Code : BFVP17BIN1	**Tentamen:** Hertentamen Bioinformatica 1	
Datum: 28-06-2018	**Tijd:** 8:30 – 10:00	**School:** SILS
Lokaal: TN/A112 & TN/A115	**Klas:** BFV1	**Duur:** 1 1/2 uur
Docent : HEMI **Tijdens het tentamen te bereiken onder nummer:** **050- 5954071**		**Aantal pagina's:** 2
Hulpmiddelen: Geen	**Overig hulpmiddelen:** Kladpapier.	
Opgave inleveren: Ja **Kladpapier inleveren: Ja**		
Bijzonderheden: **Eindcijfer = 1 + 9*(behaalde punten / totaalpunten) + evt. Bonuspunten opdrachten uit kwartaal 3**		
Naam student: **Klas:**		**Studentnummer:**

Beantwoordt de vragen zo volledig mogelijk, laat je kennis zien, dan kan ik je ook bij een incompleet antwoord zoveel mogelijk punten geven! Antwoorden mogen in het Engels of Nederlands.

1) Leg de begrippen "homologie", "global pairwise alignment", "orthologie" en "paralogie" uit en hoe aan elkaar gerelateerd zijn. (8)

Homologie betekent dat twee sequenties (eiwit of gen) afstammen van een gemeenschappelijke voorouder. Je kunt de kans hierop schatten door een global pairwiese alignment: het algoritme is zo opgezet dat een hoge score een hoge kans op homologie betekent. Orthologie is homologie tussen sequenties uit verschillende soorten; paralogie is homologie tussen sequenties binnen een soort (genoom).

2) BLAST helpt met het ontdekken van conservering en homologie van een query (dna of eiwit) met een hele grote database.
   a. Waarom zijn aminozuur-sequenties beter geconserveerd dan nucleotide-sequenties?
   Omdat de aminozuur-sequenties de functie van een eiwit mogelijk maken. En ook omdat meerdere nucleotidesequenties dezelfde aminozuursequentie kunnen coderen vanwege de verschillende codon-mogelijkheden.
   b. Welke BLAST-variant moet je gebruiken als je een gen (dna) op wilt zoeken in een database met eiwitten?
   Hiervoor gebruik je blastx.
   c. Waarom zou je zo'n vertaalde query willen doen? (2)
   Omdat je vermoedt dat een DNA sequentie een gen bevat; en dat kun je dan laten zien door te zoeken in een eiwitdatabase.
   d. Waarom heb je voor een nucleotide alignment geen BLOSUM of PAM matrix nodig? (2)
   Zo'n scorematrix is bedoeld om mogelijke substituties wat betreft biochemische functie tussen aminozuren te beoordelen; voor nucleotides wordt gewoon 1 score voor match en 1 voor mismatch gehanteerd.

3) Het BLAST algoritme staat alignment tussen een query (vraag) sequentie en een enorme database van andere sequenties toe.
   a) Wat is de "wordsize" van een BLAST query? (2)
   Dit is de grootte van de woordjes waarin je query wordt opgesplitst om te zoeken naar matches in de geindexeerde database.
   b) Als ik de "wordsize" van 11 naar 22 vergroot, wat voor effect heeft dat op de tijd die mijn BLAST query neemt? Welk ander effect heeft deze parameter vergroten? (4)
   c) *De BLAST query zal veel sneller gaan; maar het zal ook de kans vergroten op vals negatieven, je mist misschien homologe sequenties die net niet het exacte woord uit je query bevatten.*
   d) Welke andere parameter beïnvloedt de snelheid van het zoeken in de database met een word? (2)
   De andere meest effectieve parameter is de word lookup threshold; de alignment score waarboven je een hit met een woord in de database verder uitbouwt naar een heel alignment.

4) Multiple choice: omcirkel de letter van het juiste antwoord.
   Een Position Specific Score Matrix (PSSM) heeft 20 kolommen (voor elk aminozuur 1) en heeft een aantal rijen corresponderend aan de lengte van het query eiwit. Waarop zijn de individuele scores in de matrix gebaseerd? (4)
   a. Een PAM of BLOSUM matrix.
   b. De frequentie van voorkomen in een Multiple Sequence Alignment.

c. De score van de aminozuren links, linksboven en boven in de matrix.
d. De achtergrondfrequentie van het voorkomen van dat aminozuur in alle eiwitten.
Antwoord B

5) BLAST is een heuristiek algoritme; de resultaten van een query zijn niet gegarandeerd juist maar er hangt een kans aan dat de gevonden alignment klopt; de E-value. De formule is;

$$E = Kmne^{(-\lambda S)}$$

a) Kun je de waarden *K* en λ (lambda) aanpassen per query? Waarom wel, of niet? *Nee, die staan vast voor een bepaalde database waarin je zoekt.*
b) Als je de database van sequenties waartegen de query gedaan wordt verdubbelt, hoe verandert de E-value dan bij gelijke alignment score S?
Dan verdubbelt de E-value ook. (Wordt dus slechter ; grotere kans op vals positieven)
c) De E-value formule is een zogenaamde *negatieve exponentiële* functie. Wat betekent dit voor de E-value van een hele goede (hoge) alignment score's *S*?
Dit betekent dat voor heel hoge alignment scores de E-value heel laag is; met andere woorden, de kans dat het een toevalstreffer was wordt heel klein.

6) BLOSUM62 is gekozen als de default scorematrix voor BLAST. Waarom werkt een BLOSUM matrix beter bij het beoordelen van homologie in een alignment dan een PAM matrix? (5)

Omdat de BLOSUM matrix is samengesteld uit multiple sequence alignments van families verwante eiwtitten (blokken) en omdat deze nog steeds bijgehouden en uitgebreid wordt met nieuwe eiwitsequenties.

7) Multiple choice: omcirkel de letter van het juiste antwoord.
Scorematrices zijn er in verschillende mate van strengheid; bijvoorbeeld PAM10, BLOSUM45, BLOSUM90. Hoe kies je een matrix met een bepaald getal uit voor een pairwise alignment?(4)
   a. Het getal geeft de snelheid aan waarmee de matrix een score levert.
   b. Het getal geeft de mate van overeenkomst tussen de sequenties.
   c. Het getal geeft de penalty van een gap opening, naast de score per aminozuur.
   d. Het getal geeft de gemiddelde score van een aminozuur-match in de matrix.
   Antwoord B.

8) Er zijn een aantal voorgekookte databases met sequenties beschikbaar om tegen te BLASTen. Twee daarvan zijn de *nr* database en de *RefSeq* database (voor zowel nucleotide als eiwitsequenties). Leg uit wat het verschil tussen deze twee type databases is en wanneer je voor de ene of de ander zou kiezen. (5)
De nr database staat voor "non-redundant" en bevat alle bekende sequenties van dat type, zolang er geen exacte duplicaten tussen zitten. De RefSeq databases bevatten door onderzoekers doorgelichtte en goedgekeurde sequenties van vooral modelorganismen. Je gebruikt bij voorkeur altijd als eerste een RefSeq database, maar als je niet met de daar aanwezige organismen kunt werken of je kunt niet vinden wat je zoek, switch je naar de NR.

9) De active site is vaak het best geconserveerde deel van een eiwit.

   a. Leg uit waarom een PSSM gemaakt met PSI-BLAST beter is in het vinden van de

active site van een query in andere eiwitten in de *nr* database dan normale BLASTP. (3)

De PSSM (position specific scoring matrix) is juist gemaakt door PSI-BLAST om die active site (die tijdens je PSI-BLAST iteraties geconserveerd blijkt) goed te herkennen; sequenties die erop lijken krijgen hoge scores toebedeeld in een nieuw alignment. De "algemene" score matrices zoals BLOSUM en PAM waar BLASTP mee werkt nemen alle eiwitten op een hoop mee zonder te specialiseren in een bepaalde functie en bijbehorende sequenties zoals van de active site.

b. Als je eenmaal een goede PSSM hebt gemaakt, kan die gesaved worden en opgeslagen in een database, zodat je hem weer kan gebruiken. Hoe heet de database van NCBI met eerder gemaakt PSSM's? (2)

Dit is de PFAM database.

c. Welke BLAST gebruik je als je deze PSSM's wilt hergebruiken? (2)

RPS-BLAST

10) Multiple choice: omcirkel de letter van het juiste antwoord.
Welke output van een BLAST search geeft een schatting van het aantal vals positieven die de search in zijn alignments heeft? (4)

a. Bit score
b. Percentage identity
c. E-value
d. Alignment score S

Antwoord C.