Virus-Host DB

Als bioinformaticus kom je vaak bestanden tegen met redundante en inconsistente informatie. Ook is het afleiden van informatie uit platte tekstbestanden een hele opgave. Veel biologen zijn dagen bezig om uit een brei van gegevens de juiste informatie te halen.

Op http://www.genome.jp/virushostdb/ staan veel gegevens over bekende virussen. Alle data zijn te vinden in platte tekstbestanden die niet geoptimaliseerd zijn voor het doorzoeken.

De opdracht die je krijgt is om een interface te bouwen waarmee je als epidemioloog, viroloog, geneticus, bioloog of bioinformaticus snel de juiste informatie weet te halen. Veel belangrijker dan een mooie interface is het gebruik van de juiste datastructuren en algoritmes van je applicatie.

"Bad programmers worry about the code. Good programmers worry about data structures and their relationships."
– Linus Torvalds

Informatie over het bestand

Het bestand dat je ontvangt is vrij beschikbaar en wordt iedere dag aangepast met de nieuwste gegevens over virussen. Op ftp://ftp.genome.jp/pub/db/virushostdb/ staat het bestand waar we mee gaan werken. De opbouw ervan is beschreven in de readme :

```
================================================================================
This directory contains following files.

virushostdb.tsv: Tab separated file containing the following information:
                 virus tax id   ... tax ID of a virus
                 virus name     ... name of a virus
                 virus lineage  ... lineage of a virus
                 refseq id      ... RefSeq IDs of a virus
                 KEGG GENOME    ... Dblink to KEGG GENOME
                 KEGG DISEASE   ... Dblink to KEGG DISEASE
                 DISEASE        ... disease name
                 host tax id    ... tax ID of a host
                 host name      ... name of a host
                 host lineage   ... lineage of a host
                 pmid           ... PubMed ID
                 evidence       ... source of the host information
```

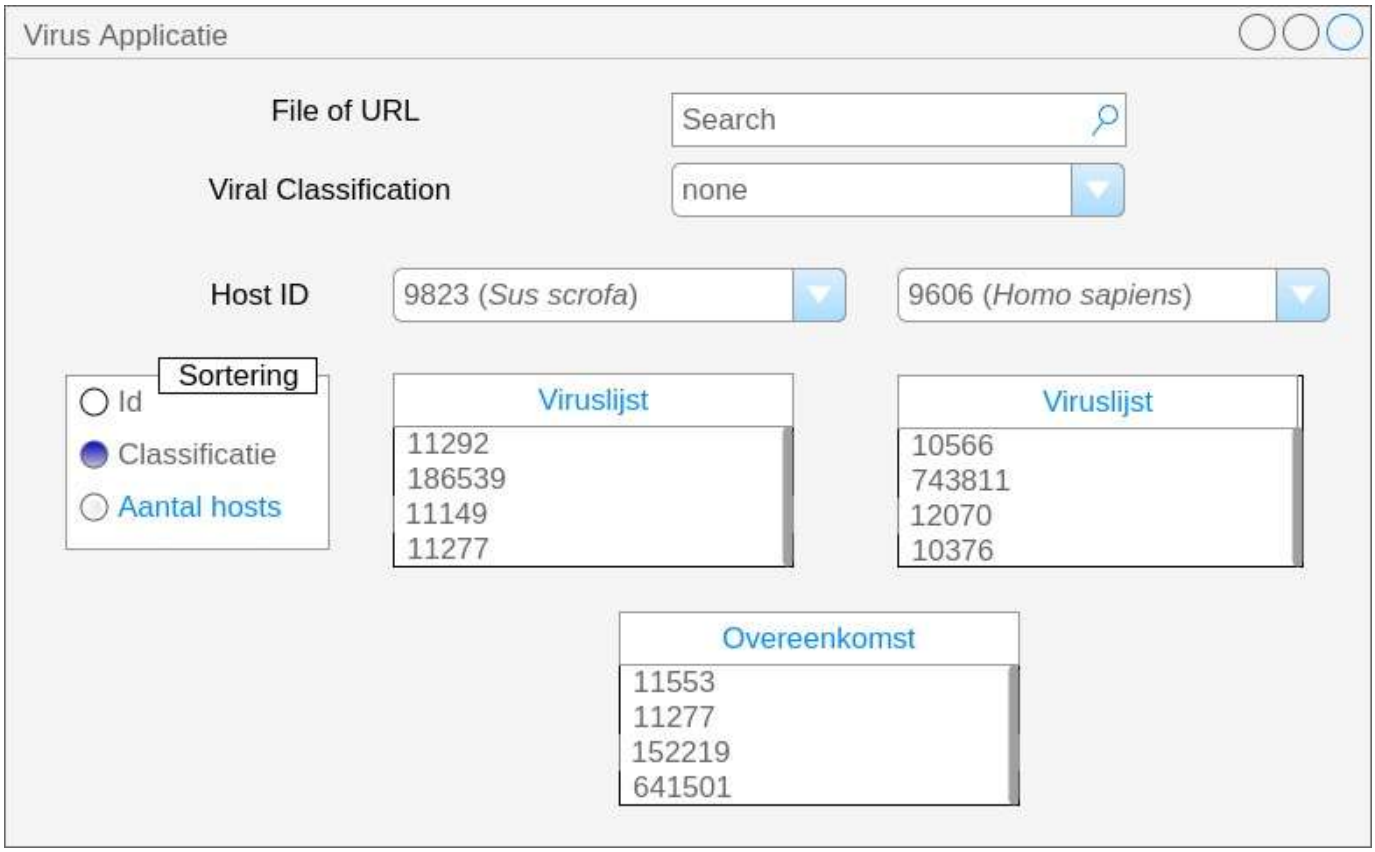
Het bestand bevat veel redundante gegevens. Zo zie je al in regel 1 en 2 en dat van een virus twee keer exact dezelfde naam, virus lineage et cetera is vastgelegd. Dat is zo opgenomen omdat er twee hosts zijn voor hetzelfde virus.

virus tax i	virus name	virus lineage	refseq id	KEGG GEN	KEGG DISE	DISEASE	host tax ic	host name	host linea	pmid	evidence
438782	Abaca bunchy top virus	Viruses; ssDNA virus	NC_010314, NC_010315, NC_010316, NC_0				46838	Musa sp.	Eukaryota	17978886	Literature, RefSeq
438782	Abaca bunchy top virus	Viruses; ssDNA virus	NC_010314, NC_010315, NC_010316, NC_0				214697	Musa acuminata A	Eukaryota	17978886	Literature
1241371	Abalone herpesvirus Victoria/A	Viruses; dsDNA virus	NC_018874				6451	Haliotidae	Eukaryota; Metazoa;		UniProt
1241371	Abalone herpesvirus Victoria/A	Viruses; dsDNA virus	NC_018874				36100	Haliotis rubra	Eukaryota; Metazoa;		RefSeq
491893	Abalone shriveling syndrome-a	Viruses; dsDNA virus	NC_011646				37770	Haliotis diversicol	Eukaryota; Metazoa;		RefSeq
11788	Abelson murine leukemia virus	Viruses; Retro-transc	NC_001499				10090	Mus musculus	Eukaryota; Metazoa;		UniProt
2025595	Abisko virus	Viruses; unclassified	NC_035470				201501	Epirrita autumnat	Eukaryota; Metazoa;		RefSeq
665102	Abutilon Brazil virus	Viruses; ssDNA virus	NC_014138, NC_014139				3630	Abutilon	Eukaryota	20349251	Literature, RefSeq
665102	Abutilon Brazil virus	Viruses; ssDNA virus	NC_014138, NC_014139				4100	Nicotiana benthan	Eukaryota	20349251	Literature
665102	Abutilon Brazil virus	Viruses; ssDNA virus	NC_014138, NC_014139				145753	Malva parviflora	Eukaryota	20349251	Literature

Figuur 1: Fragment van `virushostdb.tsv`, het bestand is tab delimited en kent een grote mate van redundante opslag. Het is virus_tax_id gecentreerd opgezet.

Opdracht

Het bestand `virushostdb.tsv` is opgebouwd rondom de eerste kolom met de virus identifier. Het is daarmee geschikt voor het ophalen van informatie als je het virus id hebt. Het is totaal ongeschikt als je informatie zoekt met de host als uitgangspunt. De centrale opdracht is om een applicatie te bouwen die de informatie vanuit host perspectief kan weergeven.



Figuur 2: Voorbeeld GUI voor de applicatie. De applicatie mag er anders uitzien maar moet dezelfde functionaliteit bevatten. (eigen werk auteur)

Toelichting GUI

i. File of URL bevat een verwijzing naar de file of URL waar het bestand met virusinformatie kan worden opgehaald door het programma.
ii. Viral classification is een dropdown menu met de mogelijkheid te filteren op de classificatie van virussen zoals beschreven op http://www.genome.jp/virushostdb/
iii. De host ID dropdown menu's bevatten alle unieke Host ID's met de benaming van de hosts.
iv. De viruslijsten zijn lijsten met scrollbars die alle virus ID's tonen van virussen die in staat zijn tot besmetting van de geselecteerde host in het dropdown menu.
v. Sortering is een radiogroup die de sortering bepaalt waarop de viruslijsten worden gesorteerd
vi. Overeenkomst is een lijst met virussen die zowel de linker als de rechter host kunnen besmetten.

Technische eisen

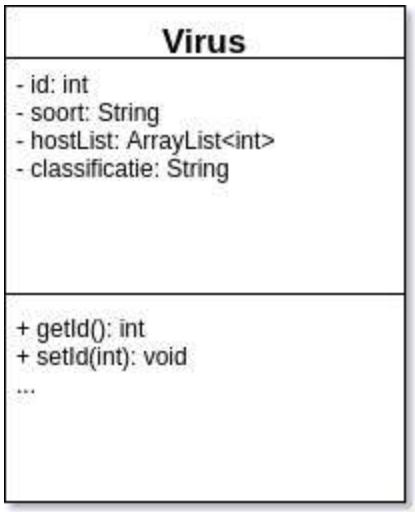
1. De applicatie is in **Java** geschreven.
2. Alle beschikbare libraries in Java zijn toegestaan mits de applicatielogica van de datastructuren zelf geprogrammeerd is.
3. **GUI logica moet apart worden geïmplementeerd van de applicatie logica.**
4. Je maakt ten minste gebruik van de classes **VirusGUI, Virus en VirusLogica** (bevat de logica voor sorteren en de datastructuren). De classes dienen ook deze naam te hebben.
5. De verzamelingen bestaan uit **Virus objecten** zoals in concept is weergegeven in Figuur 3.
6. **Big O** voor de vergelijking van virussen is zo klein mogelijk.
7. Virussen kunnen middels het implementeren van de **Comparable interface** gesorteerd worden op een van de gegevens.
8. Alle **exceptions** worden afgevangen door adequate exception handlers
9. De applicatie dient goed gedocumenteerd (**javadoc**) ingeleverd te worden als een NetBeans (of Maven) project op **GitHub**.

Functionele eisen

10. De applicatie bevat een invoerveld waar een filename of URL kan worden opgegeven.
11. In de applicatie is een keuzemenu (dropdown) waar de gebruiker een keuze maakt voor een virus classification (None, dsDNA, ssDNA et cetera)
12. Na selectie van de virus classification maakt de gebruiker een keuze voor twee hosts. Bijvoorbeeld Sus scrofa en Homo sapiens.
13. De viruslijsten eronder tonen daarna alle virus identifiers die bekend zijn voor de gegeven hosts.
14. Middels bijvoorbeeld een radiogroep kan een keuze gemaakt worden voor sortering van de virusidentifiers.
15. Van de twee hosts word weergegeven welke virussen zij gemeenschappelijk hebben.

Would have

1. De applicatie is in staat om ook een URL te gebruiken die rechtstreeks de data van de FTP-server haalt.
2. Wanneer geklikt wordt op een virus id in de lijst toont de applicatie een pop-up met de informatie over het virus: id, gastheer, classificatie en verwijzingen naar Pubmed artikelen
3. Maak in plaats van een NetBeans project een Maven project aan.



Figuur 3: class definitie van een virus object. De attributen zijn private en de methodes public. De methodes zijn niet uitputtend beschreven.

Beoordelingsformulier Eindopdracht BI6a

Naam: …….. Klas: …… Datum: ……

Competentie 2.2 (1 t/m 12)		**Beoordeling door**	
Onderdeel	Score	Student	Docent
1. Algemeen			
a. Code is generiek geschreven (onder andere opdeling in meerdere classes (1 pt) en korte compacte functies (1 pt).	2		
b. Exception handling: adequaat en volledig.	2		
2. Verwerking data			
a. Verwerking van te lezen data: efficiënt lezen en schrijven van databestanden.	2		
b. Testplan van data aanwezig en controle is uitgevoerd.	2		
3. Toepassing datastructuur			
a. Gebruik van juiste geneste datastructuur.	2		
b. Toepassing van datastructuur is efficiënt en optimaal: juiste Big-O.	2		
4. Algoritme			
a. Goede analyse van biologisch probleem en juiste vertaling naar programma.	2		
b. Opzet en werking van gevraagde algoritme. (gebruik van gevraagde objecten (2 pt), complexiteit (2 pt))	4		
5. GUI			
a. Flexibiliteit gevraagde GUI elementen: GUI is flexibel en reageert op wijzigingen run time.	2		
Eindoordeel	**Max: 20**		
Feedback:			
Naam en paraaf beoordelaar:			

Inleveren

Geschreven code en een eigen beoordeling van je geschreven code aan de hand van bovenstaande beoordelingscriteria. Vul daartoe de kolom student in. N.B. Het eindoordeel van de docent is hetgeen geadministreerd wordt.

Het cijfer is het aantal behaalde punten delen door het maximaal aantal te behalen punten maal 10.

Gebruikte bronnen:

1. http://www.genome.jp/virushostdb/