



Universiteit  
Leiden  
The Netherlands

## Exploring the chemical space of post-translationally modified peptides in *Streptomyces* with machine learning

Kloosterman, A.M.

### Citation

Kloosterman, A. M. (2021, May 12). *Exploring the chemical space of post-translationally modified peptides in Streptomyces with machine learning*. Retrieved from <https://hdl.handle.net/1887/3170172>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3170172>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <https://hdl.handle.net/1887/3170172> holds various files of this Leiden University dissertation.

**Author:** Kloosterman, A.M.

**Title:** Exploring the chemical space of post-translationally modified peptides in *Streptomyces* with machine learning

**Issue Date:** 2021-05-12

## Nederlandse samenvatting

### Secundaire metabolieten als bron van antibiotica

Planten, dieren en micro-organismen produceren een grote diversiteit aan metabolieten en andere natuurstoffen [1, 222]. Sommige van deze moleculen zijn essentieel voor elk organisme. Suikers, vetten en nucleotiden, bijvoorbeeld, vormen de bouwstenen voor het leven en worden de primaire metabolieten genoemd. Onder secundaire metabolieten worden alle overige natuurstoffen verstaan. Meestal bieden deze moleculen organismen een voordeel onder specifieke omstandigheden of helpen ze de communicatie met andere soorten. Dit leidt tot een enorme chemische diversiteit aan moleculen. Tegelijkertijd zijn de meeste varianten relatief zeldzaam, vooral vergeleken met primaire metabolieten, aangezien ze alleen in specifieke niches voordeel kunnen bieden [2].

De precieze functie van secundaire metabolieten verschilt erg van moleculair tot moleculair, maar vele ervan werken als moleculaire wapens, die in staat zijn bacteriën, schimmels of virussen te doden of hun groei te remmen [2]. Antivirale en antibacteriële metabolieten kunnen bescherming bieden tegen infecties, en hun productie door hogere organismen is daarmee geen verassing. Toch zijn het de micro-organismen zelf die het grootste deel van deze metabolieten produceren, vermoedelijk als wapens en bescherming tegen andere micro-organismen. Daarnaast kunnen sommige van deze stoffen dienen als immunosuppressiva of de groei van tumoren remmen. Het zijn deze eigenschappen die zulke moleculen interessant maken voor gebruik in de klinische omgeving, met toepassing in de behandeling van onder meer infectieziekten, tumoren of immuunziekten. Sinds de ontdekking van penicilline in de jaren '20 [7] is men erin geslaagd een groot aantal verschillende antibiotica te isoleren uit bacteriën en schimmels, die nog steeds in de kliniek gebruikt worden. Dit was een grote revolutie voor de geneeskunde: bacteriële infecties die eerder een doodsvonnis betekenden, konden opeens effectief worden behandeld met een simpele kuur. Het grote succes van antibiotica heeft geleid tot hun wereldwijde gebruik en heeft talloze mensenlevens gered [227].

Helaas zit er een keerzijde aan het wijdverbreide gebruik van antibiotica. Bacteriën in de natuur worden steeds meer blootgesteld aan antibiotica die in het milieu terecht zijn gekomen. Resistente varianten van deze bacteriën hebben daardoor een evolutionair voordeel. Dit heeft als gevolg dat genen en mutaties die resistentie verschaffen tegen antibiotica zich steeds wijder verspreiden [14]. Multiresistente pathogenen zoals MRSA (methicillin-resistent of multi-resistent *Staphylococcus Aureus*) en MDR-TB (multi-drug resistant tuberculosis) veroorzaken infecties die moeilijk te behandelen zijn en ons technologisch terugbrengen naar de tijd van voor de antibiotica [17]. Een rapport van O'Neill, opgesteld in opdracht van de Britse regering, voorspelt dat in het jaar 2050 meer dan 10 miljoen mensen jaarlijks zullen overlijden aan ziektes veroorzaakt door multiresistente bacteriën, meer dan aan de gevolgen van kanker [21]. Om deze scenario's te voorkomen zijn nu meer dan ooit nieuwe soorten antibiotica nodig, waartegen nog geen resistentie is ontwikkeld.

De meeste antibiotica zijn ontdekt in de jaren '50 en '60 van de vorige eeuw, een periode waar vaak naar wordt verwezen als de Gouden Eeuw voor ontdekking van antibiotica. De methoden waren conceptueel simpel: de potentiële producenten van antibiotica, meestal bacteriën en schimmels, werden op grote schaal en in diverse groeiomstandigheden opgekweekt en chemisch geëxtraheerd. Het extract werd daarna toegevoegd aan groeiende bacteriën, om te zien of de groei daarvan beïnvloed werd. Als dit het geval was, konden de verantwoordelijke stoffen gezuiverd worden uit het extract, en onderzocht worden op hun potentie als nieuw antibioticum [158]. Veel van deze stoffen werden geïsoleerd uit Actinomyceten, een fyllum van bacteriën, en met name uit het genus *Streptomyces*. De groei van streptomyceten lijkt op die van schimmels: in de aarde vormen ze netwerken van langwerpige hyfen, het bacteriële equivalent van schimmeldraden. Als de voedselbronnen opraken rondom dit netwerk, begint de tweede levensfase van de streptomycete. Hierin breekt hij zijn eigen netwerk van cellen af en gebruikt de bouwstoffen om sporen te vormen, die weer verspreid kunnen worden, waarna deze tot een nieuw netwerk kunnen uitgroeien [32]. Streptomyceten blijken uitzonderlijke producenten van secundaire metabolieten: meer dan twee derde is afkomstig van dit genus [158].

Hoe succesvol de methode voor het vinden van nieuwe antibiotica andere medicijnen ook was, de laatste decennia is het aantal nieuwe antibiotica dat hiermee gevonden is drastisch gedaald. Dit komt mede door de herontdekking van veel antibiotica: veelvoorkomende antibiotica wordt telkens opnieuw gevonden, wat de ontdekking van nieuwe en meer zeldzame antibiotica belemmert. Onderzoek wordt daardoor steeds minder rendabel voor de farmaceutische industrie, waardoor er minder wordt geïnvesteerd en het aantal ontdekkingen nog sneller daalt. Om het rampscenario van snelle toename van resistentie en een gebrek aan goede medicijnen te voorkomen is duidelijk dat er nieuwe methodes nodig zijn voor de ontdekking van nieuwe antibiotica [24, 25].

## De analyse van de genomen van antibioticaproductanten: een nieuwe revolutie

Een grote doorbraak werd bereikt toen de sequentie van het DNA van *Streptomyces coelicolor*, een modelorganisme voor alle streptomyceten, werd bepaald [27]. Uit het genoom bleek dat deze bacterie in staat was tot het produceren van wel dertig verschillende secundaire metabolieten, terwijl er toen nog maar enkele uit deze stam gezuiverd waren. Het aantal producten dat door één stam geproduceerd kan worden, wordt bepaald aan de hand van het aantal biosynthetische genclusters (BGCs) dat gevonden wordt [26]. Een BGC is een verzameling genen die naast elkaar op het genoom liggen, en die vertaald worden in eiwitten die allen betrokken zijn bij de productie van één groep secundaire metabolieten, of zelfs één specifiek metaboliet. Nu het bepalen van de DNA-sequentie aanzienlijk goedkoper is geworden door de ontwikkeling van nieuwe technieken, worden steeds meer genoomsequenties bepaald, en steeds meer BGCs gevonden [264, 275]. Van de meeste hiervan is het product niet bekend. Deze worden ook wel de cryptische BGCs genoemd [31], en zij representeren een potentiële bron van nieuwe secundaire metabolieten en dus nieuwe antibiotica.

Nieuwe BGCs kunnen gevonden door te zoeken naar genen die lijken op genen uit al gekarakteriseerde BGCs. Hoewel de meeste secundaire metabolieten grote chemische verschillen tonen, zit er vaak wel overlap in de

manier waarop ze geproduceerd worden. Deze overlap in productiemethoden leidt ook tot overlap op genetisch niveau, waarvan gebruik gemaakt wordt door software die BGCs in genomen detecteert. De beperking van deze methode is dat er dus altijd een overlap moet zitten tussen bekende BGCs en nieuw gedetecteerde BGCs, en dat compleet nieuwe klassen niet gevonden zullen worden. Als zulke klassen bestaan, zullen we ze alleen kunnen vinden door nieuwe methoden toe te passen om BGCs te detecteren, waarbij we buiten gebaande wegen moeten treden [276].

Een subklasse van metabolieten die goed geschikt lijkt voor het vinden van nieuwe varianten, is die van de ribosomaal gesynthetiseerde en posttranslationaal gemodificeerde peptiden (RiPPs) [42, 48]. RiPPs zijn chemisch divers, maar delen een bepaalde biosynthetische logica [44]. De basis van een RiPP wordt gelegd door een klein eiwit (vaak korter dan 100 aminozuren), dat net als andere eiwitten door het ribosoom wordt geproduceerd. Dit eiwit, ook wel de precursor genoemd, wordt vervolgens uitgebreid gemodificeerd door andere eiwitten waarvan de coderende genen ook in het BGC aanwezig zijn. Na modificatie wordt er een groot deel van afgeknipt, en het complete product, de RiPP, wordt geëxporteerd. Opvallend aan RiPPs is de grote diversiteit van hun producten. Zowel de sequentie van de precursor, als de modificaties kunnen sterk uiteenlopen, en beiden bepalen de structuur van het uiteindelijke product. Bovendien hebben veel RiPPs, net als andere secundaire metabolieten, ook antibacteriële of antivirale eigenschappen.

RiPPs worden onderverdeeld in subklassen, die elk hun eigen specifieke modificaties hebben. Zo bevatten lanthipeptiden thioetherbruggen tussen cysteïnes en serines of threonines [47] en staan lassopeptiden bekend om hun structuur die een knoop vormt [45]. De BGCs die horen bij de verschillende subklassen coderen elk voor verschillende modificerende enzymen en tonen grote verschillen. Dit heeft gevolgen voor de zoekstrategie op genetisch niveau. Over het algemeen geldt dat met informatie van een BGC wel andere BGCs van die subklasse gevonden worden, maar geen BGCs van een andere subklasse. Desalniettemin worden er nog wel vaak nieuwe subklassen ontdekt: in de afgelopen zes jaar is het aantal bekende subklassen verdubbeld van 20 naar 40 [42, 48]. Een zoekstrategie voor BGCs van RiPPs die juist nieuwe subklassen

zoekt lijkt daardoor veelbelovend, maar tot nog toe bestaan zulke strategieën niet. In het kader van Syngeno pep, een onderzoeksvorstel gericht op de ontdekking van nieuwe antimicrobiële peptiden, is hier verder onderzoek naar gedaan. Juist door nieuwe methoden te ontwikkelen voor de detectie van RiPP BGCs, die onafhankelijk zijn van de RiPP subklasse, wordt er gericht op de ontdekking van nieuwe subklassen en dus ook nieuwe soorten peptiden. Hiervoor wordt onder andere gebruik gemaakt van kunstmatige intelligentie en exploratief genetisch onderzoek. Het resultaat van dit onderzoek staat in dit proefschrift gepresenteerd.

## Identificatie van nieuwe RiPP-subklassen via detectie van de precursors

De meeste RiPP-subklassen worden gekarakteriseerd door sterk uiteenlopende modificaties en precursoreiwitten. Desalniettemin zijn er elementen die overlappen tussen verschillende RiPP-subklassen, waarvan gebruik gemaakt kan worden voor de detectie van nieuwe subklassen. De belangrijkste hiervan is ongetwijfeld het gen dat codeert voor de precursor, die de basis legt voor het uiteindelijke product. Omdat de sequentie van de precursors zo sterk varieert, wordt er steeds meer gebruik gemaakt van kunstmatige intelligentie om precursors van andere eiwitten te onderscheiden. Hierbij worden niet alleen de sequenties van de precursors vergeleken, maar ook berekende eigenschappen zoals lading, hydrofobiciteit, lengte en frequentie van verschillende aminozuren. Verschillende modellen zijn al eerder gerapporteerd, zoals NeuRiPP en NLPPrecursor [88, 89]. In Hoofdstuk 3 wordt het model van **decRiPPter** (Data-driven Exploratory Class-independent RiPP TrackER) beschreven. Dit model is gebaseerd op een Support Vector Machine (SVM) en kan precursors identificeren van vele verschillende subklassen, soms ook als deze niet in de trainingsset voorkomen. Dit suggereert dat er eigenschappen zijn die deze precursors gemeenschappelijk hebben, ongeacht de subklasse waar ze toe behoren. Dat impliceert dat de precursors van compleet nieuwe subklassen ook met dit model te vinden moeten zijn.

DecRiPPter borduurt voort op het bovenstaande idee om nieuwe RiPP-subklassen te vinden. Het is daarmee het eerste programma dat precursors

gebruikt als basis voor de detectie van RiPP-subklassen in plaats van de omliggende genen die coderen voor eiwitten betrokken bij de verdere productie van de RiPP. Een obstakel aan deze methode is het aantal kleine genen dat mogelijk voor precursors codeert. Uit de analyse van 1.295 *Streptomyces* genomen werden meer dan 71 miljoen kleine genen gevonden, terwijl een ruime schatting (10 per genoom) niet meer dan 13 duizend precursors voorspelt. Zelfs als het model maar in 0,1 procent van de gevallen een valspositief resultaat zou geven, zou dat betekenen dat het aantal valspositieven vele malen groter zou zijn dan het aantal echte precursors (71 duizend tegenover 13 duizend). In werkelijkheid werden iets meer dan 832 duizend mogelijke precursors gevonden, waarvan het merendeel waarschijnlijk valspositief is. Om deze reden moeten de resultaten verder gefilterd worden.

Er zijn vele methoden denkbaar om precursors te filteren, bijvoorbeeld op basis van precursorsequentie. Een eigenschap die de precursors van een aantal subklassen hebben is de aanwezigheid van meerdere kernsequenties [51, 239, 267]. Elk van deze sequenties wordt verwerkt tot een RiPP, terwijl de rest eraf geknipt wordt door een protease. Deze kernsequenties lijken binnen één precursor wel vaak op elkaar en de aanwezigheid hiervan zou een goede aanwijzing zijn dat het eiwit daadwerkelijk een precursor is. Dit kenmerk is alleen niet universeel voor alle RiPPs en het gebruik van deze filter zou dus betekenen dat precursors van veel RiPPs verwijderd zouden worden. Dit is de afweging die typerend is voor het exploratieve onderzoek beschreven in dit proefschrift: geen enkele filter is op zichzelf perfect, maar elke filter kan wel bepaald voordeel bieden en het deel van de data laten zien waar de gebruiker interesse in heeft.

### Het gebruik van de genetische context voor prioritering van nieuwe RiPP BGCs

Alleen de precursor gebruiken om nieuwe RiPP-subklassen te vinden is niet nauwkeurig genoeg. DecRiPPter maakt gebruik van de genetische context van de precursorgenen om de resultaten verder te filteren. Hierin wordt gekeken of er genen aanwezig zijn naast de precursors, die typisch zijn voor RiPP BGCs en waarvan de producten betrokken kunnen zijn bij RiPP biosynthese. Dit soort



genen moeten dus coderen voor modifierende eiwitten, transporteiwitten, regulerende eiwitten en peptidases, die de precursor knippen. Daarnaast wordt er een eis gesteld aan de frequentie van de genen: als deze in het merendeel van de geanalyseerde genomen voorkomen, is het waarschijnlijk dat ze betrokken zijn bij het primaire en niet bij het secundaire metabolisme. Prioriteit wordt daarom gegeven aan genclusters die maar zelden voorkomen, maar wel zoveel mogelijk elementen bevatten die typerend zijn voor RiPPs. Een laatste eis is dat de genclusters niet moeten lijken op die van al bekende RiPP-subklassen. Hiervoor worden de genomen ook geanalyseerd met antiSMASH, software die bekende RiPP BGCs detecteert [39].

Een grondige analyse van 1.295 *Streptomyces* genomen resulteerde in grofweg 700.000 potentiële RiPP BGCs, die niet allemaal handmatig geëvalueerd konden worden. Door steeds strengere eisen te stellen aan de eiwitten die door een gencluster gecodeerd moesten worden, kon de lijst kandidaten worden teruggebracht naar een werkbare hoeveelheid. Helaas bevat niet elk RiPP BGC elk kenmerk (bijvoorbeeld dat sommige RiPP BGCs geen gen voor een regulatie-eiwit of voor een peptidase bevatten), dus door de strengere filters werden deze ook gefilterd. Wel nam het percentage bekende RiPP genclusters toe naarmate strengere filters werden gebruikt, een signaal dat onder de overgebleven kandidaten waarschijnlijk steeds meer nieuwe subklassen gevonden kunnen worden. De dataset werd teruggebracht tot een lijst van enkele honderden genclusters, die handmatig geëvalueerd werden. Dit resulteerde in 151 genclusters gegroepeerd in 42 nieuwe RiPP-subklassen. Eén van deze subklassen is experimenteel gevalideerd en hiervan bleek het inderdaad om een nieuwe RiPP-subklasse te gaan, een nieuwe variant van de lanthipeptiden. Een andere lid van deze subklasse werd rond dezelfde tijd ontdekt via meer traditionele methoden. Hieruit bleek dat ook deze subklasse kandidaten bevat die antimicrobiële activiteit hebben [204]. Het is nog onduidelijk hoeveel van de andere kandidaten daadwerkelijk RiPP-subklassen zijn, maar zelfs als dat voor de helft zou gelden, zou dit als nog een significante bijdrage zijn voor het aantal bekende RiPP-subklassen.

Veel andere methoden en invalshoeken zijn denkbaar om de resultaten te filteren aan de hand van de genetische context. Een ander element dat in de

modificerende enzymen van veel RiPP-subklassen voorkomt is een RiPP Recognition Element (RRE). Dit zijn kleine elementen, die onderdeel maken van enzymen en als een grijp-arm functioneren om de precursors te herkennen en vast te houden, terwijl de modificatie wordt aangebracht [109]. Opvallend genoeg is de secundaire structuur van dit element vaak hetzelfde, ook tussen verschillende RiPP-subklassen. Deze RREs kunnen consistent gevonden worden door HHPred, een programma dat onder andere de secundaire structuur vergelijkt. HHPred is alleen ontworpen voor de vergelijking van secundaire structuur in het algemeen, niet alleen voor RREs, en neemt daarom veel tijd in beslag. Hierdoor kan de methode niet op grote schaal worden toegepast om RREs van mogelijke nieuwe subklassen te vinden. **RRE-Finder**, beschreven in Hoofdstuk 2, is ontwikkeld om de zoekfunctionaliteit van HHPred toe te spitsen op alleen RREs. RRE-Finder kan in twee modi gerund worden: een conservatieve modus, die razendsnel bekende RREs met bekende sequenties vindt; en een exploratieve modus, die meer tijd kost, maar ook RREs kan vinden die qua sequentie minder vergelijkbaar zijn. In Hoofdstuk 2 laten we zien dat met de exploratieve modus van RRE-Finder er RREs gevonden worden in modificerende enzymen die niet eerder geassocieerd werden met RiPP biosynthese en die dus kunnen leiden tot nieuwe RiPP-subklassen. De exploratieve modus levert wel een groter aantal valspositieven op in vergelijking met de conservatieve modus. De meeste valspositieven worden gevonden in regulatie-eiwitten, waarvoor dit domein vermoedelijk dient als grijp-arm om DNA te herkennen. Aangezien dit geen modificerende enzymen zijn, zijn deze makkelijk te onderscheiden van mogelijke RiPP modificerende enzymen en het ingebouwde filter kan deze verwijderen. Het combineren van RRE-Finder met decRiPPter zou bovendien tot goede resultaten kunnen leiden en het totaal aantal valspositieven nog verder doen dalen. De kans dat een gencluster zowel een valspositief van RRE-Finder als van decRiPPter bevat, is namelijk aanzienlijk kleiner dan dat één ervan dat is.

## Nieuwe RiPPs en classificatie ervan

Uit de resultaten van decRiPPter's analyse op de *Streptomyces*-genomen zijn twee kandidaten geselecteerd voor experimentele analyse. Eén van de genclusters bleek inderdaad de machinerie voor een RiPP te coderen. Dit gencluster, dat onder andere werd gevonden in *Streptomyces pristinaespiralis*,

bevat drie precursorgen en produceert daarmee drie RiPPs, die de **pristinins** zijn genoemd. Van één hiervan, pristin A3, is de structuur en de modificaties bepaald met behulp van LCMS-MS, NMR en chemische labeling (zie Hoofdstuk 4). Pristin A3 bevat thioether bruggen, gedehydrateerde aminozuren en verscheidene serines zijn omgebouwd tot alanines. Dit zijn modificaties die eerder gevonden zijn in lanthipeptides, één van de meest veel voorkomende subklassen van RiPPs. Lanthipeptides hebben vaak antimicrobiële activiteit en zijn daarmee een uitgelezen kandidaat voor verder onderzoek als nieuwe antibiotica.

Net zoals bij andere RiPP-subklassen, kan de karakterisatie van dit BGC leiden tot de identificatie van nog meer nieuwe RiPP BGCs van dezelfde klasse. Hiervoor werd onderzocht welke enzymen die gecodeerd worden door het BGC de modificatie aanbrengen die typerend is voor deze subklasse. De modificatie die elke lanthipeptide heeft is een thioether brug tussen een cysteine en een serine of threonine. Tot nu toe waren er vier verschillende sets enzymen bekend die deze thioether brug konden installeren, wat heeft geleid tot de onderverdeling van lanthipeptides in vier subklassen [47]. De genen van deze enzymen konden niet teruggevonden in het BGC van pristinins, wat hoogstwaarschijnlijk betekent dat het hier om een vijfde subklasse van lanthipeptides gaat. De producten van twee genen van het pristin BGC leken uitgelezen kandidaten om de thioether brug aan te brengen, niet alleen voor pristinins, maar in alle type V lanthipeptides. Het verwijderen van deze genen uit het pristin BGC stopte de productie van pristinins, wat suggereert dat hun producten inderdaad betrokken zijn bij de productie van deze RiPPs. Bovendien worden deze twee genen altijd samen teruggevonden, wat suggereert dat ze van elkaar afhankelijk zijn voor hun functie. De aanwezigheid van dit genenpaar in diverse genetische contexten en in de genomen van allerlei bacteriën toont aan dat er nog veel type V lanthipeptides te ontdekken en karakteriseren zijn, die eerder altijd uit het zicht lagen.

Een tweede gencluster dat onderzocht is staat beschreven in Hoofdstuk 5. Het gaat hier om een gencluster dat verdeeld is over twee delen, die naast elkaar in tegengestelde richting liggen op het genoom en elk beginnen met een sterk geconserveerd precursorgen. Deze genen bevatten ook meerdere

herhalende patronen met dezelfde sequentie (TTGWQ). Dit soort sequenties wordt vaak teruggevonden in de precursors van andere RiPPs, waarbij elke herhaalde sequentie in een RiPP wordt omgezet na modificaties. Het gencluster codeert daarnaast een scala aan verschillende modifierende enzymen, waarvan sommigen tot families behoren die betrokken kunnen zijn bij RiPP biosynthese, zoals de radical S-adenosyl methionine (rSAM) enzymen en de ATP-grasp ligases. Desalniettemin is deze overlap erg klein en werd dit gencluster daarom ook niet gevonden door RiPP-identificatiesoftware zoals antiSMASH of BAGEL. Activatie van het gencluster leidde tot de identificatie van een paar moleculen in de chemische extracten van de producerende stam, die niet meer aanwezig waren als het gencluster was verwijderd. Helaas kon de structuur van deze moleculen niet worden opgehelderd en is dus nog niet zeker of deze producten daadwerkelijk RiPPs zijn. Verder onderzoek is nog nodig om de structuur te bepalen en om te zien of deze producten inderdaad van de precursors zijn afgeleid.

Deze BGCs en die van andere recent ontdekte RiPP-subklassen brengen een interessante discussie op gang met betrekking tot RiPP classificatie. RiPPs worden meestal geclassificeerd op basis van één typerende modificatie, terwijl de rest als secundaire modificaties worden gezien. Door het toenemende aantal RiPP-subklassen dat wordt ontdekt, wordt het steeds duidelijker dat genen vaak uitgewisseld worden tussen de BGCs van verschillende subklassen, met als gevolg dat veel modificaties overlappen tussen verschillende RiPPs. Of die modificaties worden gezien als primair of als secundair hangt af van de RiPP-subklasse. Zo worden bijvoorbeeld gedehydrateerde aminozuren gezien als een primaire modificatie in linaridins, maar als secundair in lanthipeptides.

Hetzelfde zal mogelijk het geval zijn bij het gencluster dat in Hoofdstuk 5 wordt besproken. Dit gencluster bevat genen die al eerder geassocieerd zijn met RiPP BGCs en mogelijk zal het product van dit gencluster ook al bekende modificaties bevatten. Desalniettemin wordt dit gencluster niet gevonden door RiPP-detectie software, die er vooral op gericht is alleen BGCs van bekende BGCs te vinden met duidelijk gedefinieerde regels. Hoewel deze methoden effectief zijn, is het belangrijk dat de gebruiker begrijpt waarop de methoden zijn gebaseerd en wat voor belemmeringen dat met zich meebrengt. Een bredere

zoektocht, waarbij gezocht wordt naar RiPP modificerende enzymen in diverse genetische contexten, ongeacht primair of secundair, kan leiden tot interessante ontdekkingen en had mogelijk al kunnen leiden tot de ontdekking van de BGCs die in dit proefschrift besproken staan.

## Conclusie

BGCs en hun producten komen in vele soorten en maten, wat leidt tot een enorme genetische diversiteit om te onderzoeken. In dit proefschrift zijn verschillende methoden onderzocht die proberen een meer verkennende invalshoek te geven aan deze zoektocht door te zoeken naar nieuwe RiPP BGCs middels kunstmatige intelligentie. Hoewel RiPP BGCs erg verschillend zijn, blijft de biosynthetische logica behouden, waarvan gebruik wordt gemaakt bij de ontwikkeling van decRiPPter en RRE-Finder. Een groot scala aan kandidaten wordt geprioriteerd aan de hand van de precursorgen en hun genetische context. Twee van deze kandidaten zijn onderzocht en van één is aangetoond dat het inderdaad om een nieuwe RiPP-subklasse gaat. Dit laat zien dat zulke methodes veel potentie hebben, door rekening te houden met een grotere kans op het aantal vals-positieven. Verdere ontwikkeling van de modellering, aangescherpt door nieuwe biologische en chemische kennis, zullen kunnen leiden tot een doorgaande stroom van nieuwe RiPP BGCs en dit zal hopelijk leiden tot de ontdekking van nieuwe antibiotica.