Eindopdracht bij 'Advanced Programming and Grid Computing'

Vak: **Advanced Programming and Grid Computing**
Afkorting: **Bapgc**
Docenten: **Said Basmagi en Martijn Jansen**
Studiejaar: **2015 – 2016**
Datum: **16 november 2015**

Voor het behalen van het vak 'Advanced Programming and Grid Computing' (Bapgc) dien je deze eindopdracht tot een goed einde te volbrengen.

Deze eindopdracht moet je individueel uitvoeren.

Eindopdracht:

Inleiding:

Het genoom van onze soort *Homo sapiens* is volledig gesequentieerd en al veelvuldig gebruikt in diverse wetenschappelijke onderzoeken. Voor dit onderzoek moet er ook gebruik worden gemaakt van ons genoom.

Van ons humane genoom is al veel bekend, maar de wetenschap blijft niet stilstaan en ontdekt steeds meer over ons genoom. Zo is er naast het humane genoom ook meer bekend over zijn metabolische routes.

Metabolische routes zijn ketens van enzymen in een cel of een compartiment van een cel, waarbij het product van het vorige enzym dient als substraat voor het volgende enzym. Elk enzym is hierbij getranscribeerd en getransleerd uit één of meerdere genen. Al deze genen zijn dus verantwoordelijk voor die ene metabolische route.

De transcriptie van deze genen wordt aangestuurd door zogenaamde transcriptiefactoren, die zich meestal binden op en rond hun promotorsequenties. Een transcriptiefactor is een eiwit dat zich bindt aan het DNA met een bepaalde consensussequentie; ook wel een motif genoemd. Vaak kunnen er ook meerdere bindingen van één of meerdere transcriptiefactoren bij de promotorsequentie van een gen nodig zijn om de daadwerkelijke transcriptie in gang te zetten. Een vereiste hierbij is dat deze transcriptiefactoren in de juiste volgorde en op de juiste bindingsplaatsen moeten zijn gebonden. Het blijkt nu dat deze volgorde en posities van transcriptiefactoren overeenkomt bij bijna alle promotorsequenties van genen in dezelfde metabolische route. Dit wordt een framework genoemd.

Het blijkt dat genen die zich in de metabolische route bevinden vaak door dezelfde transcriptiefactor worden gereguleerd. Maar zo'n transcriptiefactor kan naast de genen die hij reguleert in de desbetreffende metabolische route ook andere genen reguleren.

Onderzoeksvraag:

Wij vragen ons af:

Zijn er genen die worden gereguleerd door dezelfde transcriptiefactor(en), maar die niet behoren tot de desbetreffende metabolische route? Zo ja, welke genen zijn dat dan? Verklaar waarom deze ook zouden kunnen worden gereguleerd door dezelfde transcriptiefactor(en) door een aantal biologische vragen te beantwoorden.

De opdracht:

Je krijgt 3 metabolische routes van het humane genoom. Uit 1 van die 3 wordt door jou 1 gekozen. Hiermee moet je de onderzoeksvraag beantwoorden door de volgende stappen te doorlopen.

Stap 1: Vinden van de regulerende transcriptiefactoren in je metabolische route
Je moet aan de hand van de motifs van transcriptiefactoren op zoek gaan naar transcriptiefactor-bindingsplaatsen (TFBS'en) op en rond promotorsequenties van de genen van het humane genoom in jóuw metabolische route. Op die manier bepaal je door welke transcriptiefactor(en) de genen in jouw metabolische route worden gereguleerd.

Stap 2: Vinden van door de transcriptiefactor(en) medegereguleerde genen buiten jouw metabolische route
Je moet met de door jou gevonden transcriptiefactor(en) gaan onderzoeken welke andere nieuwe genen door deze transcriptiefactoren worden gereguleerd buiten jouw metabolische route. De genen die worden gevonden moeten natuurlijk wel de beste genen zijn van alle gevonden genen. De manier waarop je bepaalt welke genen de beste zijn, dien je in je code en documentaties te omschrijven.

Stap 3: Meer biologische achtergrondinformatie vinden van door jouw gevonden medegereguleerde genen
Na stap 2 heb je nieuwe genen gevonden die ook worden gereguleerd door dezelfde transcriptiefactor(en) van jouw metabolische route. Het is belangrijk dat je over deze nieuwe genen meer informatie probeert te vinden door de volgende biologische vragen erover te kunnen beantwoorden.

De biologische vragen voor de gevonden medegereguleerde genen:

1.	Hoeveel van de 100 best gevonden medegereguleerde genen zijn enzymen? Hoeveel van deze enzymen hebben een bekende functie? Geef deze verhoudingen in een Venndiagram weer.
2.	Doe een multiple sequence alignment (MSA) van de 50 best gevonden medegereguleerde genen. Hoeveel procent van het langste gen is geconserveerd? Geef de naam van dit gen en het percentage van zijn geconserveerde regio ten opzichte van het gehele gen weer.
3.	Bepaal van de 10 best gevonden medegereguleerde genen hun eiwitten. Elk eiwit heeft een percentage aan hydrofiele, hydrofobe en neutrale aminozuren. Toon deze 10 eiwitten in staafdiagram voor hun aminozuurverhoudingen, waarbij elke staaf 1 afzonderlijk eiwit betreft.
4.	Vind van elk nieuw (maximaal 10) gevonden medegereguleerde gen de 4 meest verwante paraloge genen. Geef al deze genen vervolgens weer in een fylogenetische boom.
5.	Bepaal voor maximaal 20 gevonden medegereguleerde genen de cumulatieve intron- en exonlengte in aantal nucleotiden. Toon de lengtes van deze genen vervolgens in een scatterplot. Hierbij is de x-as de cumulatieve intronlengte en de y-as de cumulatieve exonlengte van elk gen.

De resultaten van de stappen en biologische vragen moeten uiteindelijk automatisch gegenereerd kunnen worden in een PDF-document.

Benodigdheden:

Voor deze stappen en vragen zul je alle motifs van de transcriptiefactoren en de promotorsequenties van het humane genoom nodig moeten hebben. Je moet zelf uitzoeken hoe je aan die motifs en promotorsequenties komt.

Waar je zeker gebruik van zult moeten maken, is Bioconductor. In Bioconductor zijn packages te vinden die je zullen helpen bij het oplossen van de stappen en vragen. Deze packages moet je zelf bestuderen. Vervolgens moet je zelf kiezen welke packages je zou kunnen gebruiken en hoe je ze dan moet gebruiken. Een andere manier is door gebruik te maken van andere packages in andere programmeertalen, zoals BioPython, of van reeds bestaande programma's. Net zoals bij de Bioconductor-packages, zul je zelf moeten uitzoeken wat je kunt en uiteindelijk wilt gebruiken.

Let op: De keuzes van packages en programma's die je wilt gaan gebruiken voor het maken deze eindopdracht moet je op tijd doorgeven aan Said en Martijn. Op het gridcluster moeten namelijk alle nodige packages en programma's nog geïnstalleerd worden. Martijn en Said zullen dan alles voor je installeren. Bij het inleveren van de scripts behoort er een losse README.txt file ingeleverd te worden. Deze file bevat alle niet standaard packages en programma's. De standaard packages en programma's kunnen gevonden worden in het document "Packages.txt" op Elo

Deze eindopdracht behoort opgelost te worden op een Linux systeem. Je mag zelf kiezen om of gebruik te maken van het gridcluster of niet. Het bouwen van de pipeline gebeurt volgens de regels van 'Organized scripting in building pipelines' (Organized_Scripting.pdf) van Martijn Jansen

Gebruik van het gridcluster:

Dit gridcluster bestaat 5 computers, waarvan 1 de manager is en de andere de 4 workers. Op elke computer staat het programma HTCondor geïnstalleerd. Als je daarvan gebruik wilt gaan maken, moet je daarvoor je eigen jobscripts schrijven en die vervolgens laten draaien op het gridcluster. Elk jobscript moet in Linux kunnen draaien en genereert zijn eigen resultaten. Deze resultaten moeten daarna (op je eigen computer) door middel van een enkelvoudige pipeline samengevoegd en verwerkt worden tot 1 geheel. Uiteindelijk wordt automatisch het PDF-document gegenereerd.

Geen gebruik maken van het gridcluster:

Het is ook toegestaan om geen gebruik te maken van het gridcluster.

Je pipeline moet bij het starten ervan de volgende vragen stellen:

Welke metabolische route van het humane genoom wil je onderzoeken?

Opleveringen:

1. Resultaten in een groot PDF-document:

Het uiteindelijke resultaat is een PDF-document, waarin de volgende resultaten verwerkt staan.

Het PDF-document moet opgebouwd zijn uit de volgende onderdelen:

Voorkant
- De gekozen metabolische route als titel
- De gekozen metabolische route als plaatje
- Auteurs (Correcte auteurs)
- Studentnummer
- Maand en jaar

Inleiding
- Aanleiding voor het maken van dit document
- Waar gaat het document over?
- Achtergrondinformatie over de gekozen metabolische route
- Onderzoeksvraag gesteld gezien vanuit de metabolische route

Materialen en methoden
- Uitleg waarop de transcriptiefactor(en) die de gekozen metabolische route reguleert/reguleren, wordt/worden bepaald.
- Uitleg waarop de best medegereguleerde genen worden bepaald.
- Het maximaal te vinden medegereguleerde genen buiten je metabolische route
- *Voor elke biologische vraag volgt uitleg wat er precies wordt onderzocht.*

Resultaten
- De transcriptiefactor(en) die de gekozen metabolische route reguleert/reguleren
- Alle gevonden medegereguleerde genen
- *Alle biologische vragen.*

Referenties

Correcte verwijzing van/naar figuren en tabellen

2. Alle scripts van je.

Deze moeten in de terminal van Linux draaien en worden aangeroepen. Op die manier zal er ook worden nagekeken.

3. Je documentaties in 1 groot PDF bestand.

Het PDF-document, al je scripts en je documentaties moeten in Nederlands geschreven zijn en dus niet in het Engels. Ze mogen verder niet meer dan 5 taalfouten per 500 woorden bevatten.

Cesuur:

De eindopdracht is behaald, als 75 % (245 punten) of meer van het totale aantal van 327 punten behaald is én aan alle voorwaardelijke eisen is voldaan.

Inlevermomenten:

1e kans: Zaterdag 30 januari 2016 vóór 23.59 uur

Puntenverdeling:

Hieronder lees je wat de voorwaardelijke eisen zijn en waarop je overal punten kan scoren.

	Voorwaardelijk:
	Ingeleverd via Ephorus: Alle scripts als 1 groot tekstbestand
	Ingeleverd via Ephorus: Alle documentatie van alle scripts en functies in 1 document
	Ingeleverd via Ephorus: Het gegenereerde PDF-document
	Ingeleverd via mail: Alle scripts als zip-bestand
	Ingeleverd via mail: De README.txt bestand
	Ingeleverd via mail: Het gegenereerde PDF-document
	De scripts werken naar behoren (Geen foutmelding tijdens runnen)
	Er wordt in ieder geval een PDF-document gegenereerd.
	De documentatie van alle scripts en functies mag niet meer dan 5 taalfouten per 500 woorden bevatten.
	Er is voor de scripts gebruik gemaakt van de Skeleton-scripts die op ELO staan.
	Documentaties van Hogeschool Leiden
	alle code van zijn/haar scripts dient volgens de regels van documentatie (van de eigen taal) geschreven te zijn.
	PDF-document
	Algemeen deel: (verplicht)
	Voorkant

	De gekozen metabolische route als titel
	De gekozen metabolische route als plaatje
	Auteurs (Correcte auteurs)
	Studentnummer
	Maand en Jaar
	Inleiding
	Achtergrondinformatie over de gekozen metabolische route
	Onderzoeksvraag gesteld vanuit de metabolische route
	Materialen en methoden
	Uitleg waarop de transcriptiefactor(en) die de gekozen metabolische route reguleert/reguleren, wordt/worden bepaald.
	Uitleg waarop de best medegereguleerde genen worden bepaald.
	Het maximaal te vinden medegereguleerde genen buiten de metabolische route
	Voor elke biologische vraag volgt uitleg, discussie en conclusie wat er precies wordt onderzocht.
	Uitleg
	discussie
	Conclussie
	Resultaten
	De transcriptiefactor(en) die de gekozen metabolische route reguleert/reguleren
	Alle gevonden medegereguleerde genen
	Referenties
	Correcte verwijzing van/naar figuren en tabellen
	biologische vraag over de gevonden medegereguleerde genen:
	Biologische vraag 1: Hoeveel van de 100 beste gevonden medegereguleerde genen zijn enzymen? Hoeveel van deze enzymen hebben een bekende functie? Geef deze

	verhoudingen in een Venndiagram weer.
	De 100 best gevonden genen worden bepaald.
	De eiwitten van de 100 best gevonden genen worden bepaald.
	De enzymen van de 100 worden bepaald
	De enzymen met een bekende functie worden bepaald.
	Er wordt een Venndiagram gemaakt
	met de juiste verhoudingen getoond
	met de juiste labels
	Biologische vraag 2: Doe een multiple sequence alignment (MSA) van de 50 best gevonden medegereguleerde genen. Hoeveel procent van het langste gen is geconserveerd? Geef de naam van dit gen en het percentage van zijn geconserveerde regio weer.
	De 50 best gevonden genen worden bepaald.
	Het langste gen wordt bepaald.
	De MSA is gedaan.
	Het percentage van de geconserveerde regio van het langste gen / een gen is bepaald.
	De naam van het langste gen / een gen wordt getoond.
	Het percentage van de geconserveerde regio van het langste gen / een gen wordt getoond.
	Biologische vraag 3: Bepaal van de 10 best gevonden medegereguleerde genen hun eiwitten. Elk eiwit heeft een percentage aan hydrofiele, hydrofobe en neutrale aminozuren. Toon deze 10 eiwitten in staafdiagram voor hun aminozuurverhoudingen, waarbij elke staaf 1 afzonderlijk eiwit betreft.
	De 10 best gevonden genen worden bepaald.
	De eiwitten van de 10 best gevonden genen worden bepaald.
	De hydrofiele aminozuren van de eiwitten worden bepaald.
	De hydrofobe aminozuren van de eiwitten worden bepaald.

	De neutrale aminozuren van de eiwitten worden bepaald.
	Een staafdiagram wordt gegenereerd.
	correcte verhouding
	Elk eiwit eigen staaf in de staafdiagram.
	Biologische vraag 4: Vind van elk nieuw (maximaal 10) gevonden medereguleerd gen de 4 meest verwante paraloge genen. Geef al deze genen vervolgens weer in een fylogenetische boom.
	De 10 best gevonden medegereguleerde genen worden bepaald.
	Van elk gen worden 4 paraloge genen gevonden.
	Er wordt een fylogenetische boom gegenereerd
	met daarin de 4 paraloge gevonden genen
	Biologische vraag 5: Bepaal voor maximaal 20 gevonden medegereguleerde genen de cumulatieve intron- en exonlengte in aantal nucleotiden. Toon de lengtes van deze genen vervolgens in een scatterplot. Hierbij is de x-as de cumulatieve intronlengte en de y-as de cumulatieve exonlengte.
	De 20 best gevonden medegereguleerde genen worden bepaald
	Voor elk gevonden medegereguleerd gen worden de juiste intronen en exonen bepaald.
	Voor elk gevonden medegereguleerd gen worden de juiste cumulatieve intron- en exonlengte bepaald.
	Een scatterplot wordt gegenereerd
	x-as: cumulatieve intronlengte
	y-as: cumulatieve exonlengte
	correcte positie van de dots