Inleiding Bioinformatica 2 29 maart 2012

Naam:
Studentennummer:

Het tentamen bestaat uit:
- deel A met 20 meerkeuzevragen (2 punten per vraag)
- deel B met 4 open vragen (10 punten per vraag).

In totaal kunnen er dus 80 punten verdiend worden, 40 voor deel A en 40 voor deel B.

N.B. er zijn extra vellen achteraan bijgevoegd om de antwoorden verder uit te werken. Vermeld duidelijk welke vraag je beantwoordt.

Deel A

---

1) Welke van de onderstaande procedures komt **niet** voor bij *NGS* methodes, maar wel bij *Sanger sequencing*

a) Het willekeurig opknippen van DNA in kleine fragmenten
b) Het sorteren op lengte van de fragmenten om de sequentie te lezen
c) Het verlengen van het DNA met fluorescerende nucleotiden
d) Het toevoegen van een primer om DNA replicatie mogelijk te maken

2) Problemen bij de het *assembly* proces van NGS data kunnen gaten tussen de *contigs* creëren; zulke problemen kunnen **niet** worden veroorzaakt door:

a) te weinig *coverage* van reads
b) fouten die optreden tijdens het sequencen
c) *repeats* in het genoom
d) *introns* en *exons*

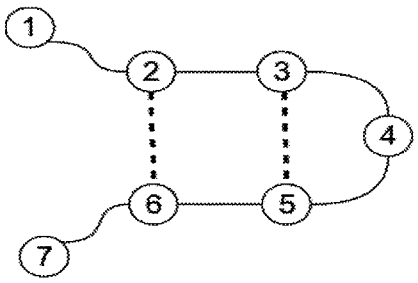
3) Gegeven de formule voor *Hydrogen bond Distance* (HD):

$$HD = \frac{1}{H} \sum_i^N \sum_j^N H_{ij} * L_{ij}$$

Waarbij:
$H_{ij}=1$ als er een waterstofbrug (*Hydrogen bond*) tussen residu i en j is
$H_{ij}=0$ als er geen waterstofbrug is.
L_{ij} de lengte aangeeft van de sequentie tussen i en j.
Bijvoorbeeld: voor residu i=3 en j=10, $L_{ij}=7$.
N is het totale aantal residuen in de structuur, H is het totaal aantal waterstofbruggen

Wat is de HD van de onderstaande structuur (waterstofbruggen zijn aangegeven door stippellijntjes)?

a) 0
b) 2
c) 3
d) 6



4) Gegeven dat HD (zie vorige vraag) gecorreleerd is aan de tijd die nodig is om te vouwen. Welke van de volgende structuren zal het snelste vouwen?

a) een alpha-helix
b) een parallelle beta strand
c) een anti-parallelle beta strand
d) ze vouwen allemaal even snel

5) Welke van de volgende uitspraken over *RNA-seq* (RNA sequencing met NGS) is juist

a) in *RNA-seq* wordt via het *cRNA* de hoeveelheid van een transcript bepaald
b) aan de hand van *cDNA* complementair aan *mRNA* kun je het *transcriptome* bepalen
c) via het *rRNA* kan de hoeveelheid van een eiwit geschat worden
d) Hybridisatie van *cDNA* wordt gebruikt voor het bepalen van de hoeveelheid transcript bij zowel *RNA-seq* als *Micro-arrays*

6) De *Rosetta Stone* is een methode om eiwit-interacties te verifiëren of te voorspellen. Tussen welke van onderstaande eiwitten (A, B, C) zou je op basis van deze methode een interactie voorspellen? Gegeven de lijst van volgende genomen, waarbij een 'A----A' aangeeft dat eiwit A op die plaats in het genoom gecodeerd wordt.

Genoom 1 A---AB-----B C-----CA---A
Genoom 2 A---A C------CA---A
Genoom 3 A---AB-----B B-----B

a) tussen A en B
b) tussen A en C
c) tussen B en C
d) tussen A en B, en tussen A en C

7) Het algoritme van Galey-Shapley om een *Stable Marriage* te vinden geeft een *Male optimal - Female pessimal* oplossing. Bekijk de volgende twee stellingen:

(1) Male optimal betekent dat alle "males" de partner van hun eerste keuze krijgen
(2) Female pessimal betekent dat alle "females" nooit de partner van eerste keuze krijgen

a) (1) is juist en (2) is onjuist
b) (1) is onjuist en (2) is juist
c) (1) en (2) zijn beiden juist
d) (1) en (2) zijn beiden onjuist

8) Telomeren zijn belangrijk voor de celdeling in de mens omdat ze:
a) de celdeling bevorderen
b) zowel de groeisnelheid als de celdeling van de cel bevorderen
c) de cel informatie verstrekken over hoeveel maal de cel gedeeld is
d) de cel informatie verstrekken over telomerase concentratie

9) Voor *Phylogentic footprinting* voor het voorspellen van regulatoire elementen is de volgende input nodig

a) een set van eiwit-interacties voor regulatoire elementen
b) een set van homologe eiwit-sequenties
c) een set van *biding sites* op het DNA
d) een set van homologe DNA sequenties

10) Gene prediction: het voorspellen van open reading frames (ORFs) in eukaryote genomen wordt bemoeilijkt doordat

a) eukaryote genen *splicing* vertonen
b) eukaryote RNAs *capping* ondergaan
c) eukaryote genen een hoge gen dichtheid hebben
d) eukaryote mRNAs transposons bevatten

11) Gen duplicatie en preferential attachment worden gezien als mechanismen voor de evolutie van protein-protein interaction networks. Deze twee mechanismen geven aanleiding tot het ontstaan van:

a) scale free networks
b) bipartite networks
c) random networks
d) directed networks

12) Het algoritme van Galey-Shapley om een *Stable Marriage* te vinden, kan gebruikt worden om twee groepen objecten paarsgewijs aan elkaar te linken. Het algoritme kan alleen gebruikt worden wanneer:

a) de groepsgrootten ongelijk zijn
b) er een volledige preferentielijst is voor beiden groepen objecten
c) de objecten uit de vrouwelijke groep NIET *female pessimal* zijn
d) de objecten uit de mannelijk groep *male optimal* zijn

13) *Phylogenetic profiling* is een techniek die functionele verbanden tussen genen kan voorspellen. Dit gebeurt op grond van:

a) het vinden van homologe eiwitproducten waar gene-fusion is opgetreden
b) het vergelijken van phylogetische bomen op grond van eiwit-interacties
c) gezamenlijke aan- of afwezigheid van orthologe genen in genomen
d) conserveringspatronen in een alignment gemaakt de gen sequenties

14) Een verschil tussen S*econdary Structure Assignment* (SSA) en *Secondary Structure Prediction* (SSP) is:

a) bij een SSA methode bepaal je de secondaire structuur per residu en bij een SSP methode niet
b) een SSA methode krijgt een PDB file als input, terwijl een SSP methode getraind wordt op PDB files
c) er is geen verschil tussen SSA en SSP methodes
d) bij een SSA methode houd je rekening met de patronen van tussen verschillende residuen en bij een SSP methode niet

15) Wat is de N50 waarde, als de contigs in je *assembly* de volgende lengtes hebben: [1,1,1,2,3,5,5,15,30]

a) 2
b) 3
c) 5
d) 15

16) Micro-RNAs (miRNAs) beinvloeden eiwit-expressie door

a) te binden tussen de ribosomale grote en kleine subunit, en daarmee eiwit translatie te verhinderen.
b) te hybridiseren met mRNA, hetgeen leidt tot onderdrukking van eiwit translatie.
c) te hybridiseren met tRNA, om zodoende de binding van het tRNA met het bijbehorende type aminozuur te verhinderen, hetgeen eiwit translatie onderbreekt.
d) te binden in kleine eiwit-bindingsplaatsen als een inhibitor, en daarmee eiwit-activiteit te verhinderen.

17) *De novo* voorspelling van RNA secundaire structuur kan gedaan worden m.b.v. dynamisch programmeren. Met dit algoritme kan de optimale RNA structuur berekend worden door

a) het maximale aantal baseparen in de mogelijke structuur te bepalen
b) een alignment met een homologe RNA sequentie te maken en het aantal identieke nucleotiden te tellen
c) een alignment met een niet-homologe RNA sequentie te maken en het aantal uitgelijnde nucleotiden te bepalen
d) een alignment met het coderende DNA te maken, om zodoende de intron-exon structuur te bepalen

18) Welke van de volgende beweringen over *metagenomics* is juist:

a) Om verschillende species te vergelijken wordt het rRNA gesequenced
b) De *depth of coverage* voor een metagenomics sample is belangrijk voor *genome assembly*
c) De aanwezigheid van verschillende genomen zorgen ervoor dat de reads ongelijke lengtes hebben
d) De *GC content* van een genoom is nodig om een *assembly* van de reads te maken en het toe te wijzen aan het juiste referentie genoom

19) Als een algoritme een *time complexity* van O(N) heeft, dan:

a) is er een lineair verband is tussen de hoeveelheid input en de hoeveelheid output van het programma
b) heeft de rekentijd een lineair verband met de hoeveelheid input
c) heeft de rekentijd een lineair verband met de hoeveelheid output
d) is geen van bovenstaande beweringen juist.

20) Welke van de volgende uitspraken of T*andem Affinity Purification* (TAP) en *Yeast two-Hybrid* (Y2H) is **onjuist**

a) TAP gebruikt de transcriptie van een reporter gen om een eiwit-eiwit interactie aan te geven
b) in TAP meerder purificaties rondes worden gebruikt om false positives te voorkomen
c) Y2H kan false positives genereren doordat genen in de cel niet op hetzelfde moment 'aan' staan
d) TAP kan false positives genereren doordat genen in de cel niet op dezelfde locatie aanwezig zijn

DEEL B

B1 - NGS and de Bruijn graphs

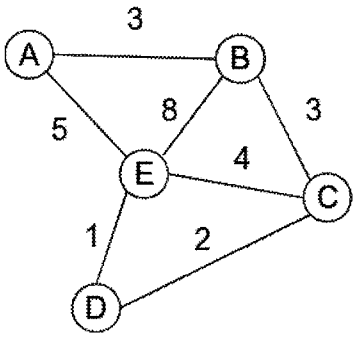
a) Maak een "de Bruijn" graaf van de volgende reads, met 3-mer als nodes en 4-mers als edges. Geef aan bij elke node en edge hoe vaak die in je reads aanwezig is.[6]

1 TGGTA
2 ACTGG
3 GTACT

b) Welke sequentie heeft je genoom? [1]

c) Stel dat read 3 **niet** aanwezig is, wat gebeurt er dan met je netwerk? Leg uit waarom je depth of coverage belangrijk is.[3]

B2 - Prim's & Dijstra's algorithm



a) Geef de kortste afstand aan tussen node A en D in het netwerk hierboven. Beschrijf duidelijk langs welke *nodes* het pad loopt. Geef de stappen aan die je genomen hebt om de kortste afstand te vinden.
[3]

b) Vind een *Minimum Spanning Tree* met Prim's algoritme voor hetzelfde netwerk [4]

c) Geven dat the *nodes* metabolen voorstellen en de *edges* een reactie gekatalyseerd door een specifiek enzym. In andere woorden, een *edge* tussen A en B betekent dat de stof A kan worden omgezet in stof B in een enkele katalytische stap.
Wat is de biologische betekenis van een kortste pad in zo'n netwerk? Wat is de biologische betekenis van een *Minimum Spanning Tree* [3].

B3 - Structure prediction and structure comparison

a) Een collega heeft biologie gestudeerd en is niet bekend met eiwitstructuur voorspelling. Zij heeft een eiwitsequentie en vraagt je om advies hoe ze de structuur kan voorspellen. Schrijf je advies op over een strategie die ze kan volgen en geef een indicatie hoe betrouwbaar de voorspelling zal zijn. Maak bij je antwoord gebruik van de volgende termen:

BLAST, sequence alignment, homology modelling, PDB, fold recognition, ab initio modelling, structure refinement, model evaluation, BLAST e-value, template [7].

b) Er is een ander eiwit, waarvan ze de structuur wel kent, en die ze graag wil opsplitsen in verschillende structurele domeinen. Welke strategie voor het opsplitsen in domeinen zou je haar adviseren, in het geval dat ze geen nieuwe experimenten wil doen, maar bestaande technieken uit de Bioinformatica wil gebruiken? [3]

B4 - Phylogenetic Profiling

a) Hieronder zie een Phylogenetic Profile van 4 eiwitten:

	genoom →
Eiwit A	0111001111
Eiwit B	0111001111
Eiwit C	0111001010
Eiwit D	0000001010

Een 1 geeft aan dat een eiwit op een genoom voorkomt, en een 0 geeft aan dat een eiwit afwezig is. In dit voorbeeld mag je ervan uitgaan dat je een interactie kan voorspellen als 60% van de absenties / presenties identiek zijn. Bijvoorbeeld: 50% van de presenties en absenties tussen eiwit A en D komen overeen.

a) Geef voor alle mogelijke eiwitparen het percentage identieke absenties/presenties aan. [3]

b) Teken een graaf van het netwerk dat ontstaat door de interacties die je nu kan voorspellen.[3]

c) Geef per eiwit aan wat de *degree* van het eiwit in het interactie netwerk is.[2]

d) Beschrijf een voorbeeld van evolutionaire gebeurtenissen die het gelijktijdig aanwezig en afwezig zijn van interacterende eiwitten op een genomen in de hand werkt [2]