Gebundelde krachten

van Kampen, A.H.C.

Publication date
2011
Document Version
Final published version

Gebundelde krachten

Gebundelde krachten

Rede

uitgesproken bij de aanvaarding van
het ambt van bijzonder hoogleraar Biologische en Biomedische
informatiewetenschappen
in het bijzonder de Medische bioinformatica
aan de Universiteit van Amsterdam
op 20 mei 2011

door

A.H.C. van Kampen

Dit is oratie 405, verschenen in de oratiereeks van de Universiteit van Amsterdam.

Opmaak: JAPES, Amsterdam

Geachte mevrouw de Rector Magnificus,
Geachte mijnheer de Decaan van de Faculteit der Natuurwetenschappen, Wiskunde en Informatica,
Geachte mijnheer de Decaan van de Faculteit der Geneeskunde,
Geachte leden van het Curatorium,
Geachte leden van het bestuur van het Genootschap ter Bevordering van Natuur-, Genees- en Heelkunde,
Beste collega's, studenten, scholieren, vrienden en familie,

Het is al weer vier jaar geleden dat mijn benoeming als bijzonder hoogleraar heeft plaatsgevonden. Mijn leerstoel is een van de acht leerstoelen die ingesteld zijn vanwege het Genootschap ter bevordering van Natuur-, Genees- en Heelkunde[1] bij de Faculteit der Natuurwetenschappen, Wiskunde en Informatica (FNWI). Het genootschap belichaamt het interdisciplinaire karakter van de geneeskunde en ik zie het als een belangrijke en uitdagende opdracht om dit karakter verder vorm te geven voor mijn vakgebied, de medische bioinformatica, en in de relatie tussen de FNWI en het Academisch Medisch Centrum (AMC).

Laat ik toch niet gelijk met de deur in huis vallen. Zoals een aantal van u weet ben ik 31 jaar geleden begonnen met bespelen van een elektronisch orgel, een toen nog populair instrument. Daarna volgde de synthesizer, piano en de Hammond. De laatste 25 jaar heb ik met uiteenlopende gezelschappen muziek gemaakt en opgetreden. Deze lange 'carrière' in de muziek is duidelijk geen maat voor virtuositeit of succes, anders had ik vandaag niet hier gestaan. Het spelen in een band heeft veel overeenkomsten met het doen van onderzoek. Door individuele bijdragen en experimenten van muzikanten en hun samenwerking ontstaat een muzikaal resultaat. Dankzij onze manager (valorisatie officer) zijn wij in staat financiering te verwerven om dit resultaat bij een breder publiek bekend te maken. De muzikanten publiceren hun werk op cd en presenteren bij uiteenlopende bijeenkomsten. Het publiek geeft terugkoppeling en staat garant voor de evaluatie. Dus Gebundelde Krachten, net zoals in mijn onderzoek. De belangrijkste overeenkomst is echter dat ik zowel muziek als onderzoek met veel plezier beoefen. Ik hoop u vandaag dan ook te laten zien hoe leuk medische bioinformatica is, en dat u na vandaag beter begrijpt wat een bioinformaticus doet.

Van data naar kennis

Explosie van data

De stelling 'Things are only impossible until they're not'[2] (Jean-Luc Picard, kapitein van het Galaxy-klasse ruimteschip USS Enterprise) geeft goed weer wat we met wetenschappelijk onderzoek proberen te bereiken. In biomedisch onderzoek wordt de moleculaire basis van ziekten steeds verder ontrafeld door een nauwe samenwerking tussen kliniek, experimentele laboratoria, bioinformatici, informatici en statistici. Vraagstellingen die vaak hun oorsprong hebben in de kliniek kunnen mede door het doen van laboratoriumexperimenten worden opgelost. Tegenwoordig zijn onderzoekers in staat om in het laboratorium op allerlei niveaus de moleculaire toestand van een cel te karakteriseren. Onderzoekers zijn niet langer beperkt tot het meten van de toestand van individuele genen, eiwitten of metabolieten, maar nieuwe experimentele technologieën hebben het mogelijk gemaakt om zogenaamde genoombrede metingen uit te voeren. Bijvoorbeeld, het bepalen van de lettervolgorde van DNA gebeurt met zogenaamde sequencingtechnieken. In 2001 publiceerden zowel *Nature* als *Science* de eerste, nagenoeg volledige, DNA-sequentie van de mens.[3,4] Dit was het resultaat van het Human Genome Project, ongeveer dertien jaar werk, waarvan de kosten 437 miljoen US dollar bedroegen.[5] Met de huidige Next Generation Sequencing-technologieën bepalen we tegenwoordig een volledig menselijk genoom bijna routinematig. De DNA microarray-technologie maakt het mogelijk om de expressie (activiteit) van nagenoeg alle genen in een cel te meten. In één meting kan eenvoudig de activiteit van alle, ongeveer 20.000, humane genen worden bepaald voor uiteenlopende condities. Het meten van eiwitten en metabolieten gebeurt met massaspectroscopie en ook dit levert grote en complexe dataverzamelingen op. Al deze data wordt 'omics'-data genoemd.[6]

Door omics en andere technologieën is de hoeveelheid data die geproduceerd wordt als onderdeel van biomedisch onderzoek de afgelopen jaren explosief gestegen. In relatie hiermee wil ik vandaag de volgende stelling poneren: 'Het is een goede regel om niet te veel waarde te hechten aan biomedische data totdat door toepassing van de juiste statistiek en de juiste informatica blijkt dat deze data informatie bevat.'

Medische bioinformatica

Mijn vakgebied betreft de biologische en biomedische informatiewetenschappen, in het bijzonder de medische bioinformatica. Informatiewetenschap is

een interdisciplinaire wetenschap die zich bezighoudt met het verzamelen, opslaan, classificeren, manipuleren, analyseren en verspreiden van informatie.[7] Informatiewetenschap benadert vraagstellingen vanuit het perspectief van de vraagsteller (bijvoorbeeld de bioloog) en niet zozeer vanuit individuele technologieën. In de medische bioinformatica[8] gaat het om biomedische informatie, met name omicsdata, maar ook om medische beelden en klinische data. Deze informatie wordt gebruikt om onze kennis over fysiologische en pathofysiologische processen te vergroten, maar ook als onderdeel van diagnostiek of voor selectie en evaluatie van therapie.

Wetenschappelijk onderzoek in medische bioinformatica is cruciaal voor het ontwikkelen van methoden voor het omzetten van data naar informatie en kennis. Allereerst draagt bioinformatica bij aan het definiëren van een optimale experimentele opzet zodat achteraf de gewenste informatie uit metingen kan worden verkregen. Een tweede rol is de analyse, visualisatie en interpretatie van experimentele data. Een derde rol is weggelegd voor informatiemanagement ten behoeve van dataopslag, -organisatie, -annotatie, -curatie, en -integratie. Hierbij wordt een steeds prominentere rol gespeeld door e-Bioscience, de discipline die zich specifiek bezighoudt met data- en rekenintensief biomedische onderzoek in multidisciplinaire samenwerkingen met behulp van geavanceerde informatietechnologie.

Het Bioinformatica Laboratorium

Mijn groep, het Bioinformatica Laboratorium,[9] werkt aan biomedische vraagstellingen vanuit bioinformatica en e-Bioscience door middel van data-analyse en informatiemanagement. Deze biomedische vraagstellingen zijn zeer verschillend van aard en betreffen uiteenlopende ziekten, biologische systemen en organismen. Omdat mijn onderzoek in belangrijke mate wordt gedreven door deze vraagstellingen heb ik ervoor gekozen om met mijn groep een brede expertise te ontwikkelen in plaats van een sterke focus op een deelgebied. We hebben expertise opgebouwd voor de analyse van alle typen omicsdata en gebruiken en ontwikkelen daarvoor uiteenlopende methoden afkomstig uit de informatica en statistiek. Ook dragen we bij aan de ontwikkeling van geavanceerde onderzoeksinfrastructuur ten behoeve van de analyse van omicsdata en medische beelden. Samen met mijn groep heb ik laten zien dat deze brede aanpak succesvol is en we uiteenlopende wetenschappelijke vraagstellingen kunnen beantwoorden. Bovendien ontstaan kruisbestuivingen en daardoor meerwaarde. Gebundelde Krachten dus. Ik zal vandaag vier van onze aan-

dachtsgebieden kort belichten. Het DNA microarray- en metabolomicsonderzoek laat ik buiten beschouwing.

Sequentieanalyse: van nucleotide naar kennis

De analyse van DNA- en RNA-sequenties loopt al dertien jaar als een rode draad door mijn onderzoek. Dit onderzoek heeft recent weer een grote impuls gekregen vanwege de zogenaamde Next Generation Sequencing-technologieën[10] die het mogelijk maken om op grote schaal sequentiedata te produceren. In één experiment van enkele dagen worden miljoenen korte DNA- of RNA-sequenties geproduceerd, voldoende voor honderden gigabytes aan data. Wereldwijd wordt er met Next Generation Sequencing meer data geproduceerd dan door de Large Hadron Collider (CERN, Genève) die jaarlijks vijftien petabyte voor de hoge-energiefysica produceert.[11] De grote aantallen sequenties plaatsen bioinformatica voor nieuwe uitdagingen. Bovendien heeft Next Generation Sequencing een groot aantal toepassingen zoals expressieanalyse, het bepalen van volledige DNA-sequenties van patiënten en micro-organismen, het identificeren van DNA-variaties tussen individuen, identificatie van micro-RNA's, analyse van bindingsplaatsen voor transcriptiefactoren en analyse van DNA-methylatiepatronen. Data voor de meeste van deze toepassingen worden in mijn groep geanalyseerd en voor veel van deze projecten worden nieuwe methoden ontwikkeld. Ik zal u een aantal voorbeelden geven. Samen met Paul Klarenbeek en Niek de Vries (Klinische Immunologie en Reumatologie) werken we aan methoden om het humane T-cel- en B-cel-repertoire te bepalen om zo de rol van van deze cellen in auto-immuunziekten te doorgronden. Met Jurgen Piet, Diederik van Beek (Neurologie) en Arie van der Ende (Medische Microbiologie) hebben we volledige bacteriële sequenties van pneumococcen geanalyseerd om genetische factoren te identificeren die bijdragen aan de virulentie van bacteriële meningitis (hersenvliesontsteking). Samen met Katja Ritz en Frank Baas (Laboratorium voor Neurogenetica en Genoomanalyse) hebben we isovormen geïdentificeerd van het SGCE-gen die in de hersenen tot expressie komen en die implicaties kunnen hebben voor de pathogenese van myoclonus dystonie. Hiernaast participeren we samen met verschillende AMC afdelingen in het exoomsequencinginitiatief waarvan Raoul Hennekam (Kindergeneeskunde) de kwartiermaker is. In dit initiatief richten we ons op de identificatie van weinig voorkomende genvarianten bij zeldzame genetische aandoeningen zoals Nicolaides-Baraitser, pseudo-TORCH syndroom, hyperplastisch polyposissyndroom en Body Integrity Identity Disorder. Het ontrafelen van de genetische basis van zeldzame aan-

doeningen zal bijdragen aan diagnostiek en mogelijk aan de selectie of ontwikkeling van therapie.

In de sequentieanalyseprojecten zullen we ons nadrukkelijker gaan richten op het ontwikkelen van methoden voor de identificatie en verwijdering van artefacten die vanwege de experimentele procedure of vanwege de dataverwerking zijn geïntroduceerd. Deze artefacten veroorzaken bias in de data en dit kan leiden tot foutieve conclusies zoals fout-positieve en fout-negatieve bevindingen in exoomsequencing of onnauwkeurige genexpressiemetingen in expressieanalyse met RNA-Seq. Een andere uitdaging betreft de verwerking van de enorme datavolumes, maar hier kom ik later op terug.

Systeem Genomics: van molecuul naar kennis

De afgelopen jaren is steeds meer interesse ontstaan voor het verwerven van gedetailleerde kennis over de werking van complete biologische systemen zoals een celorganel, een cel, een orgaan of een organisme. Dit is het domein van de systeembiologie[12] waar wij een holistische benadering (de top-downmethode) nastreven via het gebruik van omicstechnologieën die informatie geven over een groot aantal genen, eiwitten en metabolieten. Dit noemen we 'Systeem Genomics' en dit onderzoek wordt geleid door Perry Moerland. Ons doel is om bioinformaticamethoden te ontwikkelen voor het analyseren van omicsdata en om zodoende de rol en relatie van genen, eiwitten en metabolieten te bepalen in gezondheid en ziekte. Hierbij willen we zoveel mogelijk gebruikmaken van biologische voorkennis die aanwezig is in publieke biologische databases.

Publieke biologische databases in Systeem Genomics

Het gebruik van deze biologische databases is echter niet zonder valkuilen. Ik zal dit illustreren aan de hand van een voorbeeld betreffende databases met hierin informatie over het humane metabolisme. Metabolisme, of stofwisseling, is de verzameling van biochemische processen die plaatsvinden in de cel. Metabolisme speelt een belangrijke rol in gezondheid maar kan verstoord raken bij ziekte. Een nauwkeurige beschrijving van het metabolisme is noodzakelijk voor de analyse en interpretatie van omicsdata, maar ook omdat ze het uitgangspunt vormt voor de bottom-upaanpak van de systeembiologie, namelijk het ontwikkelen van mechanistische modellen die het metabolisme beschrijven. Het humane metabolisme is beschreven in een aantal publieke biologische databases. Biologen, bioinformatici, artsen en andere gebruikers zijn zich er soms van bewust dat er verschillen bestaan tussen deze databases, dat

deze databases niet compleet zijn en fouten kunnen bevatten. Toch is waarschijnlijk een slechts zeer beperkt groepje van onderzoekers op de hoogte van alle problematiek en verschillen betreffende deze databases. Er is nooit een systematische vergelijking uitgevoerd om dit in kaart te brengen terwijl dit wel van belang is omdat de keuze voor een specifieke database de uitkomsten van een analyse of interpretatie kan beïnvloeden. U moet zich realiseren dat de technische, conceptuele en inhoudelijke verschillen tussen metabolismedatabases momenteel zo groot zijn dat integratie tot één consensusnetwerk onmogelijk is. Ook een uitwisselingsstandaard zoals BioPAX[13] lost dit niet op.

Een vergelijking van de vijf belangrijkste metabolismedatabases (KEGG,[14] Reactome,[15] HumanCyc,[16] BiGG[17] en EHMN[18]) heeft ons vier jaar werk gekost! Dit maakt de complexiteit al direct duidelijk. De vergelijking vond plaats op het niveau van metabole paden, genen, enzymen (EC-nummers), metabolieten en chemische reacties. De vergelijking tussen de gekozen databases laat zien dat de overlap op alle niveaus dramatisch laag is. Bijvoorbeeld, afhankelijk van de gekozen manier om de vergelijking uit te voeren, ligt de overlap op genniveau tussen de 11% en 39%. Op het niveau van de reacties is de overlap maximaal 19%. Dit gebrek aan overlap heeft verschillende oorzaken en ook de gekozen aanpak om tot deze vergelijking te komen heeft hierop invloed, maar de resultaten geven duidelijk aan dat grote voorzichtigheid geboden is bij het gebruik van deze databases. Men zou verwachten dat een vergelijking voor een veel en uitgebreid beschreven metaboolpad zoals de citroenzuurcyclus een veel grotere overlap zou opleveren. Dit metaboolpad is in 1937 voor het eerst beschreven door Hans Krebs[19] en sindsdien zijn vele wetenschappelijke publicaties over dit deel van het metabolisme verschenen. Ook vinden we de citroenzuurcyclus terug in de meeste studieboeken over metabolisme. Desondanks levert een vergelijking van de vijf databases nog steeds een erg lage overlap op. Bijvoorbeeld, over slechts 25 van de in totaal 37 genen bestaat consensus. Samen met Ronald Wanders en Sander Houten (Genetische Metabole Ziekten) proberen we nu tot een betere, en misschien definitieve, beschrijving te komen.

Om de uitkomsten van onze vergelijkingen breed toegankelijk te maken voor domeinexperts en daarmee curatie te ondersteunen hebben we samen met de groep van Morris Swertz (Universiteit van Groningen) de zogenaamde 'Consensus en Conflict Cards' (C^2Cards) ontwikkeld op basis van Molgenis,[20] een softwareapplicatie voor de ontwikkeling van informatiesystemen. Deze kaarten brengen de verschillen duidelijk in beeld zodat deze kunnen worden opgelost waardoor de consistentie tussen de databases zal toenemen en integratie mogelijk wordt. We zijn in overleg met Ines Thiele (Centrum voor Systeembiologie, Universiteit van Reykjavik) om deze kaarten te gebruiken bij de

eerstvolgende internationale jamboree van de BiGG-database om te komen tot een betere definitie van het humane metabolisme.

Dit was slechts één voorbeeld, maar er bestaan minstens 1330[21] publieke biologische databases met hierin uiteenlopende informatie. De kwaliteit van deze databases is zeer verschillend. Het vervelende is echter dat het soms zeer lastig is om tekortkomingen van een database in kaart te brengen en op te lossen. We moeten ons hier echter wel bewust van zijn omdat het de analyse en interpretatie van data kan beïnvloeden. Dit zal daarom een aandachtspunt blijven in ons onderzoek.

Analyse van genen, eiwitten en metabolieten in biologische systemen

Een tweede aspect van het Systeem Genomics-onderzoek betreft het gebruik en de ontwikkeling van bioinformaticamethoden voor de analyse van genen, eiwitten en metabolieten die verstoord zijn in niet-fysiologische condities zoals bij ziekte of onder experimentele omstandigheden. Bijvoorbeeld, samen met Milka Sokolović, Arthur Verhoeven (Medische Biochemie), Theo Hakvoort en Wout Lamers (AMC Tytgat Instituut) hebben we gekeken naar de adaptieve respons van het metabolisme en de interorgaancoördinatie hiervan in muizen tijdens (langdurig) vasten.

Om beter in staat te zijn de verschillen en overeenkomsten tussen verschillende fysiologische en niet-fysiologische condities en organismen te identificeren en te begrijpen zullen we meer gebruik gaan maken van experimentele omicsdata uit publieke databases. In zogenaamde compendia zullen we deze publieke data integreren met data afkomstig van samenwerkingspartners. Een systeem voor het opzetten van ziekte-specifieke compendia is inmiddels ontwikkeld en dit zal gebruikt gaan worden voor het ontrafelen van verschillen tussen diermodellen en de mens.

Diermodellen, zoals muizen en ratten, zijn essentieel om kennis te vergaren over humane ziekten en ook worden ze gebruikt in preklinische trials voor het testen van de veiligheid van nieuwe geneesmiddelen. In het verleden is echter gebleken dat bevindingen gedaan met diermodellen niet altijd toepasbaar zijn op de mens. De oorzaak hiervan is lang niet altijd duidelijk en verschillende aspecten, zoals verschillen in fysiologie, omgevingsfactoren en moleculaire biologie, dragen hieraan bij. In ons onderzoek willen we kijken of we verschillen kunnen identificeren op het niveau van DNA, genexpressie, eiwitten of metabolieten. Om dit te doen zullen we bioinformaticamethoden ontwikkelen om te bepalen welke deelverzamelingen van genen en condities geconserveerd zijn tussen het modelorganisme en de mens. Door deze deelverzamelingen te kop-

pelen met functionele informatie uit metabolisme-, fenotype-, genetische en literatuurdatabases zullen we hopelijk in staat zijn een aantal verschillen tussen diermodellen en de mens voor specifieke ziekten in kaart te brengen en deze informatie te gebruiken voor de opzet en analyse van nieuwe diermodelstudies.

Semantische biosystemen: van kennis naar toepassing

Voor de analyse en interpretatie van biomedische data maken we gebruik van een deel van de enorme hoeveelheid biomedische kennis die vandaag de dag beschikbaar is. In de praktijk is het echter niet eenvoudig om kennis over een bepaald domein goed te organiseren, te representeren, te integreren en te presenteren zodat deze op eenvoudige wijze kan worden toegepast in biomedisch onderzoek. Het onderzoek van semantische biosystemen, geleid door Andrew Gibson, richt zich daarom op de ontwikkeling van een kennismanagementsysteem gebaseerd op het semantisch web.[22] Dit systeem gebruiken we voor de implementatie van specifieke biomedische kennisdatabases voor toepassing in onderzoek en kliniek. Ik zal u een aantal aspecten van dit onderzoek schetsen.

Het semantisch web

Het semantisch web wordt gepropageerd door het World Wide Web-consortium[23] (W3C), een internationale standaardisatieorganisatie voor het Web. Om te begrijpen wat het semantisch web is (of moet worden) is een vergelijking met het huidige Web (het internet) illustratief. Het Web zal door de meesten van u dagelijks gebruikt worden om informatie te zoeken via bijvoorbeeld de Googlezoekmachine. Het huidige Web is uitermate geschikt voor gebruik door mensen, maar veel minder geschikt voor gebruik door computers. Laat ik u hiervan een voorbeeld geven. Stel, u hebt de ochtend na de oratie van een collega een kater en zoekt via Google op 'behandelen van kater'. Ik acht de kans klein dat u vervolgens aan de slag gaat met inentingen, antibiotica en ontwormingskuren. Voor een computer zal het echter niet zo makkelijk zijn om onderscheid te maken tussen een kat en een kater vanwege overmatig drankgebruik. We moeten daarom de betekenis (semantiek) van informatie eenduidig beschrijven. Een ander probleem is het feit dat het huidige Web, een web van documenten (de webpagina's) is en geen web van data. Hoewel webpagina's data kunnen bevatten zijn deze data vaak moeilijk te benaderen door computerprogramma's. Technologieën, zoals RDF, OWL en SKOS, die de basis vormen voor het semantisch web maken het mogelijk om kennis een-

duidiger te modelleren, te representeren, te annoteren, te integreren en op een gestandaardiseerde manier beschikbaar te maken voor mens en computer. Deze technologieën zijn daarom ook uitermate geschikt voor kennismanagement. Het is echter goed te beseffen dat het semantisch web nog in ontwikkeling is en dat beschikbaarheid van semantisch web-technologieën het nog niet eenvoudig maakt om deze op een juiste manier te gebruiken en om toepassingen te ontwikkelen. Samen met de groep van Ronald Wanders (Genetische Metabole Ziekten) hebben we de afgelopen jaren intensief gewerkt aan een peroxisoomkennisdatabase.[24] Een peroxisoom is een celorganel dat een belangrijke rol speelt in het humane metabolisme. De peroxisoomkennisdatabase bevat een uitgebreide beschrijving van het peroxisomale metabolisme en gerelateerde ziekten. Ik zal u kort uitleggen hoe wij semantisch web-technologieën gebruiken voor de representatie van kennis.

Kennisrepresentatie

De basisbouwstenen van onze kennisdatabase worden gevormd door 'concepten' die gezamenlijk een kennisdomein afbakenen. Voorbeelden van deze concepten zijn 'peroxisoom', 'vetzuuroxidatie' en 'Zellwegersyndroom'. Op basis van deze concepten construeren we een vocabulaire waarin we deze concepten op een hiërarchische manier organiseren. Zo kunnen we bijvoorbeeld het concept 'vetzuuroxidatie' onderbrengen bij het generiekere concept 'biochemisch proces'. Ook kunnen we aan elk concept verdere definities, synoniemen en verwijzingen naar andere databanken, vocabulaires en ontologieën toekennen. Het resultaat is een specifiek en hoogwaardig peroxisoomvocabulaire dat de basis vormt voor een consistente kennisrepresentatie op hogere niveaus. Op een volgend niveau worden concepten gecombineerd in zogenaamde RDF-tripletten die bestaan uit een onderwerp, een predicaat en een object. Een voorbeeld van een triplet is 'fytaanzuur (onderwerp) hoopt zich op in (predicaat) Zellwegersyndroom (object)'. Aan deze tripletten kan een verdere betekenis worden toegekend, zoals 'fytaanzuur' is een metaboliet en 'Zellwegersyndroom' is een ziekte. Ook kunnen we de relatie 'hoopt zich op in' verder specificeren en definiëren. Het relateren en semantisch annoteren van concepten resulteert dus in kleine stukjes biologische kennis (RDF-tripletten) die op een nog hoger niveau met elkaar worden verbonden tot grotere netwerken van tripletten die een kennisdomein representeren. Wij volgen ook het 'Linked Data'-principe. [25] Linked Data is een manier om gestructureerde en semantisch geannoteerde data te publiceren zodat deze kunnen worden opgevraagd en gekoppeld met andere databronnen. Dit principe zal in belangrijke mate een web van data mogelijk maken en daarmee bijdragen aan de interoperabili-

teit en integratie van data, informatie en kennis. In onze kennisdatabase zijn alle concepten en hun beschrijvingen beschikbaar als Linked Data.

Toepassing in de kliniek

Samen met Peter Barth en Bwee Tien Poll-Thé (Kinderneurologie) proberen we te komen tot toepassing van de peroxisoomkennisdatabase in de kliniek. Dit is niet eenvoudig mede omdat wij moeten leren wat een arts nodig heeft en, vice versa, moet de arts onze aanpak begrijpen. We zijn gestart met het toevoegen van klinische informatie en patiëntgegevens met betrekking tot Zellweger Spectrum Disorders (ZSD), een groep van ernstige en erfelijke peroxisomale biogeneseaandoeningen. Dit doen we met een aantal ideeën in het achterhoofd. Nieuwe patiënten kunnen worden toegevoegd aan de kennisdatabase en worden daarmee geplaatst in de context van bestaande klinische kennis en andere patiënten. Dit zou kunnen leiden tot een verbeterde en geactualiseerde karakterisering van ZSD-patiënten. Informatie uit de kennisdatabase kan in potentie diagnostiek of selectie van behandeling ondersteunen. Integratie met beslissingsondersteunende computersystemen in de kliniek zou dit kunnen bewerkstelligen. In het algemeen kan de peroxisoomkennisdatabase als een uitstekende eerste bron van informatie dienen voor artsen die niet gespecialiseerd zijn in peroxisomale aandoeningen. Verder kan door het toevoegen van klinische informatie en patiëntdata aan de kennisdatabase een automatische koppeling worden gemaakt met de al aanwezige informatie over peroxisomale biochemische processen. Dit geeft mogelijkheden voor translationeel onderzoek. Dus, de ambities zijn groot maar er zal nog veel werk moeten worden verricht om ze te realiseren.

Naast de peroxisoomkennisdatabase werken we aan een aantal andere kennisdatabases. Recent zijn we gestart met kennisdatabases voor systeembiologie ten behoeve van definitie, uitwisseling, hergebruik en disseminatie van kennis over biologische systemen. Hiernaast werken we aan de integratie van een kennisdatabase met multivariate statistische modellen voor de analyse van metabolomicsdata, en polyphenolen in het bijzonder, om te bepalen wat de consequenties hiervan zijn voor de opzet van onze kennisdatabases.

De kennisacquisitie die gepaard gaat met onze kennisdatabases is een tijdsintensieve en uitdagende aangelegenheid. Als onderdeel van het onderzoek ontwikkelen we dan ook software en strategieën om dit proces verregaand te vereenvoudigen. Dit is nodig omdat kennisacquisitie momenteel een knelpunt is

terwijl de kennisdatabases pas echt meerwaarde gaan krijgen als verschillende experts eenvoudig kunnen bijdragen aan de inhoud.

e-Bioscience: datagedreven en rekenintensief biomedisch onderzoek

Naast behoefte aan kennismanagement in biomedisch onderzoek kunnen we ook constateren dat dit onderzoek in toenemende mate multidisciplinair, datagedreven en rekenintensief is. De benodigde grootschalige dataopslag- en rekencapaciteit evenals databases en software zijn verregaand gedistribueerd en heterogeen. De behoefte om data en software te delen, te hergebruiken en te integreren neemt toe. Daarom wordt er steeds meer informatietechnologie ingezet om een onderzoeksinfrastructuur te creëren die deze gedistribueerde en heterogene faciliteiten ontsluit voor biomedisch onderzoek. Uit het feit dat in de dagelijkse praktijk nog vaak 'quick-and-dirty' oplossingen gekozen worden blijkt dat de huidige onderzoeksinfrastructuren nog niet voldoen. Dit heeft als gevolg dat ontwikkelde softwareapplicaties niet kunnen worden hergebruikt, geen optimaal gebruik kunnen maken van beschikbare rekenkracht en dataopslagcapaciteit, moeilijk te koppelen zijn met andere software en databases, en lastig te onderhouden zijn. Hoewel zulke oplossingen voor individuele onderzoekers vaak voldoen lossen ze de echte problemen niet op. Hierdoor wordt het wiel steeds weer opnieuw uitgevonden en wordt de vooruitgang in biomedisch onderzoek uiteindelijk geremd. In het enhanced-Bioscience (e-Bioscience)-onderzoek, [26] geleid door Silvia Olabarriaga, richten we ons niet alleen op het gebruik van onderzoeksinfrastructuur maar, in nauwe samenwerking met veel betrokken partijen, ook op de verdere ontwikkeling van zo'n infrastructuur. Dit doen we door de ontwikkeling van het zogenaamde e-BioInfraplatform dat de biomedische onderzoeker in staat stelt op eenvoudige wijze gebruik te maken van gedistribueerde faciliteiten zoals computersystemen en softwareapplicaties. E-BioInfra is als het ware de besturingsconsole van een onderzoeksinfrastructuur en de uitdaging is om te zorgen voor een grote mate van interoperabiliteit, standaardisatie, schaalbaarheid, robuustheid, kwaliteit en onderhoudbaarheid van dit platform.

Software- en hardware-infrastructuur voor biomedisch onderzoek

Grofweg bestaat een IT-onderzoeksinfrastructuur uit 'middleware' (de software) en computersystemen (de hardware) zoals nationale en internationale gridnetwerken, cloud computing, supercomputers, en lokale systemen zoals

desktopgrids en computerclusters. Via SARA Reken- en Netwerkdiensten[27] zijn een aantal van deze systemen beschikbaar. Bovendien hebben we, in tegenstelling tot de meeste andere Europese landen, in Nederland dankzij Big Grid[28] een goed georganiseerd gridnetwerk tot onze beschikking waar universiteiten, academische ziekenhuizen en de industrie deel van uitmaken en dat momenteel ongeveer 7000 processors en 5000 terabyte aan opslagcapaciteit omvat. Samen met onder anderen Jan Just Keijser (Big Grid) hebben we het gebruik van dit Nederlandse e-Science Grid voor onze toepassingen mogelijk gemaakt. Bijvoorbeeld, in een recente studie die wij samen met Michel de Vries en Lia van der Hoek (Laboratorium voor Experimentele Virologie) hebben uitgevoerd is Next Generation Sequencing gebruikt als onderdeel van een methode om bekende en onbekende virussen in klinische monsters te identificeren. Een onderdeel van de bioinformatica-analyse betrof de vergelijking van ongeveer vijf miljoen sequenties met een deel van de GenBank database. [29] We waren in staat om de rekentijd hiervoor terug te brengen van zeventien dagen naar veertien uur. In een andere studie met Matthan Caan (Radiologie) is Diffusion Tensor Imaging gebruikt om wittestofbanen te analyseren die in de hersenen gebieden met elkaar verbinden. De gebruikte algoritmes zijn zeer rekenintensief maar we waren in staat om de rekentijd terug te brengen van één jaar naar zeven dagen.

Het is belangrijk te onderkennen dat het realiseren van een onderzoeksinfrastructuur slechts deels een hardwareprobleem is. De aanwezigheid van computersystemen betekent niet dat hiermee de problemen zijn opgelost. Een belangrijkere rol wordt gespeeld door de middleware die generieke softwarecomponenten met uiteenlopende functionaliteiten bevat. Deze functionaliteiten zorgen onder andere voor informatiebeveiliging, distributie van rekenopdrachten, basaal datamanagement, workflowmanagement, monitoring en gebruikersinterfaces. De middelware stelt de gebruiker in staat om data-analyse- en datamanagementtaken op gedistribueerde computersystemen uit te voeren. In ons onderzoek ontwikkelen we deze middleware, het eerder genoemde e-BioInfraplatform, aan de hand van concrete biomedische vraagstellingen. Een aantal aspecten hiervan zal ik verder toelichten.

E-BioInfraonderzoek

In het e-BioInfraplatform wordt een centrale rol ingenomen door een workflowsysteem. Zo'n systeem specificeert complexe taken op basis van simpelere componenten. Workflows abstraheren de opzet van een ingewikkeld computationeel experiment en kunnen bijdragen aan de parallellisatie van zo'n experiment. Wij gebruiken workflows als generiek mechanisme om de gebruiker

transparant toegang te geven tot het grid. Hierbij kijken we naar aspecten zoals workflowfunctionaliteiten, compatibiliteit met verschillende computerinfrastructuren, fouttolerantie, interoperabiliteit van workflowsystemen en de geschiktheid voor specifieke biomedische toepassingen.

Het ligt voor de hand om naast het grid ook andere infrastructuren te gebruiken zoals cloud computing of lokale computerclusters, afhankelijk van bijvoorbeeld de benodigde rekenkracht, de opslagcapaciteit, het computergeheugen en de beschikbaarheid. Echter, deze infrastructuren zijn niet met elkaar gekoppeld en hun gebruik is niet uniform. Ze vereisen specifieke interfaces, aanpassingen aan softwareapplicaties, kennis van de hardware en/of middleware, of hebben individuele procedures voor authenticatie en autorisatie. Dit vereist daarom verdere ontwikkeling van het e-BioInfraplatform. We onderzoeken of ook hier workflows gebruikt kunnen worden als een generiek en transparant mechanisme om verschillende infrastructuren te ontsluiten.

Om de resultaten van een grootschalig computationeel experiment te kunnen begrijpen en te interpreteren moeten we de stappen kunnen volgen die hebben geleid tot het uiteindelijke resultaat. Dit betekent dat we invoer- en uitvoerdata, executie van workflowcomponenten en hun parameters en eventuele fouten moeten kunnen registreren. Een traditioneel labjournaal zal hiervoor niet volstaan. In plaats daarvan hebben we een zogenaamde provenancearchitectuur nodig die zowel prospectieve provenance, de workflowspecificatie, als retrospectieve provenance, de daadwerkelijke executie van de workflow, registreert. Dit maakt het mogelijk om experimenten te reconstrueren, te volgen in de tijd, te vergelijken en te optimaliseren. Een provenancearchitectuur zal ook bijdragen aan de verbetering, de optimalisatie en het beheer en de ondersteuning van middelware en computersystemen. Samen met Ammar Benabdelkader (Big Grid) hebben we een start gemaakt om deze architectuur te ontwikkelen.

Een specifiek aandachtspunt betreft het ontwikkelen van strategieën en methoden om met grote datavolumes binnen een gedistribueerde computerinfrastructuur om te kunnen gaan. We zien deze problematiek bijvoorbeeld bij het Genoom van Nederland-project[30] waar de volledige DNA-sequenties van 250 Nederlandse trio's (vader, moeder, kind) in kaart worden gebracht. De omvang van de sequentiedata is ongeveer 25 terabyte en (tussentijdse) data-analyseresultaten zullen dit significant doen toenemen. We onderzoeken welke strategieën het meest geschikt zijn om, gegeven de data-analyseapplicaties, de overhead van datatransport te minimaliseren.

Inmiddels kunnen onderzoekers een aantal van onze gridapplicaties ook via een webbrowser gebruiken. Dit laat zien dat we echt in staat zijn de complexi-

teit van een IT-onderzoeksinfrastructuur te verbergen zodat de onderzoeker zich kan concentreren op onderzoek. Desalniettemin is verder onderzoek noodzakelijk om het e-BioInfraplatform verder te ontwikkelen en te optimaliseren en te blijven aansluiten bij internationale ontwikkelingen en initiatieven. Tegelijkertijd zullen we gezamenlijk moeten nadenken over nieuwe biomedische vraagstellingen die dankzij steeds geavanceerdere onderzoeksinfrastructuur binnen bereik komen.

Onderwijs

Ik denk dat iedereen het erover eens is dat bioinformaticaonderwijs aan belang toeneemt of zou moeten toenemen gezien het belang van bioinformatica voor biomedisch onderzoek. Een probleem is dat relatief weinig studenten zich specialiseren in de bioinformatica. Dit heeft te maken met de onbekendheid van dit vakgebied bij zowel middelbare scholieren als academische bachelorstudenten. Daarom zijn initiatieven zoals 'Bioinformatica in de Klas' van het Nederlands Bioinformatica Centrum (NBIC)[31] samen met het Centrum voor Moleculaire en Biomoleculaire Informatica (CMBI, Radboud Universiteit) belangrijk. Hier worden middelbare scholen bezocht om leerlingen maar ook hun leraren kennis te laten maken met bioinformatica. Sinds de start van dit initiatief in 2006 hebben meer dan 11.000 leerlingen verdeeld over 274 scholen hieraan deelgenomen. Dit zal zeker bijdragen aan de verdere bewustwording dat biologie en bioinformatica onlosmakelijk met elkaar verbonden zijn. Dit programma verdient mijns inziens brede steun van de universiteiten.

De opleiding Medische Informatiekunde (AMC) kan mijns inziens verder versterkt worden door de disciplines medische informatica, bioinformatica, e-Bioscience, gezondheidszorg en biomedisch onderzoek beter ten opzichte van elkaar te positioneren omdat deze elkaar aanvullen en versterken, maar ook gezien nieuwe ontwikkelingen op het gebied van bijvoorbeeld translationeel onderzoek, personalized medicine, biobanking en het toenemend belang van genetica en commerciële genoomservices in patiëntenzorg.

Voor bioinformatica, e-Bioscience en systeembiologie lijkt mij een vijfjarige opleiding gerechtvaardigd net zoals dit nodig is voor de opleiding van chemici, biologen, (medisch) informatici en natuurkundigen. Als alternatief kunnen we het profiel van de bioloog heroverwegen gezien het feit dat hij of zij steeds meer te maken krijgt met datagedreven en rekenintensief onderzoek. We zouden eigenlijk 50% van het huidige biologiecurriculum moeten vernieuwen met informatica en statistiek. Dit is, vandaag letterlijk, vloeken in de kerk. Beide opties zijn op dit moment niet haalbaar, en voor deze ideeën bestaat ook wei-

nig enthousiasme en draagvlak. Dit werpt de vraag op hoe we dan wel de broodnodige bioinformatici moeten opleiden. Misschien moeten we de Berlijnse muren tussen de verschillende FNWI-instituten laten vallen en op basis van bestaand onderwijs gezamenlijk een vijfjarig bioinformaticacurriculum implementeren.

Tot slot

Ik hoop dat ik u enthousiast heb kunnen maken voor bioinformatica, als u dat al niet was. Ik heb in deze rede vele namen genoemd, maar nog veel meer namen niet genoemd. Wees gerust als u niet bent genoemd, dit was slechts een random steekproef.

Ter overpeinzing wil u meegeven dat het mij onverantwoord lijkt om geld te blijven spenderen aan het produceren van omicsdata zonder ook de bijbehorende bioinformatica en/of e-Bioscience te financieren. Bijvoorbeeld, voor slechts 0,00002 euro per sequentie kunnen wij uw NGS-data analyseren.

Bioinformaticaprojecten vergen vaak langdurige softwareontwikkeling – die soms weken tot maanden in beslag neemt – alvorens de eerste onderzoeksresultaten worden verkregen. Dit komt vaak op de schouders van de onderzoeker en gaat 'ten koste' van wetenschappelijke publicaties. Ik wil ervoor pleiten om ook software-, database- en infrastructuurontwikkeling mee te nemen in onderzoeksevaluaties gezien het belang en gebruik hiervan in biomedisch onderzoek. Ook subsidiegevers doen er goed aan om programma's, coördinatie en financiering hierop af te stemmen. Het Nederlands Bioinformatica Centrum heeft dit in een zeer vroeg stadium onderkend en dit met succes meegenomen in de implementatie en evaluatie van hun programma's.

Ten slotte wil ik u erop wijzen dat deze redevoering één woord bevat dat 35 keer fout gespeld is. Ik laat het aan u om dit woord te vinden, maar wil nu wel stellen dat als een hoogleraar hetzelfde woord zo vaak verkeerd spelt het niet langer een spelfout kan zijn.

Een woord van dank

Hiermee ben ik gekomen aan het eind van mijn rede. Allereerst wil ik alle toehoorders bedanken voor hun aanwezigheid en belangstelling. Ik wil het College van Bestuur, het bestuur van de Faculteit der Natuurwetenschappen, Wiskunde en Informatica, en de Raad van Bestuur van het AMC dankzeggen

voor hun in mij gestelde vertrouwen. Ook het Genootschap wil ik bedanken voor mijn benoeming op deze leerstoel.

Mijn promotieonderzoek in de chemometriegroep van Lutgarde Buydens (Radboud Universiteit) heeft de basis gelegd voor mijn verdere wetenschappelijke carrière. Een crisis in de chemie zorgde dat ik in 1997 belande in het Cluster Software Engineering (AMC) onder leiding van Berend de Vries. De vrijheid die Berend mij gaf heeft geleid tot de bioinformaticagroep waaraan ik nu leiding geef. Berend, het is jammer dat je niet meer in het AMC werkt, maar het gebouw was te klein voor je ideeën. Initiatie van de bioinformaticagroep was niet mogelijk geweest zonder steun van mijn toenmalige collega's Michèle Huijberts en Angela Luyf die nog steeds in mijn groep werkt. Uiteindelijk is mijn groep terechtgekomen bij de afdeling Klinische Epidemiologie, Biostatistiek en Bioinformatica (KEBB). Ondanks een wat moeilijke start vanwege mijn geloof dat mijn groep beter paste bij de laboratoriumdivisie, hebben Patrick Bossuyt en Koos Zwinderman mij altijd steun gegeven wanneer die nodig was. Daarvoor mijn dank. Koos, met jou aan het stuurwiel ben ik er van overtuigd dat we samen het methodologisch onderzoek van de KEBB verder zullen versterken en positioneren als onmisbare schakel in modern biomedisch en klinisch onderzoek.

Inmiddels was ik ook aangesteld als wetenschappelijk directeur van het Nederlands Bioinformatica Centrum (NBIC) waar ik dankzij mijn NBIC-collega's met veel plezier heb gewerkt maar die door hun ondersteuning ook zorgde dat ik mijn werk op het AMC en de FNWI kon blijven doen. Ik wil in het bijzonder Bob Hertzberger bedanken die in mij zijn opvolger zag als wetenschappelijk directeur en met wie ik vier jaar lang constructief en zeer prettig heb mogen samenwerken. Ook wil ik Ruben Kok, managing director van NBIC, bedanken voor de samenwerking en ondersteuning tijdens de soms wel erg drukke periodes. En natuurlijk wil ik Els Natzijl-Visser hier ook niet vergeten. Ik mis je ondersteuning nog dagelijks. Dan Gert Vriend. Zijn unieke en controversiële stijl inzake wetenschapsmanagement, maar ook zijn tomeloze inzet voor NBIC in tijden dat het echt nodig was, hebben mij soms enorm geholpen.

De afgelopen jaren heb ik met veel plezier gewerkt in de Biosystems Data Analysis groep van Age Smilde. Age, ik ben je erkentelijk voor je gastvrijheid. Onze gezamenlijke interesses en enthousiasme zullen ook in de toekomst tot nieuwe projecten leiden.

Collega's uit het bioinformatica- en e-Scienceveld. Jullie aanwezigheid is een hele eer. Collega's uit de natte laboratoria en de kliniek. Jullie zijn met te veel om individueel te bedanken, maar veel van het werk dat ik vandaag heb ge-

presenteerd is in samenwerking met jullie tot stand gekomen. Collega's van de KEBB- en BDA-groep, dank voor jullie aanwezigheid en interesse.

Collega's en oud-collega's van het Bioinformatica Laboratorium. Mede dankzij jullie enthousiasme en inspanning mag ik hier vandaag staan. Ik wil met name Silvia Olabarriaga, Perry Moerland, Andrew Gibson en Barbera van Schaik bedanken voor hun inzet en ondersteuning, maar wil zeker de andere mensen in mijn groep niet tekortdoen.

Dan Recover, mijn huidige band. Mijn dank voor jullie komst. Ik hoop dat we nog vele decennia muziek zullen maken. Rijk zullen we er niet van worden, misschien een dertiende maand. Na vandaag zullen we serieuzer over bioinformatica kunnen praten. Dus, zet hem op!

Dierbare familie, vrienden en vriendinnen. Fijn dat jullie gekomen zijn. Lieve Bas en Kim, ik weet dat jullie soms te weinig tijd van mij krijgen en soms zelfs via Skype met mij moeten communiceren, maar nu weet je wat papa doet. Lieve Karin, zonder jou was dit allemaal veel moeilijker geweest.

Ik heb gezegd.

Referenties

1. http://www.science.uva.nl/ngh/home.cfm
2. http://www.quotegarden.com/star-trek.html
3. Lander E.S. et al., Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860-921
4. Venter J.C. et al., The sequence of the human genome. *Science*. 2001; 291 (5507):1304-51
5. http://www.ornl.gov/sci/techresources/Human_Genome/project/budget.shtml
6. Schneider M.V., Orchard S. Omics technologies, data and bioinformatics principles. *Methods Mol Biol*. 2011; 719:3-30
7. http://en.wikipedia.org/wiki/Information_science
8. Zvelebil M., Baum J.O., *Understanding Bioinformatics*, Garland Science, New York, 2008
9. http://www.bioinformaticslaboratory.nl
10. Schuster S.C. Next-generation sequencing transforms today's biology. *Nat Methods*. 2008; 5(1):16-8
11. http://public.web.cern.ch/public/en/LHC/Computing-en.html
12. Bruggeman F.J., Westerhoff H.V. The nature of systems biology. *Trends Microbiol*. 2007; 15(1):45-50
13. Demir E., et al., The BioPAX community standard for pathway data sharing. *Nat Biotechnol*. 2010; 28(9):935-42
14. Kanehisa M., Goto S., Furumichi M., Tanabe M., Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. 2010; 38(Database issue):D355-60
15. Croft D., O'Kelly G., Wu G., Haw R., Gillespie M., Matthews L., Caudy M., Garapati P., Gopinath G., Jassal B., Jupe S., Kalatskaya I., Mahajan S., May B., Ndegwa N., Schmidt E., Shamovsky V., Yung C., Birney E., Hermjakob H., D'Eustachio P., Stein L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011; 39(Database issue):D691-7
16. Romero P., Wagg J., Green M.L., Kaiser D., Krummenacker M., Karp P.D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*. 2005; 6(1):R2
17. Schellenberger J., Park J.O., Conrad T.M., Palsson B.Ø. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*. 2010; 11:213
18. Ma H., Sorokin A., Mazein A., Selkov A., Selkov E., Demin O., Goryanin I. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*. 2007; 3:135
19. Krebs H.A., Johnson W.A. Metabolism of ketonic acids in animal tissues. *Biochem J*. 1937; 31(4):645-60
20. Swertz M.A., Dijkstra M., Adamusiak T., van der Velde J.K., Kanterakis A., Roos E. T., Lops J., Thorisson G.A., Arends D., Byelas G., Muilu J., Brookes A.J., de Brock E.O., Jansen R.C., Parkinson H. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics*. 2010; 11 Suppl 12:S12

21. Galperin M.Y., Cochrane G.R. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res*. 2011; 39 (Database issue):D1-6
22. Antoniou G., van Harmelen F., *A semantic web primer*, The MIT Press, Cambridge, 2008
23. http://www.w3.org
24. http://www.peroxisomekb.nl
25. http://linkeddata.org
26. Hey T., Trefethen A.E. Cyberinfrastructure for e-Science. *Science*. 2006; 308 (5723):817-21
27. https://www.sara.nl
28. http://www.biggrid.nl
29. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. GenBank. *Nucleic Acids Res*. 2011; 39(Database issue):D32-7
30. http://www.bbmri.nl/nl-nl/activiteiten/projecten/130-genoom-van-nederland
31. http://www.nbic.nl