Samenvatting

Bioinformatica is een interdisciplinair onderzoeksveld waarbij methoden uit de computer wetenschappen, wiskunde en statistiek worden gebruikt met het specifieke doel betekenis te geven aan grote hoeveelheden biologisch experimentele gegevens. In de moleculaire genomica is bioinformatica onontbeerlijk; de grote hoeveelheden data die worden gegenereerd verdiepen pas ons inzicht juist door grondige computationele analyse.

Genen zijn de bouwstenen voor de machinekamer van de cel. In feite zijn genen regios in het DNA die kunnen worden overgeschreven naar boodschapper RNAs (mRNA) die vervolgens kunnen worden vertaald naar eiwitten. De eiwitten zijn de belangrijkste actoren in de cel. Sommigen RNA moleculen worden gecontrolleerd door kleine RNA-achtige strucuturen die microRNA worden genoemd. Deze, recentelijk ondekte, microRNAs (miRNA) zijn in feite hele kleine mRNAs en worden net als mRNA ook uit het DNA overgeschreven. Echter, in plaats van de normale vertaling naar eiwit binden deze korte RNA fragmenten aan mRNA en op deze manier kunnen ze het aflezen van het eiwit (het doel) verhinderen.

Het complete erfelijke materiaal van een organisme wordt ook wel het genoom genoemd. Teneinde het genoom te begrijpen moeten we alle functionele dele labelen. Dit proces van labelen wordt annotatie genoemd en de annotatie komt tot stand door bioinformatica. In dit proefschrift wordt annotatie gerealiseerd door middel van het gebruiken en integreren van verschillende, i.e. heterogene, bronnen. De nadruk ligt op het verkrijgen van annotaties voor genen en moleculaire structuren waarmee genen worden gecontrolleerd, de micro RNAs. Het gebruik van heterogene bronnen vergroot daarbij de mogelijkheden voor het redeneren over de data.

Het hoofddoel van dit proefschrift is efficiente en nauwkeurige annotatie van miRNAs en DNA sequenties waarvan tot nu toe geen functie bekend is. Gen-annotatie is het proces waarmee de structuur en functie uit ”ruwe” DNA sequenties wordt verkregen. In een genoom-project is dit het meest tijdrovende deel van de analyse. Wat betreft miRNA annotatie is, in het huidige onderzoek, de belangrijkste taak de doel-RNAs (target) te kunnen vaststellen van een miRNA. Dit omdat miRNA de expressie van een gen controlleert door het binden aan een specifiek doel-mRNA. Teneinde deze verschillende annotatie taken te kunnen realiseren zijn een verscheidene complexe werkschemas (workflows) opgesteld

waarin de op dit moment relevante bron data als ook de analyse technieken worden geintegreerd.

In hoofdstuk 2 wordt een integratie method behandeld waarmee een aantal aspecten worden onderzocht van de relaties tussen miRNAs en het doel-mRNA met de bedoeling om uit de poel van mogelijke doel-mRNAs juist die te selecteren die met een hoge mate van waarschijnlijkheid juist zijn. De technieken die hierbij zijn toegepast omvatten onder andere statistische testen, clustering en zogenaamde associatie regels. Het onderzoek dat in dit hoofdstuk wordt beschreven omvat ook een "case" studie voor twee specifieke miRNAs, te weten, dre-miR-10 en dre-miR-196. Voor deze twee miRNAs werden zeven candidaat mRNAs voorspeld met een hoge waarschijnlijkheids score; alle voorspelde kandidaten behoren tot de hox-gen familie. De voorspelde candidaten delen karakteristieken met genen die reeds gevalideerd zijn.

In hoofdstuk 3 wordt een strategie gepresenteerd voor het analyseren van relaties tussen miRNA's en daarbij wordt vervolgens aangegeven hoe deze relaties gebruikt kunnen worden in de voorspelling van doel-genen zoals die in de mens gevonden kunnen worden. Om dit te ondersteunen is een proces-koppeling ontwikkeld teneinde patronen van kenmerken die tussen bekende miRNAs bestaan, te onthullen. De patronen die gevonden zijn, zijn vervolgens toegepast op miRNAs waarvan de doel-genen nog niet bekend zijn om op die manier mogelijke voorspellingen te kunnen doen over doel-genen van deze miRNAs. Deze methode draagt bij aan de verbeteringen die noodzakelijk zijn voor het identificeren van de doel- genen waarbij de specificiteit vergroot is en de beperking die wordt opgelegd vanwege het principe van conservering van evolutie, niet behoefd te worden toegepast.

In hoofdstuk 4 hebben zijn de verschillende algoritmes die worden toegepast om doelgenen te voorspellen aan een evaluatie onderworpen; daarbij hebben we gebruik gemaakt van methodes van integratie om de nauwkeurigheid van de voorspelling te kunnen vergroten. Daarbij zijn zowel integratie methodes toegepast aan de bovenkant van het spectrum, i.e. combinaties van algoritmen en ranking aggregatie technieken, als ook methodes aan de onderkant van het spectrum, i.e. Bayesiaanse Netwerk classificaties. Al deze methodes zijn getest op miRNA-doel interacties die experimenteel gevalideerd zijn als ook op datasets die samengesteld zijn als negatieve controle data. Uit de resultaten komt naar voren hoe ieder van de voorspellings algoritmen zijn eigen voordelen heeft. Bovendien blijkt dat het gebruik van de Bayesiaanse Netwerk classificatie zoals toegepast op de ken-

merken die berekend zijn uit de verschillende voorspellings algoritmen een significante verbetering geven op de nauwkeurigheid van de voorspelling van het doel-gen.

In hoofdstuk 5 ligt de nadruk op het maken van een assemblage van het genoom van de karper en het annoteren van functies binnen dat genoom. De gewone karper is een nieuw model systeem dat uitermate geschikt is voor zogenaamde ”high-throughput” screenings van verzamelingen van farmaceutisch actieve stoffen. In dit hoofdstuk beschrijven we hoe de genoom assemblage is gerealiseerd en hoe we een proces-koppeling voor annotatie van het genoom maken waarbij we vooral gericht zijn op het het identificeren van de genen verantwoordelijk voor de aangeboren immuunrespons; dit zijn met name de genen die domeinen bevatten voor de Toll/Interleukin-1 receptor (TIR). De analyse is gebaseerd op zogenaamde volgende generatie sequentie gegevens. De proces-koppeling voor genoom assemblage bestaat uit het opschonen van de data, pre-assemblage en assemblage gebruik makend van CLCBio, ABySS en SOAP-denovo software. Een recht toe recht aan gen voorspelling gebaseerd op een proteine gebaseerd gen model voorspellingsmodel tesamen met een vergelijkignsannotatie gebaseerde annotatie vormen een werkbare proces-koppeling voor het annoteren van genen in dit nieuwe genoom. In de vergelijkingsannotatie wordt gebruik gemaakt van de ortologe genen in het genoom van de zebravis.

Zoals eerder vermeld, heterogene data integratie is het centrale thema in dit proefschrift. In de hoofdstukken 2 en 3 worden verschillende kenmerken zoals afstand op het genoom, sequentie similariteit, vrije energie en concepten uit de Gene Ontology zorgvuldig gecombineerd om tot een eindoordeel te komen of een voorspelling omtrent een doel-RNA goed of fout is. Integratie wordt gerealiseerd door een combinatie van data mining technieken zoals, beslisbomen, relatieve subgroup discovery, een lineaire classifier en een kwadratische classifier. In hoofdstuk 4 worden de intermediaire kenmerken, gegenereerd uit de drie geselecteerde voorspellingsmethoden, vastgelegd en geintegreerd met een Bayesiaanse Netwerk Classifier. In hoofdstuk 5, in de sectie die handelt over genoom annotatie, worden verschillende typen van data zoals DNA fragmenten, RNA-Seq fragmenten en RNA motieven geintegreerd op een sequentieele wijze. Elke stap in het werkschema voegt een nieuw type data toe dat werkt als een filter om te zoeken naar de TIR-domein bevattende kandidaat sequenties in het karper genoom.

Het doel van de integratie is het verbeteren van de sensitiviteit en/of de specificiteit van het

systeem. Deze twee maten karakteriseren de prestatie van het systeem. Sensitiviteit wordt gedefinieerd als de ratio van het aantal positieven en het aantal correct geidentificeerde positieven. De specificiteit drukt de waarschijnlijkheid uit dat de negativen correct geindentificeerd worden. Context van predictie is het wenselijk dat een algoritme zowel een hoge sensitiviteit als een hoge specificiteit kan realiseren. Er is echter een afweging tussen deze maten, een hoge sensitiviteit zal de specifiteit benadelen omdat de maat van foute positieven toeneemt en vice versa. In hoofdstuk 2 wordt een kenmerk voor genomische afstand tussen miRNA en hun doel genen toegevoegd en de analyse wordt verder verrijkt met aanvullende informatie. Hierdoor kan het aantal potentieele doel genen voor de miRNAs die werden onderzocht, dre-miR-10 and dre-miR-196, worden teruggebracht tot minder dan 10 per miRNA. In hoofdstuk 3 worden functioneel vergelijkbare miRNAs gebruikt voor de predictie van doel genen van miRNA waarvan de functie nog niet bekend is. Op deze wijze zijn 6 nieuwe doel-genen geindentificeerd voor 5 van de miRNAs uit het experiment. Op basis van heterogene data bronnen zijn we in staat geweest het aantal potentieele candidaat doel-genen terug te brengen tot een omvang waarbinnen de bioloog een validatie experiment kan opzetten. In de hoofdstukken 3 en 4 was het doel te onderzoeken hoe de specifiteit kan worden verbeterd. Door het toepassen van een strategie van data integratie zijn we in staat geweest dit te realiseren met slechts een kleine vermindering van de sensitiviteit. Een hoge specificiteit versnelt het proces van het vinden van de doel-genen aanzienlijk. In hoofdstuk 4 zijn 3 doel voorspellings methoden geintegreerd waarbij 3 verschillende integratie strategieen zijn gebruikt, hierbij is specifiek gelet op het behalen van de best mogelijke prestatie. De prestatie is uitgedrukt in een maat waarin zowel specificiteit als sensitiviteit worden meegenomen. Op basis van de resultaten hebben we een concept kunnen uitwerken hoe een correcte integratie de prestatie aanzienlijk kan verbeteren in vergelijking met ieder van de methoden afzonderlijk. In hoofdstuk 5 hebben we gestuurd naar het maximaliseren van de waarschijnlijkheid om de TIR-domein bevattende genen te vinden in het genoom en in RNA fragmenten van de karper. Hierdoor heeft de focus vooral gelegen op de sensitiviteit.

Met het toepassen van de hier genoemde methoden wordt ons inzicht in de miRNA regulatie als ook de structuur en functie van nieuwe genen bevorderd. In deze studies zijn daarmee ook nieuwe inzichten in de biologie verkregen.

Het mechanisme van miRNA regulatie in dieren wordt gezien als zeer verfijnd, echter

tot op heden is het nog niet geheel doorgrond; getuige het feit dat veel doel-genen nog niet zijn geindentificeerd en daarbij nog veel fout positieven overblijven. In onze studies hebben we een aantal interessante nieuwe kenmerken kunnen vaststellen. In hoofdstuk 3 hebben we het verband opgehelderd tussen doel-genen en miRNA door te laten zien dat veel van deze doel-genen fysiek dicht gelokaliseerd zijn bij het miRNA dat ze reguleert. Uitgaande van dit feit hebben we in hoofdstuk 3 kunnen laten zien dat functioneel vergelijkbare miRNAs gelokaliseerd zijn in clusters. Uit deze bevindingen kunnen we conlcuderen dat afstand op het genoom een belangrijke rol speelt in miRNA-doel interactie. Als twee miRNAs of een miRNA en zijn doel-gen dicht bij elkaar op het genoom zijn gelokaliseerd, dan is de waarschijnlijkheid dat ze tegelijk worden afgelezen groot. Dit gezamelijk voorkomen impliceert dat ook de functies vergelijkbaar zijn of dat ze in een interactie voorkomen. Door de kenmerken van reeds gevalideerde humane miRNA-doel gen relaties te bestuderen hebben we kunnen vaststellen dat sommige miRNA's binden op hun doel juist aan het begin of aan het eind van de 3' UTR sequentie.

Het annoteren van genen is een tijdrovende en arbeidsintensieve taak. Voor niet-model organismen waarvoor geen genoom assemblage beschikbaar is, moet eerst dit bewerkstelligd worden voordat men in staat kan zijn het genoom te annoteren. In hoofdstuk 5 wordt gedemonstreerd hoe een niet-model organisme, i.e. de gewone karper, kan worden geannoteerd. Een grote hoeveelheid data wordt gegenereerd voor de genoom en RNA sequentie analyse. Uit deze data is een voorlopig karper genoom samengesteld met een N50 contig lengte van 2260 base paren. De inschatting is dat het karper genoom een lengte heeft van ongeveer 1.23 giga base-paren. In een vergelijking met de aangeboren immuno genen van de zebravis kunnen we schatten dat er 39 TIR-domein bevattende genen en afschriften zijn in de gewone karper.

Samenvattend, annotatie is een veelomvattend onderwerp en een belangrijk onderzoeksthema voor de biologie, en daarmee voor de bioinformatica, voor nu en voor de toekomst. In dit proefschrift laten we zien hoe bioinformatica en integratie kan worden gebruikt, in het bijzonder voor de annotatie van microRNA's en genomen. Dit is slechts een klein deelgebied van annotatie, desalniettemin hebben we laten zien hoe bioinformatica een leidraad kan zijn voor laboratorium experimenten door te voorzien in kandidaten die gevalideerd kunnen worden.