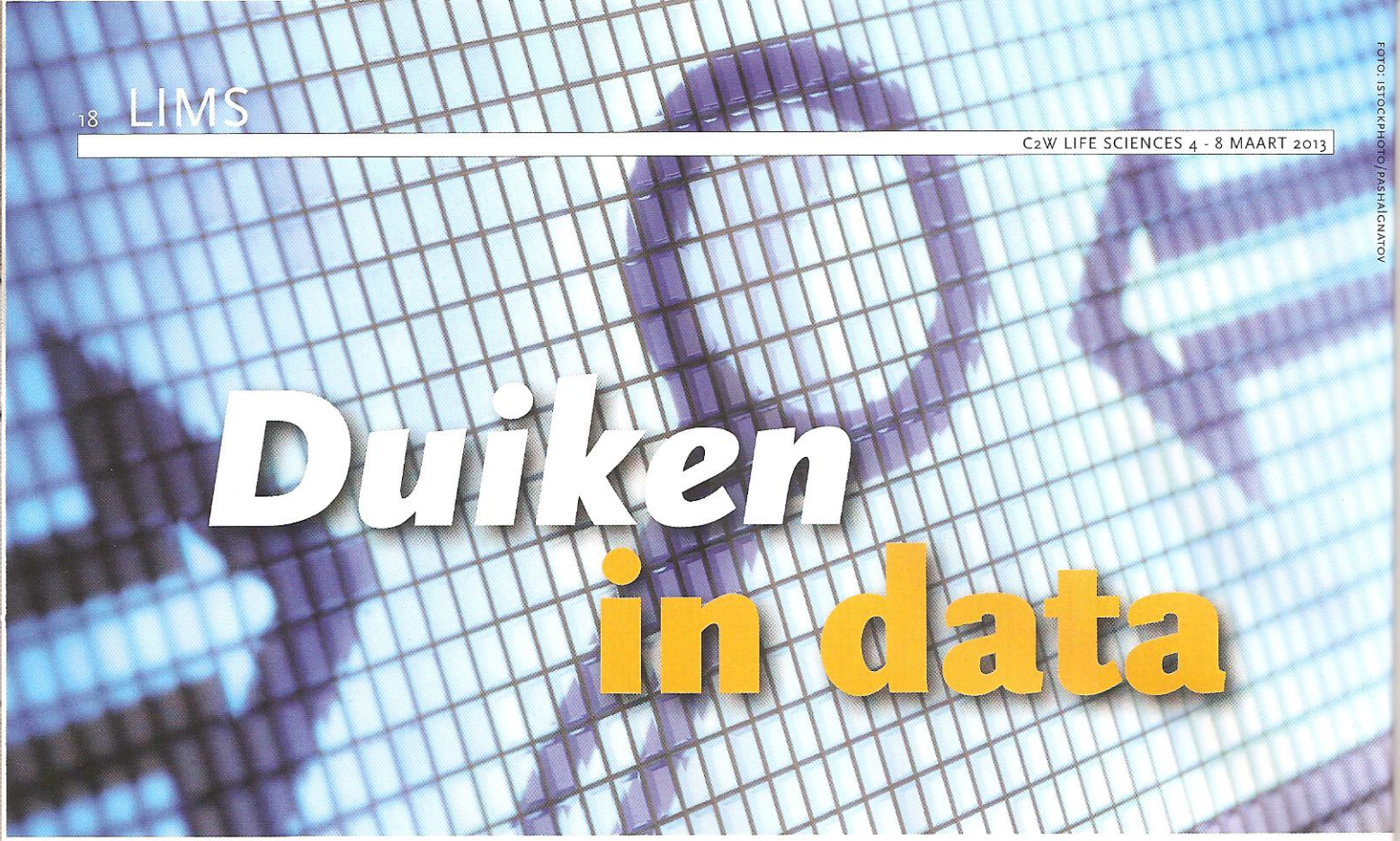


FOTO: ISTOCKPHOTO/PASHAIGNATOV

Duiken in data

Oude data bewees recent zijn nut in drug design. Nieuwe data – *big data* – moet dat ook gaan doen voor de personalized medicine.

SHURAILA STOPPEL

Spitten in bestaande databases biedt perspectieven voor de geneesmiddelenontwikkeling. Amerikaanse en Britse onderzoekers lieten in *Nature* zien hoe zij een bestaand medicijn, Donezepil, – een geheugenverbeteraar bij alzheimer – vervolmaakten. Met een lijst bekende molecuulgegevens bij de hand, vijzelden ze receptorbindingseigenschappen op en hielden daarbij rekening met de farmacologische eigenschappen, zoals opname door de hersenen. Het onderzoeksteam gebruikte de ChEMBL-database. Een open-dataregister van het European Bioinformatics Institute (EBI) dat sinds 2009 bestaat (zie kader).

BIOACTIEVE MEETWAARDEN

Herman van Vlijmen, hoogleraar *computational drug discovery* aan de Universiteit Leiden, heeft naar aanleiding van de *Nature*-publicatie, een duidelijke mening over de ChEMBL-database: "Als je er zulk mooi onderzoek mee kunt doen dan blijkt toch maar weer eens hoe belangrijk die database is."

Een centrale verzamelplaats voor big data, uit het biomedische veld, is veel gecompliceerder dan de, volgens Van Vlijmen op dit moment betrekkelijk simpele, ChEMBL-database. Biomedische data komt uit laboratoria met massaspectrometers, NMR-machines, sequencers en microarrayscanners, al dan niet gekoppeld aan een laboratoriuminformatiemanagementsysteem (LIMS) en bestaat uit grote digitale bestanden. Een recent voorbeeld van een centrale biomedische database is de humane genomische encyclopedie van het ENCODE-consortium (*Nature* richtte hiervoor een aparte site op). Daarnaast leveren moderne medische onderzoeksmethoden zoals PET-, MRI-, CT-scans en digitale pathologie beeldmateriaal op met steeds hogere resoluties. Biomedische big data die men kan gebruiken voor *personalized medicine*.

> 'Het blijkt hoe belangrijk die database is'

Biomedische onderzoeksgegevens die zijn afgestemd op de individuele patiënt vormen het uitgangspunt bij personalized medicine. Een voorbeeld daarvan zijn genomische patiëntgegevens die kunnen voorspellen of een patiënt baat heeft bij een bepaalde therapie. Biomedische big data dient ook als basis voor verbeterde diagnostische tests, moderne prognostische parameters en innovatieve therapieën.

COMBINEREN

Wil je de juiste conclusies trekken, dan moet je resultaten kunnen combineren. Door het feit dat die data in elk onderzoeksveld en voor elk experiment anders zijn, kan die combinatie echter lastig

CHEMBL-DATABASE

ChEMBL bevat een verzameling gegevens van bioactieve geneesmiddelachtige kleine moleculen, 2D-structuren, berekende eigenschappen zoals molecuulgewicht, fysische parameters en experimenteel gemeten eigenschappen zoals bindingsconstanten, farmacologische en farmacokinetische eigenschappen.

De ChEMBL-database bevat momenteel 5,4 miljoen bioactieve meetwaarden voor meer dan een miljoen stoffen en 5.200 doeleiwitten. De ChEMBL-groep bij EBI voegt nieuwe data, uit gepubliceerde literatuur, handmatig en op regelmatige basis toe.

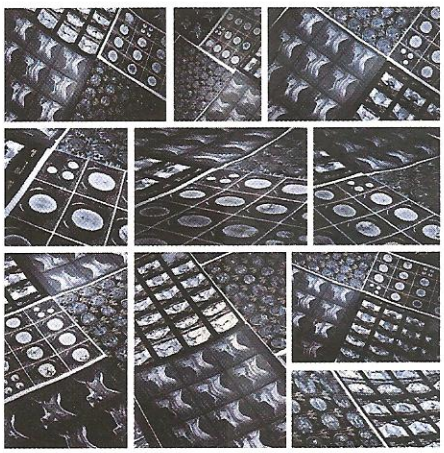
www.ebi.ac.uk/chembl

worden. Wil je bijvoorbeeld achter het ontstaan komen van een complex ziektebeeld zoals alzheimer, dan heb je als onderzoeker informatie nodig die varieert van de cognitieve status en MRI-scans van patiënten tot informatie over genetische of moleculaire defecten en farmacologische studies. Daarnaast wil je al die biomedische informatie van verschillende personen kunnen vergelijken.

VERZAMELEN

Om dat proces te vergemakkelijken gaan de onderzoekers van het publiek-private samenwerkingsverband Commit een datamanagementsysteem aanleggen. Hiermee gaan ze digitaal materiaal van verschillende patiënten voor bepaalde ziektes verzamelen. Commit noemt deze datasystemen virtuele weefselbiobanken. Een virtuele weefselbiobank bevat datasets met beelden van de moleculaire histologie van bijvoorbeeld tumorweefsel van verschillende patiënten. Voor de moleculaire histologie gebruikt men verschillende *imaging*modaliteiten. De ene techniek vertelt iets over lipidensamenstelling in het weefsel, de ander over eiwittenexpressie, biomarkers of DNA. Naast alle informatie over weefselsamenstelling heb je vaak de beschikking over analytische informatie uit patiëntenbloed en/of -urine. Door deze informatie aan elkaar te koppelen, ontstaan voor verschillende ziektes enorme gegevensbronnen. Daaraan kan een arts de toestand van de individuele patiënt toetsen en een behandelplan opstellen dat op het individu gericht is.

Ron Heeren van het FOM-instituut AMOLF is projectleider bij Commit en werkt voornamelijk met moleculaire beelden. Die beelden zijn afkomstig van massaspectrometrische imaging (MSI). Een techniek die met behulp van massaspectrometrie



allerlei biologische en farmacologische bestanddelen in weefselcoupes zichtbaar kan maken. De databestanden zijn groot door hun hoge resolutie. Heeren: "Een enkel MSI-experiment kan bijna 500 gigabite aan data kan opleveren. Een kleine patiëntenset van 100 mensen bevat dus al snel tientallen terabytes." Uit die informatiebrij identificeert Heeren herkenbare (eiwit-)patronen.

> 'Een kleine dataset bevat al snel tientallen terabytes'

Ook Jan-Willem Boiten van het Center for Translational Molecular Medicine (CTMM) vindt het koppelen van data belangrijk. Als projectleider van het consortium TraIT werkt hij aan een IT-infrastructuur voor toegepast diagnostisch onderzoek op het gebied van alzheimer, oncologie en cardiovasculaire aandoeningen. Zijn project heeft als doel de digitale samenwerking tussen verschillende grote partijen, zoals ziekenhuizen en onderzoeksinstituten, te verbeteren. Daarvoor biedt TraIT een stelsel van hardware- en softwareoplossingen om ervoor te zorgen dat grote instituten de gegevens van experimenten en klinische studies op dezelfde manier verwerken en opslaan.

Boiten: "Door nieuwe IT-technologieën kunnen we data – zoals MRI-scans, CT-scans en sequencingdata – uit verschillende bronnen combineren, iets dat vroeger niet kon."

Van Vlijmen is voorstander van vrij toegankelijke databases. Hij werkt zelf met de eerder genoemde ChEMBL-database en is blij met de toegankelijkheid die het biedt. "Voorheen was er buiten de informatie van farmaceutische bedrijven maar heel weinig van dit soort informatie verkrijgbaar. Dit soort databases zijn essentieel in drug-designonderzoek."

Heeren vindt het zelfs niet meer van deze tijd om data voor jezelf te houden. "Al het biochemisch onderzoek en farmaceutisch onderzoek vraagt om het samenkomen van verschillende informatiestromen. Gegevens van genomics- en proteomicsstudies, beelddata, patiëntendatabases en klinische data zijn nodig om antwoorden te geven op vragen over het type kanker van een patiënt en hoe we die kunnen behandelen. Ik zou graag zien dat al dit soort data vrij toegankelijk wordt, maar er zou wel een goede bronvermelding voor de leveranciers van de gegevens moeten zijn."

CAPACITEIT

Data delen is dus belangrijk voor patiëntgericht onderzoek en de ontwikkeling van personalized medicine. Door het enorme volume kan het delen van big data moeilijker zijn, maar er ontstaan de laatste tijd veel samenwerkingsverbanden zoals Commit en TraIT om de juiste dataverwerkingsmethoden en goede datamanagementsystemen van de grond te krijgen.

In Amerika leidde een samenwerkingsverband tussen IBM en Memorial Sloan-Kettering Cancer Center (MSKCC) in New York al tot een beginnend succes. De oncologen van MSKCC hebben sinds februari de beschikking over de rekenkwaliteit van Watson, de supercomputer van IBM. Deze zet men in als hulpmiddel bij de behandeling van kanker. Watson kan rekenen met de data van de 1,2 miljoen patiëntrapporten die MSKCC de laatste decennia in een database verzamelde. Bovendien bevat de computer nu al meer dan twee miljoen pagina's uit medische tijdschriften en klinische trials. Watson kan deze gegevens combineren en vergelijken, om zo tot het beste behandelingsvoorstel voor een patiënt te komen. Zo maakt men in New York nu optimaal gebruik van biomedische big data.