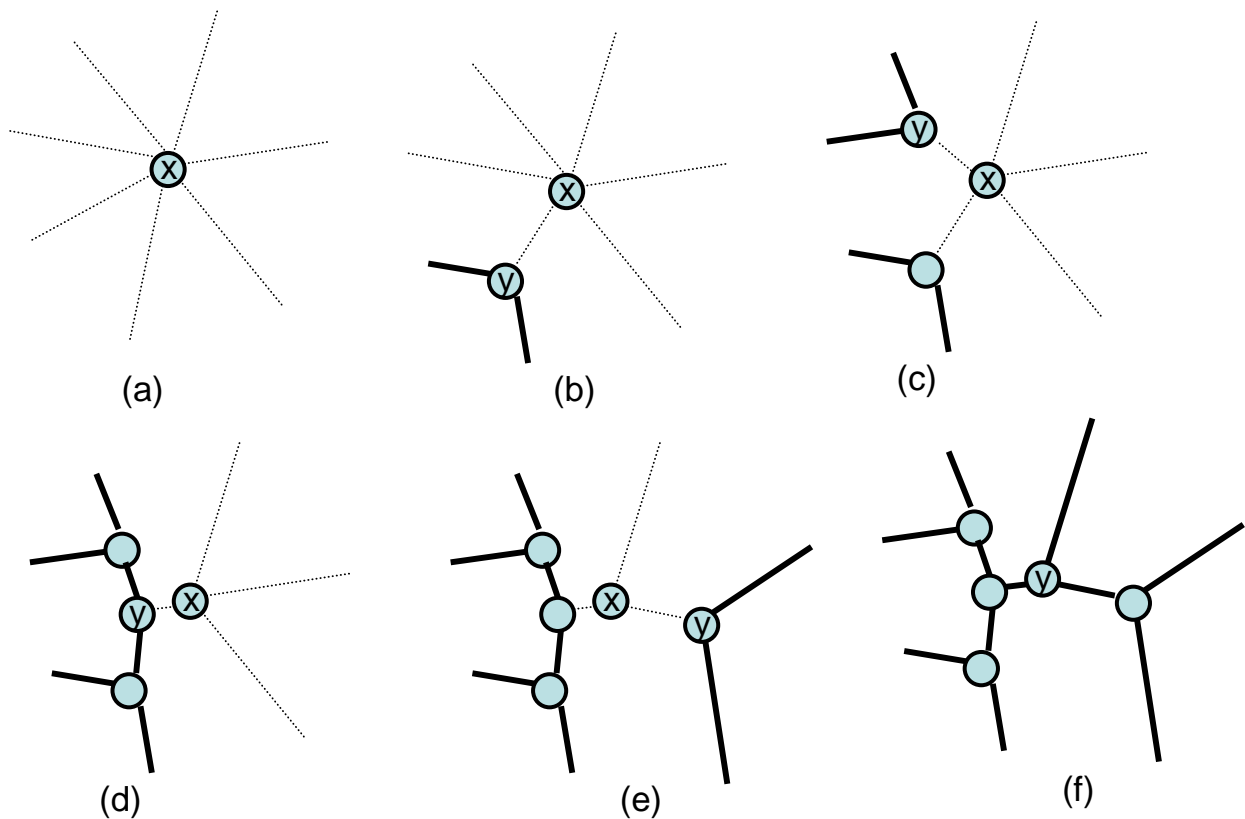
Hieronder staan de *periodicity patterns* zoals verwacht voor de α -helix en twee soorten β -strand:



Vraag: voorspel de secundaire structuur van de twee bovenstaande stukken multiple alignment (één secundaire structuur symbool per alignment kolom) met gebruikmaking van de *periodicity patterns*. Gebruik de letters 'H' voor helix, 'B' voor burried β -strand, 'E' voor edge β -strand, en 'C' voor coil.

Hint: Let op de conserveringspatronen in de multiple alignments. Wat voor aminozuren staan in idere alignment kolom?

B.6. Wat is speciaal aan de *Neighbor Joining Method* (NJ method -- Saitou en Nei, 1987) als hierarchische clustermethode? Hieronder staat een schematisch voorbeeld van een NJ clustering protocol als hint (dit is een slide van college 4). De dataset bestaat uit 7 objecten. Het schema laat zien dat er steeds gebruik wordt gemaakt van een extra *virtual node* x en een *internal node* y . Figuur (a) laat zien dat initieel een ster-vormige boom wordt gemaakt (via de extra *virtual node* x) van de 7 objecten. Hierna worden in (b) t.m. (f) stapsgewijs steeds twee groepen bij elkaar genomen door een *internal node* y te maken (neighbor joining), zodanig dat de totale boomlengte (d.w.z. alle taklengten opgeteld) minimaal blijft.



Het NJ cluster protocol wordt wel een **globaal** algoritme genoemd terwijl de andere cluster methoden (single linkage, UPGMA, complete linkage, ...) allemaal **lokale** hierarchische methoden zijn.

Vraag:

(a) Verklaar waarom de NJ methode **globaal** genoemd wordt, op grond van wat je weet over de andere clustermethoden hierboven.

(b) Aangenomen dat de objecten in de NJ boom eiwitsequenties zijn, wat is dan de biologische betekenis van het zo klein mogelijk proberen te houden van de totale boomlengte (zie hierboven)?

Hint: Het antwoord op (b) heeft met evolutie te maken.

B.6 (vervolg)

B.7. Alignment methoden voor moleculaire sequenties worden ook gebruikt om sequenties die uit verschillende laboratoria komen met elkaar te vergelijken. Stel je voor dat een zelfde stuk DNA is opgehelderd in twee verschillende laboratoria. Dit stuk DNA codeert voor een eiwit.

Bij het controleren van de twee DNA sequenties blijkt dat er één foutje in het sequencen van de DNA sequentie in één van de laboratoria is geslopen, want er wordt een verschil op één alignment positie aangetroffen.

Bij het alignen van de eiwitsequenties die gecodeerd worden door ieder van de DNA sequenties blijkt dat de eerste 11 aminozuren in beide sequenties hetzelfde zijn, maar dan worden de aminozuren opeens compleet verschillend.

Vraag: Wat is hier misgegaan? Raad de fout in de DNA sequentie (en op welke positie) en verklaar waarom dit leidt tot het bovenstaande verschil in de eiwitsequenties.

B.8. *Multiple sequence alignment (MSA)*: de volgende twee MSA kolommen worden met elkaar vergeleken op de manier van average linkage:

A		A
A		P
V	en	P
V		A
L		P

a) Maak de profile voor ieder van deze twee alignment kolommen (ga ervan uit dat ieder van de twee kolommen een positie uit een multiple alignment van 5 sequenties is). Gap penalties kunnen achterwege blijven.

b) Bereken de score voor het matchen van deze twee alignment posities, gebruik makend van average linkage (de gemiddelde gewogen score) en de twee profile kolommen gemaakt in a). **NB:** De benodigde *residue exchange matrix* voor de aminozuren die voorkomen in de alignment kolommen is als volgt (waarden volgens PAM250):

A	2			
L	-2	6		
P	1	-3	6	
V	0	2	-1	4
	A	L	P	V

c) Bepaal de score voor het matchen van de twee profile posities wanneer gebruik gemaakt wordt van de minimale afstand (hoogste *residue exchange matrix* score) tussen de twee kolommen.

Extra vel 1

Extra vel 2

Extra vel 3