**Samenvatting**

Het retrovirus HIV(Human Immunodeficiency Virus) bestaat in twee vormen: HIV1en HIV2. Beide vormen kunnen AIDS(Acquired Immune Deficiency Syndrome) veroorzaken. Virussen gerelateerd aan HIV zijn in veel niet-humane primate gevonden en worden SIV(Simian Immunodeficiency Virus) genoemd. **Achtergrond info** In dit project is er onderzocht hoeveel het HIV-1 virus verschilt van het SIV en HIV-2 virus. Dit is onder andere gedaan door te kijken naar de kenmerken van de oppervlakteproteïnen en of deze kenmerken verschillen tussen de 4 virussequenties. **Context** Om te kunnen bepalen waarom HIV1 zo gevaarlijk is worden de oppervlakteproteïnen van verschillende HIV-typen vergeleken. Het doel was dan ook om de verschillen tussen de virussequenties op basis van de oppervlakteproteïnen te bepalen. **Doel** Van de HIV1, HIV2, SIV en SIVmnd-2 virussen zijn proteïnesequenties in FASTA format gedownload, afkomstig van NCBI. Er is een python script geschreven dat voor elke sequentie 100 random sequenties genereert, in totaal 404 sequenties. De 404 sequenties worden vanuit een .txt bestand ingeladen in de multiple alignment tool: ClusterOmega. De resultaten zijn een multiple sequence alignment en een Percent Identity Matrix. De identity percentages worden in een nieuw python script geplot, om de resultaten gestructureerd weer te geven. **Data.** Uit de gegevens van de pairwaise alignment is gebleken dat HIV2 en SIV het meest overeenkomen,47.23% en de alignment van HIV1 en SIV2 het minst, 32.69%. **Resultaten** De resultaten die uit de pairwaire alignment kwamen waren niet terug te zien in het plot van de random sequenties, waaruit de conclusie getrokken kon worden dat het geen toeval is dat de 4 proteïnesequenties sterk overeenkomen. **Implicatie** Voor dit onderzoek is de multiple alignment tool ClustelOmega gebruikt. Er is geen controle uitgevoerd met een andere multiple alitnment tool. B**eperkingen** Een vervolgonderzoek zou kunnen zijn: ‘Waar zitten de genetische verschillen tussen de HIV1,HIV2,SIV en SIVmnd-2 virussen?’. **Vervolgonderzoek**

**Aantal woorden: 290**

**Materiaal&Methode**

**Pairwise alignment**

Om vast te kunnen stellen wat de verschillen zijn tussen de oppervlakteproteïnen van HIV1, HIV2, SIV en SIVmnd2 worden de oppervlakteproteïnen paarsgewijs met elkaar vergeleken. Dit wordt gedaan door de tool Align (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). De alignments worden gebruikt om de overeenkomst van twee sequenties op sequentieniveau te bepalen. Per alignment is het aantal identity’s, positives en gaps genoteerd.

**Verificatie**

Om er zeker van te zijn dat eventuele overeenkomsten tussen de proteïnesequenties niet toevallig zijn is er de random verwachte overeenkomst bepaald. Dit is gedaan d.m.v. een python script dat de proteïnesequenties van HIV1, HIV2, SIV en SIVmnd-2 ‘toevallig’ verandert. Verondersteld wordt dat het aantal identieke aminozuren bij alignments van de random proteïnesequenties normaal verdeeld is. Voor elk virus worden er 99 random sequenties gegenereerd daar komt de originele sequentie bij, in totaal genereert dit script 400 sequenties.

**Multiple alinment**

Met de tool ClustalO (<http://www.ebi.ac.uk/Tools/msa/clustalo/>) zijn paarsgewijs de random sequenties met elkaar maar ook met de originele sequenties vergeleken. ClustalO accepteert als input alleen een tekst file met daarin álle sequenties, met steeds een witregel tussen de sequenties, in FASTA format. Het resultaat is een multiple aligment en een file met de ‘Percent Identity Matrix’.

**Percent identity**

De Percent Identities uit de Matrix worden door een tweede Python script in groepen verdeeld. Per groep wordt het aantal keren dat een percentage past in deze groep geteld , deze aantallen worden uitgezet in een histogram. Verder wordt het gemiddelde aantal identities en de standaarddeviatie berekent. Hoe groter het aantal vergeleken sequenties en het aantal groepen, hoe duidelijker de normaalverdeling te zien is. Het gemiddelde en de standaarddeviatie worden berekent om te testen of de aantallen identieke aminozuren bij de alignments met de originele sequenties significant verschillen met de aminozuren bij de random gegenereerde alignments. Het aantal identieke aminozuren tussen de twee originiele sequenties zal buiten de normaalverdeling van de overeenkomst tussen de random sequenties liggen.

**Resultaten**

Om te bepalen waarom HIV1 zo’n gevaarlijk virus is zijn de oppervlakteproteïnen van verschillende HIV-typen vergeleken. Er is gekeken naar de kenmerken van de oppervlakteproteïnen van HIV1,HIV2,SIV en SIVmnd-2 om zo vast te kunnen stellen wat de verschillen zijn en hier eventueel een conclusie uit te kunnen trekken. Dit is gedaan d.m.v. alignments, er is hierbij gebruik gemaakt van een pairwise alignment zie materiaal en methode) om de 4 originele sequenties met elkaar te kunnen vergelijken. Deze resultaten worden in figuur 1 weergegeven.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | HIV1 | HIV2 | SIV | SIVmnd-2 |
| HIV1 | 100% |  |  |  |
| HIV2 | 34.96% | 100% |  |  |
| SIV | 36.69% | 47.23% | 100% |  |
| SIVmnd-2 | 32.69% | 36.33% | 36.40% | 100% |

Figuur 1: Resultaten van de pairwise alignment tool: Align Proteïne Blast. In deze tabel zijn de overeenkomsten tussen de genoemde virussen op sequentie niveau te zien. Wat opvalt is SIV en HIV2 met de hoogste identity van 47.23% en HIV1 en SIVmnd-2 met het laagste identity van 32.69%

Verder is er gebruik gemaakt van een multiple alignment tool(zie materiaal en methode), om 400 random gegenereerde sequenties en 4 originele sequenties te kunnen alignen. De resultaten hiervan zijn in figuur 2 weergegeven.

hoiiuh


Figuur 2: Resultaat van de multiple alignment tool: ClustalOmega, geplot met een Pythonscript. Het aantal sequenties op de y-as uitgeschreven tegenover het percentage identities op de x-as. Wat opvalt is dat vrijwel alle identities tussen de 0 en 10% liggen. De originele sequenties hadden identies tussen de 32.60% en 47.23%, deze worden echter niet weergegeven omdat deze identities maar een enkele keer voorkomen.

**Bronnen**