imec Editie | mei 2017

ExaScience Life Lab helpt farmabedrijven kandidaat-geneesmiddelen opsporen in bestaande data

Uiterst krachtige computers worden ingezet voor de verwerking van de enorme hoeveelheid testgegevens die gegeneerd worden bij de ontwikkeling van geneesmiddelen, om zo tot nieuwe inzichten te komen.

Farmaceutische bedrijven zijn voortdurend op zoek naar geschikte molecules voor de ontwikkeling van geneesmiddelen tegen bepaalde ziektes. Daarvoor onderzoeken ze telkens de effecten van molecules op specifieke biologische processen die met die ziekte te maken hebben. Maar het is misschien ook mogelijk om dezelfde tests veel ruimer op te zetten en veel meer biologische processen te bekijken. Eén van de knelpunten daarbij is de verwerking en datamining van de enorme hoeveelheid testgegevens. Roel Wuyts, senior scientist van het ExaScience Life Lab, legt uit hoe hij met zijn collega's dergelijke knelpunten bij life sciences-toepassingen probeert op te lossen. Aan de hand van een recent project laat hij zien hoe het expertisecentrum de levenskwaliteit van mensen probeert te verbeteren met de uiterst krachtige computers die ons vandaag ter beschikking staan.

Miljoenen beelden doorploegen

Om hun onderzoekspijplijn met nieuwe kandidaat-geneesmiddelen te vullen, gaan farmabedrijven na welke effecten molecules hebben op cellen, de basisbouwstenen van ons lichaam. Zij gebruiken daarvoor schalen met honderden kleine putjes. In elk putje deponeren ze eerst een beetje van de cellencultuur die ze willen onderzoeken en vervolgens één van de molecules waarvan ze het effect willen nagaan. Dan laten ze beide

een tijdje met elkaar reageren, voegen contrast- en kleurstoffen toe en maken één of meer microscopische hogeresolutiebeelden van de putjes.

Die beelden worden dan automatisch verwerkt, waarbij er wordt gekeken naar een aantal morfologische kenmerken van de cellen en hun organellen. Op die manier proberen de onderzoekers het effect van de toegevoegde molecules te achterhalen. Is de cel gegroeid of gekrompen? Is de celwand nog intact of beschadigd? En hoe zit het met de celkern, die het genetisch materiaal bevat?

“Dit proces wordt high-throughput cell imaging (HTI) genoemd,” zegt Roel Wuyts, “en meestal wordt de open-source software CellProfiler gebruikt om de beelden te doorzoeken. CellProfiler werd ontwikkeld door het Broad Institute. De bedoeling was om biologen zonder specifieke opleiding in computervisie of -programmering in staat te stellen om automatisch grote aantallen beelden te verwerken en daarbij specifieke celkenmerken te kwantificeren. Een CellProfiler-script creëert een pijplijn waarin een reeks programma’s na elkaar worden uitgevoerd, waarbij elk programma de output van het vorige als input gebruikt.”

Farmabedrijven noemen deze uitgebreide tests ook wel ‘assay’s’. Ze worden meestal opgezet om de effecten van een biblioteek aan molecules op slechts één specifiek biologisch proces te onderzoeken. Er wordt bijvoorbeeld gekeken waar een bepaalde actieve proteine in een cel aanwezig is hoe dat wordt beïnvloed door de molecule toe te voegen. “Door slechts naar een handvol van alle beschikbare celkenmerken te kijken, maakt men de verwerking van de data eenvoudiger,” zegt Roel Wuyts. “Maar aan deze methode kleeft een groot nadeel: slechts een klein gedeelte van de potentieel beschikbare informatie in deze dure beelden wordt benut.”

Alle informatie aanboren

Onderzoekers van Janssen Pharmaceutica, het Broad Insitute, en een aantal onderzoekspartners hebben recent een project opgezet om te kijken hoe ze met de hoogperformante computers van vandaag meer informatie uit de beelden kunnen halen.

In elke cel vinden vele duizenden biochemische processen plaats. In een assay worden die allemaal blootgesteld aan de geteste chemische stoffen. En veel van de effecten hebben een impact op de celmorfologie en kunnen dus aan de hand van de beelden worden bestudeerd.

De onderzoekers wilden daarom nagaan of ze op basis van de beelden een neutrale vingerafdruk van elk putje konden maken: een beschrijving van de celeigenschappen onafhankelijk van een bepaald biologisch proces. Een dergelijke vingerafdruk zou vervolgens bruikbaar zijn om de activiteiten van alle geteste molecules bij een groot aantal processen te voorspellen.

Maar er is een belangrijk verschil tussen een assay voor één welbepaalde doelstelling en een neutrale vingerafdruk: elk hogeresolutiebeeld - en in één enkele assay kan het om miljoenen dergelijke beelden gaan - moet dan niet op een handvol kenmerken worden gescand, maar op honderden of zelfs duizenden kenmerken. Qua rekenkracht een hele krachttoer! Precies dit knelpunt hebben Roel Wuyts en zijn collega’s helpen oplossen.

“In deze studie,” zegt Roel Wuyts, “keken we op een nieuwe manier naar de beelden die waren genomen om na te gaan welke invloed een half miljoen verschillende molecules hadden op H4-neurogliacellen, een specifiek type kankercellen in de hersenen. De oorspronkelijke bedoeling van het assay was om te onderzoeken welke impact de molecules hebben op de verplaatsing van glucocorticoïdereceptoren van het celcytoplasma naar de celkern. Maar nu keken we dus veel breder.”

“Wij ontvingen beeldvormingsdata van bijna 2.000 schalen, elk met 384 putjes. Dat leverde miljoenen beelden op, in totaal meer dan 10 terrabyte aan data. Het was de bedoeling om CellProfiler te gebruiken om kwantitatieve gegevens voor ongeveer 1.400 kenmerken per beeld te verkrijgen. Maar de CellProfiler-scripts van het Broad Institute werden niet ontwikkeld om zoveel data te analyseren. Ze slagen er met name niet in om voordeel te halen uit een computerinfrastructuur die vandaag de dag kan bestaan uit tientallen computers met een array van hoogperformante processoren, die allen bovendien nog een hele reeks rekenkernen hebben. Wie het CellProfiler-programma daarom ongewijzigd wil inzetten voor een

dergelijke analyse, heeft enorm veel rekentijd en een (te) groot computerbudget nodig. Dit is een van de redenen waarom een dergelijke uitgebreide analyse nooit eerder heeft plaatsgevonden."

De flessenhals wegwerken

Het ExaScience Life Lab heeft een krachtige computercluster met in totaal 32 processoren van elk 36 rekenkernen. Het datacenter van het lab is bovendien gecertificeerd door Janssen Pharmaceutica, dat zowel voor het ExaScience Lab als voor dit project als partner optreedt. Zo'n certificatie is een voorwaarde om dergelijke gevoelige biologische gegevens te mogen verwerken.

"De uitdaging bestaat erin om de berekeningen zo efficiënt mogelijk over alle processoren en nodes van de computercluster te verdelen", zegt Roel Wuyts. "Je kunt dit bijvoorbeeld doen door de programma's te herschrijven, zodat ze maximaal geschikt zijn voor parallelle verwerking. Wij hadden dit eerder al voor een aantal andere processen gedaan, met mooie resultaten. Maar hier hebben we beslist om CellProfiler te gebruiken en aan te passen. Onze experts moesten bijgevolg met diverse kunstgrepen alle processen zo efficiënt mogelijk verdelen over de beschikbare cores, bijvoorbeeld met scripts die CellProfiler in headless mode op alle beelden losliet zonder interactie met een gebruiker, en door het totale werk in kleine pakketjes op te splitsen die door aparte processor cores konden worden uitgevoerd. Op deze manier slaagden wij erin om de rekentijd op onze computer met bijna twee derde in te korten. Wij zien bovendien mogelijkheden om nog meer winst te boeken."

Natuurlijk is beeldverwerking maar één - weliswaar belangrijke - stap in het proces om de resultaten voor nieuwe doeleinden te gebruiken. De onderzoekers van het project kwamen tot de conclusie dat hun "resultaten erop wijzen dat beelden van HTI-screeningprojecten kunnen worden gebruikt om de slaagpercentages in andere projecten op te trekken, zelfs in projecten die geen enkele relatie lijken te hebben met de oorspronkelijke bedoeling van de HTI-screening. Dit zou het mogelijk kunnen maken om specifieke assay's te vervangen door algemenere beeldvormingstechnieken in combinatie met artificiële intelligentie."

"Ons lab was in staat om een flessenhals voor de berekeningen weg te werken en daardoor de praktische haalbaarheid van dergelijke projecten aanzienlijk te verbeteren", zegt Roel Wuyts. "Wij hebben laten zien dat het met zorgvuldige en doelgerichte ingrepen mogelijk is om de computertijd voor het verwerken van zo'n grote hoeveelheid beelden aanzienlijk in te korten. Aangezien precies computertijd en de bijbehorende kosten vaak een remmende factor zijn bij dergelijke life sciences-projecten, zijn wij ervan overtuigd dat wij projecten kunnen mogelijk maken die de gezondheid en het welzijn van veel mensen ten goede zullen komen."

Meer weten?

Het Exascience Life Lab (www.exascience.com) is een expertisecenter voor veeleisende big data gegevensverwerking in de sector van de life sciences. Het laboratorium, dat ondergebracht is bij imec, werd in 2013 opgericht als een gezamenlijk initiatief van Intel, Janssen Pharmaceutica, imec en alle Vlaamse universiteiten. Het heeft als kernopdracht om knelpunten in rekenkracht bij softwaretoepassingen weg te werken, zodat ze kunnen worden ingezet om problemen in de sector van de life sciences op te lossen. Voor dit specifieke project werd steun verleend door IWT130405 ExaScience Life Pharma, IWT130406 ExaScience Life HPC en IWT150865 Exaptation van VLAIO, het Vlaamse Agentschap Innoveren en Ondernemen.

Biografie Roel Wuyts

Roel Wuyts is wetenschapper bij imec en part-time hoogleraar bij imec- DistriNet – KU Leuven. Zijn belangrijkste onderzoeksdomein is de runtime-managementlaag van toekomstige hoogpresterende computerhardware. Vóór hij in dienst trad bij imec was hij geassocieerd professor aan de ULB (Université Libre de Bruxelles). Hij behaalde zijn doctoraat in de computerwetenschappen aan de VUB (Vrije Univeriteit Brussel). Roel Wuyts maakte deel uit van diverse programmacomités van conferenties zoals ECOOP, OOPSLA, SC, Net.ObjectDays of ESUG, organiseerde workshops zoals de DATE'08 Workshop over Software Engineering for Embedded Systems en beoordeelde papers voor TOPLAS en TOSEM.