

DE UNIVERSELE TAAAL VAN HET MOLECUUUL

Al het leven spreekt een universele taal in geuren, kleuren en smaken: de taal van de moleculen, aldus Justin van der Hooft, universitair docent bij de Bioinformatics Group. Het mag dan universeel zijn, begrijpen doen we het allerminst. Van der Hooft wil daar verandering in brengen met zijn onderzoek in metabolomics.

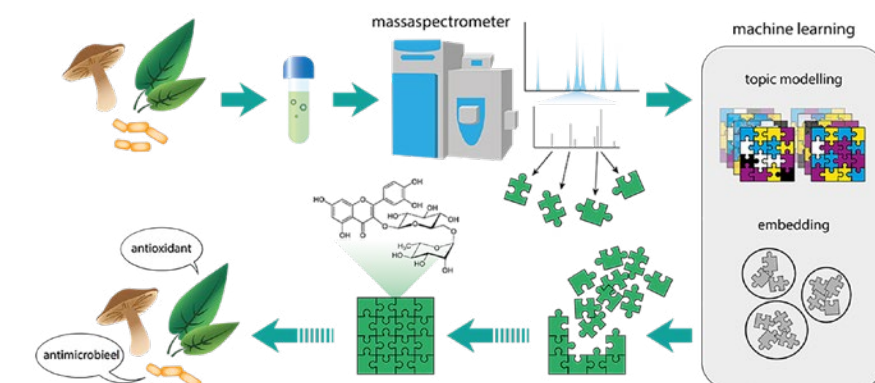
Tekst en illustratie Stijn Schreven • Foto Eric Scholten

Wat is metabolomics?

Van der Hooft: 'Het is de vierde in het kwartet van omics, naast genomics, transcriptomics en proteomics. De eerste bestudeert DNA, de tweede de transcrip-ten van DNA, oftewel RNA, en de derde bekijkt de eiwitten die daaruit gemaakt worden. Metabolomics karakteriseert mengsels van kleine moleculen zoals glucose. Bij mengsels kun je denken aan wat bacteriën uitscheiden, maar ook aan plantenextracten zoals koffie en thee en aan urinemonsters. Het doel is dat we uiteindelijk de taal van die mengsels begrijpen: de functies, de werkzame stoffen en in geval van urine bijvoorbeeld kunnen aflezen wat iemands dieet of gezondheid is.'

Wat boeit je aan metabolomics?

'De verscheidenheid van moleculaire vormen fascineert me. Dat zulke kleine verbindingen zo'n impact kunnen hebben. Soms verschilt de structuur van twee moleculen enkel in één groep die een andere kant op wijst, maar maakt dat een enorm verschil in werking of geur. Een fascinerend voorbeeld zijn (-) en (+)-menthol: de eerste vorm ruikt als pepermunt,



de andere is bitter. Dat wil ik graag leren begrijpen. Verder is dit veld multidisciplinair: je bent bezig met analytische chemie, statistiek, *machine learning* en chemische informatica.'

Je wilt de taal van kleine moleculen ontcijferen. Hoe?

'Mijn groep werkt aan *computational metabolomics*: wij ontwikkelen de tools om metabolomics-data te analyseren. Die datasets komen uit speciale apparaten zoals massaspectrometers. Daarin gaat een mengsel van moleculen die binnen botsen met edelgas en daarbij uiteenvallen in brokstukken. Die fragmenten zien we terug als pieken in een spectrum. De plek en hoogte van de piek in het

spectrum zeggen iets over de vorm van het brokstuk en diens hoeveelheid in het molecuul. Vervolgens is de vraag welke brokstukken dat zijn en welk molecuul zij samen vormen. We kunnen dus niet direct zien welke moleculen er in het mengsel zaten, maar moeten eerst de brokstukken bij elkaar puzzelen en begrijpen hoe deze in elkaar passen. Een molecuul kan bestaan uit verschillende brokstukken en een brokstuk kan als bouwsteen in verschillende moleculen voorkomen.'

Hoe los je die puzzels op?

'Eerst gebeurde dit handmatig. Tijdens mijn PhD zocht ik precies uit welke fragmenten bij welk molecuul hoorden. Zulke

studies vormen wereldwijd inmiddels een grote databank van 16.000 verbindingen die we volledig hebben uitgezocht, een verzameling opgeloste puzzels. Die verzameling groeit gestaag maar langzaam, omdat het onderzoek tijdrovend is. Sinds kort gebruiken we *machine learning*. Hiermee automatiseren en versnellen we het proces. Je geeft de computer de data en de labels van wat het is en dan wens je hem succes. De computer leert zelf de patronen herkennen. We gebruiken twee methoden van *machine learning* die beide geïnspireerd zijn op tekstanalyse.’

Tekstanalyse? Om welke methoden gaat het?

‘*Topic modelling* probeert de onderwerpen uit een tekst te halen aan de hand van de woorden die het meest voorkomen. In de metabolomics achterhaalt het de brokstukken van moleculen aan de hand van fragmenten die vaak samen voorkomen in spectra. Daarnaast ontwikkelen we een nieuwe techniek gebaseerd op *word embedding*, die kijkt naar de context van woorden en zo beslist of zinnen op elkaar

lijken. Bijvoorbeeld “ik hou van koffie en een koekje” en “ik hou van cappuccino en cake” – de woorden verschillen maar de zinnen betekenen bijna hetzelfde.

Vergelijkbaar proberen we in de metaboloomdata op basis van de brokstukken (de woorden) de chemische klassen (betekenis) te achterhalen zonder de volledige moleculen (de zinnen) in elkaar te hoeven zetten. Chemische klassen zijn bijvoorbeeld flavonoïden en alkaloiden. Het is als het vinden van de hoekpunten en randen van de puzzel: de structuur van het molecuul staat dan grotendeels vast en helpt bij het oplossen van de rest van puzzel.’

Hoe ver zijn jullie in het ontwikkelen van deze tools?

‘Begin dit jaar publiceerden we een eerste methode die *word embedding* toepast in metabolomics. Momenteel schieten de *machine learning*-studies als paddenstoelen uit de grond, elke maand verschijnt wel een nieuwe publicatie. Voor eiwitten werd onlangs AlphaFold 2 gelanceerd, een *machine learning*-techniek die 3D-structuren van eiwitten 15 tot 20 procent nauwkeuriger kan voorspellen.

In plaats van maanden labwerk, kost het soms slechts tien minuten om te weten hoe een eiwit eruitziet. Het is een kwestie van tijd voordat er een soortgelijke doorbraak komt in de metabolomics.’

Waar wil je uiteindelijk heen met dit onderzoek?

‘Mijn groep richt zich op de structuren en functies van natuurlijke producten om bijvoorbeeld nieuwe antibiotica te vinden. Uiteindelijk wil ik die moleculaire puzzels oplossen om te begrijpen waarom een ecosysteem werkt zoals het werkt, welke taal er wordt gesproken. Bijvoorbeeld, welke functies heeft een plantenextract met bepaalde flavonoïden? Als we dat eenmaal weten, kunnen we ook gaan sturen. Zo zou je met de inbreng van de juiste bacteriën en schimmels een bodem tolerant kunnen maken voor zoutstress, droogte of hitte, zodat het de functies behoudt en planten blijven groeien. Dat is nu nog ver weg, dat is echt toekomstmuziek.’ ■

