Modèles bio-informatiques pour les peptides non-ribosomiques et leurs synthétases

Maude Pupin

▶ **To cite this version:**

Maude Pupin. Modèles bio-informatiques pour les peptides non-ribosomiques et leurs synthétases. Bio-informatique [q-bio.QM]. Université des Sciences et Technologie de Lille - Lille I, 2013. tel-00918918

HAL Id: tel-00918918

https://theses.hal.science/tel-00918918

Submitted on 16 Dec 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modèles bio-informatiques pour les peptides non-ribosomiques et leurs synthétases

Mémoire présenté le 3 décembre 2013

pour l'obtention de

l'Habilitation à Diriger des Recherches

en informatique

par

Maude Pupin

Composition du jury

Rapporteurs :	Muriel Gugger, responsable de laboratoire, Institut Pasteur de Paris Frédérique Lisacek, group leader, Swiss Institute of Bioinformatics Pierre Tufféry, DR INSERM, INSERM UMR-S 973
Examinateurs :	Gilbert Deléage, Professeur, Université de Lyon Pierre Boulet, Professeur, Université Lille 1
Directeur :	Hélène Touzet, DR CNRS, UMR CNRS 8022 et Inria Lille-Nord Europe

UNIVERSITÉ Lille 1
Laboratoire d'Informatique Fondamentale de Lille — UMR CNRS 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX

À ma famille qui a accepté que je passe beaucoup de temps à travailler, au lieu d'être avec elle
À Frédéric qui partage mes bonheurs et soulage mes peines
À mes filles, Eléonore et Capucine, mes rayons de soleil

Merci

Pour commencer, merci à Hélène Touzet sans qui le travail présenté dans ce mémoire n'aurait pu se faire. Elle a construit l'équipe de recherche en bio-informatique sur Lille, Bonsai. Elle a soutenu dès le début le projet de recherche sur les peptides non-ribosomiques et a permis qu'il se développe en le soutenant politiquement et financièrement. Enfin, elle m'a fait confiance pour mener à bien ce projet, aidée de mes collaborateurs.

Merci à Muriel Gugger, Frédérique Lisacek et Pierre Tufféry d'avoir accepté d'être les rapporteurs de mon HdR, ainsi que Gilbert Deléage et Pierre Boulet d'avoir accepté de faire partie de mon jury.

Merci à Mathieu Giraud, Laurent Noé, Aïda Ouangraoua, Hélène Touzet et Jean-Stéphane Varré d'avoir relu attentivement mon mémoire et pour leurs conseils avisés qui m'ont permis de l'améliorer.

Merci à Valérie Leclère, ma coéquipière sur le sujet de la bio-informatique pour les peptides non-ribosomiques, de transmettre sa bonne humeur. Ensemble, nous avons surmonté les épreuves et remporté des victoires.

Merci à Philippe Jacques d'avoir apporté les recherches concernant les peptides non-ribosomiques sur Lille et d'avoir œuvré pour que la bio-informatique dédiée à ces molécules se développe.

Merci à Ségolène Caboche d'avoir choisi le sujet de master recherche puis de thèse que nous avions proposé et d'avoir brillamment donné vie à Norine.

Merci à Areski Flissi, Stéphane Janot et Laurent Noé d'avoir pris en charge les aspects techniques de Norine. Un merci particulier à Areski d'avoir choisi de rejoindre l'équipe Bonsai et de consacrer une partie de son temps aux peptides non-ribosomiques et à Laurent de se lancer avec moi dans la chemo-informatique.

Merci à Yoann Dufresne d'avoir choisi le sujet de master recherche puis de thèse que nous avons proposé et d'être motivé pour mener à bien ses recherches.

Merci aux étudiants, ingénieurs, doctorants et postdoctorant avec qui j'ai travaillé pour leur contribution à la bio-informatique pour les peptides non-ribosomiques et leurs enzymes.

Merci à Aïda Ouangraoua, Jean-Stéphane Varré et Stéphane Janot, mes voisins de bureau, pour les discussions que nous avons eues.

Merci à Jean-Frédéric Berthelot, Mikaël Salsom et quelques autres pour les bons plats qu'ils amènent lors des repas du mardi midi.

Merci aux membres de l'équipe Bonsai que je n'ai pas cités pour les discussions lors des repas, des poses ou des pots.

Merci aux habitués de la salle cafet' du M3 de me changer les idées lors des repas de midi.

Merci à tout le personnel technique et administratif de faire fonctionner l'Université et Inria.

Bonjour aux collègues enseignants, aux membres du laboratoire ProBioGEM que je côtoie et à tous les collègues que je n'ai pas cités ici.

Table des matières

Avant-propos 1

Introduction 3

1 Stratégies d'annotation des synthétases peptidiques non-ribosomiques 7

1.1 Introduction et motivations . . . 7

1.2 Mécanisme de synthèse non-ribosomique . . . 8

1.2.1 Incorporation des acides aminés . . . 9

1.2.2 Liaison entre acides aminés . . . 10

1.2.3 Incorporation d'acides aminés sous la forme D . . . 10

1.2.4 Incorporation d'acides aminés modifiés . . . 12

1.2.5 Incorporation d'autres monomères . . . 14

1.2.6 Libération du peptide . . . 14

1.3 Bio-informatique pour l'annotation des synthétases . . . 15

1.3.1 Prédiction de gènes et annotation de protéines . . . 15

1.3.2 Logiciels d'annotation des synthétases peptidiques non-ribosomiques . . . 17

1.4 Exploration du potentiel de synthèse de genres bactériens . . . 23

1.4.1 Protocole d'annotation de synthétases peptidiques non-ribosomiques . . . 23

1.4.2 Annotation à grande échelle de synthétases . . . 27

1.4.3 Encadrements et collaborations . . . 29

2 Norine, plate-forme d'analyse bio-informatique des peptides non-ribosomiques 33

2.1 Introduction et motivations . . . 33

2.2 Diversité des peptides non-ribosomiques . . . 34

2.2.1 Diversité de composition . . . 35

2.2.2 Diversité de liaisons chimiques . . . 37

2.2.3 Diversité de structures bidimensionnelles . . . 41

2.2.4 Peptides et plus encore . . . 44

2.2.5 Diversité d'activités 45
2.2.6 Diversité des organismes producteurs 47
2.3 Modèles bio-informatiques pour les peptides non-ribosomiques 49
2.3.1 Bases de données contenant des PNR ou d'autres peptides. 49
2.3.2 Outils pour la spectrométrie de masse pour les PNR 52
2.4 La plate-forme NORINE 52
2.4.1 Les données de NORINE 53
2.4.2 Les algorithmes de recherche de peptides selon leur structure 57
2.4.3 Encadrements et collaborations 62

Conclusion et perspectives **65**

Bibliographie **67**

Curriculum vitæ **77**

article Norine **85**

article Motif **91**

article Monomères **101**

article Fingerprint **109**

article Kurstakin **117**

Avant-propos

Ce document retrace les travaux de recherche que j'ai menés au sein de l'équipe Bonsai du LIFL et du centre Inria Lille Nord-Europe depuis 2006.

L'année 2006 a été pour moi une année charnière dans ma carrière. À mon arrivée sur Lille en 2000, moins d'un an après la fin de ma thèse, j'ai choisi d'investir mon temps dans l'enseignement. Le défi était de concevoir des cours et supports de TP *ab initio* car, à l'époque, la bio-informatique n'était pas enseignée à l'Université Lille 1 et peu à l'échelle nationale. Avec mes collègues Hélène Touzet et Jean-Stéphane Varré, j'ai mis en place les enseignements d'introduction à la bio-informatique pour les étudiants en filières biologiques. D'ailleurs, nos matériels pédagogiques sont disponibles sur Internet et ont été repris par d'autres enseignants en France. J'ai également créé le DESS Bio-informatique de Lille avec Hélène Touzet et nous l'avons co-dirigé pendant 5 ans, entre 2001 et 2006. Les cinq promotions ont rassemblé entre 12 et 18 étudiants issus d'un cursus en informatique dont la moitié ne provenait pas de l'Université Lille 1. Environ un quart des diplômés occupe en emploi dans le domaine de la bio-informatique aussi bien en France qu'à l'étranger et certains ont fait une thèse. Les autres travaillent dans le domaine de l'informatique «traditionelle». En parallèle de mon investissement pédagogique, j'ai continué les recherches débutées pendant ma thèse sur la comparaison de séquences nucléiques, en collaboration avec Claudine Devauchelle, Ivan Laprevotte, Alex Grossmann et Alain Hénaut du laboratoire Statistique et Génome de l'Université d'Évry, et Gilles Didier de l'Institut de Mathématiques de Luminy. Nous avons publié ensemble trois articles sur cette période. Le premier [68], écrit en 2001, présente les observations réalisées sur les séquences LTR (*Long Terminal Repeat*) des virus de l'immunodéficience humaine (VIH) grâce à notre méthode d'aide à l'alignement, appelée N-écriture. Le deuxième article [32], publié en 2006, formalise notre méthode qui est alors présentée comme le décodage local à l'ordre N d'une séquence. En 2007, nous avons publié un calcul de distances basé sur le décodage local, accompagné d'une application à l'identification des sous-types de VIH [31]. Sur cette même période, j'ai également co-signé un article sur l'annotation des génomes de différentes souches d'*Helicobacter pilori* [14], présentant des recherches réalisées pendant ma thèse. J'ai choisi de ne pas inclure ces travaux dans ce mémoire car je ne contribue plus désormais à ces thèmes.

Les rencontres avec mes collègues biologistes dans le cadre des enseignements de bio-informatique, ont été l'occasion d'apprendre à connaître les recherches menées dans les laboratoires de biologie de la métropole lilloise. En particulier, Valérie Leclère, maître de conférences au laboratoire ProBioGEM, qui enseignait la microbiologie pour le DESS bio-informatique, a proposé des projets étudiants concernant les peptides non-ribosomiques (PNR). Le suivi de ces projets m'a permis de découvrir ces molécules particulières dont le mode de synthèse et les spécificités sont peu connus des biologistes et encore moins des bio-informaticiens. Le laboratoire ProBioGEM (Procédés Biologiques, Génie Enzymatique et Microbien) de l'Université Lille 1 est spécialisé dans l'obtention de peptides bioactifs qu'ils soient d'origine non-ribosomique ou obtenus par hy-

drolyse[1]. Avec Hélène Touzet, nous avons remarqué que les peptides non-ribosomiques avaient des spécificités qui rendaient inutilisables les outils de bio-informatique dédiés aux peptides et protéines ribosomiques. De plus, ces peptides peuvent être représentés sous la forme de graphes, induisant des problèmes algorithmiques intéressants. Ainsi, en 2006, nous avons suggéré à Valérie Leclère et Philippe Jacques de proposer un sujet de stage de master recherche. Il se trouve que cette année là, Ségolène Caboche, une étudiante ayant suivi un cursus pluridisciplinaire en informatique et biologie, recherchait un stage de master recherche. Ségolène a fait son stage avec nous puis a continué en thèse grâce à l'obtention d'un co-financement Inria et région. Grégory Kucherov, qui a été recruté directeur de recherche CNRS dans l'équipe cette même année, a rejoint le groupe de travail sur les PNR. Ségolène Caboche a concrétisé notre idée de départ en une plate-forme d'analyse des peptides non-ribosomiques appelée NORINE[2]. Elle est composée d'une base de données recensant plus de 1100 peptides, associée à des outils graphiques, ainsi que des algorithmes de recherche et de comparaison de structures et d'un logiciel de prédiction d'activités. La collaboration avec des membres du laboratoire ProBioGEM a initié une nouvelle activité de recherche dans l'équipe Bonsai dont la montée en puissance s'est faite progressivement. Maintenant, je dirige le groupe NRP au sein de l'équipe Bonsai et des collègues m'ont rejoint : Stéphane Janot, maître de conférences en informatique, et Areski Flissi, ingénieur de recherche en informatique, sur les développements de la plate-forme NORINE ainsi que Laurent Noé, maître de conférences en informatique, sur le nouveau projet de recherche concernant la chémo-informatique pour les PNR. Entre 2006 et 2013, j'ai co-encadré deux thèses et une troisième vient de commencer, j'ai dirigé trois ingénieurs qui se sont succédés pour améliorer l'interface d'interrogation de NORINE, j'ai encadré une stagiaire et co-encadré trois autres en master 2 recherche en bio-informatique, j'ai guidé quatre doctorants du laboratoire ProBioGEM dans leurs analyses bio-informatiques (dont un est toujours en thèse) et j'ai recruté et dirigé un postdoctorant. Les recherches et développements menés ont conduit au succès de la plate-forme NORINE, l'unique ressource dédiée aux peptides non-ribosomiques comportant une base de données avec des annotations de qualité car validées manuellement, des outils d'interrogation et de visualisation des informations concernant les peptides, ainsi que des algorithmes de comparaison de ces molécules représentées sous la forme de graphes. Les résultats obtenus ont été décrits dans 5 articles publiés dans des journaux à audience internationale, avec comité de lecture. Ils ont également été présentés sous forme de communications orales dans 8 conférences nationales et internationales, de posters dans 17 autres et d'un article de vulgarisation scientifique rédigé en français.

1. Découpage d'une substance par l'action de l'eau.
2. Disponible à l'adresse : `http://bioinfo.lifl.fr/norine`

Introduction

Les micro-organismes, tels que les bactéries, les levures ou les moisissures, sont capables de coloniser les milieux terrestres, aquatiques et même d'autres êtres vivants. Pour cela, ils doivent optimiser les interactions avec leur environnement et développer des stratégies d'attaque et de défense contre les espèces concurrentes car les ressources sont limitées. Par exemple, ils optimisent leur capacité à absorber les nutriments et sources d'énergie disponibles ; ils envahissent le milieu en formant un biofilm [3] ; ils détruisent les micro-organismes rivaux par la production d'antimicrobiens ou de toxines et se protègent contre les attaques des autres espèces. Parmi les métabolites [4] synthétisés pour accomplir ces tâches, les peptides non-ribosomiques (PNR) ont la particularité de couvrir, grâce à leur grande diversité de structures chimiques, les différentes stratégies énoncées précédemment. Ce large spectre de structures et donc d'activités est dû à leur mode de synthèse alternatif à la voie de synthèse ribosomique pour les peptides et protéines. Leur production est mise en œuvre par de grands complexes enzymatiques appelés synthétases peptidiques non-ribosomiques (SPNR).

Parmi les nombreuses applications possibles des peptides non-ribosomiques, celle concernant leur utilisation dans le domaine pharmaceutique est la plus développée. Il se trouve que les PNR sont parmi les premiers antibiotiques naturels à avoir été découverts. L'un des événements fondateurs de l'ère moderne des antibiotiques est la découverte de la pénicilline par l'écossais Alexander Flemming en 1928 [41], même si l'action antibactérienne de la moisissure *Penicillium notatum* a été utilisée dès l'Antiquité par les égyptiens, les chinois et les indiens. Les premières étapes de la synthèse de cette molécule sont effectuées par une SPNR qui produit le peptide ACV, pour δ-(L-α-aminoadipoyl)- L-cysteinyl-D-valine. D'autres peptides non-ribosomiques sont commercialisés tels que la daptomycine (nom commercial : Cubicin®) ou la vancomycine qui sont des antimicrobiens ; la cyclosporine A qui est utilisée pour réduire les risques de rejet de greffes ou encore l'actinomycine D utilisée dans le traitement de certains cancers. Dans un registre proche, les PNR peuvent être utilisés comme agents phytosanitaires. D'autres applications sont possibles comme la décontamination des eaux polluées par les métaux lourds en employant des sidérophores [5], ou polluées par des composés organiques en employant des biosurfactants [6] ; ou bien encore la fabrication de cosmétiques, toujours par l'intermédiaire des biosurfactants.

Les PNR ne sont pas les seules molécules naturelles à entrer dans la composition de médicaments. La recherche de nouveaux produits naturels à activités thérapeutiques est un domaine toujours actif comme en témoignent les nombreux articles publiés sur le sujet (voir [30, 70, 82, 84] pour les plus récents). Mais, l'identification expérimentale de principes actifs et de leur mode d'action reste un processus long et délicat, qui n'est que l'étape préliminaire à l'étude clinique

3. Colonie de micro-organismes qui forme un film sur une surface.
4. Petits composés organiques produits par les cellules.
5. Molécules capables de capter le fer.
6. Molécules capables de modifier la tension d'un liquide en formant un film à sa surface.

préalable à une mise sur le marché. Heureusement, de nouvelles stratégies ont vu le jour en s'appuyant sur l'analyse bio-informatique du flux de l'information génétique. La figure 1 schématise ce flux : l'ADN, molécule qui porte l'information génétique, est transcrit par l'ARN-polymérase en ARN messager qui est lui-même traduit en protéine par les ribosomes. Les protéines ainsi produites réalisent de nombreuses fonctions dans la cellule dont la catalyse de réactions chimiques [7]. L'enchaînement de réactions chimiques, appelé voie de synthèse, aboutit à la production de métabolites.

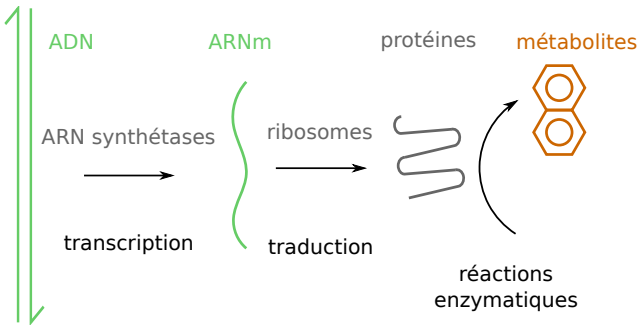


FIGURE 1 – Le flux de l'information génétique, appelé aussi dogme central de la biologie.

Les progrès des techniques de séquençage associés à des logiciels d'annotation de séquences génomiques toujours plus performants procurent les molécules potentiellement produites par un organisme [119, 67, 40]. Dans un premier temps, la séquence de son génome est déterminée, puis des analyses bio-informatiques prédisent les gènes présents sur ce génome et la fonction des protéines pour lesquelles ils codent. Lorsque ces protéines sont des enzymes, il est possible de déterminer quelles sont les molécules chimiques qu'elles produisent. Ensuite, les propriétés et activités de ces molécules sont étudiées grâce à d'autres outils bio-informatiques. Ainsi, les principes actifs synthétisés par un organisme peuvent être prédits à partir de son génome. Pour finir, ces analyses bio-informatiques doivent être validées expérimentalement et d'autres investigations sont nécessaires avant de pouvoir exploiter industriellement une nouvelle molécule.

Les peptides non-ribosomiques ont le double avantage d'avoir de nombreuses applications potentielles et de pouvoir être identifiés à partir de séquences génomiques, en passant par l'analyse des protéines qui les produisent. C'est pourquoi les PNR sont de plus en plus étudiés comme en témoigne la rapide augmentation du nombre d'articles concernant ces molécules visible dans l'histogramme de la figure 2. Les auteurs de ces publications sont soit des biochimistes ou des pharmaciens qui extraient le peptide puis analysent ses propriétés, soit des biologistes qui étudient sa voie de synthèse et ses activités. Quelque-soit leur approche, les scientifiques cherchent à collecter le maximum d'informations concernant leur peptide d'intérêt ou des peptides similaires. L'objectif de mon travail est de répondre à ce besoin en fournissant une base de données et une palette d'outils bio-informatiques permettant l'analyse *in silico* des peptides non-ribosomiques. Il est atteint au travers de la ressource NORINE.

Je décris dans ce mémoire tout le travail réalisé depuis 2006 sur les peptides non-ribosomiques et les enzymes qui les produisent. Le premier chapitre porte sur les synthétases peptidiques non-ribosomiques pour introduire des notions utiles à la compréhension du chapitre suivant. Le deuxième chapitre présente les peptides non-ribosomiques, l'objet principal de mes recherche, et

7. Action d'accélération de la vitesse d'une réaction chimique.

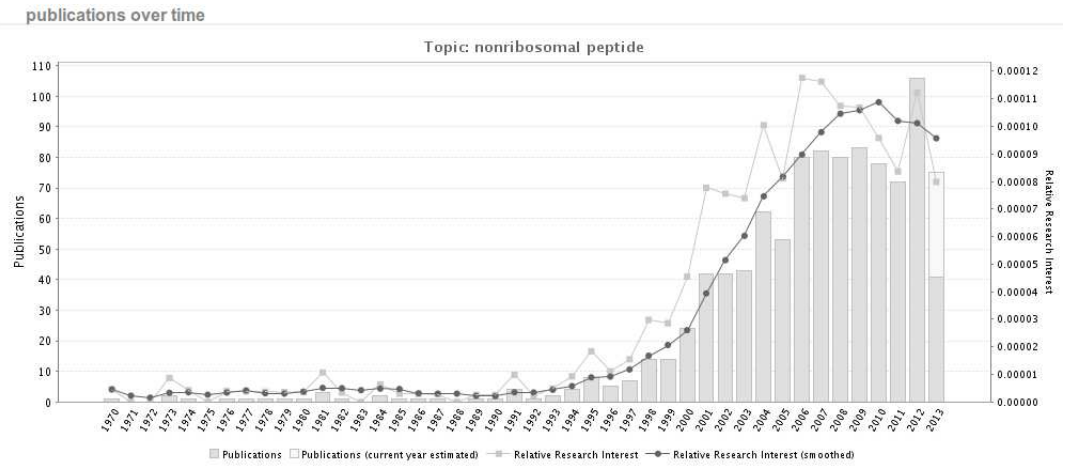


FIGURE 2 – Histogramme du nombre de publications par an contenant l'expression «nonribosomal peptide», généré sur le site de gopubmed®.

leurs spécificités. Les deux chapitres sont construits selon le même plan, à savoir une première partie introduisant les notions biologiques, une deuxième présentant les outils bio-informatiques développés par d'autres équipes de recherche et la troisième partie décrivant les travaux auxquels j'ai participé. Dans le premier chapitre, le mode de fonctionnement des synthétases et des enzymes auxiliaires est décrit en prenant comme fil conducteur les particularités des peptides. Puis, les étapes nécessaires à la conception des outils bio-informatiques d'analyse de ces enzymes sont passées en revue. Un tableau comparatif de ces outils est fourni. Je présente ensuite mes contributions qui portent à la fois sur des développements informatiques et de l'analyse de données. Dans le deuxième chapitre, je décris toute la diversité des PNR des points de vue biologique et chimique, telle que la ressource NORINE m'a permis de l'appréhender. J'ai complété ces informations à l'aide de recherches bibliographiques poussées. Les quelques bases et outils bio-informatiques dédiés ou non aux PNR sont présentés. Puis, je décris la ressource NORINE dédiée aux peptides non-ribosomiques, dont je suis un des membres fondateurs, et les outils et algorithmes associés. Une partie des résultats est exposée dans les articles co-écrits avec mes collaborateurs et qui sont insérés dans ce manuscrit.

Chapitre 1

Stratégies d'annotation des synthétases peptidiques non-ribosomiques

1.1 Introduction et motivations

Mon activité de recherche principale porte sur la bio-informatique pour les peptides non-ribosomiques (PNR). Cependant, je travaille aussi sur leurs synthétases, c'est-à-dire les enzymes qui les produisent. Ma participation au travail d'exploration du potentiel de synthèse des peptides non-ribosomiques a été motivée par deux raisons principales. La première est que j'ai apporté ma maîtrise des outils d'analyse de séquences, mon approche de bio-informaticienne et ma capacité à identifier les développements informatiques réalisables. J'ai donc accompagné mes collègues du laboratoire ProBioGEM dans leurs analyses bio-informatiques. J'ai participé à la mise au point d'un protocole de prédiction puis d'annotation de synthétases à partir d'une séquence génomique ; à l'analyse de génomes bactériens et à la conception de logiciels répondant à leurs besoins. Mes connaissances des outils généralistes d'analyse de séquences étaient d'autant plus nécessaires au début de ce travail que les outils spécifiques aux synthétases peptidiques non-ribosomiques (SPNR) étaient peu nombreux et ne proposaient pas l'analyse d'un génome complet. Le protocole a beaucoup évolué en quelques années puisque de nouveaux logiciels ont été développés.

La deuxième raison est la nécessité de maîtriser les différentes étapes d'analyse des synthétases afin de pouvoir intégrer au mieux Norine, la ressource dédiée aux peptides non-ribosomiques développée à Lille, dans le processus d'identification de nouveaux PNR. Je réponds aux besoins spécifiques des biologistes et autres utilisateurs des outils que je conçois, parce que je suis moi-même une utilisatrice avertie de ces outils.

Avant de présenter les outils d'analyse bio-informatique des synthétases peptidiques non-ribosomiques existants dans la partie 1.3, je vais décrire le mode de synthèse non-ribosomique dans la partie 1.2. Ensuite, je décrirai mes contributions dans la partie 1.4. Il s'agit de la mise au point du protocole d'exploration du potentiel de synthèse de PNR à partir de données de séquences génomiques ou protéiques avec ma collègue Valérie Leclère, des analyses bio-informatique réalisées sur plusieurs genres bactériens et la conception de logiciels répondant à des besoins spécifiques aux SPNR. Enfin, je terminerai en présentant mes collaborateurs et les doctorants, ingénieurs et étudiants que j'ai encadrés ou co-encadrés.

Les résultats de certains travaux présentés dans ce chapitre ont été publiés dans une mini-revue concernant la kurstakine [11], un peptide non-ribosomique produit par les bactéries du

genre *Bacillus*. D'autres ont été soumis pour présenter le pipeline d'annotation de synthétases FLORINE [22]. D'ailleurs, un poster présentant ce pipeline a reçu le prix du meilleur poster du congrès de la Société Française de Microbiologie (SFM) lors de son édition 2013.

1.2 Mécanisme de synthèse non-ribosomique

La synthèse non-ribosomique est effectuée par de grands complexes enzymatiques appelés synthétases peptidiques non-ribosomiques (SPNR). Leur nom vient du fait que ces enzymes produisent des peptides par une autre voie que le flux de l'information génétique (voir figure 1) et donc ne passent pas par les ribosomes. Les enzymes impliquées fonctionnent comme des chaînes de montage qui captent les acides aminés dans le milieu cellulaire et les assemblent pour former des peptides. Dans le cas de la *synthèse linéaire*, qui est la plus fréquente, les synthétases sont composées d'autant de modules qu'il y a de monomères dans le peptide final. Chaque module est divisé en domaines responsables d'une étape d'incorporation d'un monomère via la catalyse d'une réaction chimique. Ces domaines correspondent à une fonction enzymatique particulière. La figure 1.1 représente la synthétase de l'ACV avec ses différents domaines. Dans la suite de ce mémoire, j'utilise toujours les mêmes conventions pour représenter l'enchaînement des domaines dans les protéines. J'ai choisi d'utiliser le symbole | pour représenter un domaine qui est à l'extrémité d'une protéine (début ou fin) et les symboles < et > pour représenter un domaine qui peut être précédé ou suivi d'autres domaines. Pour la synthétase de l'ACV qui n'est composée que d'une protéine, la structure en domaines est la suivante : |A T C A T C A T E Te| où les abréviations représentent les domaines suivants dont la fonction sera décrite par la suite :

- A : adénylation
- T : thiolation
- C : condensation
- E : épimérisation
- Te : thioestérase

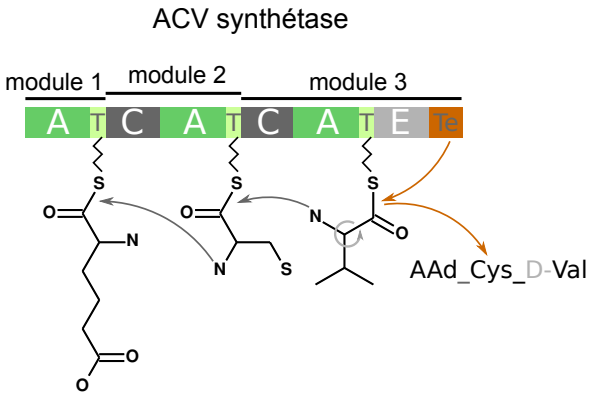


FIGURE 1.1 – La synthétase de l'ACV. Cette enzyme est composée de 3 modules, elle incorpore donc trois monomères, reconnus par les domaines d'adénylation, notés A. Puis ces monomères sont fixés de façon covalente à l'enzyme par les domaines de thyolation (T). Enfin, les domaines de condensation (C) forment des liaisons peptidiques entre les monomères. Le peptide ainsi formé est libéré par le domaine de thioestérase (Te). Cette enzyme possède un domaine optionnel, le domaine d'épimérisation (E) qui modifie l'isomérie du monomère porté par le module 3. Ces domaines seront décrits plus en détail par la suite.

Lors de la *synthèse itérative*, les modules sont utilisés plus d'une fois pour incorporer plusieurs exemplaires d'un même monomère dans le peptide final. Par exemple, lors de la synthèse de la gramicidine S deux exemplaires du même décapeptide (10 monomères) sont assemblés en un peptide circulaire via un domaine particulier de thioestérase (Te) [66]. Autre exemple, la synthétase de l'enniatine est composée de deux modules et se termine par un domaine de condensation : |C A T C A NM T C| (NM est le domaine de N-méthylation) [47]. Elle produit un peptide cyclique composé de trois exemplaires d'un dipeptide (2 monomères). Enfin, la *synthèse non linéaire* est effectuée par des synthétases peptidiques non-ribosomiques dont les domaines ne sont pas agencés selon les règles habituelles. Certaines protéines ne portent qu'un seul domaine qui peut intervenir à différentes étapes de la synthèse. Par exemple, la synthétase de la vibriobactine [62] est composée de 4 protéines dont trois sont des domaines isolés. La structure de la synthétase est la suivante : |A||T||C||Cy Cy A C T C| (où Cy est un domaine de cyclisation). Le peptide produit est composé de répétitions d'acides aminés et d'une polyamine [8], la norspermidine.

Il est à noter que le terme de synthèse linéaire ne sous-entend pas que les peptides produits soient linéaires. De même, la synthèse non linéaire peut produire des peptides linéaires. Ces termes décrivent le mode de fonctionnement des synthétases qui peut ou non suivre l'ordre des domaines présents sur les protéines.

Puisque mes recherches portent principalement sur les peptides non-ribosomiques, j'ai choisi de présenter la synthèse non-ribosomique en allant des peptides vers les synthétases et les enzymes externes associées afin de mettre en avant les différentes voies menant à une même configuration dans le peptide final. Pour fabriquer un PNR, la cellule doit d'abord fabriquer les briques de base que sont les monomères. Je ne décrirai pas en détail les voies de synthèse des différents types de monomères mentionnés dans la partie 2.2.1. Ensuite, les monomères sont assemblés, et éventuellement modifiés, par les modules des synthétases ou par des enzymes externent qui agissent *a posteriori*. La composition en domaines d'un module détermine son action au sein de la voie de synthèse. Je présenterai au fur et à mesure les différentes configurations de domaines observées au sein des modules et les actions associées. De la même façon, l'enchaînement des modules n'est pas dû au hasard, mais est influencé par les domaines qui les composent.

Les principaux articles qui m'ont aidé à décrire la synthèse non-ribosomique et les synthétases sont les suivants [103, 77, 53, 102].

1.2.1 Incorporation des acides aminés

Deux domaines sont nécessaires à l'incorporation d'un acide aminé, ou un composé chimique de nature proche, dans un peptide non-ribosomique.

Le domaine d'adénylation (A) sélectionne un acide aminé spécifique et l'active en formant un aminoacyl-O-AMP (acide aminé lié à une adénosine monophosphate) en consommant un ATP (adénosine triphosphate). Il est le domaine NRPS le plus étudié car il est déterminant pour prédire la composition monomérique du peptide qui est produit par la synthétase. Les principaux résultats obtenus à propos de ce domaine sont décrits dans la partie 1.3.2 et concernent la détermination de l'acide aminé sélectionné en fonction des acides aminés présents à certaines positions du site actif.

8. Molécule organique composée de plus d'un groupement amine NH_2.

Le domaine de thiolation (T) aussi appelé PCP pour *peptidyl carrier protein* fixe de manière covalente l'acide aminé préalablement activé par le domaine A. Ainsi, lorsqu'un acide aminé est sélectionné, il est conservé par la synthétase en attendant, si besoin, que les autres monomères à incorporer soient également présents.

Ces deux domaines sont les constituants de base d'un module. D'ailleurs, la plupart des modules d'initiation sont de la forme : |A T> car ils incorporent le premier acide aminé du peptide (voir le premier module de l'ACV synthétase sur la figure 1.1).

1.2.2 Liaison entre acides aminés

Le domaine de condensation (C) est responsable de la formation d'une liaison peptidique entre deux monomères consécutifs portés chacun par un domaine de thiolation. Des sous-familles ayant des activités spécifiques ont été identifiées [94, 130]. Par convention, les domaines C sont considérés comme étant au début des modules. Les domaines C sont spécifiques de l'isomérie [9] des monomères qu'ils lient (voir figure 1.2).

Le domaine de condensation ($^{L}C_{L}$) est la sous-famille du domaine C la plus répandue. Il effectue une liaison peptidique entre l'acide aminé de forme L porté par le module précédant à un autre acide aminé de configuration L porté par le module qui commence par ce domaine $^{L}C_{L}$ (voir module 4 de la bacitracine dans la figure 1.2). Ce domaine est entre deux modules qui n'incorporent pas d'acide aminé sous la forme D : <C A T $^{L}C_{L}$ A T>

Le domaine de condensation ($^{D}C_{L}$) effectue une liaison peptidique entre l'acide aminé de forme D porté par le module précédant à un autre acide aminé de forme L porté par le module qui commence par ce domaine $^{D}C_{L}$ (voir module 5 de la bacitracine dans la figure 1.2). Ce domaine est entre un module qui incorpore un acide aminé sous la forme D et un autre : <C A T E $^{D}C_{L}$ A T>

Il n'existe pas de domaine C spécifique liant un acide aminé de forme L à un acide aminé de forme D pris dans cet ordre (domaine $^{L}C_{D}$), car le passage de la forme L à la forme D est réalisé sur le monomère qui est déjà lié au peptide en cours de synthèse, soit le module situé à gauche du domaine C.

1.2.3 Incorporation d'acides aminés sous la forme D

Les synthétases incorporent fréquemment des acides aminés sous la forme D dans les peptides non-ribosomiques (voir partie 2.2.1). Il s'agit d'une configuration alternative à la configuration la plus courante qui est appelée L. Ces deux configurations correspondent à des molécules dites énantiomères. Par abus de langage, elles sont aussi appelées stéréoisomères car elles en sont un cas particulier. Ces molécules possèdent la même formule chimique, la même structure bidimensionnelle, mais des structures tridimensionnelles différentes, images l'une de l'autre dans un miroir. Il faut casser une liaison et permuter certains groupements pour passer d'une configuration à une autre. Dans le cas des acides aminés, la configuration la plus courante est appelée L

9. Configurations L et D des molécules, c'est-à-dire les structures chimiques non superposables d'un même composé.

et l'autre D. Le changement est effectuée par la réaction d'épimérisation [10], aussi appelée racémisation. Plusieurs stratégies sont utilisées par les synthétases pour incorporer des conformères sous la forme D.

Le domaine d'épimérisation (E) épimérise l'acide aminé qui est fixé sur le domaine de thiolation qui le précède. Sauf si le module précédant effectue l'initiation de la synthèse, le domaine E agit sur l'acide aminé lorsqu'il est déjà lié à une chaîne peptidique en cours d'élongation. Un tel domaine est nécessairement suivi par un domaine $^{D}C_{L}$ (voir figure 1.2). L'agencement des domaines est donc le suivant : <C A T E $^{D}C_{L}$>. Ce mode de modification est le plus répandu parmi les synthétases non-ribosomiques.

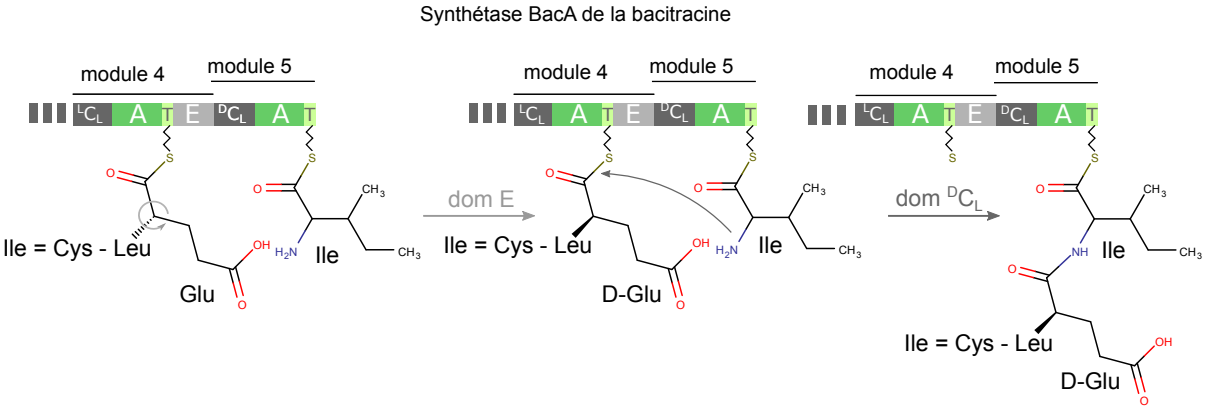


FIGURE 1.2 – Modules 4 et 5 de la synthétase de la bacitracine. L'acide glutamique est épimérisé puis le domaine de condensation $^{D}C_{L}$ forme la liaison avec l'isoleucine. L'acide glutamique est déjà lié aux monomères incorporés par les modules précédents. Par convention, les liaisons qui sont dessinées avec des hachures partent à l'arrière du plan et celles en triangle plein vers l'avant. Cela permet de distinguer les formes D et L.

Le domaine dual condensation-épimérisation (C/E) est un domaine qui possède à la fois la faculté de condenser les acides aminés fixés sur deux domaines de thiolation voisins, et d'épimériser l'acide aminé situé sur le premier module. La structure en domaines est donc la suivante : <C' A' T' C/E A T> avec le domaine C/E qui épimérise l'acide aminé porté par le domaine de thiolation T'. Ces domaines duaux n'ont été observés que dans le cadre de la synthèse de lipopeptides cycliques et chez très peu de genres bactériens. Ils ont étés mis en évidence chez les *Pseudomonas* qui possèdent dans leur génome à la fois des synthétases avec des domaines E qui produisent, par exemple, des pyoverdines [11] et des synthétases avec des domaines duaux C/E qui produisent des lipopeptides [9]. Récemment, des domaines duaux ont également été observés chez un *Streptomyces* [128] et chez un *Lysobacter* [50]. Il est à noter que ces bactéries colonisent toutes le sol qui est un environnement dans lequel les échanges de matériel génétique sont fréquents.

Le domaine d'épimérisation en β (Eβ) est un domaine d'épimérisation qui agit sur un carbone de la chaîne latérale, appelé carbone β. Le domaine qui a été observé modifie une

10. Action de changer la conformation d'un monomère.
11. Chromopeptides capables de fixer le fer.

thréonine en allo-thréonine dans la lysobactine [50]. Ce domaine est localisé dans un module <C Eβ A T>.

NORINE contient également des isoleucines et quelques dérivés épimérisés en β. Deux études expérimentales ont montré que les domaines A des synthétases de la la nostophycine [38] et de la coronatine [28] avaient une préférence pour l'incorporation directe d'une allo-isoleucine. Celle-ci serait donc synthétisée par une enzyme externe à la synthétase.

Le domaine A spécifique aux D-Ala de certaines synthétases incorpore directement le monomère sous sa forme D. L'épimérisation n'est donc pas effectuée par la synthétase, mais par une autre enzyme impliqué dans la voie de synthèse du monomère, appelée racémase externe. Celle-ci change la conformation du monomère avant qu'il ne soit sélectionné par la synthétase. A ma connaissance, seule l'incorporation directe de D-alanine a été observée chez quatre synthétases, celles produisant la cyclosporine [49] et l'HC-toxine [24] chez des *Fungi*[12], et celles produisant la fusaricidine [69] et la leinamycine [110] chez les bactéries. L'épimérisation de cet acide aminé est effectuée par l'alanine racémase (Alr ou EC 5.1.1.1).

1.2.4 Incorporation d'acides aminés modifiés

Les synthétases incorporent dans les peptides non-ribosomiques non seulement les 20 acides aminés présents dans les protéines ribosomiques, mais aussi d'autres acides aminés qui sont parfois dérivés des 20 principaux (voir partie 2.2.1). Ces dérivés peuvent être synthétisés par des domaines supplémentaires présents dans les synthétases. Nous avons déjà évoqué le domaine d'épimérisation, il en existe d'autres.

Le domaine de N-méthylation (NM) ajoute un groupement méthyle CH_3 à l'atome d'azote de l'acide aminé fixé sur le domaine de thiolation du module suivant (voir figure 1.3). Ce domaine est localisé après le domaine de thiolation : <C A T NM>.

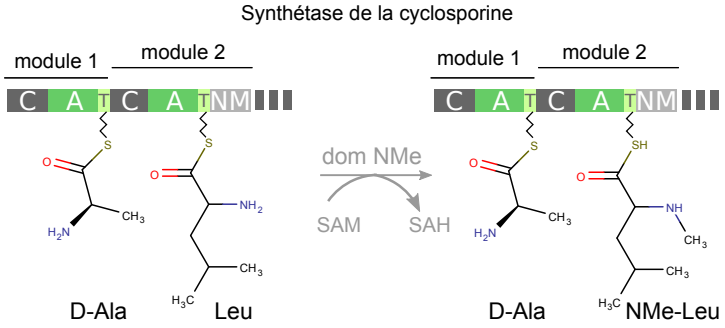


FIGURE 1.3 – Les deux premiers modules de la synthétases de la cyclosporine. Le domaine A du premier module reconnaît une D-alanine. Le domaine NM du deuxième module ajoute un méthyle à une leucine, associé au coenzyme (S)-adenosyl méthionine (SAM).

Le domaine de formylation (F) ajoute un groupement formyle C(=O)H, aussi appelé aldéhyde, au groupement amine NH_2 d'un un acide aminé (voir figure 1.4). Ce domaine est localisé avant le domaine A et, *a priori* dans des modules d'initiation : <F A T>.

12. Terme scientifique pour champignons.

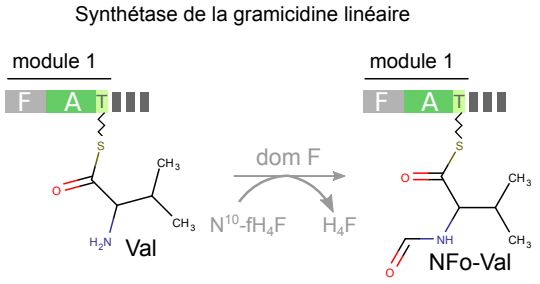


FIGURE 1.4 – Les premiers module de la synthétase de la gramicidine linéaire. Le domaine F ajoute un groupement formyle à une valine, associé au coenzyme N^{10}-formyltetrahydrofolate

Le domaine de cyclisation (Cy) forme un cycle interne à partir de deux acides aminés dont un possède un groupement réactif de type OH ou SH. Le domaine de cyclisation remplace le domaine de condensation. Les deux modules adjacents ont la structure suivante : <C A T Cy A' T'>. Si une cystéine est liée au domaine T', alors une thiazoline est formée (voir la figure 1.5). S'il s'agit d'une sérine ou d'une thréonine, alors une oxazoline est formée.

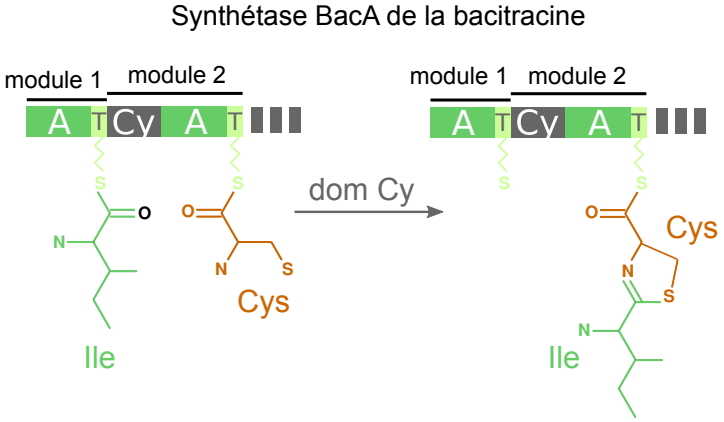


FIGURE 1.5 – Le domaine Cy de la bacitracine et les domaines voisins. Le module 1 incorpore une isoleucine et le module 2 une cystéine. Le domaine de cyclisation (Cy) forme une thiazoline, c'est-à-dire une molécule qui contient un cycle comprenant un atome d'azote N et un atome de soufre S, alors appelé hétérocycle du fait qu'il n'est pas composé que de carbones.

Le domaine d'oxydation (Ox) agit après le domaine de cyclisation pour transformer la thiazoline en thiazole ou l'oxazoline en oxazole par oxydation (ajout d'une double liaison dans l'hétérocycle). La position de ce domaine varie selon les enzymes. L'article [101] présente deux configurations. Dans la protéine EpoB, membre de la synthétase de l'épothilone, le domaine Ox est enchâssé dans le domaine A, formant un module : <Cy A Ox A T>. Dans la protéine BImIII, membre de la synthétase de la bléomycine, le domaine Ox suit le domaine T, formant la structure : <A T Ox>.

Le domaine de réduction (Rn) a une action opposée au domaine d'oxydation puisqu'il agit également après le domaine de cyclisation, mais transforme la thiazoline en thiazolidine ou l'oxazoline en oxazolidine par réduction (suppression de la double liaison de l'hétérocycle). La

réduction est peu fréquente chez les synthétases peptidiques non-ribosomiques. Un seul domaine est cité en exemple. Il est isolé sur une protéine et participe à la synthèse de la pyochéline [88].

Les halogénases externes ajoutent des halogènes à certains monomères. Les halogènes sont les atomes de fluor F, de chlore Cl, de brome Br, dâiode I et dâastate At. Seul le dernier atome, l'astate, n'est pas présent dans les monomères car il est radioactif.

1.2.5 Incorporation d'autres monomères

Les monomères qui ne sont pas des acides amino-carboxyliques sont soit incorporés directement par la synthétase dans le cas des lipides et des chromophores, soit par des enzymes externes aux synthétases, une fois que le peptide est formé dans le cas des monosaccharides et des polykétides. En général, les gènes qui codent ces enzymes sont localisés à proximité des synthétases peptidiques non-ribosomiques sur le génome. Cet ensemble de gènes participant à une même voie de synthèse et localisés dans une même région d'un génome est appelé cluster. Lors du processus de prédiction d'un PNR, il est donc important d'analyser également les gènes localisés à proximité des synthétases afin de compléter la structure avec les monomères autres que les acides aminés.

Le domaine de condensation ($C_{starter}$) incorpore les acides gras en début de chaîne peptidique. Comme son nom l'indique, ce domaine est localisé dans le premier module de la synthétase, le module d'initiation, qui est alors composé des domaines $C_{starter}$, A et T. Ce domaine particulier a la faculté d'effectuer une liaison entre un acide gras et l'acide aminé sélectionné par le domaine A du module. Il se trouve donc dans un module du type : $|C_{starter}$ A T>

Les modules spécifiques aux chromophores sont des protéines composées de domaines de type SPNR qui condensent des acides aminés. Ces derniers subissent alors une suite de réactions oxydatives [34], effectuées par des enzymes dédiées [35]. Par exemple, la tyrosine et l'acide 2,4-diaminobutyrique sont associés pour former le chromophore des pyoverdines.

Les glycosyltransférases ajoutent les monosaccharides au peptide [114]. L'équipe de recherche dirigée par D. Mohanty, National Institute of Immunology, New Delhi, India, a développé l'outil SEARCHGTr de prédiction de ces enzymes [60].

Les synthases de polyketides (SPK) sont d'autres enzymes modulaires qui synthétisent des polyketides (PK). Pour la synthèse de composés hybrides PNR-PK, les domaines SPK sont présents soit dans des protéines hybrides portant aussi des domaines SPNR, soit dans des protéines uniquement SPK. Pour plus d'informations concernant les SPK, voir la revue [127]. La plupart des outils d'annotation des SPNR sont également capables d'annoter les SPK, indiquant ainsi la nature hybride du produit final.

1.2.6 Libération du peptide

Pendant l'incorporation des monomères par les synthétases, le peptide en formation reste lié à l'enzyme en passant d'un domaine de thiolation à un autre lors des réactions de condensation successives. La dernière étape de la synthèse est donc la libération du peptide.

Le domaine de thioestérase (Te) de type I est généralement situé à la fin du dernier module de la synthétase : <C A T Te|. Il libère le peptide soit par hydrolyse pour former un peptide linéaire, soit par macrocyclisation pour former un peptide cyclique ou partiellement cyclique. N. Roongsawang *et al.* [97] ont observé une forte corrélation entre la répartition des séquences des domaines de thioestérase dans un arbre phylogénétique et leur fonction. Ils ont distingué cinq sous-types spécifiques aux peptides produits.

Le domaine réductase (Re) est également situé à la fin des synthétases : <C A T Re|. Ce domaine libère le peptide en réduisant l'extrémité C-terminale du peptide sous la forme d'un groupement alcool COH et non d'un groupement carboxyle C(=O)OH. Il est présent dans les synthétases des peptaibols (voir partie 2.2.4), des peptides non-ribosomiques produits par les *Fungi* qui se terminent par un alcool aminé [76]. Il est également présent dans les synthétases de lipopeptides et lipoglycopeptides dans la famille des mycobactéries. L'extrémité alcool libérée peut se lier à un monosaccharide [13][25].

Le domaine de condensation terminal (Ct) est un domaine de condensation capable de libérer le peptide en étant situé à la fin des synthétases peptidiques non-ribosomiques : <C A T Ct>. Dans la littérature, différents domaines Ct sont décrits de façon indépendante. Chez les *Fungi*, les Ct libèrent le peptide par macrocyclisation [43]. Dans l'embranchement des protéobactéria, deux bactéries possèdent un domaine de condensation terminal qui lie le peptide en cours de formation à une polyamine non liée à la synthétase. Ainsi, le peptide est libéré. La synthétase de la vibriobactine [62] et celle de l'ornibactine [4] suivent ce processus.

Le domaine de thioestérase (Te) de type II n'est, en fait, pas impliqué dans la libération du peptide. Ils réparent les bras des domaines de thiolation impliqués dans la fixation des monomères à la synthétase [73] afin qu'ils soient de nouveau fonctionnels. Ces domaines sont observés en tandem, à la suite d'un Te de type I dans les enzymes produisant les lipopeptides cycliques.

1.3 Bio-informatique pour l'annotation des synthétases

Maintenant que j'ai décrit le fonctionnement des synthétases peptidiques non-ribosomiques, je vais décrire les différents outils bio-informatiques qui ont été développés pour analyser les synthétases. Je commence par évoquer le problème de l'identification des gènes codant des SPNR, en particulier dans le contexte du séquençage à haut débit. Ensuite, j'ai choisi de ne pas décrire les outils un par un, mais reprendre les grandes étapes communes à ces outils et de mentionner les spécificités de certains logiciels. J'ai écrit cette partie à l'aide des articles dédiés aux logiciels en question et de trois revues écrites sur le sujet [7, 56, 57].

1.3.1 Prédiction de gènes et annotation de protéines

Avant de présenter les programmes d'annotation des synthétases, je vais introduire le contexte dans lequel ils se situent.

Ces dernières années, les techniques de séquençage ont progressé et permettent maintenant d'obtenir rapidement et pour un coût raisonnable la séquence nucléique complète d'un génome.

13. Brique de base des sucres.

Cependant, ce gain en rapidité est, pour l'instant, réalisé au détriment de la qualité des séquences qui contiennent des erreurs ou qui ne couvrent pas la totalité du génome. En effet, seuls des fragments de quelques dizaines à plusieurs centaines de nucléotides peuvent être séquencés. Afin de reconstruire la séquence génomique complète, il est nécessaire d'obtenir la séquence d'un grand nombre de fragments chevauchants. La reconstruction du génome est ensuite réalisée à l'aide d'outils bio-informatiques, cette étape est appelée l'assemblage. Selon la qualité et la quantité des séquences produites, l'assemblage peut aboutir à un génome morcelé, composé de fragments plus ou moins longs, appelés contigs, et pouvant être mal reconstitués. De plus en plus de projets de séquençage ne prévoient pas l'étape de finalisation du génome car elle est délicate à réaliser et nécessite de nouvelles expériences. Les génomes restent alors à l'état de brouillon permanent. La base de données GOLD (Genomes Online Database) [87] référençait, en avril 2013, 4325 génomes complets et 20 210 génomes en cours de séquençage. Parmi les génomes dit complets, 1855 (soit 43 %) sont à l'état de brouillon permanent. Une source majeure de difficulté lors de la finalisation de l'assemblage est la présence de répétitions dans les génomes. Or, du fait de leur structure modulaire, donc répétitive, les synthétases sont intrinsèquement une source de mauvais assemblage. Les conséquences touchent la justesse de l'enzyme ainsi reconstituée. Les domaines et modules risquent de ne pas être reconstruits dans le bon ordre, voire des parties de protéines différentes peuvent être associées à tord. Ce manque de qualité est un vrai problème pour la caractérisation des synthétases et autres enzymes modulaires de grande taille.

Une fois la séquence reconstituée, un autre travail d'analyse bio-informatique commence, à savoir la prédiction des gènes présents sur le génome et la détermination de la fonction des protéines pour lesquelles ils codent. Les gènes sont prédits soit par comparaison de la séquence nucléique traduite en protéines avec des protéines connues, soit par apprentissage à partir d'un ensemble initial de gènes connus.

La fonction des protéines est déduite grâce aux connaissances acquises expérimentalement sur des protéines modèles. Les protéines inconnues sont comparées aux protéines modèles selon différentes méthodes. Du fait de la grande quantité de séquences produites, l'annotation des génomes est majoritairement réalisée de manière automatique, à l'aide de programmes généralistes d'annotation qui aboutissent à des annotations erronées ou peu précises. Les protéines ayant une fonction synthétase peptidique non-ribosomique sont retrouvées sous différentes dénominations dans les banques de données telles que :

- «NRPS», «putative NRPS», «probable peptide synthetase protein» ou «Nonribosomal peptide synthetase» qui sont peu précises, mais désignent bien des proteines qui sont des NRPS ;
- «amino-acid adenylation domain protein», «AMP-dependent synthetase and ligase» ou «thiotemplate mechanism natural product synthetase» qui correspondent à la présence de domaines particuliers caractéristiques des NRPS ;
- «gramicidin S synthetase 2», «NRPS similar to ACV synthetase», «Similar to bacitracin synthetase 1 (BA1), protein contains NRPS modules» ou «EntE» qui sont des noms précisant le produit de la synthétase qui, très souvent, n'est pas celui indiqué car une analyse fine des synthétases est nécessaire pour le déterminer ;
- «hypothetical protein» qui montre que le programme n'est pas parvenu à identifier la protéine en tant que synthétase.

En fait, il est presque impossible d'annoter automatiquement correctement une synthétase, c'est-à-dire de déterminer ses modules et son produit, par une simple comparaison de séquences telle qu'elle est faite par les programmes généralistes d'annotation. La cause est la suivante. Les synthétases ont toutes en commun le fait d'être composées de modules eux-mêmes divisés en

domaines. Les séquences des domaines pris individuellement sont similaires, par contre, l'enchaînement de ces derniers est propre à chaque synthétase et est caractéristique de son produit.

NORINE référence plus de mille peptides, il existe donc au moins mille synthétases différentes. C'est pourquoi il est nécessaire d'utiliser des programmes dédiés à l'annotation des synthétases afin de prédire correctement l'enchaînement de leurs domaines, ainsi que les monomères incorporés dans le peptide.

1.3.2 Logiciels d'annotation des synthétases peptidiques non-ribosomiques

Les programmes d'annotation de synthétases reprennent la stratégie des autres programmes d'annotation à savoir la comparaison avec des protéines modèles, ou plus précisément, des domaines modèles.

Ils suivent tous les mêmes étapes principales, leurs originalités résident dans la collecte et la définition des données modèles, ainsi que dans les techniques d'assignation des fonctions enzymatiques. Je vais décrire ces étapes en précisant les spécificités de certains programmes. Le tableau 1.1 reprend les principales caractéristiques des programmes d'annotation de synthétases peptidiques non-ribosomiques. Il est à noter que Clusean [125] n'est pas présenté dans ce texte car il s'agit d'un programme qui fonctionne en version installée sur un ordinateur et dont antiSMASH [80, 13] est une version améliorée et utilisable via un formulaire Web. De même, je ne présente pas le logiciel SMURF [63] car il n'annote pas les protéines, mais se contente d'extraire des clusters de gènes produisant des métabolites à partir d'une liste d'annotations fournie en entrée du logiciel.

Étape 1 : collecte des séquences de synthétases. Les synthétases peptidiques non-ribosomiques servant de modèles sont extraites des banques généralistes de séquences. Il est important de constituer un jeu de données représentatif des domaines étudiés, contenant suffisamment de données. Au fur et à mesure des années, les auteurs bénéficient des données de leurs prédécesseurs et ajoutent le résultat de leurs recherches au jeu existant. Mais, les annotations de synthétases basées sur des analyses expérimentales, sont encore rares dans les banques protéiques. Par exemple, UniProt ne compte qu'une soixantaine de synthétases peptidiques non-ribosomiques ayant le statut «reviewed», c'est-à-dire dont les annotations ont été vérifiées manuellement.

Étape 2 : extraction des domaines. Pour extraire les domaines, des connaissances préliminaires sont nécessaires. Dans l'idéal, elles sont apportées par l'expérimentation. Par exemple, une région de la synthétase est isolée et son fonctionnement est étudié, ou bien sa structure 3D est établie. En l'absence de données expérimentales, les régions conservées dans les alignements multiples témoignent de l'importance de celles-ci pour la fonction enzymatique. Un cercle vertueux part de quelques séquences bien caractérisées pour identifier de nouvelles séquences de la même famille et ainsi mieux définir les domaines fonctionnels. Ce processus est décrit dans le paragraphe suivant en ce qui concerne les synthétases peptidiques non-ribosomiques.

Étape 3 : détermination de signatures représentatives des domaines. Dès 1997, M. Maraheil *et al.* [78] ont défini des motifs caractéristiques des domaines A (adénylation), C (condensation), E (épimérisation), NM (N-méthylation), T (thiolation) et Te (thioestérase). Ils ont croisé des informations issues d'alignements de séquences et d'articles étudiants les mécanismes biochimiques de fonctionnement des domaines catalytiques. Les motifs ainsi définis sont de courtes

Tableau 1.1 – Principales caractéristiques des programmes d'annotation de synthétases

	Interrogation		Domaines étudiés		Prédiction	Format		
Logiciel	Mode	Seq	Type	Format	spécificité A	peptide	Principal point négatif	Réf.
NRPS-PKS...	Web	prot	tous	seq	3D vs SCC	mono	Résultats éparpillés sur plusieurs pages	[6]
NRPSPredictor	Web	prot	A	pHMM	TSVM+SCC (1)	mono	Seulement sur spécificité des domaines A	[95, 99]
Y. Minowa *et al.*	Privé	Ø	tous	pHMM	QET+pHMM (2)	mono	Non accessible	[81]
Clustscan	Install	ADN	tous	IPR	Ø	Ø	Pas de prédiction des monomères	[106]
PKS/NRPS...	Web	prot	tous	pHMM	BLAST vs SCC	mono	Erreurs pour certaines séquences	[7]
np.searcher	Web	ADN	A	seq	BLAST vs SCC	mono+SC	Sortie peu informative et difficilement lisible	[71]
antiSMASH	Web	ADN|prot	tous	pHMM	(1)+(2)	mono+SC	Moins performant sur les fungi	[80, 13]
NRPSsp	Web	prot	A	pHMM	pHMM	mono	Seulement sur spécificité des domaines A	[90]
NaPDos	Web	ADN|prot	C|KS	IPR+seq	Ø	Ø	Analyse un seul domaine C ou KS à la fois	[130]
NRPS/PKS...	Web	prot	A|AT	Ø	pHMM	1 mono	Analyse un seul domaine A ou AT à la fois	[64]

La colonne «format des domaines» renseigne sur la façon dont les domaines sont représentés par l'outil :
– seq : banque de séquences modèles, interrogée via BLAST ;
– pHMM : profils HMM construits spécifiquement pour l'outil
– IPR : profils HMM issus de la banque de domaines InterPro
La colonne «prédiction spécificité A» précise la méthode utilisée pour prédire le monomère reconnu par les domaines A identifiés :
– SCC : *Specificity Confering Code*
– TSVM : *Transductive Support Vector Machines*
– QET : *Quantitative Evolutionary Trace*

Dans la colonne «format peptide», la valeur «mono» signifie qu'une liste de monomères est fournie et «mono+SC» signifie qu'une structure chimique au format SMILES est également proposée, ce qui implique que les protéines de la synthétase ont été ordonnées les unes par rapport aux autres.

séquences (4 à 20 aa) avec des positions variables réparties le long des domaines. La liste des motifs a été étendue aux domaines Cy, Ox et R, dans un article de 2003 [103]. Depuis, plus de séquences de NRPS sont disponibles et les outils bio-informatiques ont progressés, facilitant la localisation des domaines sur une séquence inconnue.

Parmi les programmes d'annotation NRPS, deux (NRPS-PKS [6] et np.searcher [71]) utilisent directement les séquences brutes des différents domaines. La synthétase à annoter est comparée à ces séquences à l'aide de BLAST [14] [5]. Les régions qui s'alignent avec un domaine sont alors identifiées comme ayant la fonction enzymatique correspondante. Les auteurs de ces outils affirment que les performances sont meilleures que celles obtenues avec des profils HMM [15].

Le programme ClustScan [106] utilise principalement les profils fournis par les banques de domaines protéiques. Parmi ces dernières, InterPro [52] compile les données d'autres banques de domaines dont PFAM [91] et TIGRFAMs [48] qui contiennent des profils HMM représentant les principales fonctions enzymatiques des NRPS (voir tableau 1.2). Cependant, ces profils concernent souvent une famille de séquences plus vaste que celle des NRPS car les fonctions enzymatiques en question sont partagées par différentes enzymes.

Tableau 1.2 – Profils HMM des domaines NRPS dans InterPro

domain	InterPro id	database id
adenylation	IPR000873	PF00501
	IPR010071	TIGR01733
thiolation	IPR009081	PF00550
condensation	IPR001242	PF00668
thioesterase	IPR001031	PF00975

Les programmes restants (NRPSPredictor2 [99], Y. Minowa *et al.* [81], «PKS/NRPS Analysis Web-site» [7], antiSMASH [80, 13] et NRPSsp [90]) construisent leurs propres profils HMM à partir des séquences qu'ils ont collectées. Le programme HMMER [37] recherche chacun des profils sur la synthétase à annoter et localise ainsi les différents domaines.

Deux travaux de recherche se sont focalisés sur le domaine C (condensation) car il est divisé en sous-familles. Une première étude phylogénétique [94] a confirmé non seulement la spécialisation des domaines C selon leur contexte dans la synthétase (voir tableau 1.3), mais aussi une origine évolutive commune entre les domaines C, E, et Cy. Des motifs discriminants les sous-familles ont été déterminés, à partir des alignements multiples, à l'aide des programmes FRPred [39] et SDPred [59], ainsi que des motifs communs aux séquences des sous-familles avec le programme meme [8]. Les motifs issus de cette étude sont utilisés par le programme d'annotation antiSMASH [80, 13]. Une étude plus récente [130] étend le nombre de domaines pris en compte. Les auteurs ont construit un arbre phylogénétique à partir de séquences protéiques provenant des différentes familles et sous-familles connues, regroupées dans une banque de données. La séquence soumise à l'outil est alors comparée, à l'aide de BLAST, aux séquences de la banque afin de lui assigner une fonction potentielle. Elle est également insérée dans l'arbre à l'aide de FastTree [89], afin que l'utilisateur puisse visualiser sa ressemblance avec les séquences de fonction connue.

14. BLAST, *Basic Local Alignment Search Tool*, compare une séquence requête aux séquences d'une banque et retourne celles qui s'alignent le mieux avec la requête

15. Un profil HMM, *Hidden Markov Models*, modélise un alignement multiple en donnant des probabilités d'apparition des acides aminés à chaque position ainsi que celles des insertions-délétions

Tableau 1.3 – Sous-familles du domaine de condensation

Nom	Fonction
${}^{L}C_{L}$	condense deux L-monomères
${}^{D}C_{L}$	condense le D-monomère en fin de chaîne avec un L-monomère
C-starter	condense un monomère avec un acide gras (en début de chaîne)
Dual E/C	épimérise le dernier monomère de la chaîne et le condense au suivant
glyco-C	condense deux monomères, uniquement dans les NRPS de glycopeptides

Étape 4 : spécificité des domaines d'adénylation. En 1997, la structure 3D du domaine d'adénylation de la phénylalanine dans la synthétase de la Gramicidine S, appelée PheA, a été déterminée par E. Conti *et al.* [26]. Elle a permis d'identifier 10 acides aminés impliqués dans la reconnaissance du monomère. Sur 24 séquences de domaines A, E. Conti *et al.* ont observé une relation entre la nature des 10 aa du site actif et celle du substrat. Deux ans plus tard, la même équipe a confirmé cette relation sur 160 séquences de domaines A [105]. De plus, elle a obtenu des mutants capables de sélectionner le monomère prédit à l'aide du code formé par les 10 aa, appelé *Specifity Confering Code* (SCC), c'est-à-dire code conférant la spécificité du domaine A, ou encore code Stachelhaus, du nom de son auteur. Une étude équivalente a été menée par G. Challis *et al.* [23]. Ce code est encore intégré dans nombreux programmes d'annotation de NRPS qui prédisent les monomères sélectionnés. Le jeu de données utilisé pour construire le code est enrichi au fur et à mesure des années par les nouveaux domaines dont les substrats ont été déterminés expérimentalement. Les acides aminés du code ne sont pas consécutifs sur la séquence des synthétases. Leur extraction est faite par alignement du domaine étudié avec celui de PheA. Le programme NRPS-PKS utilise le programme GenThreader [58] pour replier le domaine inconnu selon la structure 3D de PheA. Ainsi, les positions impliquées dans la reconnaissance du substrat sont identifiées avec une meilleure précision.

En 2005, C. Rausch *et al.* ont proposé une nouvelle méthode de prédiction des monomères sélectionnés par les domaines A, implémentée dans NRPSPredictor [95]. D'une part, les acides aminés du site actif pris en compte ont été étendus à 34 positions, soit les acides aminés situés à 8 Å du substrat au lieu de la distance de précédente 5,5 Å. D'autre part, la technique d'apprentissage TSVM (*Transductive Support Vector Machines*) a servi pour la prédiction. Cette méthode prend en entrée des vecteurs de tailles quelconques représentant les données d'apprentissage. Ici, il s'agit d'indices physico-chimiques décrivant les 34 aa du site actif. Les SVM permettent de calculer un hyperplan capable de séparer les vecteurs du jeu de données positives de celui des données négatives. Dans cette étude, les données positives sont les domaines A sélectionnant un même acide aminé et les négatives, tous les autres domaines. Les TSVM permettent de prendre en compte des données non annotées, en se basant sur le principe que des vecteurs proches vont appartenir au même jeu (positif ou négatif). Ainsi, la quantité de domaines A utilisés pour l'apprentissage a pu être significativement augmentée avec 397 domaines dont la spécificité est connue et 833 dont elle est inconnue. Enfin, les auteurs ont formé des petits et grands groupes de monomères, basés sur les propriétés physico-chimiques, toujours pour palier la faible quantité de données. NRPSPredictor2 est une amélioration du premier programme basée sur un ensemble d'apprentissage plus grand (79 domaines de spécificité connue chez les bactéries et 100 chez les fungi et, respectivement, 4282 et 814 pour les domaines de spécificité inconnue), de nouveaux indices pour construire les vecteurs et une stratégie d'apprentissage plus fine [99].

Une troisième stratégie a été proposée par Y. Minowa *et al.* pour prédire les monomères

sélectionnés par les domaines A [81]. Celle-ci n'utilise pas la structure 3D de GrsA, mais se base sur les positions conservées au sein des alignements multiples réalisés à partir des séquences de domaines A reconnaissant les même monomères. La mesure appelée *Quantitative Evolutionary Trace* (QET) ou trace quantitative de l'évolution, permet d'extraire les résidus informatifs des alignements qui sont ensuite modélisés à l'aide de profils HMM. Chaque profil représente donc les domaines A reconnaissant un monomère particulier. Le monomère prédit pour un domaine A inconnu est celui dont le profil s'aligne le mieux avec celui-ci.

Il est à noter que le programme d'annotation antiSMASH [80, 13] combine les 3 méthodes de prédiction présentées et considère que le monomère prédit majoritairement est le plus plausible.

Dernièrement, les auteurs de deux outils, NRPSsp [90] et «NRPS/PKS substrate predictor» [64] ont choisi de prédire les acides aminés reconnus par les domaines A à l'aide de profils HMM. Ils annoncent de meilleurs taux de prédiction que d'autres logiciels, qui sont peut-être dus à un jeu de données d'apprentissage plus important.

Exemple d'utilisation d'outils. Pour compléter cette présentation des outils d'analyse des synthétases, je commente les résultats obtenus pour le cluster de gènes de la synthétase de la bacitracine qui contient un domaine de cyclisation et des domaines d'épimérisation. La page de description de la bacitracine, avec sa structure monomérique est présentée dans la figure 2.20.

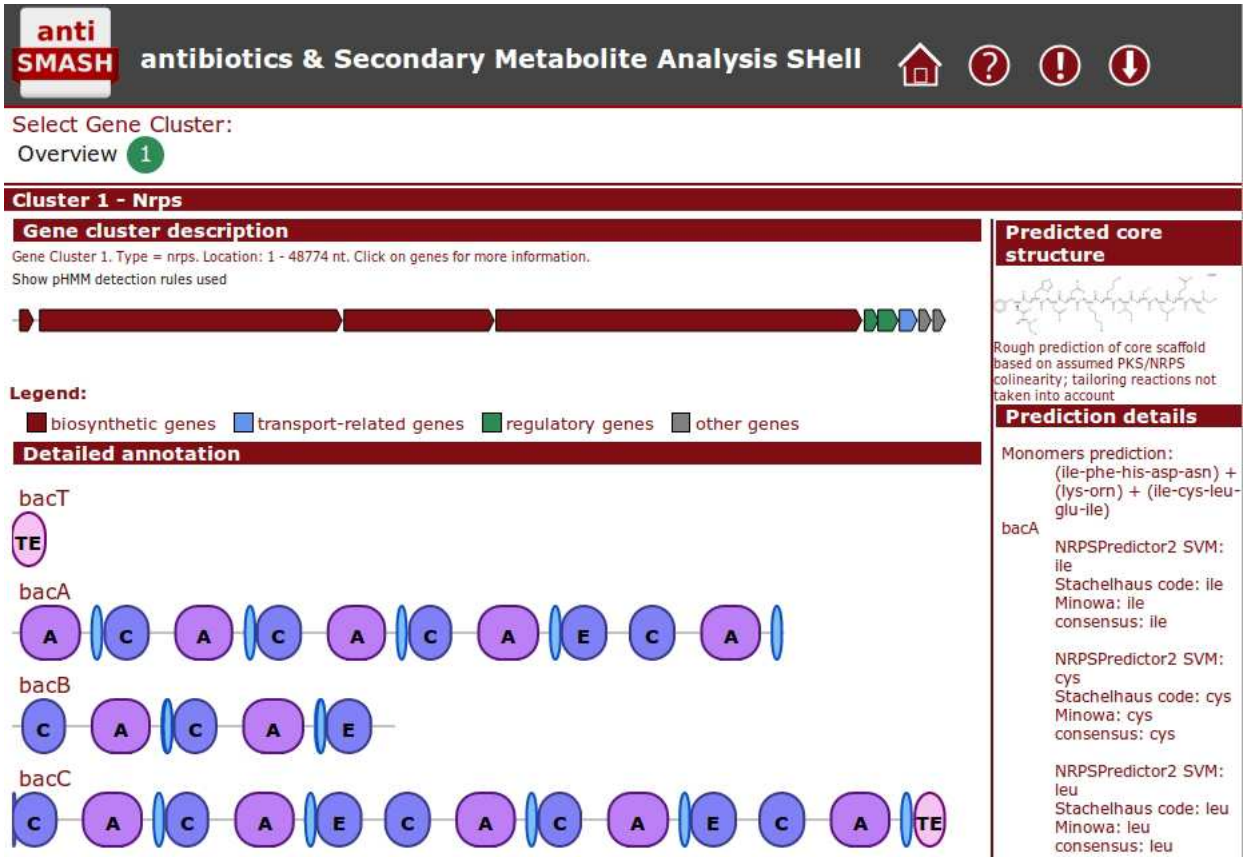


FIGURE 1.6 – Résultats du logiciel antiSMASH appliqué au cluster de gènes de la synthétase de la bacitracine.

Le logiciel antiSMASH identifie 4 gènes codant des SPNR et 4 autres gènes impliqués dans

la régulation (voir figure 1.6). La synthétase est composée de 12 modules répartis sur 3 gènes plus un domaine de thioestérase (Te) isolé. Les monomères prédits pour chaque domaine A sont indiqués dans le cadre de droite. Ils sont regroupés selon les protéines qui les incorporent. Les acides aminés indiqués correspondent bien à ceux qui sont observés dans le peptide réel. La configuration D de certains monomères peut être déduite de la présence de domaines E et des domaines ${}^{D}C_{L}$ associés (information visible lorsque l'on clique sur le dessin du domaine concerné). La position des domaines correspond bien aux D-acides aminés présents dans la bacitracine. Mais, l'énantiomérie des monomères n'est pas reportée sur la prédiction affichée dans antiSMASH. Enfin, le domaine de cyclisation qui se trouve au début du deuxième module est bien reconnu (annotation visible lorsque l'on clique sur le domaine). Seule la formation d'un cycle partiel n'est pas déductible de ces résultats.

Les résultats obtenus pour la protéine bacA de la synthétase de la bacitracine avec l'outil «PKS/NRPS Analysis Web-site» sont représentés dans la figure 1.7. Cet outil prend en entrée des protéines, seul le résultat pour une protéine est représenté dans ce manuscrit. Le domaine Cy et le domaine E sont également bien identifiés. Par contre, les prédictions pour deux des acides aminés incorporés par les domaines A sont erronées. Les isoleucines sont annoncées comme étant des leucines. Les résultats pour la protéine bacB sont exacts. Dans la protéine bacC, l'outil n'est pas capable de prédire l'acide aspartique (asp). La qualité de prédiction est donc satisfaisante.

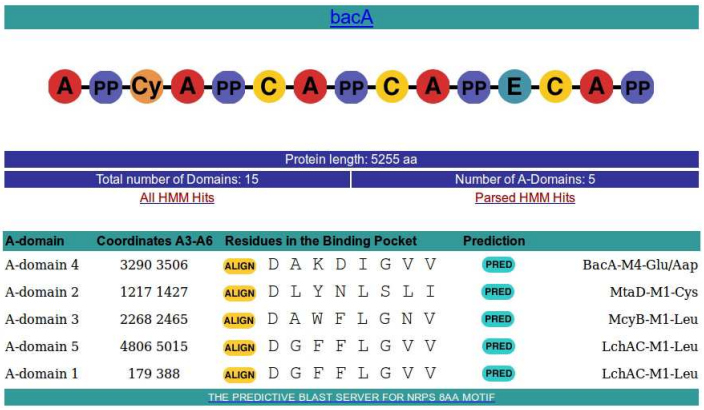


FIGURE 1.7 – Résultats du logiciel «PKS/NRPS Analysis Web-site» appliqué à la protéine bacA de la synthétase de la bacitracine.

Un bilan sur les outils existants met en avant leurs avantages et inconvénients et, surtout, permet d'identifier les manques globaux qui restent à combler. Dans le tableau 1.1, j'ai déjà mentionné la principale limite de chacun des programmes, qui n'est parfois qu'un manque de convivialité ou de facilité d'utilisation. D'ailleurs, il est conseillé d'utiliser plusieurs logiciels afin de pouvoir comparer leurs résultats et ainsi, maximiser les chances d'aboutir à une annotation proche de la réalité. Aucun logiciel ne donne des résultats très mauvais, ils ont chacun leurs faiblesses sur certaines données. Pour ma part, j'utilise «PKS/NRPS Analysis Web-site» qui a l'avantage de donner un résultat en quelques secondes à partir de la séquence protéique d'une synthétase et antiSMASH qui donne les résultats les plus fiables et les plus détaillés directement à partir d'une séquence génomique. D'après les observations que j'ai pu faire, ce dernier outil retrouve bien toutes les synthétases d'un génome bactérien et les prédictions des acides aminés sélectionnés sont exactes dans la plupart des cas. Le protocole complet d'analyse sur lequel j'ai travaillé est décrit plus en détails dans la partie 1.4.1.

Malgré la qualité de prédiction atteinte par les outils récents, il reste des fonctionnalités qui ne sont, pour l'instant, proposées par aucun outil. Notamment, tous prédisent les domaines optionnels qui modifient le monomère reconnu par le domaine A. Mais ils ne prennent pas en compte leur présence lors de la reconstruction de la liste des monomères incorporés par la synthétase. De la même façon, la fonction des protéines non SPNR qui appartiennent aux clusters de gènes prédits par certains outils, np.searcher et antiSMASH, n'est donnée qu'à titre indicatif. Cependant, leur présence pourrait être utilisée pour prédire, par exemple, l'incorporation de monomères autres que des acides aminés (voir 1.2.5). Une autre limite est la différence importante entre la capacité des outils à prédire la spécificité des domaines A (une quarantaine d'acides aminés différents dans le cas de antiSMASH), et les centaines d'acides aminés référencés dans Norine. Même en enlevant les dérivés obtenus par modification de l'acide aminé reconnu par le domaine A, il reste de nombreux acides aminés pour lesquels aucune, ou très peu de, synthétase(s) les incorporant n'ont de séquence disponible. Enfin, très peu de données, même expérimentales, ne sont disponibles concernant les mécanismes impliqués dans la formation de peptides non linéaires. Ainsi, les outils ne peuvent prédire la structure 2D acquise par le peptide.

1.4 Exploration du potentiel de synthèse de genres bactériens

Je présente maintenant mes contributions à la bio-informatique pour l'analyse des synthétases peptidiques non-ribosomiques. Je commence par le travail de conception de logiciels et de protocoles bio-informatiques que j'ai réalisé en collaboration avec mes collègues du laboratoire ProBioGEM et du LORIA. Ensuite, je décris les analyses bio-informatiques que j'ai moi-même réalisées ou dont j'ai encadré l'opérateur.

1.4.1 Protocole d'annotation de synthétases peptidiques non-ribosomiques

Je vais décrire le protocole, mis au point en collaboration avec Valérie Leclère, tel qu'il a été optimisé avant la mise à disposition de l'outil antiSMASH en 2011 et préciser les étapes qui ne sont plus nécessaires à présent. Ce protocole a pour objectif d'identifier de façon exhaustive les peptides non-ribosomiques produits par un organisme donné à partir de la séquence de son génome. L'avantage de travailler avec un laboratoire de micro-biologie, est que l'analyse bio-informatique peut être validée par des résultats expérimentaux. Certaines expériences seront également présentées car elles nécessitent des outils bio-informatiques dédiés, dont certains ont été développés sous ma responsabilité.

Recherche des synthétases au sein des génomes et proéomes. Cette première étape était fastidieuse avant la publication du logiciel antiSMASH car il n'existait pas de méthode satisfaisante et simple d'utilisation. En effet, les logiciels de comparaison de séquences tels que BLAST ne sont pas adaptés à l'identification de toutes les synthétases d'un organisme du fait de la structure modulaire, donc unique, des SPNR. L'exhaustivité des résultats dépend de la séquence requête qui est comparée soit à la séquence génomique traduite (utilisation de tblastn), soit directement au protéome de l'organisme étudié (utilisation de blastp). La structure en domaines de la synthétase requête contraint fortement la structure des synthétases identifiées par comparaison de séquences. Ainsi, il est nécessaire de compléter la stratégie de recherche avec d'autres méthodes afin d'être le plus exhaustif possible. Nous utilisions la recherche par mots-clés a sein des annotations disponibles et l'extraction des protéines les plus longues codées par

un génome, les synthétases étant, en général, très grandes.

J'ai proposé de compléter ce dispositif avec la recherche non plus d'une protéine modèle à l'aide de BLAST, mais directement des domaines caractéristiques des SPNR à l'aide du logiciel HMMER [37] de recherche de profils HMM. Cette technique permet de palier aux faiblesses de BLAST puisque la recherche se focalise sur les domaines, quelque soit leur agencement, et non sur une protéine entière qui est unique ou presque.

Une fois un premier ensemble de synthétases identifié, il est intéressant d'étudier également les gènes voisins afin de reconstituer le cluster complet. Ces gènes peuvent coder soit des synthétases non détectées par les autres méthodes, soit des protéines qui participent à la voie de synthèse du peptide. Comme évoqué dans les parties 1.2.5 et 1.3.2, la connaissance de la présence de ces gènes dans un cluster facilite l'identification du peptide produit.

Toutes ces étapes sont maintenant intégrées dans antiSMASH qui utilise justement la recherche de profils HMM afin d'identifier les gènes codant des SPNR et les protéines accessoires afin de reconstituer des clusters complets.

Analyse bio-informatique des synthétases. Elle est effectuée à l'aide des outils d'annotation présentés précédemment (voir 1.4) qui fournissent le découpage des enzymes en domaines et les acides aminés potentiellement incorporés par les domaines A. Nous croisons les résultats obtenus par différents outils [16] afin d'améliorer la qualité des prédictions, en particulier pour les domaines optionnels. Avant, nous complétions cette analyse avec NRPSPredictor pour améliorer la qualité de prédiction des acides aminés reconnus par les domaines A. Maintenant, antiSMASH suffit car il confronte les résultats des trois méthodes les plus fiables.

Une analyse experte reste indispensable afin de reconstruire au mieux les peptides produits par les clusters identifiés. Des connaissances sur la synthèse non-ribosomique sont nécessaires afin d'arbitrer entre les résultats des différents outils et obtenir une interprétation réaliste. Par exemple, la présence d'un domaine C en début de synthétase est représentatif de l'incorporation d'un lipide au début du peptide, ou encore, la présence de domaines E ou d'autres domaines optionnels implique la modification des acides aminés incorporés. Enfin, la présence d'enzymes non SPNR doit être prise en compte pour prédire la présence de monomères tels que des monosaccharides, des polykétides ou des chromophores.

Nous avons formalisé cette analyse avec nos collègues Marie-Dominique Devignes et Malika Smaïl-Tabonne de l'équipe Orpailleur de Nancy et proposé un pipeline d'annotation des synthétases appelé Florine. Pour l'instant, notre contribution a essentiellement porté sur la prédiction de la configuration D des monomères incorporés. Nous avons mis en évidence une région de 150 acides aminés située à la fin du domaine C et spécifique des sous-famille de ce domaine et des domaines E. J'ai participé à la constitution des jeux de données et à l'élaboration de la stratégie d'analyse des domaines. Puis, j'ai réalisé des arbres phylogénétiques sur les domaines complets et les régions que nous avons mises en évidence. Mais, ces régions sont trop courtes pour obtenir un signal significatif. J'ai également étudié les domaines A qui reconnaissent les alanines sous leur forme L ou leur forme D. En effet, l'alanine est le seul acide aminé qui est parfois incorporé directement sous sa forme D. Pour cela, j'ai extrait la séquence des quatre domaines connus pour incorporer directement une D-alanine (voir 1.2.3) et les autres domaines incorporant une L-alanine qu'elle soit ou non épimérisée par la suite. Le faible échantillonnage n'a pas permis

16. antiSMASH, «PKS/NRPS Analysis Web-site» et PKS-NRPS quand il est accessible (car il est régulièrement indisponible).

de trouver une signature spécifique aux domaines incorporant directement une D-alanine. Nous avons soumis ces résultats a journal PLOS ONE [22].

Caractérisation d'un peptide via des analyses bio-informatiques. Elle est possible grâce aux outils de comparaison de peptides développés dans NORINE. Le peptide peut être obtenu soit par analyse bio-informatique de la synthétase, soit par caractérisation expérimentale de la molécule. Les outils de NORINE, ainsi que leur utilisation sont présentés en détail dans la partie 2.4.2.

Dans les deux cas, les peptides obtenus grâce à une recherche par structure dans la base de données NORINE permet d'aller plus loin dans la prédiction du peptide. Si au moins un peptide est identique ou très proche de la requête, quelque soit le mode de recherche utilisé, il est fort probable que la synthétase étudiée produise ce peptide. Les différences entre la requête et le peptide de la base peuvent être dues à soit des erreurs ou des imprécisions dans la prédiction, soit la découverte d'un variant. Si la différence se situe au niveau du choix du dérivé incorporé ou de l'ajout d'un monomère qui n'est pas un acide aminé, alors il est fort possible que la prédiction ait été incomplète. Je vous conseille d'analyser les fonctions des protéines codées par les gènes du cluster étudié, afin de confirmer ou infirmer la différence observée en vous aidant des informations suivantes :

- un acide aminé sous la forme D a plusieurs origines possibles (voir partie 1.2.3) qui peuvent être vérifiées par analyse des domaines qui composent le module qui l'incorpore et de la sous-famille du domaine C du module suivant ;
- un groupement méthyle CH_3 lié à l'atome d'azote d'un acide aminé peut être confirmé par la présence d'un domaine de N-méthylation dans le module concerné (voir la partie 1.2.5) ;
- un groupement aldéhyde C(=O)H est dû à un domaine de formylation (voir la partie 1.2.5) ;
- une thiazoline, une oxazoline ou un de leurs dérivés sous la forme réduite ou oxydée est obtenu à l'aide d'au moins un domaine de cyclisation suivi d'un domaine A reconnaissant une cystéine, une sérine ou une thréonine (voir la partie 1.2.5) ;
- un lipide doit être corrélé à la présence d'un domaine $C_{starter}$ en début de synthétase (voir la partie 1.2.5) ;
- un chromophore est associé à la présence des gènes codant les enzymes effectuant les réactions oxydatives (voir la partie 1.2.5) ;
- les monosaccharides sont ajoutés par les glycosyltransférases (voir la partie 1.2.5) ;
- un monomère du type polykétide doit être corrélé à la présence de domaines de type SPK (voir la partie 1.2.5). Cependant, NORINE ne contient pas beaucoup d'hybrides PNR-PK.

Dans le cas d'une différence sur au moins un monomère, il n'est pas toujours facile de discerner entre les erreurs de prédiction et la découverte d'un variant. L'utilisation de plusieurs logiciels de prédiction, une analyse du contexte telle que les organismes producteurs du peptide prédit et du peptide de la base, l'existence de variants connus ou les articles décrivant le peptide de la base peuvent aider à statuer.

Au delà de la composition en monomères, la détection de ressemblances entre une requête et des peptides de la base peut aider à déterminer la structure 2D du peptide, si celle-ci n'est pas connue. En effet, si la composition de la requête est identique ou proche de celle d'un peptide ou d'une famille de NORINE, alors il est fort probable que le peptide prédit possède la même structure que cette famille. Pour l'instant, je n'ai pas de corrélation à proposer entre une structure possible pour un peptide et les gènes présents dans le cluster producteur. Établir des liens entre les deux est une des perspectives que je propose.

Enfin, la possibilité d'afficher certaines annotations concernant les peptides obtenus suite à une recherche par structure dans NORINE, y compris sous forme graphique, permet de prédire la fonction du peptide étudié. En effet, si la plupart des peptides qui ressemblent à la requête ont la même activité, alors ce peptide a de fortes chances d'avoir également cette activité. Un outil de prédiction d'activité des peptides non-ribosomiques permettra bientôt de compléter ces résultats. Ammar Abdo Hasan a obtenu des taux de prédiction intéressants [3], mais des études complémentaires sont nécessaires avant de pouvoir incorporer l'outil dans NORINE.

Détermination expérimentale de la structure des peptides par spectrométrie de masse. Elle est indispensable pour compléter le protocole. Elle peut être réalisée avant ou après les prédictions bio-informatiques. Dans le premier cas, une recherche bio-informatique des peptides potentiellement produits par un organisme met en évidence des peptides intéressants qui sont ensuite recherchés et étudiés expérimentalement. Dans le deuxième cas, l'étude expérimentale d'un peptide nécessite une étude bio-informatique complémentaire pour aider à la détermination de sa composition en monomères. En effet, la caractérisation de la structure 2D des PNR est souvent effectuée par spectrométrie de masse. Or, de part leurs compositions et structures particulières, les résultats obtenus pour les peptides non-ribosomiques avec ces appareils sont délicats à interpréter. Si la séquence de la synthétase ou du génome de l'organisme produisant ce peptide est connue, il est possible de prédire les monomères qui composent ce dernier. La connaissance des monomères peut être exploitée lors de l'analyse des spectres de masse. Les outils dédiés à la spectrométrie de masse pour les PNR sont présentés dans la partie 2.3.2.

Dans le cadre de la thèse d'Aurélien Vanvlassenbroeck, nous avons conçu le logiciel Pyomass d'aide à l'interprétation des spectres de masse dédié aux pyoverdines qui sont des PNR particuliers composés d'une chaîne peptidique constituée à partir d'une quarantaine d'acides aminés différents, d'un chromophore et d'une chaîne latérale qui varie pour un même peptide. Le logiciel propose les peptides possibles à partir d'une masse donnée et réciproquement il calcule un ensemble de masses à partir d'un peptide. Cet ensemble correspond aux différentes chaînes latérales possibles et ions émis lors de l'analyse par spectrométrie. Le logiciel facilite également l'analyse de résultats d'expériences dites d'apport en excès d'un acide aminé. Il s'agit de fournir aux bactéries un acide aminé en quantités supérieures à leurs besoins pour les inciter à utiliser cet acide aminé au dépend d'un autre et de l'intégrer dans un peptide non-ribosomique. Ce changement d'acide aminé est observé pour certains domaines A qui sont dits permissifs, c'est-à-dire capables d'incorporer différents acides aminés. Pyomass propose, à partir de la composition d'un peptide, de l'indication de l'acide aminé fournit en excès et d'une masse expérimentale, les acides aminés qui ont pu être remplacés par celui en excès. Aurélien a utilisé Pyomass pour interpréter les résultats des expériences qu'il a réalisées au cours de sa thèse.

Identification expérimentale de synthétases. Il s'agit de déterminer la présence d'une synthétase donnée sur un génome de façon expérimentale et non via la bio-informatique. La technique utilisée est la PCR (*Polymerase Chain Reaction*). Je ne vais pas décrire la réaction de PCR en détails, l'information importante pour comprendre mes propos est le fait qu'elle utilise des amorces. Ce sont de courtes séquences d'ADN mesurant entre 10 et 20 nucléotides et capables de s'hybrider avec les séquences génomiques. Elles sont complémentaires[17] aux extrémités de la région à amplifier. Plusieurs stratégies existent pour les déterminer. Marlène Chollet et Athur Tapi ont mis au point un protocole pour définir des amorces capables de détecter des

17. Dans l'ADN, le A est complémentaire au T et le G au C.

synthétases peptidiques non-ribosomiques spécifiques [112, 2]. Ils commencent par extraire les séquences nucléiques des domaines de thiolation et d'adénylation de synthétases produisant des peptides de la même famille. Avant le développement de PrimerDeg, ils prédisaient les domaines sur les séquences protéiques puis calculaient les coordonnées correspondantes sur la séquence du gène et, enfin, extrayaient les régions d'intérêt. Ces étapes sont maintenant automatisées. Ensuite, les domaines sont alignés afin de déterminer des régions suffisamment conservées pour définir des amorces. Comme les synthétases produisant des peptides de la même famille ne sont pas strictement identiques, il est préférable d'utiliser des amorces dégénérées, c'est-à-dire un ensemble d'amorces ayant des séquences légèrement différentes les unes des autres. Les alignements multiples aident à la conception de cet ensemble en tenant compte de la diversité observée des séquences. Marlène Chollet et Arthur Tapi ont défini des règles et un coefficient de conservation des colonnes de l'alignement qui guident la détermination des amorces. Une partie de ces critères ont été intégrés dans PrimerDeg qui propose à l'utilisateur différentes paires d'amorces dégénérées.

Les applications sont variées. Par exemple, la synthétase d'un peptide extrait d'un milieu de culture peut être identifiée en se basant sur des synthétases produisant un peptide similaire. De même, les espèces d'un échantillon environnemental peuvent être caractérisées en se basant sur l'amplification de régions caractéristiques telles que des domaines de synthétases qui sont souvent spécifiques à une espèce voir une souche.

J'ai dirigé la conception et le développement du logiciel PrimerDeg qui facilite la détermination de ces amorces pour les SPNR, selon le protocole mis au point par Marlène Chollet et Arthur Tapi.

1.4.2 Annotation à grande échelle de synthétases

Le protocole que je viens de présenter a été perfectionné grâce à sa mise en application lors de l'annotation de génomes bactériens et de l'étude de peptides non-ribosomiques. J'ai moi-même participé à ces annotation ou encadré des étudiants les réalisant (voir 1.4.3). Je vais décrire les résultats marquants de ces annotations et expliciter ma contribution.

Exploration du potentiel de synthèse des PNR chez les bactéries de trois genres de la famille des *Enterobacteriaceae*. En 2010, Zohra Saci a étudié le potentiel de synthèse de bactéries appartenant à la famille des *Enterobacteriaceae*. Les trois genres étudiés, *Erwinia*, *Pectobacterium* et *Dickeya* ont été choisis parce qu'ils ont été classés dans le même genre, appelé *Erwinia*, en 1920 par Winslow *et al.* (pas d'article associé) du fait qu'ils étaient des entérobactéries causant des maladies aux plantes. Récemment, des études expérimentales et phylogénétiques ont abouti à séparation des trois genres [100, 75]. En 2010, 10 séquences complètes de génomes appartenant à ces trois genres étaient disponibles. Nous avons réalisé une phylogénie de ces organismes en alignant leurs génomes à l'aide du logiciel MAUVE [29] et prédit les synthétases présentes dans ceux-ci. Nous avons, ainsi, mis en évidence une forte corrélation entre la phylogénie des espèces et le potentiel de synthèse NRPS (voir figure 1.8).

Les *Erwiniae* tels qu'ils étaient définis avant le changement de nomenclature étaient connus pour produire deux peptides non-ribosomiques : l'indigoïdine, un pigment bleu qui protège la bactérie contre le stress oxydatif généré par la plante infectée, et la chrysobactine, un sidérophore qui permet de capter le fer de cette plante. Nous avons recherché les clusters responsables de la synthèse de ces molécules. Le cluster de la chrysobactine a été identifié chez tous les *Pectobacterium* et *Dickeya*, avec toutefois une différence dans les gènes composant ce cluster entre les deux

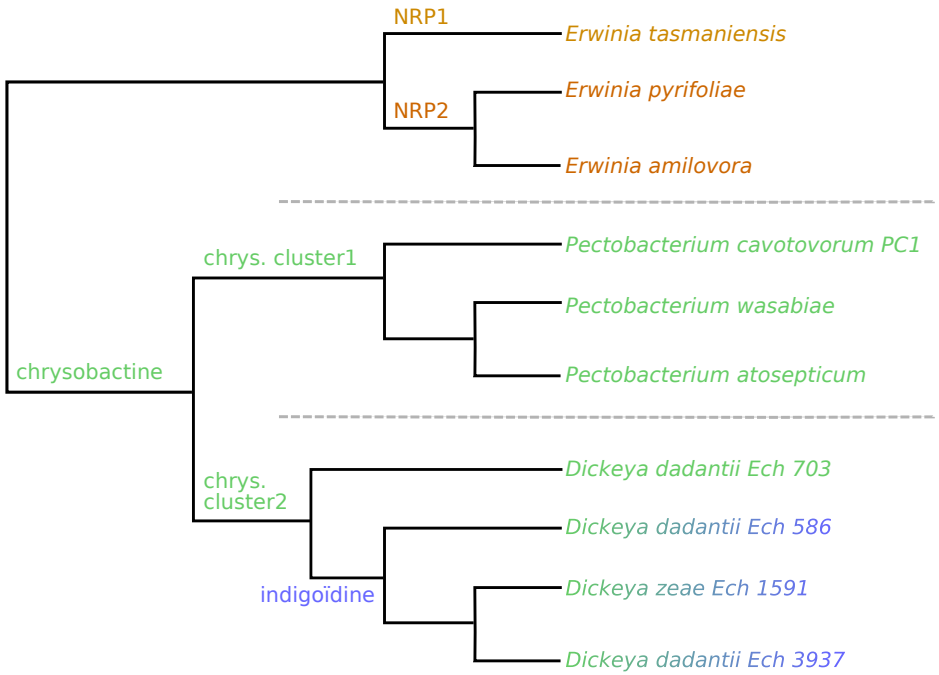


FIGURE 1.8 – L'arbre des *Erwinia*, *Pectobacterium* et *Dickeya*, annoté avec leurs PNR caractéristiques.

genres. Nous pouvons ainsi supposer que la chrysobactine était produite par l'ancêtre commun aux deux genres et que la voie de synthèse a évolué indépendamment après la spéciation. Le cluster de l'indigoïdine a lui été identifié chez 3 espèces de Dickeya (*D. dadantii Ech 586*, *D. zeae Ech 1591*, *D. dadantii Ech 3937*). D'après l'arbre phylogénétique construit, ces trois espèces ont bien un ancêtre commun différent de la quatrième espèce de *Dickeya* étudiée. Enfin, aucune des trois espèces d'*Erwinia* étudiée ne possède de cluster de synthèse de l'indigoïdine ou de la chrysobactine ou de tout autre PNR connu. Par contre, un cluster commun entre les deux espèces les plus proches a été trouvé. En conclusion, l'arbre obtenu à partir de la séquence des génomes complets, ainsi que le potentiel de synthèse des PNR sont congruents avec la nouvelle taxonomie des trois genres étudiés.

Exploration du potentiel de synthèse des PNR chez les bactéries du groupe des *Bacilli*. En 2011, j'ai travaillé avec Walaa Hussein, Valérie Leclère et Philippe Jacques sur l'annotation des synthétases présentes dans les 39 génomes du groupe des *Bacilli* et espèces proches disponibles à l'époque. Nous avons ainsi pu établir un catalogue des peptides non-ribosomiques produits par ces organismes. Nous avons pu observer que le spectre de production des PNR varie beaucoup d'une espèce à l'autre, au sein d'un même groupe. Il y a même des souches dans lesquelles nous n'avons pas trouvé de synthétases. Il semblerait que les différences de production sont liées à l'habitat des souches et non à leur filiation évolutive.

Après le travail exhaustif sur les peptides produits par les *bacilli*, nous nous sommes focalisés sur la kurstakine, un lipopeptide produit par ce groupe bactérien. J'ai participé à la recherche des synthétases produisant ce peptide dans les génomes bactériens disponibles. Une mini-revue

fait le point sur les connaissances actuelles sur cette molécule et apporte une liste des organismes producteurs identifiables par analyse bio-informatique [11].

Exploration du potentiel de synthèse des PNR chez les bactéries *Pseudomonas* fluorescents. Entre 2009 et 2012, dans le cadre de la thèse d'Aurélien Vanvlassenbroeck, nous avons inventorié les clusters de gènes NRPS présents chez les 20 génomes de *Pseudomonas* disponibles à l'époque. Nous avons également étudié la spécificité de domaines d'adénylation présents chez les bactéries de ce genre. Nous avons observé une légère permissivité de certains domaines capables de sélectionner une sérine à la place d'une thréonine, deux acides aminés ayant des propriétés physico-chimiques proches. Par contre, nous n'avons pas déterminé d'acides aminés spécifiques aux sites actifs de ces domaines par rapport à ceux des domaines non-permissifs. Nous avons également mis en évidence des acides aminés particuliers au niveau du site actif des domaines d'adénylation reconnaissant certains dérivés de l'ornithine et d'autres monomères présents dans les pyoverdines qui sont des peptides sidérophores produits par les *Pseudomonas*.

1.4.3 Encadrements et collaborations

Pour finir ce chapitre, je vais présenter les scientifiques avec qui j'ai travaillé. J'ai collaboré avec des collègues du laboratoire ProBioGEM, un laboratoire de microbiologie de l'Université Lille 1, deux autres collaborations non locales sont en cours à propos des outils et protocoles d'analyse des synthétases et une collaboration a été abandonnée. *Philippe Jacques*[18] et *Valérie Leclère*[19] sont présentés dans le chapitre 2 car ils font partie des membres fondateurs de Norine.

Marlène Chollet, maître de conférences en microbiologie et membre du laboratoire ProBioGEM, travaille sur l'identification expérimentale de nouvelles synthétases. J'ai collaboré avec elle sur l'automatisation du protocole de détermination d'amorces dégénérées spécifiques aux SPNR, grâce à l'outil PrimerDeg, développé par des étudiants en informatique ;

Marie-Dominique Devignes et Malika Smaïl-Tabonne, de l'équipe Orpailleur de Nancy, collaborent avec Valérie Leclère et moi pour mettre au point un pipeline d'annotation des synthétases, appelé Florine, qui permet une prédiction plus détaillée de la structure finale du peptide produit par un cluster de gènes SPNR ;

Tilmann Weber, responsable du groupe «Secondary Metabolite Genomics» de l'Université de Tuebingen en Allemagne, est l'un des auteurs de l'outil de prédiction de synthétases appelé antiSMASH. Nous travaillons avec lui depuis 2012 pour mettre en place des liens directs entre son outil et Norine. En réalité, nous sommes en discussion depuis plusieurs années déjà avec des membres de son Université en tant que partenaires de projets européens (demandes sélectionnées ou non pour être financées). De plus, nous avons organisé avec lui un workshop intitulé «Bioinformatics tools for NRPS discovery, from genomic data to the products» [20] qui s'est déroulé à Lille du 10 au 12 juillet 2013. Nous avons présenté nos deux logiciels et les interactions entre eux lors de conférences, couplées à des séances de manipulation en salle informatique. Nous avons accueilli 24 doctorants et jeunes chercheurs en biologie et biochimie venant d'Europe, d'Afrique et même du Brésil ;

Antonio Starcevic et Daslav Hranueli, du department bio-informatique de «Faculty of Food Technology and Biotechnology of Zagreb » en Croatie, ont développé l'outil d'annotation

18. Professeur en microbiologie à Polytech'Lille.
19. Maître de conférences HdR en biologie à l'UFR de Biologie.
20. http://www.lifl.fr/~pupin/workshopnrps/2013/

de synthétases ClustScan. En 2009, ils sont venus une semaine en France et nous avons passé une semaine en Croatie. Mais, leur outil n'est pas assez performant pour l'analyse des SPNR car ils sont spécialistes des synthases de polykétides, d'autres molécules produites par de gros complexes enzymatiques. Nous avons donc abandonné cette collaboration.

J'ai encadré des informaticiens et bio-informaticiens qui ont développé des outils pour faciliter le travail de mes collègues du laboratoire ProBioGEM. Certains de ces outils ont été présentés précédemment. Parmi les projets étudiants, je ne présente que ceux qui ont abouti à un outil fonctionnel et réellement utilisé.

Louise Ott, ingénieure en bio-informatique payée par l'Université Lille 1 dans le cadre du PPF Bio-informatique (Plan Pluri-Formations), a été affectée au projet SPNR pendant 5 mois en 2011, sous ma direction. Elle a développé pendant 1 mois une nouvelle version de l'interface de PimerDeg. Elle a également travaillé sur une base de données de synthétases peptidiques non-ribosomiques, appelée Doris, en concevant une interface d'interrogation via le web, ainsi que des scripts pour importer des données depuis les banques Protein du NCBI et UniProt. Mais, avant que cette base ne soit finalisée et rendue publique, nous avons appris que nos collègues de l'Université de Tuebingen en développaient également une, contenant les résultats de leur outil d'analyse de synthétases, antiSMASH. Nous avons donc choisi de collaborer avec eux plutôt que de proposer une base concurrente. Louise est actuellement Ingénieure d'étude en bio-informatique à l'Institut d'Immunologie de Strasbourg ;

Uciel Pablo Chorostecki, étudiant argentin en deuxième année de master en bio-informatique de l'Université de Rosario a effectué un stage de 4 mois financé par l'opération «Internship Inria« sous ma direction lors de l'été 2010. Il a travaillé sur l'amélioration d'un outil d'annotation des domaines de synthétases interne à ProBioGEM et l'intégration de PrimerDeg dans cet outil. Uciel est actuellement en thèse à l' «Instituto de Biología Molecular y Celular de Rosario» Argentine ;

Sylvain Bialasik et Vincent Knockaert, ont été encadrés par *Valérie Leclère* et moi-même lors de leur projet de première année de master MIAGE en 2011 (travail de deux cents heures). Ils ont développé le logiciel Pyomass qui est utilisé en interne à ProBioGEM pour analyser les résultats d'analyse par spectrométrie de masse de pyoverdine. Ils travaillent maintenant en tant qu'ingénieurs informaticiens dans des entreprises privées ;

Grégory Flipo et Achille Hennion, ont été encadrés par moi-même lors de leur projet de première année de Master Informatique en 2009 (travail de deux cents heures). Ils sont les premiers à avoir travaillé sur PrimerDeg. Ils travaillent maintenant en tant qu'ingénieurs informaticiens dans des entreprises privées ;

J'ai également participé à l'encadrement d'étudiants et doctorants en biologie sur quelques projets portant sur l'identification de nouvelles synthétases au sein de séquences génomiques.

Aurélien Vanvlassenbroeck a réalisé sa thèse de doctorat en «Ingénierie des fonctions biologiques » entre octobre 2009 et juillet 2012 [115], sous la direction de *Philippe Jacques* et co-encadré par *Valérie Leclère* et moi-même. Il a étudié le potentiel de synthèse peptidique non-ribosomique des bactéries du genre *Pseudomonas*. Son travail a consisté en des expériences de caractérisation de peptides non-ribosomiques produits par les souches étudiées, couplées à une analyse bio-informatique des synthétases présentes sur les génomes de *Pseudomonas* dont la séquence était disponible ;

Zohra Saci, étudiante en master recherche de bio-informatique de l'Université de Versailles a effectué son stage de 6 mois encadrée par *Valérie Leclère* et moi lors de l'été 2010. Elle a annoté les synthétases des bactéries du groupe *Erwiniae* en explorant les séquences des génomes disponibles à l'époque. Elle est actuellement en thèse dans l'équipe «Séquence, Structure et Fonction des ARN» de l'Institut de Génétique et Microbiologie, financée par l'Institut Curie à Paris ;

Walaa Hussein, a effectué sa thèse de doctorat en «Ingénierie des fonctions biologiques » sous la direction de *Philippe Jacques* et co-encadrée par *Valérie Leclère*. Elle a étudié expérimentalement le mécanisme de synthèse et de régulation de deux lipopeptides produits par *Bacillus subtilis*. De plus, elle a prédit les synthétases présentes dans les génomes des *bacilli* dont les séquences étaient disponibles. J'ai participé à ce travail d'annotation et à l'analyse des résultats en recherchant une corrélation entre l'habitat des souches et les PNR qu'ils produisent. J'ai été membre de son jury de thèse [1]. Walaa est maintenant chercheuse au «Genetics and Cytology Departement» au sein du «National Research Centre» au Caire, en Égypte ;

Arthur Tapi, a effectué sa thèse de doctorat en «Ingénierie des fonctions biologiques » sous la direction de *Philippe Jacques* et co-encadré par *Marlène Chollet* entre 2009 et 2012. Il a mis au point une nouvelle approche moléculaire de criblage par la réaction de polymérisation en chaîne, dans le but de détecter des souches de *Lactobacillus* et de *Bacillus* capables de produire des PNR. Cette approche a été facilitée par le développement de l'outil PrimerDeg. J'ai participé à son jury de thèse [111]. Il est maintenant chef de projet au département «Plant Sciences and Propagation» au centre R&D de Nestlé en Côte d'Ivoire ;

Thibault Caradec, effectue sa thèse de doctorat en «Ingénierie des fonctions biologiques » depuis 2011 sous la direction de *Philippe Jacques* et co-encadré par *Marlène Chollet* et *Valérie Leclère*. Il participe à l'identification de l'organisme producteur d'un PNR extrait de racines de la plante *Aster tataricus*. Le producteur de ce peptide est peut-être un *Fungus* symbiote de la plante. Une approche est de rechercher via des analyses bio-informatique des domaines caractéristiques des *Fungi* puis de localiser ces domaines *in situ*, directement parmi les organismes présents dans les racines à l'aide d'amorces spécifiques. Je le guide dans ses analyses bio-informatiques.

Chapitre 2

NORINE, plate-forme d'analyse bio-informatique des peptides non-ribosomiques

2.1 Introduction et motivations

Ce chapitre de mon HdR décrit le travail de pionniers de la bio-informatique pour les peptides non-ribosomiques qui a commencé sur Lille en 2006 et a abouti à l'unique plate-forme d'analyse bio-informatique des PNR, appelée NORINE. Rapidement, ce thème est devenu mon principal sujet de recherche et il l'est encore aujourd'hui. Ma contribution sur ce projet a évoluée au cours du temps. Au début, mes connaissances en biologie m'ont permis de rapidement comprendre les particularités des PNR et mes connaissances en informatique et bio-informatique d'identifier les problèmes algorithmiques intéressants soulevés par ces molécules. J'ai favorisé le dialogue entre mes collègues biologistes, Valérie Leclère et Philippe Jacques, et mon collègue informaticien, Gégory Kucherov. J'ai pu transférer mon expérience sur les bases de données SubtiList et GenoList acquises lors de ma thèse, à la conception de la base de données NORINE et de son interface d'interrogation. Au fur et à mesure des années, mon expertise sur les peptides non-ribosomiques a progressé tant à propos de leurs propriétés, que de leurs modes de synthèse et plus récemment de leurs structures chimiques. J'ai également une bonne vision des développements bio-informatique existants et, surtout des manques qu'il reste à combler concernant ces molécules. Je suis maintenant responsable du thème appelé NRP au sein de l'équipe Bonsai. J'ai lancé de nouvelles recherches pour ce thème. J'ai proposé et encadré le travail sur la prédiction d'activités des peptides qui a commencé lors de la thèse de Ségolène Caboche et s'est concrétisé lors du séjour postdoctoral de Ammar Hasan Abdo. Dernièrement, j'ai initié avec Laurent Noé, maître de conférences en informatique de l'équipe Bonsai, les recherches en chémo-informatique concernant les PNR. Yoann Dufresne commence une thèse en informatique sur ce sujet. J'ai donc un rôle moteur au sein de la *NORINE team*, l'ensemble des scientifiques ayant travaillé sur NORINE qu'ils soient membres de Bonsai ou ProBioGEM.

Après avoir décrit les principales caractéristiques des peptides non-ribosomiques dans la partie 2.2, je présenterai les outils et bases de données existants pour les peptides ou les petites molécules organiques dans la partie 2.3. Ensuite, je détaillerai dans la partie 2.4, les travaux que j'ai dirigés à savoir la conception de la base de données NORINE, ainsi que les outils et algorithmes dédiés aux PNR.

Les résultats décrits dans ce chapitre ont été présentés dans quatre articles publiés dans des journaux internationaux avec comité de lecture, ayant pour contenu :

1. la présentation de NORINE dans le numéro spécial, édition 2008, du journal *Nucleic Acids Research* dédié aux bases de données [20]. Cet article a été cité 55 fois selon Scopus et 86 fois selon google scholar, y compris par des chercheurs reconnus du domaine tel que le Prof. Christopher T. Walsh responsable d'un laboratoire du département de chimie biologique et pharmacologie moléculaire de *Harvard Medical School* à Boston, qui travaille sur les aspects moléculaires de la catalyse enzymatique avec un focus sur les antibiotiques et sidérophores ; le Prof. Mohamed A. Marahiel, responsable d'un groupe de recherche du département de Chimie à l'Université de Marburg en Allemagne spécialisé dans l'étude du mécanisme de synthèse non-ribosomique, ou encore le Dr. Pieter C Dorrestein, responsable d'un laboratoire du département de Pharmacologie, chimie et biotechnologie au sein de l'Université de Californie, San Diego, spécialisé dans l'étude de molécules thérapeutiques, dont les PNR, via la spectrométrie de masse. Certains articles parlent de NORINE sans citer un des articles que nous avons écrit ;
2. l'algorithme de recherche d'un motif structurel représentant un peptide non-ribosomique [21]. Cet article est cité 6 fois selon Scopus et 14 selon google scholar ;
3. les observations réalisées lors d'une étude statistique de grande ampleur menée par Ségolène Caboche à la fin de sa thèse [19]. Un des résultats majeurs est la mise en évidence d'une relation directe entre la structure des peptides et leur activité. Cet article est cité 10 fois selon Scopus et 21 fois selon google scholar ;
4. la conception d'une nouvelle représentation des PNR sous la forme d'un vecteur de comptage du nombre d'occurrences des briques de base des PNR présents dans un peptide donné [3]. Cette représentation est pertinente dans le cadre de la prédiction d'activité des peptides non-ribosomiques. D'ailleurs, un article décrivant une nouvelle méthode de prédiction d'activité de molécules est en cours de révision.

2.2 Diversité des peptides non-ribosomiques

Un peptide est un polymère linéaire composé d'acides aminés. Seule la taille diffère entre un peptide et une protéine, un peptide ne compte que quelques dizaines d'acides aminés, alors qu'une protéine en compte plusieurs centaines et même milliers le plus souvent. Les êtres vivants produisent des peptides via différentes voies qui sont la traduction d'un petit ARNm via les ribosomes, l'hydrolyse[21] enzymatique d'une protéine, ou encore la synthèse non-ribosomique. Cette voie alternative offre la possibilité de produire une grande variété de molécules dont je vais vous présenter la diversité que ce soit au niveau de leur composition, que de leurs structures ou de leurs activités. Je vais décrire ces peptides tels que NORINE permet de les découvrir, en présentant de nouveaux résultats obtenus sur leurs caractéristiques chimiques.

La définition officielle d'un peptide est une molécule composée d'au moins deux acides amino-carboxyliques liés par des liaisons peptidiques. L'IUPAC indique un poids moléculaire inférieur à 10 000 pour distinguer les peptides des protéines. Il se trouve que les peptides présents dans NORINE ne dépassent pas 5033,7 g/mol, masse de la polytheonamide B qui est le peptide le plus grand de NORINE avec 49 acides aminés. Les définitions d'un acide amino-carboxylique et d'une liaison peptidique seront données dans la suite de ce chapitre.

21. Coupure chimique.

Les peptides non-ribosomiques ne contiennent pas seulement des acides amino-carboxyliques, mais aussi des composés d'autres natures chimiques que je vais décrire dans la section suivante. C'est pourquoi, le terme monomère est utilisé à la place de acide aminé pour parler des composés de base des PNR.

2.2.1 Diversité de composition

Norine compte actuellement 528 monomères différents, tous présents dans au moins un peptide de la base. J'ai travaillé sur la classification des monomères afin de respecter au mieux les définitions chimiques des différentes familles.

Les acides amino-carboxyliques sont composés d'un groupement carboxyle C(=O)OH et d'un groupement amine NH_2. Le terme de peptide est souvent employé pour les molécules composées uniquement d'acides dit α-aminés ou, plus simplement, acides aminés dont les deux groupements sont portés par le même carbone, appelé carbone α (voir figure 2.1). Ce carbone porte également une chaîne latérale, appelée R, qui est spécifique à chaque acide aminé. Par exemple, la glycine porte la chaîne latérale la plus petite qui puisse exister à savoir un atome d'hydrogène, H. Lorsque les deux groupements sont séparés par deux carbones, la molécule est appelée acide β-aminé. Lorsque ils sont séparés par trois carbones, elle est appelée acide γ-aminé.

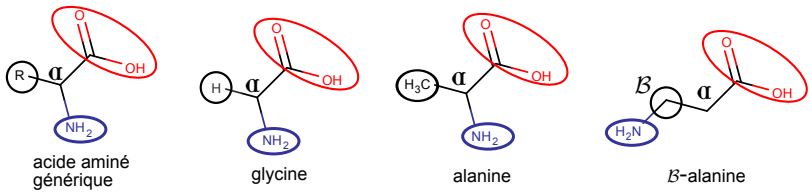


Figure 2.1 – Exemples d'acides amino-carboxyliques. Les groupements carboxyles, C(=O)OH, sont entourés en rouge, les groupements amines NH_2, sont entoutés en bleu et les chaînes latérales en noir. La position des carbones α et β est précisée

Théoriquement, il est possible de générer une infinité d'acides aminés en faisant varier la chaîne latérale. Cependant, les acides aminés incorporés dans les protéines et peptides synthétisés par la voie ribosomique ne sont que 20, plus quelques uns insérés dans des cas particuliers. Ces acides aminés sont qualifiés de protéogéniques. D'autres sont produits par les cellules comme intermédiaires de synthèse. La plus grande variété d'acides aminés observée dans la nature entre dans la composition des peptides non-ribosomiques. Norine contient 141 acides aminés au sens strict, c'est-à-dire des monomères avec les deux groupements C(=O)OH et NH_2 portés par un même carbone et 5 acides β-aminés.

Cette famille de molécules comprend également les dérivés des acides aminés, c'est-à-dire ceux qui ont subi des modifications qui peuvent être effectuées par les synthétases peptidiques non-ribosomiques elles-mêmes ou par d'autres enzymes (voir la partie 1.2).

Parmi les dérivés, les *énantiomères*[22] sont les plus courants (61 dans Norine). Dans les protéines et peptides synthétisé par la voie ribosomique, la quasi-totalité des acides aminés sont de configuration L. Une étude récente [65] a compté seulement 837 énantiomères de configuration D

22. Configurations L et D des acides aminés, c'est-à-dire les structures chimiques d'un même composé, mais non superposables.

parmi les 187 941 074 acides aminés de la banque de séquences protéiques SwissProt [23] [113], soit moins de 5 pour 10^6. Au contraire, 1920 occurrences de monomères sous la configuration D apparaissent parmi les 11206 monomères totaux de la base NORINE, soit 17%. De plus, près des deux tiers des peptides de NORINE contiennent au moins un monomère sous la configuration D.

D'autres dérivés sont générés par l'ajout de groupements. Par exemple, un groupement méthyle CH_3 est ajouté soit sur le groupement amine NH_2, soit sur le carboxyle C(=O)OH, soit sur les deux comme c'est le cas de la N,O-dimethyl-isoleucine (voir figure 2.2). Dans NORINE, 53 acides aminés sont *N-méthylés*, 7 sont *O-méthylés* et 4 sont O et N méthylés.

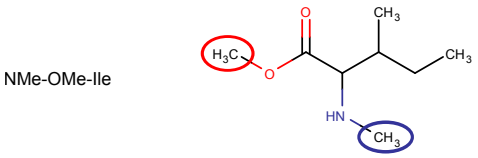


FIGURE 2.2 – La N,O-dimethyl-isoleucine porte un groupement méthyle (entouré en rouge) sur son groupement amine et un autre (entouré en bleu) sur son groupement carboxyle.

Des atomes de type *halogènes* (le chlore Cl, le brome Br, l'iode I et le fluor F) peuvent être ajoutés aux acides aminés. NORINE compte 19 acides aminés chlorés (par exemple la 3,4-dichloro-proline représentée dans la figure 2.3), 8 avec du brome et 1 seul avec du fluor.

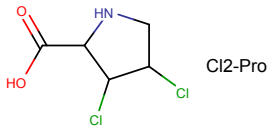


FIGURE 2.3 – La 3,4-dichloro-proline est une proline qui porte deux atomes de chlore.

Il y a aussi des dérivés particuliers qui sont observés uniquement dans les peptides faisant partie des peptaibols (voir la partie 2.2.4). Les *acides aminés acétylés* portent un groupement $C(=O)CH_3$ accroché au groupement amine NH_2 (voir Ac-Phe dans la figure 2.4). Les *alcools aminés* perdent le $= O$ de leur groupement carboxyle C(=O)OH pour ne garder que la fonction hydroxyle OH (voir Pheol dans la figure 2.4).

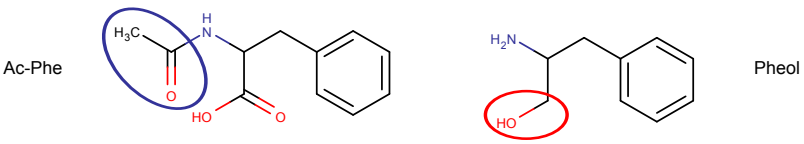


FIGURE 2.4 – Les deux dérivés de la phénylalanine, la N-acétyl-phénylalanine dont le groupement additionnel est entouré en bleu et la phénylalaninol avec sa fonction hydroxyle entourée en rouge.

Les acides gras font partie de la famille des lipides. Ce sont des acides aliphatiques monocarboxyliques, c'est-à-dire des molécules formées d'une chaîne carbonée avec au moins 4 carbones et terminées par un groupement carboxyle C(=O)OH. Ces composés sont donc capables de ne former qu'une liaison peptidique, avec une molécule portant un groupement amine NH_2. Ainsi , ils sont toujours incorporés en début de chaîne peptidique. Leurs chaînes peuvent être saturées

23. Banque internationale regroupant toutes les protéines ayant une annotation validée manuellement

(ni double, ni triple liaisons, voir C13 :0-NH2(3) dans la figure 2.5) ou insaturées (au moins une double ou une triple liaison, voir C8 :2(2.t4) dans la figure 2.5). Elles peuvent aussi contenir des groupements supplémentaires tels que méthyle CH_3, amine NH_2 ou hydroxyle OH (voir C13 :0-NH2(3) dans la figure 2.5), offrant la possibilité de former une deuxième liaison peptidique. NORINE compte 77 acides gras différents, mais nous avons choisi de ne pas référencer tous les variants d'un même peptide s'ils ne diffèrent que par leur acide gras.

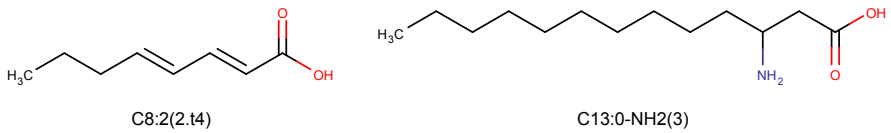


FIGURE 2.5 – Deux exemples d'acides gras ayant des longueurs de chaînes différentes. C8 :2(2.t4) est l'acide 2,trans4-octenoïque, un acide gras insaturé avec deux doubles liaisons. C13 :0-NH2(3) est l'acide 3-amino-tridecanoïque, un acide gras saturé qui porte un groupement amine.

Les monosaccharides sont les briques de base des glucides (sucres). À l'origine, les glucides ont été caractérisés par la formule chimique $Cn(H_2O)n$ (voir le glucose dans la figure 2.6), d'où leur nom anglais *carbohydrate* qui signifie hydrate (H_2O) de carbone (C). Leur formule ne respecte pas toujours strictement la définition précédente car ils peuvent avoir des groupements supplémentaires (voir l'érémosamine dans la figure 2.6). NORINE compte 17 monosaccharides différents qui forment une ou deux liaisons avec d'autres monomères. Pour les monosaccharides, la configuration D est la forme naturelle.

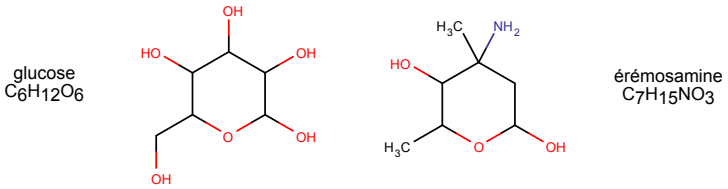


FIGURE 2.6 – Exemple de deux sucres, le glucose et l'érémosamine

Les chromophores émettent de la fluorescence par l'intermédiaire soit d'une alternance de simples et doubles liaisons chimiques, soit d'un complexe métallique. NORINE compte 5 chromophores différents. Les chromophores ChrA, ChrD, ChrI et ChrP (voir figure 2.7) forment une seule liaison peptidique via leur groupement carboxyle C(=O)OH et sont donc placés au début des peptides. Le chromophore ChrAct (voir figure 2.7) est lui impliqué dans deux liaisons peptidiques via ses deux groupements carboxyles.

Les polyketides sont des composés formés par une alternance de groupes carbonyles (C=O) et méthylènes ($R{-}CH_2{-}R$ ou $R{=}CH_2$) (voir figure 2.8). NORINE compte 7 polykétides différents.

2.2.2 Diversité de liaisons chimiques

J'ai déjà mentionné le fait que les peptides non-ribosomiques ne sont pas formés uniquement par des acides aminés liés entre eux par des liaisons peptidiques. Je vais présenter les diffé-

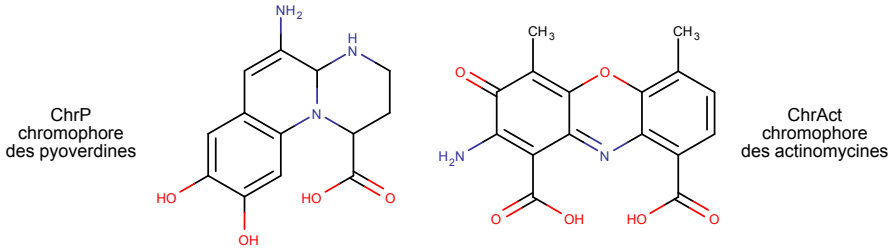


FIGURE 2.7 – Exemple de deux chromophores, celui des pyoverdines et celui des actinomycines.

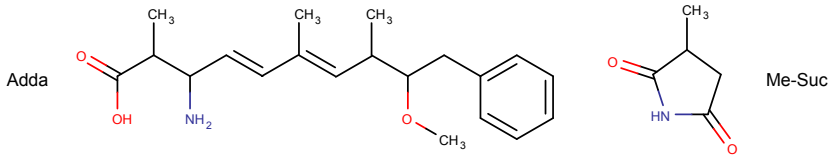


FIGURE 2.8 – Exemple de deux polyketides, l'acide 3-amino-9-methoxy-2,6,8-trimethyl-10-phenyldeca-4,6-dienoique (Adda) et la methyl-succinimide (Me-Suc).

rentes liaisons que nous avons déduites de nos observations. La formation d'une liaison chimique implique toujours la perte d'atomes et/ou d'électrons sur les monomères, je préciserai ces pertes.

La liaison peptidique est une liaison covalente [24] entre deux acides amino-carboxyliques (voir figure 2.9). La formation de cette liaison induit la libération d'une molécule d'eau issue de la fonction hydroxyle OH du groupement carboxyle et d'un atome d'hydrogène du groupement amine. Lors de la synthèse des protéines, qu'elle soit effectuée par les ribosomes ou les synthétase peptidique non-ribosomique, l'assemblage des acides aminés se fait toujours dans le même sens. Le composé en début de chaîne peptidique a un groupement amine libre (extrémité N-terminale) et celui en fin de chaîne un groupement carboxyle (extrémité C-terminale). Les nouveaux acides aminés sont toujours ajoutés à l'extrémité C-terminale. Chez les synthétases, cette liaison est formée par les domaines de condensation.

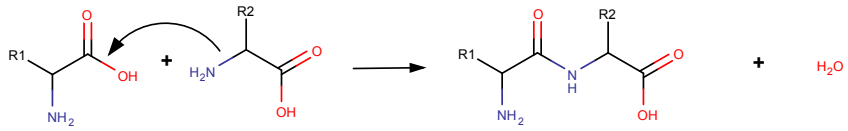


FIGURE 2.9 – La formation d'une liaison peptidique

Certains monomères ont des groupements NH_2 ou C(=O)OH supplémentaires qui leur permettent de former une troisième liaison appelée iso-peptidique. Les acides gras et chromophores sont eux aussi capables de former une liaison peptidique à l'aide de leur groupement carboxyle. Certains acides gras portent également des groupements supplémentaires qui leur permettent de former plus d'une liaison iso-peptidique. Les molécules synthétisées par la voie polykétide ont des structures très variées (voir la figure 2.8), certains possèdent des groupements carboxyles et/ou amines qui leur permettent de former des liaisons [iso-]peptidiques. D'ailleurs, certains sont des acides aminés.

24. Liaison chimique forte.

Le pont disulfure est une liaison covalente entre deux molécules portant un groupement thiol CSH. Un seul acide aminé protéogénique possède ce groupement, la cystéine, NORINE n'en contient pas d'autres, à part des dérivés de la cystéine. Six PNR de la famille des malformines et deux de la famille des triostines contiennent un pont disulfure. Le mode de formation de ces ponts n'est pas documenté dans la littérature.

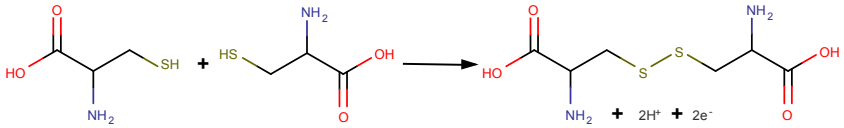


FIGURE 2.10 – La formation d'un pont disulfure

Un autre type de pont soufré est observé dans la famille des quinomycines. Dans un premier temps, un pont disulfure est formé par une oxydo-reductase (Ecm17), puis le pont dithioacétal RC(SR)SR est formé par une enzyme ressemblant à une S-adenosyl-L-methionine (SAM)-dependent methyltransferase (Ecm18) (voir figure 2.11) [123, 124].

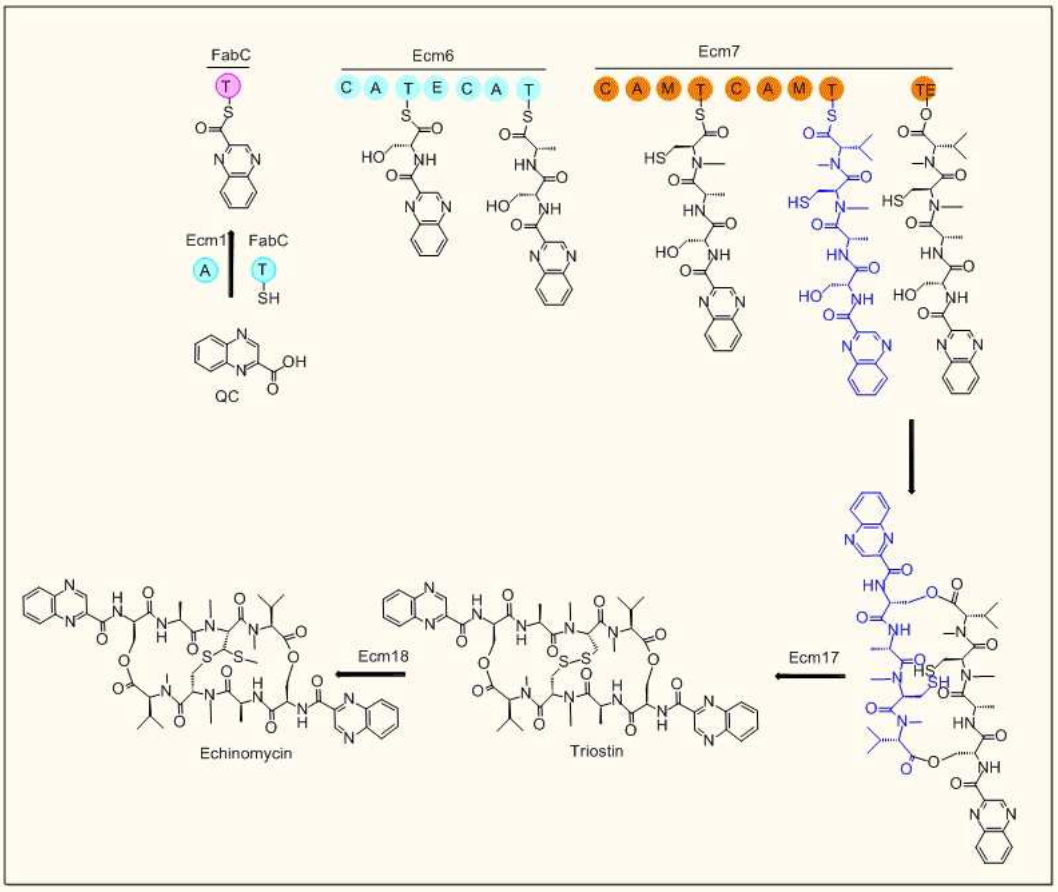


FIGURE 2.11 – La voie de synthèse de l'échinomycine. Source : Wikimedia Commons, author : Kvecchar.

Les liaisons phénolique et éther phénolique sont des liaisons covalentes formées entre monomères de la famille des phénols. Un phénol est un composé aromatique [25] portant une fonction hydroxyle OH. La présence des doubles liaisons et de l'hydroxyle en font une molécule réactive capable d'effectuer des réactions de couplage, c'est-à-dire l'association de deux molécules composées majoritairement de carbones à l'aide d'un catalyseur métallique. Les liaisons formées peuvent être de deux types, la liaison phénolique qui se fait directement entre deux carbones des cycles aromatiques (C-C) et la liaison éther phénolique qui présente un atome d'oxygène entre les deux cycles (C-O-C) (voir figure 2.12). Chez les PNR, ces réactions sont catalysées par des enzymes de la super-famille des cytochromes P450. Cette super-famille est constituée de nombreuses enzymes ayant en commun de catalyser l'oxydation d'une grande variété de molécules organiques au moyen, entre autre, d'un cofacteur appelé hème [26]. Le mécanisme précis de cette réaction n'est pas bien connu. N. Geib *et al.* ont étudié la première réaction de couplage mise en œuvre lors de la synthèse de la vancomycine [45, 107]. La présence de l'enzyme ferredoxine réductase, associée à la ferredoxine, ainsi que du co-facteur NADPH (nicotinamide adénine dinucléotide phosphate) et enfin de l'oxygène semblent nécessaires à la formation du couplage, en plus de l'enzyme OxyB, de la famille des cytochromes P450. De plus, cette enzyme agit sur le PNR en cours de synthèse, alors qu'il est encore attaché à la synthétase.

Les phénols de NORINE qui sont impliqués dans des réactions de couplage sont l'hydroxyphenyl-glycine (Hpg), la tyrosine (Tyr) et leurs dérivés. Plusieurs de ces monomères s'associent pour former des structures complexes (voir partie 2.2.3) observées principalement dans les glycopeptides [27] (voir partie 2.2.4 et figure 2.12).

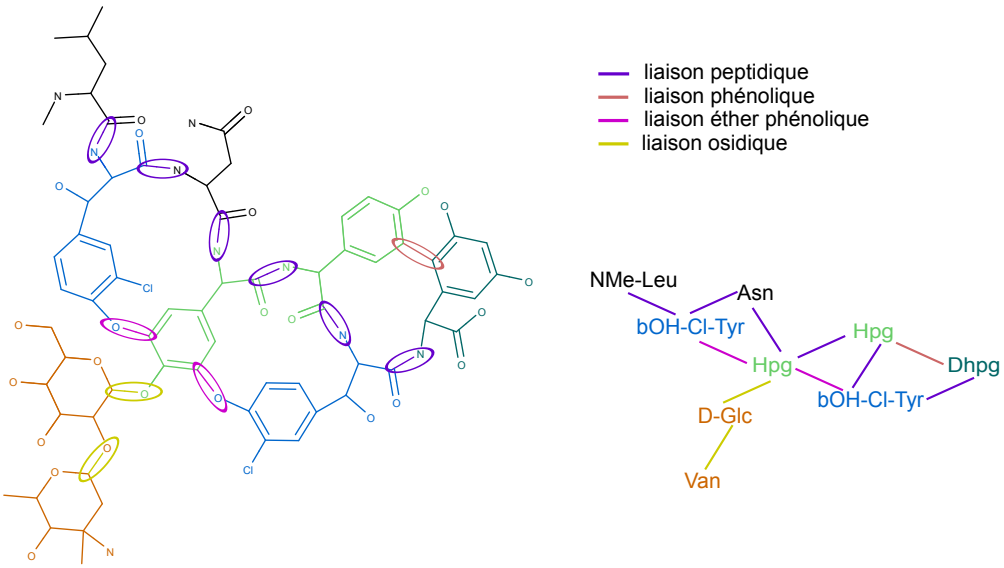


FIGURE 2.12 – Les structures chimiques et monomériques de la vancomycine, un glycopeptide.

La liaison osidique est une liaison covalente formée entre deux glucides ou un glucide et un autre composé tel qu'un acide aminé dans le cas des peptides. La liaison se fait entre une fonction hydroxyle placée à côté de l'atome d'oxygène présent dans le cycle du monosaccharide (voir la

25. Molécule formant un cycle carboné avec une alternance de simples et doubles liaisons.
26. Anneau organique entourant un atome de fer et pouvant accueillir un gaz tel que l'oxygène.
27. Molécules formées d'acides aminés et de glucides.

figure 2.12) et une autre fonction hydroxyle ou un groupement amine. Cette réaction est effectuée par des glycosyltransférases [51, 107].

2.2.3 Diversité de structures bidimensionnelles

Les protéines et peptides synthétisés par la voie ribosomique sont représentées par la succession des acides aminés qui les composent, appelée couramment séquence. Cette représentation est appelée structure 1D ou primaire. Le premier niveau de repliement d'une séquence constitue la structure 2D ou secondaire et le deuxième niveau la structure 3D ou tertiaire (voir figure 2.13). Enfin, la structure 4D ou quaternaire est l'association de plusieurs protéines, comme dans le cas des synthétase peptidique non-ribosomique (voir partie 1.2).

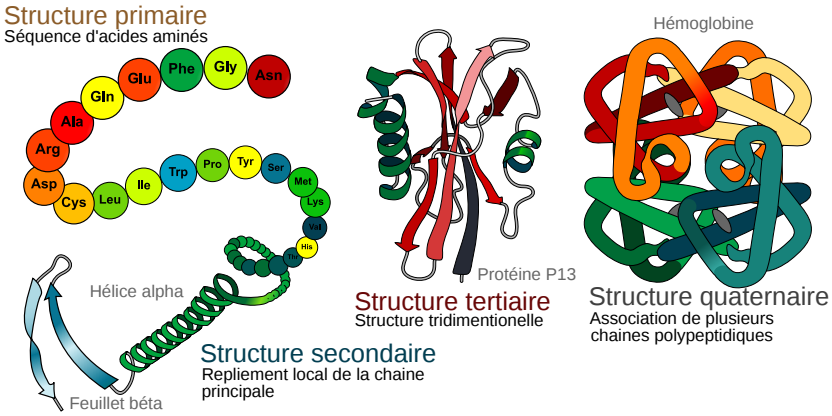


FIGURE 2.13 – Les différents niveaux de structure des protéines ribosomiques. Adapté de l'image Main_protein_structure_levels_fr, source : Wikimedia Commons, author : LadyofHats

Les peptides non-ribosomiques ne peuvent pas être représentés sous la forme d'une séquence (structure 1D) puisqu'ils ne sont pas tous linéaires. Le terme approprié pour parler de leurs structures est donc structure 2D, même s'il ne s'agit pas exactement de la même notion que celle utilisée dans le cas des peptides et protéines ribosomiques. Dans NORINE, les PNR sont représentés sous la forme de graphes dont les nœuds, aussi appelés sommets, sont les monomères et les arêtes sont les liaisons chimiques qui les relient. Ces graphes sont appelés structures monomériques par analogie avec les structures chimiques qui sont elles un graphe dont les nœuds sont des atomes et les arêtes les liaisons chimiques qui les relient (voir figure 2.12). Nous avons mis au point une syntaxe particulière, que j'appelle format NORINE, pour représenter les structures monomériques sans ambiguïté sous la forme d'une chaîne de caractères (voir figure 2.14).

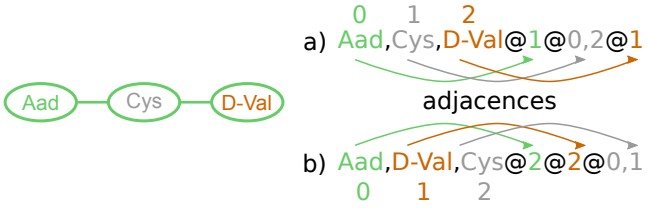


FIGURE 2.14 – Schéma représentant la structure monomérique pour le peptide ACV et la syntaxe sous forme d'une chaîne de caractères, avec deux exemples possibles (a et b).

Le format NORINE est la liste des monomères du peptide représentés par leur code[28] et séparés par des virgules, suivie des listes d'adjacences pour chaque monomère, séparées par le caractère « @ ». Une liste d'adjacences est constituée des nœuds reliés par une arête à un nœud donné. Pour énumérer la liste d'adjacences, les monomères sont numérotés de 0 à n, n étant le nombre de monomères du peptide. Par exemple, dans la ligne a) de la figure 2.14, le monomère Aad (acide 2-amino-adipique) porte le numéro 0 puisqu'il est le premier cité, puis Cys (cystéine) porte le numéro 1 et enfin D-Val (D-valine) porte le numéro 2. Le monomère Aad n'est relié qu'à Cys qui porte le numéro 1 donc sa liste d'adjacences est 1. La cystéine est reliée à Aad (numéro 0) et à D-Val (numéro 2), sa liste d'adjacences est donc 0, 2. Enfin, la liste d'adjacences de D-Val n'est composée que de 1, le numéro du seul monomère auquel il est lié : Cys.

Il est à noter que l'ordre dans lequel sont énumérés les monomères dans la liste de départ n'est pas important et peut varier. Seules comptent les listes d'adjacences pour reconstituer le graphe du peptide. Dans la figure 2.14, la ligne b) est une autre représentation possible du peptide ACV. Vous remarquerez que la liste des monomères n'est pas dans le même ordre et, par conséquent, les listes d'adjacences ne sont pas les même. Cependant, les lignes a) et b) décrivent bien le même graphe qui correspond au peptide ACV. En résumé, une représentation textuelle correspond à un et un seul graphe. Par contre, un graphe peut être décrit par plusieurs représentations textuelles équivalentes. Dans NORINE, l'ordre dans lequel les monomères sont incorporés par la synthétase est utilisé, lorsqu'il est connu.

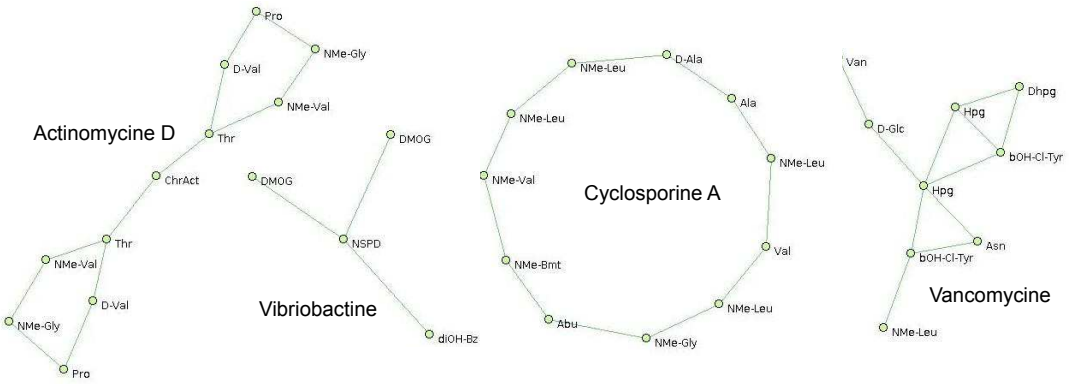


FIGURE 2.15 – Exemples de structures monomériques de peptides telles qu'elles sont dessinées dans NORINE.

Nous avons défini 6 groupes structuraux (voir aussi les figures 2.15 et 2.16 et le tableau 2.1) :

- *linéaire* : séquence traditionnelle avec un enchaînement de monomères ;
- *circulaire* : séquence se refermant sur elle-même pour former un cycle ;
- *double cyclique* : structure composée de deux cycles ayant soit des monomères en commun, soit au moins un monomère qui les relient ;
- *branchée* : séquence avec des branchements ;
- *circulaire partielle* : structure cyclique avec une chaîne reliée au cycle ;
- *complexe* : structure contenant des monomères avec plus d'un branchement et des cycles, ainsi que d'autres structures qui ne peuvent pas être classées dans les catégories précédentes.

Nous pouvons observer que la variété des structures des peptides non-ribosomiques observée dans NORINE n'explore pas toutes les possibilités. Par exemple, il n'y a pas de peptides formé

28. Abréviation communément admise ou définie par la NORINE *team*.

d'un seul cycle et de plus d'un branchement ou de trois cycles, sans branchement, alors que ces structures peuvent facilement être réalisées avec des monomères capables de former trois liaisons. Ceci est sûrement dû au mécanisme de formation des liaisons supplémentaires aux liaisons peptidiques (voir partie 2.2.2).

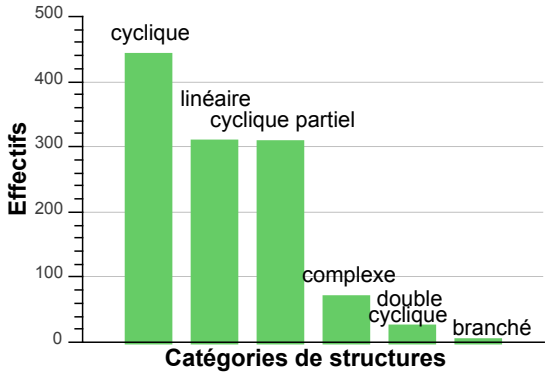


FIGURE 2.16 – Répartition des 6 groupes structuraux de NORINE.

En observant les graphes des PNR, j'ai remarqué qu'il était possible de caractériser la structure des peptides en comptant le nombre de monomères ayant un degré donné. Le degré d'un nœud dans un graphe est le nombre d'arêtes qui partent de ce nœud, c'est-à-dire le nombre de liaison qu'un monomère donné effectue avec d'autres monomères. En effet, la répartition des effectifs dans les différents degrés est représentative de la structure d'un peptide (voir tableau 2.1). Par exemple, les monomères de degré 2 forment une séquence circulaire s'ils sont tous seuls, et une séquence linéaire s'ils sont accompagnés de deux monomères de degré 1 (les extrémités de la séquence).

TABLEAU 2.1 – Catégories des structures monomériques en fonction des degrés des monomères.

	nb mono avec degré				
Catégories	1	2	3	4 ou 5	Commentaires
linéaire	2	X			
branché	Y+2	X	Y		Y branchements
circulaire		X			
cycle partiel	1	X	1		
cycle branché	Y	X	Y		Y branchements
double cycle		X	2		
multi-cycle		X	Z		(Z/2)+1 cycles dans le PNR
complexe	Y+2	X	Z	W	

Le degré d'un nœud est le nombre d'arêtes partant de ce nœud, c'est-à-dire le nombre de liaisons qu'un monomère forme avec ses voisins. La première ligne du tableau se lit : les peptides linéaires sont composés de 2 monomères ne formant qu'une seule liaison et X monomères formant 2 liaisons. X, Y et Z représentent un nombre quelconque de monomères.

Dans NORINE, 39 monomères différents sont impliqués dans au plus 3 liaisons avec d'autres monomères. Certains sont des acides aminés ou leur dérivés ayant la capacité de former une liaison en plus des deux liaisons peptidiques parce qu'ils portent sur leur chaîne latérale un groupement soit amine NH_2, soit amide $C(=O)NH_2$, soit hydroxylamine NOH, soit hydroxyle

OH, soit thiol CSH ou encore parce qu'il font partie de la famille des phénols. D'autres acides aminés forment des liaisons atypiques comme la tert-leucine (t-Leu) dans la bottromycine A2 [61]. D'autres encore ne sont pas des acides aminés mais possèdent 3 groupements réactifs.

Les 7 monomères impliqués dans au plus 4 liaisons avec d'autres monomères sont tous des acides aminés aromatiques capables d'établir des liaisons phénoliques (voir partie 2.2.2). Enfin, le seul monomère qui forme jusqu'à 5 liaisons est un phénol, l'hydroxy-phenyl-glycine (Hpg).

2.2.4 Peptides et plus encore

Comme nous l'avons vu, les peptides non-ribosomiques ne sont pas composés uniquement d'acides amino-carboxyliques. Ils ne sont donc pas des peptides *stricto sensu*, d'autres catégories de molécules sont produites (voir effectifs dans la figure 2.17), associant des acides aminés à d'autres composés. Les catégories sont déterminées à l'aide de la composition en monomères du peptide. Je vais décrire les caractéristiques de ces différentes catégories.

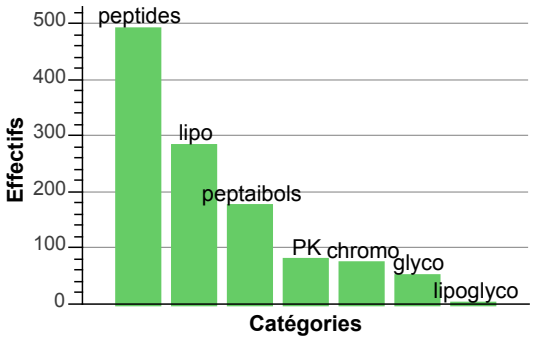


FIGURE 2.17 – Effectifs des catégories de PNR observées dans NORINE.

Les peptides représentent un peu moins de 50% des molécules de NORINE avec 492 peptides purs sur les 1164 NRP actuellement référencés.

Les peptaibols sont aussi des peptides purs, mais ont des caractéristiques particulières. Ils commencent par un acide aminé acétylé et se terminent par un alcool aminé (voir partie 2.2.1) et entre 20 et 50% des acides aminés de ces peptides sont des acides 2-aminoisobutyriques (Aib). L'isovaline (Ival) est un autre acide aminé caractéristique de ces peptides. NORINE contient 177 peptaibols. Ces peptides ont en commun d'être linéaires, antibiotiques et produits par des *fungi*. Un numéro spécial du journal Chemistry and Biodiversity leur est consacré [16].

Les lipopeptides sont formés d'acides aminés et d'un seul acide gras. Parmi les 284 lipopeptides référencés dans NORINE, 133 sont partiellement cycliques, 112 sont constitués d'un seul cycle et 39 sont linéaires. Leurs activités sont variées. Une revue complète sur ces molécules a été écrite en 2010 par N. roongsawang *et al.* [98].

Les hybrides NRPS/PKS sont formés d'acides aminés et de polyketides. Les composés hybrides sont synthétisés par des enzymes, elles aussi hybrides, possédant des modules NRPS et

des modules PKS (polyketides synthase). NORINE contient 81 hybrides NRPS/PKS, un effectif en deçà du nombre de molécules existantes, mais leur représentation sous la forme d'un graphe de monomères est délicate, nous avons donc choisi de n'intégrer que les hybrides dont la partie PKS est petite et peut être représentée sous la forme d'un monomère.

Les chromopeptides sont formés d'acides aminés et d'un seul chromophore (groupement fluorescent). Deux groupes peuvent être distingués parmi les 75 chromopeptides de NORINE du fait de leurs chromophores différents (voir figure 2.7) qui leur confèrent des structures et des activités différentes :

- les sidérophores sont des chélateurs du fer, c'est-à-dire qu'ils ont la capacité de capter le fer, qui est un cation, par l'intermédiaire de l'association du chromophore à d'autres monomères chargés négativement. Parmi les 60 sidérophores, 56 font partie de la famille des pyoverdines, des PNR produits par des bactéries du genre *Pseudomonas* [118] et 4 azotobactines ;
- les actinomycines sont des antibiotiques produits par des bactéries du genre *Streptomyces*. Elles sont toutes formées de deux cycles reliés par le chromophore et ont une taille de 11 monomères. Parmi les 15 actinomycines, 3 sont aussi des anti-tumoraux.

Les glycopeptides sont formés d'acides aminés et de glucides. Ici encore, le terme glycopeptide est très général, seuls des monosaccharides sont incorporés dans les PNR. Les 52 glycopeptides référencés dans NORINE ont tous une structure complexe composée de plusieurs cycles et de branchements (voir l'exemple de la vancomycine dans la figure 2.12). Leur taille varie entre 8 et 14 monomères. Ils ont tous une activité antibiotique et sont produits par des bactéries de la classe *Actinobacteria*.

Les lipoglycopeptides sont formés d'acides aminés, d'un lipide et de 2 glucides. Pour l'instant, seuls 3 peptides de cette catégorie sont référencés dans NORINE, ils sont des variants d'une même molécule, la ramoplanine.

2.2.5 Diversité d'activités

Au total, 11 activités différentes (voir effectifs dans la figure 2.18) sont référencées dans NORINE et il reste encore 118 peptides sur lesquels nous n'avons pas d'information. La liste des activités et leur association à un peptide sont, plus l'instant, extraites des articles scientifiques.

Les antibiotiques, aussi appelés anti-microbiens, sont des molécules qui tuent des micro-organismes. Ce terme générique comprend, entre autre, les antibactériens (dirigés contre les bactéries), les antifongiques (dirigés contre les *Fungi*) et les antiviraux (dirigés contre les virus). NORINE ne contient pas d'information précise quant aux modes d'action et aux cibles des 642 antibiotiques qu'elle contient. Cette classe comprend les peptaibols, les PNR linéaires et produits par des *Fungi*, les seuls eucaryotes avérés producteurs de PNR. Il est à noter que le monomère Hpg et ses dérivés est uniquement présent dans des antibiotiques produits par des bactéries de la classe des *Actinobacteria*. Ces organismes doivent être les seuls capables de synthétiser ces monomères particuliers dont les voies de synthèses varient selon le monomère final [107]. L'Hpg est notamment le composé majoritaire des glycopeptides qui ont une structure complexe formée de cycles et branchements, ressemblant à celle de la vancomycine (voir figure 2.12).

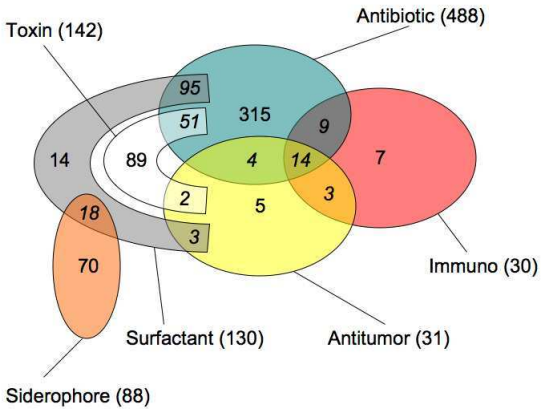


FIGURE 2.18 – Les classes d'activité avec leurs effectifs. Les intersections représentent le nombre de peptides ayant les différentes activités concernées. Schéma réalisé par Ségolène Caboche pour son mémoire de thèse (sept 2009).

Les toxines altèrent les cellules. Ce terme est également générique et regroupe 263 peptides avec des propriétés variées. Du fait de leur toxicité, ils ont d'autres activités comme antibiotique ou antitumoral.

Les surfactants modifient la tension d'un liquide en formant un film à sa surface. Ils sont constitués d'une chaîne peptidique qui est la partie hydrophile [29] et d'un acide gras qui est la partie hydrophobe [30]. Ainsi, ils se placent à la surface d'un liquide ou entre un liquide et un corps gras pour former un film. Les surfactants peuvent être utilisés comme détergents ou entrer dans la composition de cosmétiques. Les 139 surfactants sont, forcément, des lipopeptides. En fait, surfactant est plus une propriété physique qu'une activité. Cette classe sera supprimée de NORINE.

Les anti-tumoraux, comme leur nom l'indique, détruisent les tumeurs. Cette activité n'est pas celle qui est recherchée par les organismes qui produisent ces peptides, c'est pourquoi les 101 anti-tumoraux de NORINE ont souvent d'autres activités.

Les sidérophores sont sécrétés par les micro-organismes pour capter le fer qui se trouve dans leur environnement. Lorsque le fer est complexé par le sidérophore, il peut être internalisé dans la cellule pour être libéré puis utilisé dans différentes voies métaboliques. La majorité des 90 sidérophores de NORINE sont des chromopeptides car le chromophore intervient dans la captation du fer, associé à d'autres monomères tels que les dérivés hydroxylés (avec une fonction OH) de l'ornithine (Orn), de l'acide aspartique (Asp) et de l'histidine (His). Les autres sont des peptides purs ou des lipopeptides. Les sidérophores peuvent être utilisés pour décontaminer des eaux polluées par les rejets des mines de fer ou d'autres industries.

Les inhibiteurs de protéases bloquent les enzymes qui coupent les protéines, appelées protéases. Les protéases, sont, entre autre, utilisées par les virus pour couper les précurseurs des

29. Qui accepte le contact de l'eau.
30. Qui fuit le contact de l'eau et lui préfère celui des lipides.

protéines composant leurs enveloppes. Ces molécules peuvent donc être utilisées comme antiviraux. Il existe 6 familles de protéases, donc différents inhibiteurs sont nécessaires, NORINE en contient 39.

Les immuno-modulateurs agissent sur le système immunitaire pour modifier l'intensité de leur effet. Par exemple, la cyclosporine A est utilisée contre le rejet des greffes car elle diminue la réaction immunitaire due à la présence d'un organe étranger dans le corps du patient. NORINE contient 37 immuno-modulateurs.

Les antithrombotiques réduisent la formation des caillots sanguins, aussi appelés thrombus. Ils peuvent être des anticoagulants, ou agir pour dissoudre le caillot. NORINE en contient 12.

Les anti-inflammatoires luttent contre les inflammations qui sont des réactions du système immunitaire contre une agression. NORINE en contient 8.

Les anti-athéromateux agissent contre l'athérome, aussi appelé athérosclérose, qui est un dépôt situé sur la paroi interne d'une artère. Ces dépôt sont la cause de la majorité des affections cardio-vasculaires.

Les antagonistes de la calmoduline bloquent les effets de la calmoduline en prenant sa place au près des molécules avec laquelle elle interagit. La calmoduline est une protéine qui fixe le calcium et qui est impliquée dans la transmission de différents messages dans les cellules eucaryotes tels que le déclenchement d'une réaction inflammatoire ou la contraction des muscles lisses (ceux de l'estomac ou des vaisseaux sanguins, par exemple).

Vous pouvez constater que les activités des PNR sont nombreuses et peuvent avoir des applications dans différents domaines, au delà de la pharmacologie. Cette grande variété d'activités est dû à la variété de structures secondaires que ce soit les structures non linéaires, ou la diversité des monomères.

2.2.6 Diversité des organismes producteurs

Les peptides non-ribosomiques ne sont pas produits par tous les être vivants. Jusqu'à présent, des gènes codant des synthétases peptidiques non-ribosomiques ont été identifiés chez des bactéries et des *Fungi*. Ils sont donc présents dans deux des trois domaines du vivant que sont les eucaryotes, les bactéries et les archées.

Les archées sont des organismes constitués d'une seule cellule. Ils sont appelés ainsi car ils ont été découverts dans des milieux extrêmes tels que des sources chaudes et/ou soufrées. La forme de leurs cellules et leur génome, qui est généralement un chromosome circulaire, les rapprochent des bactéries ; certaines de leurs voies métaboliques [31] sont très proches de celles des eucaryotes et d'autres leur sont spécifiques. A ma connaissance, aucun gène codant une PNRS n'a été découvert dans un génome d'archée.

31. Ensemble de réactions chimiques réalisées par des enzymes au sein des cellules.

Les eucaryotes sont des organismes constitués d'une ou plusieurs cellules, selon les espèces. Ils ont en commun d'avoir leur génome réparti dans plusieurs chromosomes linéaires et protégé par une membrane, formant le noyau cellulaire. En simplifiant, les règnes participants au domaine eucaryote sont les animaux, les plantes, les protistes[32] et les *Fungi*.

Le règne des *Fungi* regroupe des organismes microscopiques tels que les levures ou les moisissures avec les champignons qui sont des organismes pluricellulaires. Ils produisent différents PNR tels que les peptaibols qui leur sont spécifiques et semblent être produits presque exclusivement par l'embranchement des Ascomycètes. Seul un peptide qui peut être assimilé à un peptaibol a été observé chez un Basidiomycète [109]. Une étude phylogénomique [17] a mis en évidence les gènes de PNRS dans les génomes des *Fungi* disponibles en 2010. Ils ont troué la présence de gènes dans les génomes d'Ascomycètes, chez quelques Basibiomycètes, et quelques gènes chez les Zygomycètes et les Chytridiomycètes. Par contre, aucun gène n'a été trouvé chez les Microsporidies. Une autre étude phylogénomique s'est focalisée sur les *Fungi* producteurs de toxines qui se sont révélés encore une fois être les Ascomycètes [42]. A priori, seuls les Fungi microscopiques sont producteurs de PNR.

Par contre, à ce jour et à ma connaissance, aucun gène codant une PNRS n'a été identifié dans le génome des autres règnes eucaryotes. Des eucaryotes non *Fungi* sont dans NORINE car des peptides ont été extraits de ces organismes. Mais, aucune preuve de leur production directe par l'organisme en question n'a été trouvée. Les espèces concernées sont des éponges, des mollusques, des tunicates[33] ou encore des plantes. Elles sont souvent accompagnées de symbiotes soit bactériens, soit *Fungi* qui sont sûrement les producteurs réels des PNR. D'ailleurs, plusieurs études ont identifié des gènes codant des SPNR dans le génome de symbiotes extraits d'organismes pluri-cellulaires [129, 104, 57]. De plus, à l'occasion de notre étude statistique [19, 18] nous avons mis en évidence le fait que la composition en monomère des peptides est significativement différente entre ceux produits par les bactéries et ceux produits par les *Fungi*. Par contre, les peptides extraits des autres eucaryotes s'apparentent aux peptides bactériens, même s'ils présentent certaines spécificités. Ceci peut être dû au fait que les producteurs de ces peptides peuvent être soit des bactéries, doit des *Fungi* symbiotes des organismes dit supérieurs. Deux peptides de NORINE, la dolastatine et la majusculamide, sont reliés à un organisme eucaryote supérieur et une bactérie.

Il est à noté que, en 2013, un gène codant une enzyme intervenant dans la synthèse des β-lactames, la familles antibiotiques qui comprend la pénicilline, a été identifié dans le génome d'une fourmi [96]. Cette enzyme n'est pas une synthétase peptidique non-ribosomique, mais transforme le peptide ACV en isopenicilline N. Cependant, la même équipe a montré la présence d'ARNm codant l'ACV synthétase dans épithélium de l'estomac de la même espèce de fourmi, mais n'a pas identifié la séquence correspondant au gène dans les fragments de génomes qu'ils ont à leur disposition. Les ARNm pourraient provenir des bactéries présentes dans l'estomac. L'hypothèse d'un transfert de gène entre la fourmi et une bactérie ou un *Fungi* est l'explication la plus probable pour la présence du gène de cette enzyme dans le génome de la fourmi.

Les bactéries sont des organismes constitués d'une seule cellule et possédant généralement un chromosome circulaire. Parmi les 26 embranchements cités dans la taxonomie du NCBI, 5 sont présents dans NORINE : les Actinobactéria, les Chlorobi (aussi appelés Bacteroidetes), les

32. Règne qui regroupe une grande variété d'organismes microscopiques eucaryotes qui ne peuvent être classés dans un autre règne.

33. Petits animaux marins qui vivent souvent accrochés à des rochers.

Cyanobacteria, les Firmicutes et les Proteobacteria. Ils sont les embranchements qui comptent le plus d'espèces. Les peptides produits varient d'une bactérie à une autre, y compris entre souches d'une même espèce, mais sont souvent caractéristiques d'un taxon donné. Leur évolution est influencée par deux facteurs, le contexte environnemental et l'évolution des espèces. En effet, les bactéries vont déployer des stratégies équivalentes pour utiliser au mieux les ressources de leur environnement et mener une compétition pour coloniser le milieu dans lequel elles sont. dans ce cas, des convergences évolutives peuvent être observées. De plus, des espèces différentes, qui partagent le même environnement, peuvent s'échanger des gènes via le transfert horizontal de gènes. Les synthétases peptidiques non-ribosomiques sont concernés par ces échanges. L'exemple le plus connu est l'histoire évolutive des gènes de la synthèse des β-lactames dont la pénicilline fait partie. Il s'agit même d'un cas de transfert entre des bactéries et des *Fungi* [15]. Évidemment, les gènes des PNRS suivent aussi l'évolution des espèces qui les portent sont transmis d'une génération à une autre. Une étude [79] a démontré l'absence de transfert horizontal pour le cluster de gènes de synthèse de la valinomycine chez les *Steptomyces* et une autre étude [93] a porté sur celle des mycrocystines.

2.3 Modèles bio-informatiques pour les peptides non-ribosomiques

Comme évoqué précédemment, il n'existait aucun outil, ni base de données dédiés aux peptides non-ribosomiques lorsque la Norine *team* s'est lancée dans le développement de Norine, une plate-forme d'analyse des PNR, en 2006. Les outils développés pour les protéines ou peptides synthétisés par la voie ribosomique ne sont pas applicables sur ces molécules, ceux pour les structures chimiques le sont, mais ne prennent pas en compte les spécificités des PNR. Nous avons donc commencé par proposer une représentation adaptée aux PNR, la structure monomérique, accompagnée du format Norine (voir partie 2.2.3) puis nous avons conçu des algorithmes et outils dédiés aux PNR.

Par la suite, seuls des algorithmes et outils d'analyse des résultats de spectrométrie de masse ont été développés spécifiquement pour les peptides non-ribosomiques. A ma connaissance, il n'existe pas d'autres développements dédiés aux PNR. Par contre, une base de données contenant des polykétides et quelques PNR a été mise en ligne récemment.

Je vais présenter les principales bases de données dans lesquelles il est possible de trouver des peptides non-ribosomiques, puis les outils pour la spectrométrie de masse dédiées aux PNR. Dans un deuxième temps, je présenterai les travaux de la Norine *team*, à savoir la base de données, les algorithmes lié à la représentation sous la forme monomérique, ceux liés à la nouvelle représentation sous la forme de vecteurs et enfin les travaux en cours sur les structures chimiques.

2.3.1 Bases de données contenant des PNR ou d'autres peptides.

Je vais présenter les principales bases de données liées aux peptides non-ribosomiques qui soit en contiennent, comme les bases généralistes de molécules organiques ; soit sont dédiées à des molécules proches, comme les bases spécialisées dans les peptides ou les métabolites secondaires et leur synthèse.

Les bases généralistes de molécules organiques contiennent des molécules ayant principalement un intérêt biologique. Du fait de leur petite taille et de leurs activités, les PNR sont

présents dans ces bases. Ces dernières proposent différents formats pour représenter les structures chimiques des molécules en 1D, 2D ou 3D, selon leur domaine d'application.

PubChem [72] est un dépôt public de structures chimiques et de leurs activités. Il est géré par le NCBI[34], et une partie des expériences sont prises en charge par le NIH[35] Molecular Libraries Program. Ce dépôt est composé de trois bases interconnectées. *Substance* collecte les molécules qui lui sont soumises avec leurs annotations ; *BioAssay* contient les résultats de tests d'activité ; *Compound* contient des structures chimiques uniques provenant de la standardisation des molécules contenues dans la base *Substance*. NORINE propose des liens vers les entrées de la base *Compound* pour 205 peptides que nous avons identifiés pour avoir une entrée dans cette base. Ainsi, nous avons pu collecter les structures chimiques des peptides correspondants.

ChEMBL [44] est une autre banque de structures chimiques, gérée par l'EBI[36]. Les données sont principalement extraites manuellement de la littérature, cependant certaines ont été directement soumises ou d'autres proviennent d'échanges avec d'autres bases telles que PubChem. En complément à ces données, ChEMBL contient également les structures et annotations concernant les médicaments approuvés par la *Food and Drug Administration* (FDA). Ainsi, ChEMBL est interrogeable via les références bibliographiques utilisées pour remplir la base ; les composés qu'elle contient ; les expériences de tests d'activité, d'interaction avec les cibles ou autre et les cibles elle-même. En effet, les annotateurs de ChEMBL construisent une base non-redondante contenant les cibles des molécules qu'elle contient afin de faciliter la comparaison des résultats obtenus lors de différentes expériences.

Ces deux bases de données proposent le format appelé *SMILES* (*simplified molecular input line specification*) [126] pour représenter les structures chimiques en 2D. Comme le format NORINE, un SMILES est une chaîne de caractères représentant un graphe. Les atomes, qui sont les nœuds de ces graphes, sont écrits sous leur forme abréviée les uns derrière les autres, sans séparateur, pour former la chaîne principale de la molécule. En général, les atomes d'hydrogène H ne sont pas précisés car ils peuvent être déduits. Les simples liaisons ne sont pas représentées. Par contre, les doubles liaisons sont représentées par le symbole = et les triples liaisons par le symbole ~. Les branchements sont isolés par des parenthèses et suivent l'atome sur lequel ils se fixent. Les cycles sont marqués par la présence de numéros ajoutés sur leur premier et dernier atome. Les atomes participants au même cycle portent le même numéro. Pour plus d'informations concernant ce format, voir l'exemple de la figure 2.19 et consulter la page dédiée[37] sur le site de Daylight, les inventeurs de ce format.

Hpg

O=C(O)C(N)C1=CC=C(O)C=C1

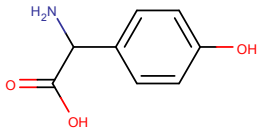


FIGURE 2.19 – Le monomère Hpg (hydroxy-phenyl-glycine) et son SMILES qui se lit de la façon suivante : Un O est relié par une double liaison à un C qui porte une fonction OH, suivi d'un C qui porte un groupement NH_2, suivi d'un autre C qui débute un cycle de 6 C (quatre C entre les deux C1) avec une fonction OH sur le $4^{ème}$ C après l'autre branchement.

Worldwide Protein Data Bank (wwPDB) [12] est l'association de quatre organisations impli-

34. *National Center for Biotechnology Information*, centre américain de bio-informatique.
35. *National Information for Health*
36. *European Bioinformatics Institute*, centre européen de bio-informatique.
37. http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

quées dans la collecte, le traitement et la distribution des données expérimentales de structures 3D de macromolécules biologiques. Les organisations sont le *Research Collaboratory for Structural Bioinformatics* (RCSB) PDB (USA), *Protein Data Bank in Europe* (PDBe), *Protein Data Bank Japan* (PDBj), and BioMagResBank (BMRB, USA). Parmi les données contenues dans ce répertoire de structures, il y a des peptides non-ribosomiques qui sont souvent associés à d'autres molécules telles que les protéines avec lesquelles ils interagissent. Nous avons mis en place un échange de données avec l'antenne européenne, PDBe. Seuls 26 peptides de NORINE ont effectivement un lien vers au moins une entrée de PDB. Ce nombre limité est notamment dû à la différence des formats utilisés par NORINE et PDB pour représenter les peptides. Les travaux de recherche en cours sur la conversion d'une structure chimique en structure monomérique faciliterons sûrement les échanges.

Cambridge Structural Database (CSD) [46] est une base de données complémentaire à PDB puisqu'elle contient les structures 3D de petites molécules déterminées à l'aide d'expériences de cristallographie. Nous n'avons pas encore de liens avec cette base, mais nous avons déjà été en contact avec eux par l'intermédiaire de PDBe puisque nous étions dans une demande de financement commune. Comme beaucoup de bases de données en chimie, la consultation de la base est payante.

D'autres bases de molécules existent, mais beaucoup sont payantes du fait de l'exploitation de ces bases dans le but de découvrir de nouveaux médicaments. Une revue récente [85] présente les banques/bases de données publiques utiles à la chimie médicale.

Les bases spécialisées dans les peptides sont elles limitées à un type de peptide particulier. Elles sont complémentaires à la ressource internationale dédiée aux protéines ribosomiques, *UniProt* [113] qui contient aussi des peptides, mais qui n'est pas dédié à ces molécules et ne propose donc pas d'annotations particulières les concernant. De plus, le format de représentation des séquences peptidiques utilisé par UniProt ne permet pas de gérer efficacement toutes les modifications observées chez certains peptides dont les structures ne sont pas toujours linéaires et peuvent contenir des acides aminés non protéogéniques. Par contre, UniProt contient les synthétases peptidiques non-ribosomiques qui produisent les PNR. NORINE associe à 249 peptides les protéines de UniProt qui les produisent. Mais, ces données n'ont pas été actualisées depuis longtemps.

Je ne vais pas faire un inventaire des bases de données concernant les peptides car elles ne concernent pas souvent les PNR, mais des peptides produits par la voie ribosomique. Par exemple, *Antimicrobial Peptide Database* (APD) [122, 121] recense, comme son nom l'indique, les peptides ayant une action contre les micro-organismes, mais aussi contre des parasites ou contre le cancer. Il n'est pas précisé si cette banque est limitée aux peptides ribosomiques, mais il semble que cela soit le cas d'après les familles annoncées. Une autre base, *Cybase* [120], contient des peptides totalement cycliques, en se limitant à ceux qui sont codés directement par les génomes.

Une base de données est dédiée aux peptaibols qui sont des PNR (voir 2.2.4), il s'agit de *The Comprehensive Peptaibiotics Database* [109]. Bien que cette base de données soit récente, elle est proposée au uniquement au téléchargement et accompagnée d'une interface d'interrogation sous MS Access. Cette base contient plus de peptaibols que nous n'en avons dans NORINE. Il faudrait que nous les contactions pour collaborer avec eux.

Les bases spécialisées dans les métabolites secondaires et leur synthèse contiennent essentiellement des PNR et des PK qui sont les métabolites secondaires les plus faciles à prédire à

partir des séquences génomiques (voir chapitre 1). Souvent, ces bases sont focalisées sur les gènes et protéines impliquées dans la synthèse des métabolites, plutôt que les métabolites eux-même, tel que *NRPS-PKS* [6], *ClustScanDB* [33] ou encore la base de données récente *DoBiscuit* [55]. Ces bases ne sont pas en concurrence avec NORINE puisqu'elles contiennent des informations sur les enzymes et non leurs produits. Enfin, une dernière base contient non seulement des annotations sur les enzymes synthétases peptidiques non-ribosomiques et polykétides synthases (PKS), mais aussi sur leurs produits, il s'agit de *Clustermine360* [27]. Cette base contient plus de 185 familles de molécules (majoritairement PK), un nombre inférieur au plus de 1100 peptides contenus dans NORINE.

En conclusion, NORINE n'a pas de concurrent direct en ce qui concerne sa base de données de peptides non-ribosomiques car nous avons un nombre inégalé de peptides et des annotations spécifiques.

2.3.2 Outils pour la spectrométrie de masse pour les PNR

Les seuls algorithmes conçus exclusivement pour les peptides non-ribosomiques sont dédiés à l'aide à l'analyse des résultats d'expérience de spectrométrie de masse. Pour commencer, il est important de rappeler que cette technique expérimentale est la principale technique disponible pour déterminer la structure d'un PNR. Cependant, les résultats expérimentaux sont difficiles à analyser du fait de la complexité des structures des PNR avec leur grande variété de monomères et la présence de liaisons non peptidiques. Les logiciels fournis avec les spectromètres de masse ne sont pas capables de gérer ces molécules. Trois projets de recherche distincts ont été menés afin de proposer des programmes adaptés aux PNR.

NRP-Annotation [10, 74, 83], anciennement appelé MS-CPA, est le premier algorithme d'analyse des résultats de spectrométrie de masse dédié aux PNR à avoir été développé, dès 2008. Les algorithmes proposés sont restreints aux PNR cycliques et utilisent les monomères de NORINE pour prendre en compte leur diversité de composition. Nous avons été en contact avec les auteurs.

iSNAP [54], pour *informatic search strategy for natural products*, a été publié en 2012. Ce programme s'appuie sur plus de 1000 peptides non-ribosomiques issus en grande partie de NORINE et dont les structures chimiques ont été déterminées à l'aide de différentes sources externes puisqu'elles ne sont pas disponibles dans NORINE. A priori, leur base leur permet de traiter les différents types de structures observées dans NORINE.

Mmass [86] est un logiciel gratuit et open-source d'analyse des spectres de masse. En 2012, il a été étendu à la prise en compte des peptides cycliques et à la diversité de composition des PNR (en utilisant les monomères de NORINE).Nous avons également été en contact avec les auteurs.

2.4 La plate-forme NORINE

Dans les parties qui vont suivre, je vais présenter les travaux auxquels j'ai participé. Je fais partie des membres fondateurs de NORINE, l'unique ressource dédiée aux peptides non-ribosomiques créée en 2006. Mon parcours pluridisciplinaire m'a apporté une vision globale sur ce

projet et a facilité l'intégration des idées de mes collègues biologistes dans une infrastructure informatique adaptée. J'ai participé à chaque étape de construction de NORINE, que ce soit la définition du schéma de la base de données, avec le souci de fournir un maximum d'annotations importantes pour les utilisateurs tout en conservant un schéma optimisé ; le choix de la représentation des peptides sous la forme d'un graphe de monomères ; le cahier des charges de l'interface de consultation des données, en m'appuyant sur mes pratiques en tant qu'utilisatrice de bases de données biologiques ; la conception d'algorithmes de comparaison de peptides représentés sous la forme de graphes en apportant mon expérience sur la comparaison de séquences ; le remplissage des données en intégrant les corrections des utilisateurs et mes propres observations. J'ai encadré tous les développeurs qui ont contribué à NORINE. Dernièrement, j'ai introduit un nouvel axe de recherche pour la thématique de l'informatique pour les PNR en proposant de s'orienter vers la chémo-informatique, tout en exploitant les spécificités de ces molécules.

2.4.1 Les données de NORINE

NORINE est une base de données de peptides non-ribosomiques conçue avec des biologistes, principalement pour les biologistes et biochimistes. L'objet central de la base est donc le peptide produit par un organisme vivant, à l'aide d'une synthétase. De plus, la représentation des peptides est au plus proche de leur mode de synthèse en choisissant comme élément de base les monomères incorporés par les enzymes. Je vais présenter le schéma de la base, le processus de collecte et enfin l'interrogation et l'affichage des annotations. Au fil des descriptions, je préciserai ma participation et les évolutions à court terme qui me semblent intéressantes à intégrer.

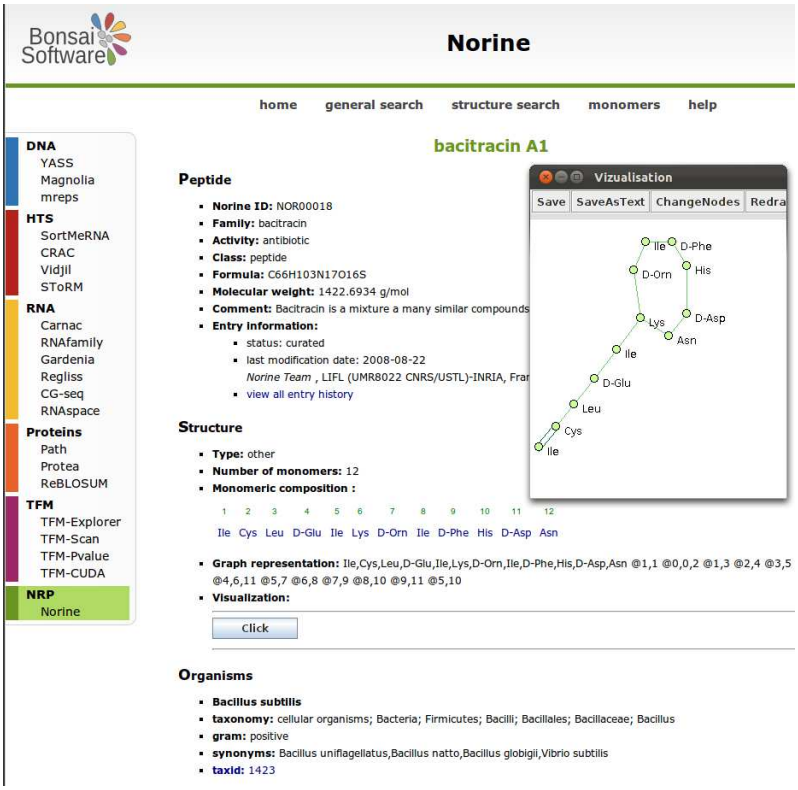


FIGURE 2.20 – Page de description de la bacitracine A1, avec le visualiseur de peptide ouvert.

L'affichage et l'interrogation des annotations de NORINE sont accessibles à tous gratuitement via le web, sans inscription préalable, à l'adresse : http://bioinfo.lifl.fr/norine NORINE offre en plus des formulaires d'interrogation, des outils spécifiques aux peptides non-ribosomiques. L'ensemble des 1164 peptides est consultable sous plusieurs formats, y compris une représentation graphique, et est même téléchargeable. Un exemple de page de présentation d'un peptide est donné dans la figure 2.20. Des informations concernant les 528 monomères sont également fournies.

Le formulaire *General search* permet d'interroger toutes les annotations disponibles dans la base, champ par champ. Le formulaire *Structure search* recherche parmi les structures des peptides celles qui ressemblent à la structure saisie par l'utilisateur. Je décris les différentes recherches de ce formulaire et leur contexte d'utilisation dans la partie 2.4.2. Ce formulaire vient d'être mis à jour car de nouvelles méthodes de recherche ont été conçues.

Le schéma de la base de données a évolué et s'est étoffé au fur et à mesure des développements et de la nécessité d'ajouter de nouvelles informations. Le schéma actuel est représenté dans la figure 2.21.

La table PEPTIDE est la table centrale de la base de données. Elle contient les informations spécifiques à un PNR telles que ses noms car certains en possèdent plusieurs, sa masse moléculaire ou encore sa structure monomérique. Certaines annotations sont dans des tables séparées pour optimiser la gestion des données en éliminant la redondance d'information et facilitant l'interrogation de la base. Elles permettent aussi d'harmoniser le vocabulaire utilisé pour décrire les peptides. Par exemple, les tables ACTIVITÉS et CATÉGORIES permettent de partager une annotation entre différents peptides et d'affecter plusieurs de ces annotations à un même peptide. Pendant la rédaction de ce mémoire, j'ai ajouté la catégorie «PK-NRP» car je me suis rendue compte que les hybrides peptides-polykétides étaient classés dans la catégorie «peptide», alors qu'ils n'en étaient pas.

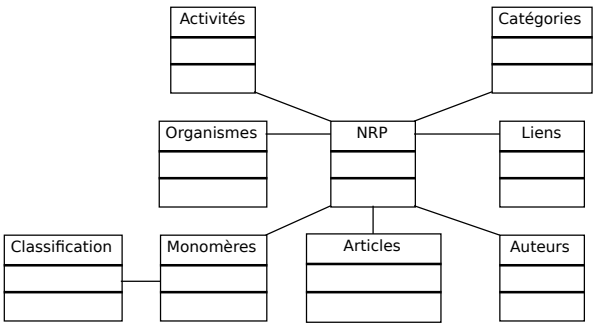


FIGURE 2.21 – Schéma simplifié de la structure de la base de données NORINE, reprenant les tables principales.

De la même façon, un organisme produit souvent plusieurs peptides et certains peptides sont produits par différents organismes. La table ORGANISME contient le nom de l'organisme, avec la souche si elle est connue, et ses synonymes. Un autre champ contient toute la taxonomie. Nous reprenons la taxonomie du NCBI lorsque l'organisme est référencé dans celle-ci et proposons un lien vers la page correspondante sur le site du NCBI. Dans le cas contraire, nous utilisons d'autres taxonomies. Il reste certains organismes pour lesquels la taxonomie n'est pas complète car elle n'a pas été trouvée ou les noms de genre et d'espèce ne sont pas précisés dans l'article de

description du peptide. Un travail de complétion des données est donc nécessaire. J'ai commencé à le faire, mais n'ai pas terminé.

Les tables concernant les ARTICLES SCIENTIFIQUES archivent les sources d'information à l'origine des annotations présentes dans NORINE. Ici encore, un travail de nettoyage est nécessaire car j'ai remarqué des erreurs dans certaines références et également des différences de syntaxe. Pour ces mises à jour, il faudrait chercher automatiquement sur le web les publications présentes dans NORINE et extraire les informations nécessaires pour remplir convenablement les tables. Nous pouvons utiliser des outils existants pour simplifier la mise en place de ce processus.

Les tables sur les AUTEURS D'UN PEPTIDE contiennent les informations concernant les scientifiques qui ont signalé un nouveau peptide et soumis des annotations le concernant, ou fourni des corrections concernant un peptide qui est déjà dans la base.

La table MONOMÈRES archive les informations relatives aux monomères contenus dans les peptides de NORINE. Un peptide est lié à tous les monomères qu'il contient. Ainsi, nous pouvons efficacement trouver tous les peptides contenant un ensemble de monomères. Nous avons différentes informations sur les monomères telles que leur formule chimique, leur structure atomique ou leur nom complet selon l'IUPAC afin d'aider à leur identification. En effet, nous utilisons des abréviations pour nommer les monomères, mais elles ne sont pas toujours universelles. Il faut donc fournir aux utilisateurs différents moyens d'identifier les monomères.

La table CLASSIFICATION regroupe les monomères de NORINE en trois niveaux hiérarchiques :

- niveau le plus bas : les monomères avec distinction des énantiomères ;
- niveau intermédiaire : regroupement des dérivés d'un même monomère (ex : les énantiomères, la forme méthylée ...) ;
- niveau supérieur : regroupement soit des acides aminés ayant des propriétés physico-chimiques proches telles que l'alanine et la glycine qui sont tous les deux petits et neutres, soit des monomères d'une même catégorie tels que les lipides ou les glucides.

Une table supplémentaire est prévue pour gérer plusieurs classifications en parallèle. Par exemple, nous pourrions également proposer la classification des auteurs du logiciel NRPSPredictor2 [99] (voir chapitre 1.3.2) ou encore une classification regroupant les monomères selon leur influence sur l'activité du peptide.

La collecte des données archivées dans NORINE est effectuée à la main, à partir d'articles ou d'ouvrages scientifiques. Parfois, la lecture de plusieurs articles est nécessaire afin de collecter toutes les annotations et de confirmer certaines informations car les études successives d'une même molécule peuvent aboutir à des résultats différents voir contradictoires. D'un autre côté, certains articles décrivent une famille de peptides. Nous avons choisi de ne citer que les articles dont l'objet principal est la structure du peptide, soit 493 articles.

Le travail d'annotation nécessite de solides connaissances en biologie et biochimie afin de comprendre les articles et un recul par rapport aux PNR afin de pouvoir juger de la vraisemblance des informations contenues dans les articles. Un premier niveau de connaissance est nécessaire puis le recul est acquis au fur et à mesure du travail d'annotation. C'est pourquoi, les données de NORINE doivent être validées par un expert de la *NORINE team* avant d'être officiellement intégrées dans la version publique.

Le processus historique d'annotation est celui qui a permis de saisir tous les peptides actuellement présents dans NORINE. La quasi-totalité des peptides ont été saisis par Ségolène Caboche dans un tableur, puis transférés dans la base à l'aide d'un script. Deux étudiants en thèse au

laboratoire ProBioGEM, Jovana Deravel et Arthur Tapi, ont également participé à l'annotation de quelques peptides. Valérie Leclère les a accompagnés et a vérifié l'exactitude de leur travail. Quand le travail d'annotation est effectué par des scientifiques de Lille, nous ne précisons pas le nom de l'auteur car il s'agit souvent d'un travail collectif, *NORINE team* est indiqué dans la base.

Les scientifiques qui ne font pas partie de la *NORINE team* envoient les informations soit via le formulaire proposé dans l'ancienne version de l'interface web, soit par l'envoi d'un message électronique. Pour l'instant, quatre autres contributeurs nous ont soumis des données :

- Berti AD de l'Université du Wisconsin-Madison aux USA a soumis 16 peptides ;
- Jin JM de l'Université Hong Kong Baptist en Chine a soumis 8 peptides ;
- des membres de l'équipe PDBe, European Bioinformatics Institute au Royaume-Uni, avec qui nous avons mis en place un échange de données, nous ont soumis 11 peptides ou modifications qui ont été intégrées dans NORINE, d'autres sont en attente car elles nécessitent encore des analyses ;
- Gessmann Renate et Petratos Kyriacos de FORTH-IMBB (Foundation of Research and Technology-Hellas, Institute of Molecular Biology and Biotechnology), Greece, ont soumis 3 peptides et une modification.

Les informations transmises par d'autres scientifiques sont au format texte et doivent être vérifiées, complétées et saisies dans le tableur afin de pouvoir être intégrées dans la base. J'ai intégré des corrections signalées par nos utilisateurs, mais les peptides complets qui ont été soumis récemment ne sont pas pas encore intégrés car la structure de la base a changé et que le mode de saisie via un tableur n'est pas convivial. Je propose donc d'intégrer les différentes étapes du processus de collecte des données dans l'interface de NORINE.

L'évolution de la collecte des données est un projet qui vient d'être commencé et qui va être réalisé par l'ingénieur de recherche Areski Flissi. Je préconise de développer un formulaire de saisie convivial et facilitant le processus d'annotation et de correction des peptides, selon le fonctionnement schématisé dans la figure 2.22. Pour accéder au formulaire, les scientifiques doivent créer un compte sur NORINE. Ce compte permet de connaître le statut et l'affiliation de l'auteur afin de vérifier son expérience dans le domaine des PNR et d'avoir son adresse électronique pour pouvoir le contacter. Une fois le compte créé, l'auteur accède à un formulaire de saisie. Ce formulaire accompagne l'auteur en cherchant certaines informations sur d'autres sites comme la taxonomie sur le site du NCBI ou les références bibliographiques sur les sites dédiés. Pendant la saisie, des vérifications internes sont réalisées pour vérifier la cohérence des annotations et éviter l'insertion de doublons dans la base. Il existe déjà un éditeur qui permet de dessiner les structures monomériques et qui génère la structure monomérique selon la syntaxe NORINE. Les recherches en cours sur la conversion des structures atomiques en structures monomériques permettront d'offrir à l'utilisateur la possibilité de soumettre directement une structure chimique.

Les informations saisies dans le formulaire sont mémorisées sous la forme d'un fichier XML. L'auteur a le choix de mémoriser ses données soit sur le serveur de NORINE, soit sur son propre ordinateur s'il veut garder son travail confidentiel. L'interface permet de reprendre un peptide en cours de saisie afin de compléter les informations, si le travail s'étend sur plusieurs jours. Il peut aussi utiliser un peptide comme modèle s'il veut saisir plusieurs peptides d'une même famille. Une fois qu'il a terminé son travail, il peut soumettre les peptides aux validateurs de NORINE. Un message électronique est alors envoyé aux validateurs avec le fichier XML contenant les données soumises. Les validateurs sont des membres de la *NORINE team* capables de vérifier

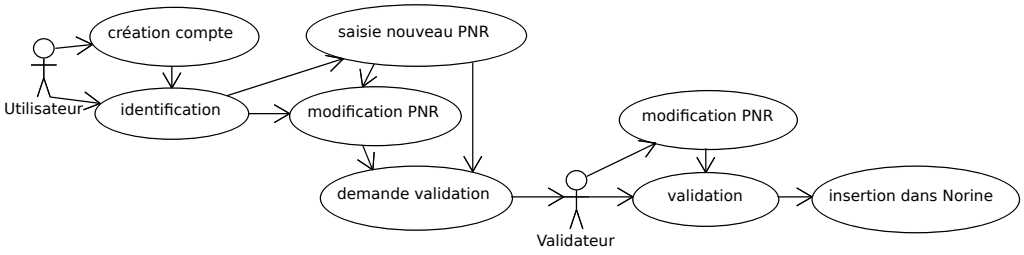


FIGURE 2.22 – Schéma reprenant le processus d'annotation ou de correction d'un peptide que je propose de mettre en place.

l'exactitude des données soumises et de les corriger ou les compléter si besoin. Pour l'instant, seuls trois scientifiques ont le rôle de validateur : Valérie Leclère, Philippe Jacques et moi-même. Une fois travail de vérification terminé, nous validons le peptide qui est alors automatiquement intégré dans la base de données publique, sous le nom de l'auteur qui a fait la saisie initiale.

Une fois que le processus d'annotation des peptides aura été mis en place et testé, je propose de solliciter la communauté afin d'augmenter le nombre de saisies effectuées par des scientifiques qui ne font pas partie de la NORINE *team*. Un moyen serait d'identifier les auteurs d'articles sur les peptides non-ribosomiques et de leur envoyer un message électronique leur présentant NORINE, expliquant le processus d'annotation et les bénéfices pour eux et la communauté d'ajouter de nouveaux peptides dans NORINE. La sollicitation des auteurs de publications par les bases de données afin qu'ils soumettent eux-même leurs données est déjà pratiqué par de nombreuses bases. Des partenariats entre certaines bases et les grandes maisons d'édition existent déjà. Les éditeurs demandent directement aux auteurs de soumettre leurs données dans des bases reconnues pour citer l'entrée de la base dans l'article.

2.4.2 Les algorithmes de recherche de peptides selon leur structure

La recherche par composition renvoie tous les peptides dont la composition correspond à celle qui est saisie dans le formulaire de recherche par structure. Cette recherche est proposée pour faciliter l'identification des résultats donnés par les logiciels d'annotation des synthétases peptidiques non-ribosomiques (voir section 1.3). Ces logiciels prédisent les monomères reconnus par les domaines d'adénylation, mais ils ne prédisent pas la structure 2D du peptide. Une recherche par composition est donc pertinente. Deux variantes ont été programmées.

La première variante a été développée par Ségolène Caboche. Les compositions sont considérées comme des ensembles et l'algorithme vérifie si l'ensemble recherché est inclus dans chacun des ensembles de NORINE, testés un à un. Le formulaire autorise de préciser un nombre maximum de monomères de la liste saisie qui ne sont pas dans le peptide. Les limites de cet algorithme sont qu'il demande à l'utilisateur de préciser le nombre d'erreurs et qu'il n'a pas de moyen de trier les peptides trouvés avec le même nombre d'erreurs.

La deuxième variante a été conçue récemment par Ammar Abdo Hasan et mise en œuvre par Antoine Engelaere. Ce travail a été publié en 2013 dans *Journal of Computer-Aided Molecular Design* [3]. Les peptides sont représentés par un vecteur possédant autant de dimensions qu'il n'y a de monomères dans NORINE. Chaque dimension contient le nombre d'occurrences d'un monomère donné dans un peptide. Les comptages sont toujours énumérés dans le même ordre. En chémo-informatique, les vecteurs représentant des molécules à l'aide de leurs caractéristiques sont appelés *fingerprint*.

Il est possible de calculer une distance entre deux vecteurs A et B, c'est-à-dire entre deux peptides. Nous avons choisi d'utiliser la mesure de similarité de Tanimoto (voir formule 2.1) qui donne des résultats satisfaisants pour de nombreux vecteurs moléculaires.

$$S_{A,B} = \frac{\sum_{m=1}^{n} \sqrt{A_m B_m}}{\sum_{m=1}^{n} A_m + \sum_{m=1}^{n} B_m - \sum_{i=m}^{n} \sqrt{A_m B_m}} \tag{2.1}$$

où A_m et B_m représentent les comptages du monomère m dans les vecteurs A et B ; et n est le nombre de monomères de NORINE.

Cette représentation a l'avantage de mesurer la ressemblance entre compositions et, ainsi, de classer les peptides selon leur ressemblance avec la composition cible. Si deux peptides ont le même nombre de monomères en commun avec la composition cible, alors le peptide ayant le plus petit nombre de monomères spécifiques aura une plus grande valeur de Tanimoto que celle de l'autre peptide.

De plus, dans la nouvelle implémentation de la recherche par composition, nous avons autorisé la possibilité d'indiquer des alternatives entre plusieurs monomères. Cela signifie que si au moins un des monomères d'une liste d'alternatives donnée est trouvé dans le peptide, un unique point commun entre la composition cible et le peptide est compté. Cette fonctionnalité est utile lorsque la prédiction des modifications subies par les monomères n'a pas été faite, comme c'est le cas dans programmes d'annotation de synthétases. Ainsi, il est possible de rechercher un monomère et tous ses dérivés.

distance	peptide
0.500	bacitracin A1
0.450	bacitracin A2
0.450	bacitracin B1
0.450	bacitracin B2
0.450	bacitracin B3
0.412	bacitracin F
0.390	bacitracin C
0.390	bacitracin D1
0.390	bacitracin D2
0.375	[Ile4,7]surfactin

~ Structure match ~

Monomer matches appear in same color

query

Ile , Cys , Leu , Glu , Ile , Lys , Orn , Ile , Phe , His , Asp , Asn

bacitracin A1

Ile , Cys , Leu , D-Glu , Ile , Lys , D-Orn , Ile , D-Phe , His , D-Asp , Asn

FIGURE 2.23 – Résultats obtenus en recherchant la composition de la ligne *query* à l'aide de la recherche via les fingerprints. Les monomères en commun entre la composition cible et la bacitracine A1 sont écrits avec la même couleur.

Si je reprends les résultats obtenus avec les outils d'analyse de SPNR appliqués au cluster de la bacitracine et que je soumets la composition obtenue à la recherche par composition, j'obtiens les résultats visibles dans la figure 2.23. Si la recherche est effectuée en autorisant les dérivés de chacun des monomères, un score de 1 est obtenu avec la bacitracine A1 car tous les monomères de la prédiction sont dans ce peptide avec ou sans modification.

La recherche d'un motif structural localise de façon exacte dans un peptide, une structure monomérique qui peut contenir des listes de monomères à certaines positions et même des jokers, c'est-à-dire des positions qui peuvent correspondre à n'importe quel monomère. L'algorithme,

défini par Ségolène Caboche, a été publié en 2009 dans *BMC Structural Biology* [21]. Je l'ai également décrit en français et dans une version compréhensible par un lectorat plus large, pour le site de culture scientifique)i(nterstices [92]. La solution que nous avons conçue se base sur les algorithmes d'isomorphisme de sous-graphe, c'est-à-dire localisation sans erreur d'un graphe dans un autre graphe plus grand. Nous l'avons optimisée en exploitant les particularités de nos graphes, à savoir le grand nombre d'étiquettes disponibles pour nommer les nœuds (les 528 monomères de NORINE) et le faible nombre d'arêtes par nœud (maximum 5), ainsi que le nombre limité de cycles (maximum 3).

Pour une description plus complète de l'algorithme, vous pouvez vous référer aux articles cités ci-dessus. Je vais décrire le contexte d'utilisation de cette fonctionnalité et donner un exemple d'utilisation. La recherche d'un motif est plus précise que la recherche par composition puisqu'elle inclut une structure 2D. Le nombre de peptides contenant un motif donné est inférieur ou égal au nombre de peptides contenant sa composition en monomères. Par contre, si la structure n'est pas bonne, le risque est de ne pas trouver le(s) peptide(s) recherché(s). Une structure partielle peut être obtenue par différents moyens : la spectrométrie de masse, l'analyse bio-informatique des synthétases peptidiques non-ribosomiques, la détermination d'une structure commune à une famille de peptides. Dans certains cas, le motif peut être composé de plusieurs fragments. Dans la nouvelle version de NORINE, la possibilité de rechercher plusieurs fragments en même temps au sein de chacun des peptides a été ajoutée. Notre algorithme est capable de retrouver les motifs y compris s'ils forment des structures différentes dans un peptide, à condition qu'aucune liaison ne soit cassée.

Afin de faciliter le processus d'analyse des synthétases et de proposer un outil complet, nous collaborons avec les auteurs du logiciel antiSMASH afin de mettre en place un lien direct entre les peptides prédits par leur logiciel et NORINE. En plus de la structure chimique du peptide et de la liste des monomères potentiellement incorporés par la synthétase, antiSMASH propose un lien vers NORINE. Pour l'instant, ce lien aboutit à notre formulaire de recherche et l'utilisateur doit le remplir manuellement. A court terme, il est prévu qu'une requête soit construite automatiquement pour interroger directement NORINE et afficher les résultats d'une recherche par composition en monomères croisés avec ceux de la recherche des fragments prédits. Ainsi, les différentes étapes d'exploration du potentiel de production des peptides non-ribosomiques à partir de la séquence d'un génome seront réalisables en soumettant simplement une séquence à antiSMASH. Toutes les analyses seront enchaînées automatiquement, y compris la comparaison avec les données de NORINE.

Pour illustrer la recherche par motifs, je vais de nouveau prendre l'exemple de la bacitracine. Sa synthétase est composée de 3 protéines qui produisent les 3 fragments peptidiques représentés ci-dessous, sous la forme de motifs linéaires qui autorisent les dérivés de chaque monomère prédit :

1. Ile*,Cys*,Leu*,Glu*,Ile*@1@0,2@1,3@2,4@3
2. Lys*,Orn*@1@0
3. Ile*,Phe*,His*,Asp*,Asn*@1@0,2@1,3@2,4@3

Si ces trois fragments sont soumis à la recherche de motifs, 35 peptides sont retournés car ils possèdent au moins un des motifs. Parmi ceux-ci, 3 bacitracines contiennent les 3 motifs, 8 en contiennent 2 et 5 n'en contiennent qu'un. Les autres peptides trouvés sont des pyoverdines qui contiennent le petit motif (le deuxième de la liste). Les 3 bacitracines qui possèdent les 3 motifs ne diffèrent que par les dérivés des monomères présents. Il est à noter que les motifs sont trouvés dans ces peptides bien que les monomères Ile*, Cys* soit liés par deux liaisons chimiques,

représentées par deux arêtes dans le graphe des monomères, et que les bacitracines forment un cycle partiel qui se ferme au niveau de Lys* (voir le graphe dans la figure 2.20).

La comparaison de structures monomériques a également été conçue par Ségolène Caboche. L'objectif est de ne plus faire de la recherche exacte d'un graphe dans un autre, mais d'autoriser des différences entre les deux graphes. L'algorithme fait partie de la classe de la recherche de sous-graphe commun maximum entre deux graphes. Concrètement, si 2 nœuds dont les étiquettes correspondent sont séparés par un même nombre de nœuds, alors ils sont inclus dans le sous-graphe commun maximum. Ainsi, les substitutions de monomères sont autorisées. Par contre, les insertions-délétion sont possibles uniquement à l'extérieur du sous-graphe commun maximum. Ensuite, une distance est calculée entre les deux peptides comparés, basée sur les nœuds et les arêtes identifiés dans le sous-graphe commun maximum.

NORINE propose de comparer une structure soumise à l'ensemble des peptides de la base et renvoie ceux qui ont une similarité non nulle avec la cible. Il est également possible de ne pas pénaliser les substitutions entre monomères d'un même groupe, les groupes étant ceux définis dans la base à l'aide de la table CLASSIFICATION. La comparaison de structures monomériques permet, par exemple, d'obtenir tous les peptides ayant des structures similaires.

La conversion des structures chimiques en structures monomériques est une nouvelle thématique de recherche qui a commencé en 2012 lors du stage de première année de Master Informatique de Yoann Dufresne et qui est encore en cours puisqu'il continue en thèse sur ce sujet. Son travail a été présenté lors de l'édition 2013 de JOBIM [36]. L'idée est d'être capable d'identifier, au sein des peptides, les monomères qui les composent et, ainsi, reconstruire de façon semi-automatique les structures monomériques à partir des structures chimiques. Les difficultés rencontrées sont les suivantes :

1. les 528 monomères comptent entre 9 et 60 atomes et sont recherchés dans des peptides comportant entre 37 et 338 atomes répartis dans au plus 26 monomères ;
2. lorsque les monomères sont incorporés dans le peptide, ils perdent des atomes du fait de la formation des liaisons chimiques. Nous avons caractérisé 5 types de liaisons fortement représentées au sein des peptides non-ribosomiques (voir partie 2.2.2), mais il en existe d'autres plus rares. Par homologie avec les acides aminés, les monomères incorporés dans un peptide et donc privés de certains de leurs atomes sont appelés résidus ;
3. les monomères peuvent subir des transformations avant, pendant ou après leur incorporation dans le peptide. Ces transformations peuvent être effectuées par des enzymes comme l'ajout de groupements par les domaines optionnels des synthétases, ou la transformation d'un pont disulfure en pont dithioacétal chez les quinomycines (voir partie 2.2.2). D'autres sont des réactions spontanées comme la tautomérisation qui est le mouvement de doubles liaisons et d'atomes au sein d'une molécule ;
4. la liste des monomères incorporés dans les PNR augmente régulièrement. Les monomères contenus dans NORINE ne couvrent pas tous les monomères possibles ;
5. les dérivés des monomères, et même certains monomères proches ont des sous-structures communes qui peuvent fausser l'identification d'un monomère à une position donnée d'un peptide.

Ces difficultés rendent difficile la localisation des monomères au sein des peptides représentés sous la forme de structures chimiques. Nous avons mis au point un premier algorithme basé sur

l'isomorphisme de sous-graphe. Dans un premier temps, les résidus des monomères sont générés en appliquant des règles de transformation que nous avons établies à partir d'observations sur les 5 liaisons les plus courantes. Cette étape permet de palier une partie de la difficulté numéro 2. Ces résidus sont classés par ordre de taille puis de score décroissant. Le score est la somme des poids correspondants aux liaisons appliquées aux monomères. Le poids est basé sur la fréquence d'apparition des différents types de liaisons dans les peptides de NORINE. Par exemple, le poids le plus élevé est celui des liaisons peptidiques. Ensuite, les résidus sont localisés dans les peptides à l'aide, pour l'instant, de fonctions proposées par la librairie OpenBabel [36]. Ils sont placés un par un, par ordre de tri et de façon définitive. Ainsi, les monomères contenant le plus d'atomes ne peuvent pas prendre la place des plus petits, alors que l'inverse peut se produire s'ils ne sont pas positionnés par ordre de taille décroissant. Cette stratégie permet de répondre en partie à la difficulté numéro 5. Notre algorithme glouton donne déjà des résultats satisfaisants puisque, sur les 204 peptides dont nous avons la structure atomique, 171 (soit près de 85%) ont l'ensemble de leurs monomères bien prédits. Les 33 peptides restants ont au moins 50% de leurs monomères bien prédits et ceux qui ne sont pas trouvés sont principalement des tautomères ou des erreurs dans NORINE. Ainsi, ce premier travail m'a également permis de commencer la correction des structures de la base.

Actuellement, Yoann Dufresne étudie des pistes d'amélioration de l'algorithme. Pour commencer, nous allons maintenant utiliser la librairie CDK [108] qui propose déjà des algorithmes d'isomorphisme de sous-graphe et de sous-graphe commun maximum. Cette dernière fonctionnalité nous aidera à mieux gérer les difficultés 3 et 4. Nous allons également travailler sur les stratégies de placement en permettant de revenir en arrière si les atomes du peptide ne sont pas suffisamment couverts par les résidus placés lors de la première itération.

L'algorithme de conversion facilitera le processus d'insertion d'un nouveau peptide dans NORINE. La détermination de structure monomérique à partir d'une structure atomique est l'étape la plus délicate. Dans certains articles, seule la structure chimique des peptides est fournie. Il est alors nécessaire de déterminer les monomères présents dans le peptide. La première étape est la localisation des liaisons peptidiques qui permettent de délimiter un certain nombre de monomères, voir tous les monomères dans les peptides linéaires ou cycliques. Lorsque d'autres types de liaisons sont en jeu, il n'est pas toujours facile de découper les monomères et de déterminer quels sont les atomes qui les composent et ceux qui ont été perdus lors de la formation de la liaison avec un autre monomère. Ensuite, il faut identifier les monomères ainsi découpés. En sachant que NORINE compte plus de 500 monomères différents, il est difficile de tous les connaître.

Le premier algorithme conçu a permis de mettre en avant des erreurs et incohérences dans les données de NORINE qui proviennent du processus de reconstruction des structures monomériques qui a été fait à la main et qui demande une grande expertise et un recul sur les données. Lors du travail d'automatisation de la conversion des structures chimiques en structures monomériques, nous avons détecté des erreurs dans la détermination de monomères présents dans quelques peptides. J'ai également observé des incohérences dans la gestion de certains monomères. Par exemple, les synthétases sont capables de former un cycle interne à partir de deux acides aminés (voir partie 1.2.4). Dans certains peptides, ces cycles internes sont représentés par deux liaisons entre deux monomères (ex : les bacitracines contiennent une isoleucine (Ile) et une cystéine (Cys) reliées par deux liaisons chimiques) ; dans d'autres ils sont représentés par un seul monomère (ex : les apramides contiennent soit le monomère Pro-Thz, soit le monomère NMe-Gly-Thz qui sont un cycle interne formé par une cystéine et, respectivement, une proline ou une N-méthyl-glycine).

Sur plus de mille peptides saisis sur une période de plusieurs années, il est impossible d'éviter les incohérences et les erreurs. Une observation globale des structures, ainsi qu'une automatisation au moins partielle de la détermination des structures monomériques à partir des structures chimiques va aider à harmoniser les données. Ce travail nécessite au préalable de définir formellement la notion de monomère. En effet, nous pouvons choisir de conserver soit un monomère qui représente le cycle interne formé par des acides aminés tel que Pro-Thz, soit deux monomères liés par deux liaisons chimiques tels que Ile avec Cys. Dans le premier cas, nous considérons que toutes les modifications effectuées par les synthétases elles-mêmes sur les monomères sont intégrées dans les monomères. Cette convention est déjà pratiquée lorsqu'il est précisé la configuration D du monomère ou sa forme méthylée par exemple. Dans le deuxième cas, nous considérons que le domaine de cyclisation est aussi un domaine de condensation qui effectue la liaison entre deux monomères, mais avec la particularité de former un cycle avec ces composés en les modifiants et réalisant deux liaisons chimiques. A ma connaissance, il n'y a pas d'autres cas particuliers de domaines posant des problèmes de définition de la notion de monomère. Par contre, les liaisons non peptidiques doivent être étudiées avec soin pour définir quels sont les atomes perdus lors de leur formation afin de déterminer avec justesse quels sont les monomères complets à l'origine du peptide.

2.4.3 Encadrements et collaborations

NORINE est le fruit d'une collaboration étroite entre des membres du laboratoire ProBioGEM et des membres de l'équipe Bonsai du LIFL. Elle est née du travail de thèse de Ségolène Caboche, encadrée par quatre [enseignants-]chercheurs, dont moi-même. Toutes les personnes décrites ci-dessous sont ou ont été membres de la *NORINE team*.

Grégory Kucherov, DR CNRS en Informatique actuellement au Laboratoire d'Informatique Gaspard-Monge de l'Université Paris-Est Marne-la-Vallée, et à l'époque membre de l'équipe Sequoia du LIFL, est le directeur de thèse de S. Caboche. Il a participé à la conception des algorithmes sur les graphes ;

Philippe Jacques, professeur en microbiologie à PolyTech'Lille, membre du laboratoire ProBioGEM, gère, notamment, la valorisation de NORINE. Il a apporté la thématique des PNR sur Lille ;

Valérie Leclère, maître de conférences HdR en microbiologie, membre du laboratoire ProBioGEM, gère, en particulier, l'annotation des peptides et de leurs synthétases. Elle a développé la thématique bio-informatique au sein de son laboratoire ;

Ségolène Caboche, actuellement ingénieure de recherche bio-informatique à l'Université Lille 2, a conçu NORINE lors de son doctorat (2006 – 2009) [18]. Elle a développé les premières versions de l'interface, incorporé plus de 1000 peptides non-ribosomiques et conçu des algorithmes de recherche d'un motif structural et de comparaison de peptides ;

J'ai encadré seule d'autres scientifiques qui ont contribué au développement de NORINE.

Ammar Abdo Hasan, actuellement enseignant à l'Université Hodeidah du Yemen, a effectué un séjour postdoctoral d'1 an (fev 2012 – fev 2013) dans l'équipe Bonsai du LIFL, au cours duquel il a mis au point une nouvelle représentation pour les peptides non-ribosomiques, sous la forme de vecteurs d'entiers, associée à des algorithmes de comparaison et de prédiction d'activités ;

Antoine Engelaere, alors étudiant en stage de 3ème année de Licence informatique à l'Université Lille 1 (avr 2013 – juin 2013), a développé et intégré dans NORINE la comparaison de peptides non-ribosomiques telle qu'elle a été définie par Ammar Abdo Hasan.

Mohcen Benmounah, ingénieur d'étude en Bio-informatique pour la plate-forme Bilille, a continué pendant 4 mois en 2012 les développements initiés par Laurie Tonon ;

Laurie Tonon, actuellement ingénieure de recherche bio-informatique à l'Institut National du Cancer, et à l'époque ingénieure jeune diplômée Inria dans l'équipe Sequoia, a travaillé pendant environ 6 mois entre 2010 et 2011 sur NORINE en structurant le code source avec une architecture Modèle-Vue-Contrôleur et en développant de nouvelles fonctionnalités ;

Enfin, les membres qui ont rejoint la *NORINE team* récemment et qui continuent à travailler dessus :

Stéphane Janot, maître de conférences en informatique à PolyTech'Lille, membre de l'équipe Bonsai, a pris en charge le code de NORINE ;

Areski Flissi, Ingénieur de recherche en Informatique au LIFL, a rejoint l'équipe Bonsai en juin 2013 et consacrera une partie de son temps à NORINE.

Laurent Noé, maître de conférences en informatique à l'UFR IEEA, membre de l'équipe Bonsai, apporte son expertise en algorithmique et co-encadre Yoann Dufresne avec moi ;

Yoann Dufresne, doctorant en informatique dans l'équipe Bonsai, a fait son stage de master recherche en informatique (fev 2013 – août 2013) sur la conversion des structures chimiques des peptides en structures monomériques et poursuit ce travail dans le cadre d'une thèse en informatique.

NORINE a acquis une reconnaissance internationale qui lui a valu d'être sollicitée par des organismes ou chercheurs étrangers pour initier des collaborations ou bénéficier d'accès particuliers à nos données.

wwPDB [12], la banque internationale de structures 3D de macromolécules, nous a choisi comme banque de référence pour les peptides non-ribosomiques, au même titre que la banque internationale UniProtKB [38] [113] l'est pour les protéines. Des liens croisés [39] entre nos deux banques sont accessibles via nos interfaces respectives. Cette liaison est citée dans l'article [117]. Nous entretenons une relation privilégiée avec le site européen de la banque wwPDB, appelé PDBe [116]. Nous avons même déposé en 2011, avec un autre partenaire européen, CCSD [40], et un partenaire des États-Unis d'Amérique, RCSB [41], un projet collaboratif dans le cadre de l'appel européen «NFRA 2012.3.2 : International cooperation with USA on common e-infrastructure for scientific data», qui n'a malheureusement pas obtenu une note suffisante pour être financé ;

le groupe de Pavel Pevzner, du *Center for Algorithmic and Systems Biology* de l'Université de Californie, utilise les données de NORINE dans leur outil d'interprétation de spectres de masse (voir 2.3.2) et nous ont contacté pour échanger sur nos thématiques de recherche respectives ;

38. Plate-forme centrale pour la collecte d'informations concernant les protéines.
39. Les entrées qui portent sur la même molécule se référencent l'une l'autre.
40. *Cambridge Crystallographic Data Centre*, fournisseur de la banque CSD (*Cambridge Structural Database*) de structures 3D de petites molécules organiques
41. *Research Collaboratory for Structural Bioinformatics*, le site américain de wwPDB

Timo Niedermeyer et Martin Strohalm, respectivement responsable du département *Natural Product Chemistry* chez Cyano Biotech GmbH à Berlin, Allemagne, et chercheur à l'Institut de Microbiologie de l'Académie des Sciences de République Tchèque, sont les développeurs de mMass, un outil open source pour la spectrométrie de masse (voir 2.3.2). Ils nous ont contacté pour savoir s'ils pouvaient diffuser librement les données de NORINE avec leur logiciel, nous avons accepté ;

les utilisateurs et contributeurs de NORINE, nous sollicitent de temps en temps à travers l'adresse de messagerie électronique que nous avons mise en place. Je suis maintenant la seule à répondre à ces messages, en sollicitant mes collègues si besoin.

Conclusion et perspectives

Au début de l'année 2006, la bio-informatique pour les peptides non-ribosomiques n'existait pas au LIFL. Sept ans après, je dirige un groupe de scientifiques travaillant sur ce thème, en collaboration étroite avec des membres du laboratoire ProBioGEM. Nous avons créé NORINE, la première ressource dédiée aux PNR. Sa base de données contient 1164 peptides annotés à partir de la littérature scientifique. Au cours des années, des outils de visualisation et d'édition des structures monomériques ont été développés, ainsi qu'une interface d'interrogation et de consultation des données. Des algorithmes de comparaison de peptides représentés sous la forme d'un graphe ou d'un vecteur de comptages de monomères ont été conçus et intégrés dans NORINE. Les algorithmes mis au point traitent des problèmes NP-complets sur les graphes. Les cinq articles scientifiques que nous avons publiés dans des journaux internationaux et qui présentent nos travaux sont cités par les scientifiques renommés dans le domaine des peptides non-ribosomiques. Nos données sont référencées par la banque de données internationale de structures 3D de protéines, wwPDB.

Il est rare de pouvoir travailler sur un sujet de recherche non encore exploré par d'autres scientifiques. J'ai saisi cette chance car j'ai compris que la bio-informatique pour les peptides non-ribosomiques est un sujet prometteur de part les spécificités des molécules concernées et les problèmes informatiques qu'elles soulèvent, ainsi que les applications possibles d'un point de vue pharmacologique ou biotechnologique. Le travail de pionniers et de qualité réalisé sur Lille par la *NORINE team* nous a permis d'acquérir une reconnaissance internationale. J'ai joué un rôle moteur lors de la conception de NORINE puis dans son évolution. Je propose maintenant de nouvelles activités de recherche originales qui s'appuient sur notre expertise concernant les PNR et sur le savoir-faire acquis à propos des méthodes algorithmiques pour manipuler les peptides représentés sous la forme de graphes de monomères. De plus, l'étude bibliographique menée pour rédiger ce mémoire a renforcé mes connaissances biochimiques sur ces peptides. Mon idée est de construire un pipeline semi-automatique qui prend en entrée une activité cible et est capable de proposer des synthétases peptidiques non-ribosomiques susceptibles de produire un peptide ayant cette activité. Ce processus nécessite plusieurs étapes et des données spécifiques dont certaines sont déjà existantes dans NORINE ou antiSMASH.

De l'activité vers les peptides. Nous avons déjà démontré une corrélation entre l'activité d'un peptide et ses monomères ainsi que sa structure 2D. Des études supplémentaires sont nécessaires pour mieux appréhender ces relations.

Pour commencer, il est intéressant de regrouper les monomères en fonction de leur influence sur l'activité du peptide. Deux techniques complémentaires peuvent être utilisées. La première est de reprendre le principe utilisé pour la constitution des matrices de substitution protéiques à savoir se baser sur les substitutions tolérées par la nature, en utilisant les variants d'un même

peptide. Mais, les volumes de données sont inférieurs à ceux des familles protéiques. De plus, certains variants d'un même peptide n'ont pas la même activité. Toute fois, les mutations qui impactent l'activité constituent des indices pour déceler des substitutions à éviter.

La deuxième technique pour regrouper les monomères est de constituer des ensembles en se basant sur les propriétés physico-chimiques. Par exemple, les acides aminés acéthylés et les amines alcools situés au début et à la fin de certains peptides permettent de discriminer les peptaibols. À partir de nos premières observations sur la relation entre les monomères et les activités et de nos connaissances sur certaines familles de peptides, j'ai déjà constitué des groupes de monomères. Il faut maintenant les tester et les affiner en regardant l'influence des groupes sur la qualité de la prédiction d'activités de PNR.

La structure 2D est importante pour l'activité du peptide, mais aussi pour sa stabilité. En effet, les formes non linéaires sont moins sensibles à 'hydrolyse, qu'elle soit chimique ou enzymatique. Il faut donc intégrer cette information dans le processus de prédiction d'activité.

Une fois que nous aurons identifié les facteurs caractéristiques de chaque activité, nous serons en mesure de construire des peptides théoriques ayant une activité donnée.

Des peptides vers les synthétases. L'étape suivante est d'établir une synthétase peptidique non-ribosomique capable de produire un peptide théorique ou observé mais dont la synthétase n'est pas connue. Il s'agit de proposer au moins un enchaînement de domaines cohérent avec les monomères désirés, avec d'éventuelles enzymes secondaires si nécessaire. Un niveau supplémentaire de prédiction est la prise en compte de la structure du peptide. Les logiciels actuels d'analyse des synthétases ne le font pas. En construisant des liens entre les annotations des clusters de gènes prédites par antiSMASH et les peptides de Norine, nous serons en mesure de découvrir, par apprentissage automatique, les fonctions enzymatiques impliquées dans la formation de la structure des peptides. Quelques unes sont déjà connues et ont été présentées dans ce mémoire.

Ensuite, la ou les synthétases théorique(s), accompagnée(s) de leurs enzymes secondaires, pourront être recherchées au sein de la base de données de SPNR que le groupe «Secondary Metabolite Genomics» de l'Université de Tuebingen constitue à l'aide de son logiciel antiSMASH. Les synthétases peptidiques non-ribosomiques réelles ressemblant à la synthétase théorique permettront d'ajuster la structure en domaines de cette dernière et, par conséquence, la structure monomérique du peptide cible.

Bibliographie

[1] W. H. A. M. Abdelwahed. *Study on the regulation and biosynthesis of fengycin and plipastatin produced by* Bacillus subtilis. ED SMRE - filière ingénierie des fonctions biologiques, Université des Sciences et Technologie de Lille - Lille I, Sept. 2011.

[2] A. Abderrahmani, A. Tapi, F. Nateche, M. Chollet, V. Leclère, B. Wathelet, H. Hacene, and P. Jacques. Bioinformatics and molecular approaches to detect NRPS genes involved in the biosynthesis of kurstakin from *Bacillus thuringiensis*. *Applied Microbiology and Biotechnology*, 92(3) :571–581, Nov. 2011.

[3] A. Abdo, S. Caboche, V. Leclère, P. Jacques, and M. Pupin. A new fingerprint to predict nonribosomal peptides activity. *Journal of Computer-Aided Molecular Design*, 26(10) :1187–1194, Oct. 2012.

[4] K. Agnoli, C. A. Lowe, K. L. Farmer, S. I. Husnain, and M. S. Thomas. The ornibactin biosynthesis and transport genes of *Burkholderia cenocepacia* are regulated by an extracytoplasmic function σ factor which is a part of the fur regulon. *Journal of Bacteriology*, 188(10) :3631–3644, May 2006.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410, Oct. 1990.

[6] M. Z. Ansari, G. Yadav, R. S. Gokhale, and D. Mohanty. NRPS-PKS : a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Research*, 32(suppl 2) :W405–W413, Jan. 2004.

[7] B. O. Bachmann and J. Ravel. Chapter 8 : Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. In David A. Hopwood, editor, *Methods in Enzymology*, volume Volume 458, pages 181–217. Academic Press, 2009.

[8] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of ISMB (International Conference on Intelligent Systems for Molecular Biology)*, volume 2, pages 28–36, 1994.

[9] C. J. Balibar, F. H. Vaillancourt, and C. T. Walsh. Generation of d amino acid residues in assembly of arthrofactin by dual Condensation/Epimerization domains. *Chemistry & Biology*, 12(11) :1189–1200, Nov. 2005.

[10] N. Bandeira, J. Ng, D. Meluzzi, R. G. Linington, P. Dorrestein, and P. A. Pevzner. De novo sequencing of nonribosomal peptides. In *Proceedings of RECOMB (Research in Computational Molecular Biology)*, page 181–195, 2008.

[11] M. Béchet, T. Caradec, W. Hussein, A. Abderrahmani, M. Chollet, V. Leclère, T. Dubois, D. Lereclus, M. Pupin, and P. Jacques. Structure, biosynthesis, and properties of kurstakins, nonribosomal lipopeptides from *Bacillus spp. Applied Microbiology and Biotechnology*, 95(3) :593–600, Aug. 2012.

[12] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley. The future of the protein data bank. *Biopolymers*, 99(3) :218–222, 2013.

[13] K. Blin, M. H. Medema, D. Kazempour, M. A. Fischbach, R. Breitling, E. Takano, and T. Weber. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research*, 41(W1) :W204–W212, Jan. 2013.

[14] I. G. Boneca, H. d. Reuse, J.-C. Epinat, M. Pupin, A. Labigne, and I. Moszer. A revised annotation and comparative analysis of *Helicobacter pylori* genomes. *Nucleic Acids Research*, 31(6) :1704–1714, Mar. 2003.

[15] A. A. Brakhage, M. Thön, P. Spröte, D. H. Scharf, Q. Al-Abdallah, S. M. Wolke, and P. Hortschansky. Aspects on evolution of fungal beta-lactam biosynthesis gene clusters and recruitment of trans-acting factors. *Phytochemistry*, 70(15-16) :1801–1811, Oct. 2009.

[16] H. Brückner and C. Toniolo. Towards a myriad of peptaibiotics. *Chemistry & Biodiversity*, 10(5) :731–733, 2013.

[17] K. E. Bushley and B. G. Turgeon. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evolutionary Biology*, 10(1) :26, Jan. 2010.

[18] S. Caboche. *Mise en place d'une plate-forme logicielle pour l'analyse des peptides nonribosomiaux*. ED SPI - spécialité informatique, Université des Sciences et Technologie de Lille - Lille I, Sept. 2009.

[19] S. Caboche, V. Leclère, M. Pupin, G. Kucherov, and P. Jacques. Diversity of monomers in nonribosomal peptides : towards the prediction of origin and biological activity. *Journal of Bacteriology*, 192(19) :5143–5150, Jan. 2010.

[20] S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques, and G. Kucherov. NORINE : a database of nonribosomal peptides. *Nucleic Acids Research*, 36(suppl 1) :D326–D331, Jan. 2008.

[21] S. Caboche, M. Pupin, V. Leclère, P. Jacques, and G. Kucherov. Structural pattern matching of nonribosomal peptides. *BMC Structural Biology*, 9(1) :15, Mar. 2009.

[22] T. Cadarec, M. Pupin, A. Vanvlassenbroeck, M.-D. Devignes, M. Smaïl-Tabbone, P. Jacques, and V. Leclère. Florine : an efficient workflow leading to accurate prediction of nonribosomal peptides, including isomery of monomers. *submitted to PLoS ONE*, 2013.

[23] G. L. Challis, J. Ravel, and C. A. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & Biology*, 7(3) :211–224, Mar. 2000.

[24] Y.-Q. Cheng and J. D. Walton. A eukaryotic alanine racemase gene involved in cyclic peptide biosynthesis. *Journal of Biological Chemistry*, 275(7) :4906–4911, Feb. 2000.

[25] A. Chhabra, A. S. Haque, R. K. Pal, A. Goyal, R. Rai, S. Joshi, S. Panjikar, S. Pasha, R. Sankaranarayanan, and R. S. Gokhale. Nonprocessive [2 + 2]e- off-loading reductase domains from mycobacterial nonribosomal peptide synthetases. *Proceedings of the National Academy of Sciences of the United States of America*, 109(15) :5681–5686, Oct. 2012.

[26] E. Conti, T. Stachelhaus, M. A. Marahiel, and P. Brick. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin s. *The EMBO Journal*, 16(14) :4174–4183, July 1997.

[27] K. R. Conway and C. N. Boddy. ClusterMine360 : a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Research*, 41(D1) :D402–D407, Jan. 2013.

[28] R. Couch, S. E. O'Connor, H. Seidle, C. T. Walsh, and R. Parry. Characterization of CmaA, an adenylation-thiolation didomain enzyme involved in the biosynthesis of coronatine. *Journal of Bacteriology*, 186(1) :35–42, Jan. 2004.

[29] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve : Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7) :1394–1403, Jan. 2004.

[30] A. L. Demain. Antibiotics : Natural products essential to human health. *Medicinal Research Reviews*, 29(6) :821–842, 2009.

[31] G. Didier, L. Debomy, M. Pupin, M. Zhang, A. Grossmann, C. Devauchelle, and I. Laprevotte. Comparing sequences without using alignments : application to HIV/SIV subtyping. *BMC Bioinformatics*, 8(1) :1, Jan. 2007.

[32] G. Didier, I. Laprevotte, M. Pupin, and A. Hénaut. Local decoding of sequences and alignment-free comparison. *Journal of Computational Biology*, 13(8) :1465–1476, Oct. 2006.

[33] J. Diminic, J. Zucko, I. T. Ruzic, R. Gacesa, D. Hranueli, P. F. Long, J. Cullum, and A. Starcevic. Databases of the thiotemplate modular systems (CSDB) and their *in silico* recombinants (r-CSDB). *Journal of Industrial Microbiology & Biotechnology*, 40(6) :653–659, June 2013.

[34] P. C. Dorrestein, K. Poole, and T. P. Begley. Formation of the chromophore of the pyoverdine siderophores by an oxidative cascade. *Organic Letters*, 5(13) :2215–2217, June 2003.

[35] E. J. Drake and A. M. Gulick. Three-dimensional structures of *Pseudomonas aeruginosa* PvcA and PvcB, two proteins involved in the synthesis of 2-isocyano-6,7-dihydroxycoumarin. *Journal of Molecular Biology*, 384(1) :193–205, Dec. 2008.

[36] Y. Dufresne, V. Leclère, P. Jacques, L. Noé, and M. Pupin. Non ribosomal peptides : A monomeric puzzle. In *Proceedings of JOBIM*, Toulouse, France, July 2013.

[37] S. R. Eddy. Accelerated profile HMM searches. *PLoS Comput Biol*, 7(10) :e1002195, Oct. 2011.

[38] D. P. Fewer, J. Österholm, L. Rouhiainen, J. Jokela, M. Wahlsten, and K. Sivonen. Nostophycin biosynthesis is directed by a hybrid polyketide synthase - nonribosomal peptide synthetase in the toxic *Cyanobacterium nostoc sp.* strain 152. *Applied and Environmental Microbiology*, 77(22) :8034–8040, Nov. 2011.

[39] J. D. Fischer, C. E. Mayer, and J. Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24(5) :613–620, Jan. 2008.

[40] C. D. Fjell, J. A. Hiss, R. E. W. Hancock, and G. Schneider. Designing antimicrobial peptides : form follows function. *Nature Reviews Drug Discovery*, 11(1) :37–51, Jan. 2012.

[41] A. Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *The British Journal pf experimental pathology*, 10 :226–236, 1929.

[42] A. Gallo, M. Ferrara, and G. Perrone. Phylogenetic study of polyketide synthases and nonribosomal peptide synthetases involved in the biosynthesis of mycotoxins. *Toxins*, 5(4) :717–742, Apr. 2013.

[43] X. Gao, S. W. Haynes, B. D. Ames, P. Wang, L. P. Vien, C. T. Walsh, and Y. Tang. Cyclization of fungal nonribosomal peptides by a terminal condensation-like domain. *Nature Chemical Biology*, 8(10) :823–830, Oct. 2012.

[44] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. ChEMBL : a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1) :D1100–D1107, Sept. 2011.

[45] N. Geib, K. Woithe, K. Zerbe, D. B. Li, and J. A. Robinson. New insights into the first oxidative phenol coupling reaction during vancomycin biosynthesis. *Bioorganic & Medicinal Chemistry Letters*, 18(10) :3081–3084, May 2008.

[46] C. R. Groom and F. H. Allen. The cambridge structural database : experimental three-dimensional information on small molecules is a vital resource for interdisciplinary research and learning. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(3) :368–376, 2011.

[47] A. Haese, M. Schubert, M. Herrmann, and R. Zocher. Molecular characterization of the enniatin synthetase gene encoding a multifunctional enzyme catalysing n-methyldepsipeptide formation in *Fusarium scirpi*. *Molecular Microbiology*, 7(6) :905–914, 1993.

[48] D. H. Haft, J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck. TIGRFAMs and genome properties in 2013. *Nucleic Acids Research*, 41(D1) :D387–D395, Nov. 2012.

[49] M. Hoppert, C. Gentzsch, and K. Schörgendorfer. Structure and localization of cyclosporin synthetase, the key enzyme of cyclosporin biosynthesis in *Tolypocladium inflatum*. *Archives of Microbiology*, 176(4) :285–293, Oct. 2001.

[50] J. Hou, L. Robbel, and M. A. Marahiel. Identification and characterization of the lysobactin biosynthetic gene cluster reveals mechanistic insights into an unusual termination module architecture. *Chemistry & Biology*, 18(5) :655–664, May 2011.

[51] A. R. Howard-Jones, R. G. Kruger, W. Lu, J. Tao, C. Leimkuhler, D. Kahne, and C. T. Walsh. Kinetic analysis of teicoplanin glycosyltransferases and acyltransferase reveal ordered tailoring of aglycone scaffold to reconstitute mature teicoplanin. *Journal of the American Chemical Society*, 129(33) :10082–10083, Aug. 2007.

[52] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats, and S.-Y. Yong. InterPro in 2011 : new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(D1) :D306–D312, Nov. 2011.

[53] G. H. Hur, C. R. Vickery, and M. D. Burkart. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Natural Product Reports*, 29(10) :1074–1098, Sept. 2012.

[54] A. Ibrahim, L. Yang, C. Johnston, X. Liu, B. Ma, and N. A. Magarvey. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47) :19196–19201, Nov. 2012.

[55] N. Ichikawa, M. Sasagawa, M. Yamamoto, H. Komaki, Y. Yoshida, S. Yamazaki, and N. Fujita. DoBISCUIT : a database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Research*, 41(D1) :D408–D414, Nov. 2012.

[56] H. Jenke-Kodama and E. Dittmann. Bioinformatic perspectives on NRPS/PKS megasynthases : Advances and challenges. *Natural Product Reports*, 26(7) :874–883, June 2009.

[57] C. Johnston, A. Ibrahim, and N. Magarvey. Informatic strategies for the discovery of polyketides and nonribosomal peptides. *MedChemComm*, 3(8) :932–937, Aug. 2012.

[58] Jones D.T. GenTHREADER : an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 286(4) :797–815, 1999.

[59] O. V. Kalinina, A. A. Mironov, M. S. Gelfand, and A. B. Rakhmaninova. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Science*, 13(2) :443–456, 2004.

[60] P. Kamra, R. S. Gokhale, and D. Mohanty. SEARCHGTr : a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites. *Nucleic Acids Research*, 33(suppl 2) :W220–W225, Jan. 2005.

[61] M. Kaneda. Studies on bottromycins. i. 1H and 13C NMR assignments of bottromycin a2, the main component of the complex. *The Journal of antibiotics*, 45(5) :792–796, May 1992.

[62] T. A. Keating, C. G. Marshall, and C. T. Walsh. Reconstitution and characterization of the *Vibrio cholerae* vibriobactin synthetase from VibB, VibE, VibF, and VibH. *Biochemistry*, 39(50) :15522–15530, Dec. 2000.

[63] N. Khaldi, F. T. Seifuddin, G. Turner, D. Haft, W. C. Nierman, K. H. Wolfe, and N. D. Fedorova. SMURF : genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology*, 47(9) :736–741, Sept. 2010.

[64] B. I. Khayatt, L. Overmars, R. J. Siezen, and C. Francke. Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden markov models. *PLoS ONE*, 8(4), Apr. 2013.

[65] G. A. Khoury, R. C. Baliban, and C. A. Floudas. Proteome-wide post-translational modification statistics : frequency analysis and curation of the swiss-prot database. *Scientific Reports*, 1, Sept. 2011.

[66] R. M. Kohli, J. W. Trauger, D. Schwarzer, M. A. Marahiel, and C. T. Walsh. Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry*, 40(24) :7099–7108, June 2001.

[67] H. Lachance, S. Wetzel, K. Kumar, and H. Waldmann. Charting, navigating, and populating natural product chemical space for drug discovery. *Journal of Medicinal Chemistry*, 55(13) :5989–6001, July 2012.

[68] I. Laprevotte, M. Pupin, E. Coward, G. Didier, C. Terzian, C. Devauchelle, and A. Hénaut. HIV-1 and HIV-2 LTR nucleotide sequences : Assessment of the alignment by n-block presentation, "Retroviral signatures" of overrepeated oligonucleotides, and a probable important role of scrambled stepwise Duplications/Deletions in molecular evolution. *Molecular Biology and Evolution*, 18(7) :1231–1245, Jan. 2001.

[69] J. Li and S. E. Jensen. Nonribosomal biosynthesis of fusaricidins by *Paenibacillus polymyxa* PKB1 involves direct activation of a D-amino acid. *Chemistry & Biology*, 15(2) :118–127, Feb. 2008.

[70] J. W.-H. Li and J. C. Vederas. Drug discovery and natural products : End of an era or an endless frontier ? *Science*, 325(5937) :161–165, Oct. 2009.

[71] M. H. Li, P. M. Ung, J. Zajkowski, S. Garneau-Tsodikova, and D. H. Sherman. Automated genome mining for natural products. *BMC Bioinformatics*, 10(1) :185, June 2009.

[72] Q. Li, T. Cheng, Y. Wang, and S. H. Bryant. PubChem as a public resource for drug discovery. *Drug discovery today*, 15(23-24) :1052–1057, Dec. 2010.

[73] U. Linne, D. Schwarzer, G. N. Schroeder, and M. A. Marahiel. Mutational analysis of a type II thioesterase associated with nonribosomal peptide synthesis. *European Journal of Biochemistry*, 271(8) :1536—1545, 2004.

[74] W.-T. Liu, J. Ng, D. Meluzzi, N. Bandeira, M. Gutierrez, T. L. Simmons, A. W. Schultz, R. G. Linington, B. S. Moore, W. H. Gerwick, P. A. Pevzner, and P. C. Dorrestein. The interpretation of tandem mass spectra obtained from cyclic non-ribosomal peptides. *Analytical chemistry*, 81(11) :4200–4209, June 2009.

[75] B. Ma, M. E. Hibbing, H.-S. Kim, R. M. Reedy, I. Yedidia, J. Breuer, J. Breuer, J. D. Glasner, N. T. Perna, A. Kelman, and A. O. Charkowski. Host range and molecular phylogenies of the soft rot enterobacterial genera *Pectobacterium* and *Dickeya*. *Phytopathology*, 97(9) :1150–1163, Sept. 2007.

[76] B. Manavalan, S. K. Murugapiran, G. Lee, and S. Choi. Molecular modeling of the reductase domain to elucidate the reaction mechanism of reduction of peptidyl thioester into its corresponding alcohol in non-ribosomal peptide synthetases. *BMC Structural Biology*, 10(1) :1, Jan. 2010.

[77] M. Marahiel and L. Essen. Chapter 13 nonribosomal peptide synthetases : Mechanistic and structural aspects of essential domains. In David A. Hopwood, editor, *Methods in Enzymology*, volume Volume 458, pages 337–351. Academic Press, 2009.

[78] M. A. Marahiel, T. Stachelhaus, and H. D. Mootz. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chemical Reviews*, 97(7) :2651–2674, Nov. 1997.

[79] A. M. Matter, S. B. Hoot, P. D. Anderson, S. S. Neves, and Y.-Q. Cheng. Valinomycin biosynthetic gene cluster in *Streptomyces* : Conservation, ecology and evolution. *PLoS ONE*, 4(9) :e7194, Sept. 2009.

[80] M. H. Medema, K. Blin, P. Cimermancic, V. d. Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling. antiSMASH : rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(suppl 2) :W339–W346, Jan. 2011.

[81] Y. Minowa, M. Araki, and M. Kanehisa. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *Journal of Molecular Biology*, 368(5) :1500–1517, May 2007.

[82] D. J. Newman and G. M. Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of Natural Products*, 75(3) :311–335, Mar. 2012.

[83] J. Ng, N. Bandeira, W.-T. Liu, M. Ghassemian, T. L. Simmons, W. H. Gerwick, R. Linington, P. C. Dorrestein, and P. A. Pevzner. Dereplication and de novo sequencing of nonribosomal peptides. *Nature Methods*, 6(8) :596–599, 2009.

[84] L. T. Ngo, J. I. Okogun, and W. R. Folk. 21st century natural product research and drug development and traditional medicines. *Natural Product Reports*, 30(4) :584–592, Mar. 2013.

[85] G. Nicola, T. Liu, and M. K. Gilson. Public domain databases for medicinal chemistry. *Journal of Medicinal Chemistry*, 55(16) :6987–7002, Aug. 2012.

[86] T. H. J. Niedermeyer and M. Strohalm. mMass as a software tool for the annotation of cyclic peptide tandem mass spectra. *PLoS ONE*, 7(9), Sept. 2012.

[87] I. Pagani, K. Liolios, J. Jansson, I.-M. A. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The genomes OnLine database (GOLD) v.4 : status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40(D1) :D571–D579, Dec. 2011.

[88] H. M. Patel and C. T. Walsh. In vitro reconstitution of the *Pseudomonas aeruginosa* nonribosomal peptide synthesis of pyochelin : Characterization of backbone tailoring thiazoline reductase and n-methyltransferase activities. *Biochemistry*, 40(30) :9023–9031, July 2001.

[89] M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3) :e9490, Mar. 2010.

[90] C. Prieto, C. García-Estrada, D. Lorenzana, and J. F. Martín. NRPSsp : non-ribosomal peptide synthase substrate predictor. *Bioinformatics*, 28(3) :426–427, Jan. 2012.

[91] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The pfam protein families database. *Nucleic Acids Research*, 40(D1) :D290–D301, Nov. 2011.

[92] M. Pupin. Des peptides à explorer. In *Interstices*, Mar. 2010.

[93] A. Rantala, D. P. Fewer, M. Hisbergues, L. Rouhiainen, J. Vaitomaa, T. Börner, and K. Sivonen. Phylogenetic evidence for the early evolution of microcystin synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2) :568–573, Jan. 2004.

[94] C. Rausch, I. Hoof, T. Weber, W. Wohlleben, and D. H. Huson. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evolutionary Biology*, 7(1) :78, May 2007.

[95] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Research*, 33(18) :5799–5808, Jan. 2005.

[96] D. Roelofs, M. J. T. N. Timmermans, P. Hensbergen, H. v. Leeuwen, J. Koopman, A. Faddeeva, W. Suring, T. E. d. Boer, J. Mariën, R. Boer, R. Bovenberg, and N. M. v. Straalen. A functional isopenicillin N synthase in an animal genome. *Molecular Biology and Evolution*, 30(3) :541–548, Jan. 2013.

[97] N. Roongsawang, K. Washio, and M. Morikawa. In vivo characterization of tandem C-terminal thioesterase domains in arthrofactin synthetase. *ChemBioChem*, 8(5) :501–512, 2007.

[98] N. Roongsawang, K. Washio, and M. Morikawa. Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. *International Journal of Molecular Sciences*, 12(1) :141–172, Dec. 2010.

[99] M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch, and O. Kohlbacher. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research*, 39(suppl 2) :W362–W367, Jan. 2011.

[100] R. Samson, J. B. Legendre, R. Christen, M. F.-L. Saux, W. Achouak, and L. Gardan. Transfer of *Pectobacterium chrysanthemi* (burkholder et al. 1953) brenner et al. 1973 and *Brenneria paradisiaca* to the genus *Dickeya gen. nov.* as *Dickeya chrysanthemi comb. nov.* and *Dickeya paradisiaca comb. nov.* and delineation of four novel species, *Dickeya dadantii sp. nov.*, *Dickeya dianthicola sp. nov.*, *Dickeya dieffenbachiae sp. nov.* and *Dickeya zeae sp. nov. International Journal of Systematic and Evolutionary Microbiology*, 55(4) :1415–1427, Jan. 2005.

[101] T. L. Schneider, B. Shen, and C. T. Walsh. Oxidase domains in epothilone and bleomycin biosynthesis : Thiazoline to thiazole oxidation during chain elongation. *Biochemistry*, 42(32) :9722–9730, Aug. 2003.

[102] G. Schoenafinger and M. A. Marahiel. Nonribosomal peptides. In N. Civjan, editor, *Natural Products in Chemical Biology*, pages 109–125. John Wiley & Sons, Inc., 2012.

[103] D. Schwarzer, R. Finking, and M. A. Marahiel. Nonribosomal peptides : from genes to products. 20(3) :275–287, June 2003.

[104] R. F. Seipke, J. Barke, C. Brearley, L. Hill, D. W. Yu, R. J. M. Goss, and M. I. Hutchings. A single *Streptomyces* symbiont makes multiple antifungals to support the fungus farming ant *Acromyrmex octospinosus*. *PLoS ONE*, 6(8) :e22028, Aug. 2011.

[105] T. Stachelhaus, H. D. Mootz, and M. A. Marahiel. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry & Biology*, 6(8) :493–505, Aug. 1999.

[106] A. Starcevic, J. Zucko, J. Simunkovic, P. F. Long, J. Cullum, and D. Hranueli. ClustScan : an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Research*, 36(21) :6882–6892, Jan. 2008.

[107] E. Stegmann, H.-J. Frasch, and W. Wohlleben. Glycopeptide biosynthesis in the context of basic cellular functions. *Current Opinion in Microbiology*, 13(5) :595–602, Oct. 2010.

[108] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (CDK) : an open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2) :493–500, Mar. 2003.

[109] N. Stoppacher, N. K. N. Neumann, L. Burgstaller, S. Zeilinger, T. Degenkolb, H. Brückner, and R. Schuhmacher. The comprehensive peptaibiotics database. *Chemistry & Biodiversity*, 10(5) :734–743, 2013.

[110] G.-L. Tang, Y.-Q. Cheng, and B. Shen. Chain initiation in the leinamycin-producing hybrid nonribosomal peptide/polyketide synthetase from *Streptomyces atroolivaceus* S-140 : discrete, monofunctional adenylation enzyme and peptidyl carrier protein that directly load D-alanine. *Journal of Biological Chemistry*, 282(28) :20273–20282, July 2007.

[111] A. Tapi. *Stratégie moléculaire de mise en évidence de peptides actifs d'origine non ribosomiale chez* Bacillus sp. *et* Lactobacillus sp. ED SMRE - filière ingénierie des fonctions biologiques, Université des Sciences et Technologie de Lille - Lille I, Jan. 2010.

[112] A. Tapi, M. Chollet-Imbert, B. Scherens, and P. Jacques. New approach for the detection of non-ribosomal peptide synthetase genes in *Bacillus* strains by polymerase chain reaction. *Applied Microbiology and Biotechnology*, 85(5) :1521–1531, Feb. 2010.

[113] The UniProt Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Research*, 41(D1) :D43–D47, 2013.

[114] C. J. Thibodeaux, C. E. Melançon, and H.-w. Liu. Unusual sugar biosynthesis and natural product glycodiversification. *Nature*, 446(7139) :1008–1016, Apr. 2007.

[115] A. Vanvlassenbroeck. *Etude expérimentale et* in silico *du potentiel de synthèse NRPS chez les* Pseudomonas *fluorescents*. ED SMRE - filière ingénierie des fonctions biologiques, Université des Sciences et Technologie de Lille - Lille I, July 2012.

[116] S. Velankar, Y. Alhroub, C. Best, S. Caboche, M. J. Conroy, J. M. Dana, M. A. Fernandez Montecelo, G. van Ginkel, A. Golovin, S. P. Gore, A. Gutmanas, P. Haslam, P. M. S. Hendrickx, E. Heuson, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L. E. Newman, T. J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-Garcia, S. Sen, R. Slowley, A. Suarez-Uruena, G. J. Swaminathan, M. F. Symmons, W. F. Vranken, M. Wainwright, and G. J. Kleywegt. PDBe : Protein Data Bank in Europe. *Nucleic Acids Research*, 40(D1) :D445–D452, 2012.

[117] S. Velankar, J. M. Dana, J. Jacobsen, G. v. Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M.-J. Martin, and G. J. Kleywegt. SIFTS : structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(D1) :D483–D489, Jan. 2013.

[118] P. Visca, F. Imperi, and I. L. Lamont. Pyoverdine siderophores : from biogenesis to biosignificance. *Trends in Microbiology*, 15(1) :22–30, Jan. 2007.

[119] C. T. Walsh and M. A. Fischbach. Natural products version 2.0 : Connecting genes to molecules. *Journal of the American Chemical Society*, 132(8) :2469–2493, Mar. 2010.

[120] C. K. L. Wang, Q. Kaas, L. Chiche, and D. J. Craik. CyBase : a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, 36(suppl 1) :D206–D210, Jan. 2008.

[121] G. Wang. *Antimicrobial Peptides : Discovery, Design and Novel Therapeutic Strategies*. CABI, 2010.

[122] G. Wang, X. Li, and Z. Wang. APD2 : the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, 37(suppl 1) :D933–D937, Jan. 2009.

[123] K. Watanabe, K. Hotta, A. P. Praseuth, K. Koketsu, A. Migita, C. N. Boddy, C. C. C. Wang, H. Oguri, and H. Oikawa. Total biosynthesis of antitumor nonribosomal peptides in *Escherichia coli*. *Nature Chemical Biology*, 2(8) :423–428, Aug. 2006.

[124] K. Watanabe, H. Oguri, and H. Oikawa. Diversification of echinomycin molecular structure by way of chemoenzymatic synthesis and heterologous expression of the engineered echinomycin biosynthetic pathway. *Current Opinion in Chemical Biology*, 13(2) :189–196, Apr. 2009.

[125] T. Weber, C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D. Huson, and W. Wohlleben. CLUSEAN : a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *Journal of Biotechnology*, 140(1–2) :13–17, Mar. 2009.

[126] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1) :31–36, Feb. 1988.

[127] K. J. Weissman. Chapter 1 introduction to polyketide biosynthesis. In David A. Hopwood, editor, *Methods in Enzymology*, volume Volume 459, pages 3–16. Academic Press, 2009.

[128] X. Yin and T. M. Zabriskie. The enduracidin biosynthetic gene cluster from *Streptomyces fungicidicus*. *Microbiology*, 152(10) :2969–2983, Jan. 2006.

[129] W. Zhang, Z. Li, X. Miao, and F. Zhang. The screening of antimicrobial bacteria with diverse novel nonribosomal peptide synthetase (NRPS) genes from south china sea sponges. *Marine Biotechnology*, 11(3) :346–355, June 2009.

[130] N. Ziemert, S. Podell, K. Penn, J. H. Badger, E. Allen, and P. R. Jensen. The natural product domain seeker NaPDoS : a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS ONE*, 7(3) :e34064, Mar. 2012.

Maude Pupin | Maître de conférences en informatique

Université Lille 1 - LIFL - Bat M3 – 59655 Villeneuve d'Ascq Cedex – France
☎ +33 (0)3 28 77 85 55 • ✉ maude.pupin@univ-lille1.fr
http://www.lifl.fr/~pupin/

Diplomes

Doctorat de Génétique — **Université Versailles Saint-Quentin**
Mention très honorable avec félicitations du jury, — *1997–2000*

DESS en informatique appliquée à la biologie — **Université Paris 6**
Mention bien, — *1995–1996*

Maîtrise de biologie cellulaire — **Université Paris 7**
Mention assez bien, — *1994–1995*

Thèse

Titre: *Étude de répétitions locales et approximatives et élaboration d'une base de données génomique*

Directeur: Antoine Danchin, DR CNRS, Institut Pasteur de Paris

Date soutenance: 6 janvier 2000

Description: J'ai travaillé avec des mathématiciens pour adapter leurs méthodes de détection de répétitions locales et approximatives aux séquences nucléiques [2, 3]. Ensuite, j'ai appliqué ces méthodes aux génomes complets disponibles à l'époque (les bactéries *E. coli* et *B. subtilis* et l'eucaryote *S. cerevisiae*). J'ai également contribué à la réalisation d'une base de données de génomes bactériens, appelée GenoList [1, 4], [12] (`http://genolist.pasteur.fr/`).

Experiences professionnelles

Délégation Inria, temps plein — **Inria Lille-Nord Europe**
équipe Sequoia, LIFL et Inria Lille-Nord Europe — *09/08–08/09*

Maître de conférences en informatique — **Université Lille 1**
équipe Bonsai, LIFL et Inria Lille-Nord Europe, UFR IEEA — *depuis 2000*

ATER en informatique — **Université Lille 1**
équipe Bioinfo, LIFL, UFR IEEA — *02/00–08/00*

Recherche

- **Sujet en continuité avec ma thèse : comparaison de séquences nucléiques**
  - adaptation de la N-écriture à la comparaison multiple de séquences nucléiques ;
  - participation à l'amélioration de l'algorithme de la N-écriture qui est devenu le décodage local à l'ordre N [5] ;

- application à la phylogénie de 70 génomes de virus immuno-déficient humains et simiens [6] ;
- encadrement de un doctorant et six étudiants en stage de master recherche sur des sujets connexes.

- **Nouveau sujet : bio-informatique pour les peptides non-ribosomiques**
  - création de la ressource Norine dédiée aux peptides non-ribosomiques [7] ;
  - conception d'une base de données consultable via Internet, associée à des outils de visualisation et édition ;
  - conception d'algorithmes de recherche de motifs structuraux et de comparaison de peptides représentés sous la forme de graphes [8] ;
  - étude de la diversité structurelle des peptides [9] ;
  - prédiction de l'activité des peptides non-ribosomiques [10] ;
  - étude des organismes producteur de la kurstakine, un lipopeptide produit par les *Bacilli* [11] ;
  - encadrement d'un postdoctorant, de 3 doctorants (dont un en cours), de 3 ingénieurs et de 4 étudiants en stage de master recherche.

Enseignement et diffusion scientifique

Bio-informatique: Création des supports pédagogiques destinés aux étudiants des filières biologie et informatique, responsabilité de six UE.

- UFR de Biologie
  - $3^{ème}$ et $4^{ème}$ années de l'IUP Génomique et Protéomique (10h cours, 24h TD)(2003–2007)
  - Maîtrise/Master 1ère année Génétique et Microbiologie (24h TD)(2002–2004)
  - DESS puis Master pro Génie Cellulaire et Moléculaire (30h TD)(2001–2012)
  - DESS/Master pro Protéomique en commun avec le Master recherche Physico-Chimie du Vivant de l'UFR de Chimie (26h cours-TD)(2001–2009)
  - DEA/Master recherche Biologie-Santé, co-habilité avec l'Université de Lille 2 (6h cours, 8h TD)(2004–2012)
  - Ecole Doctorale de Biologie-Santé (18h TD)(2007–2009)
- École d'ingénieurs Polytech'Lille
  - Option transversale proposée aux élèves ingénieurs en $3^{ème}$ année (12h cours)(2005–2008)
- UFR IEEA (Informatique, Électronique, Électrotechnique et Automatique)
  - Master MOCAD (MOdèles Complexes, Algorithmes et Données) (7h Cours-TD) (depuis 2011)
  - DESS/Master pro Bio-informatique de Lille (10h cours, 34h TD) (2001–2006)
  - DEA/Master recherche Informatique (4h cours) (2005–2010)

Informatique: Encadrement de TD/TP dans 2 UE et responsabilité de 2 UE.

- UFR IEEA
  - Chargée de TD/TP de l'UE Initiation à la programmation en $1^{ère}$ année de DEUG/Licence Sciences pour l'Ingénieur (24h TD, 24h TP) (2000–2003, 2006–...)
  - Chargée de TD/TP de l'UE Introduction aux Bases de Données Relationnelles en $3^{ème}$ année de Licence Informatique (18h TD, 18h TP) (depuis 2012)
- UFR de Mathématiques
  - Responsable de l'UE Mise à niveau en informatique en $1^{ère}$ année de Master Mathématique et Finances (depuis sa création en 2010)
- École Doctorale SPI (Sciences Pour l'Ingénieur)
  - Responsable du module Internet, présentation du langage HTML et des CSS (18h cours-TD) (2007–2011)

Insertion professionnelle: Coordination des enseignements liés à l'insertion professionnelle des étudiants en informatique.

- UFR IEEA
  - Responsable de l'UE DPP (Détermination du Projet Professionnel) en $3^{ème}$ année de Licence Informatique (9h TD) (depuis 2005)
  - Responsable de l'UE PPP (Préparer son Projet Professionnel) en $1^{ère}$ année de Master Informatique (gestion des intervenants extérieurs) (depuis 2006)
- Collège Doctoral Lille Nord de France
  - Référente insertion professionnelle pour les doctorants de l'école doctorale SPI (depuis 2012)

Vulgarisation scientifique: Interventions régulières lors de manifestations scientifiques destinées à des élèves de primaire, collège ou lycée et au grand public. Rédaction d'un article pour le site de culture scientifique)i(nterstices [13]

Encadrement d'étudiants

- **PostDoctorant**
  - **Ammar Hasan Abdo**
    Prédiction de l'activité des peptides non-ribosomiques *fev 12–fev 13*
    yéménite ayant fait sa thèse en chémo-informatique au Mali

- **Doctorants**
  - **Yoann Dufresne**: thèse en informatique *depuis 2013*
    Titre: *Modèles et algorithmes pour la gestion de la biodiversité des peptides non-ribosomiques et la mise en évidence de nouveaux peptides bioactifs*
    Directrice: Hélène Touzet, DR CNRS, puis moi
    Co-encadrant: Laurent Noé, MCF en informatique
  - **Aurélien Vanvlassenbroeck**: thèse en ingénierie des fonctions biologiques *2008–2012*
    Titre: *Etude expérimentale et in silico du potentiel de synthèse NRPS chez les Pseudomonas fluorescents*
    Directeur: Philippe Jacques, Pr en microbiologie
    Co-encadrantes: Maude Pupin et Valérie Leclère, MCF en microbiologie
  - **Ségolène Caboche**: thèse en informatique *2006–2009*
    Titre: *Mise en place d'une plate-forme logicielle pour l'analyse des peptides non-ribosomiaux*
    Directeur: Grégory Kucherov, DR CNRS
    Co-encadrants: Maude Pupin, Valérie Leclère, MCF en microbiologie et Philippe Jacques, Pr en microbiologie
  - **Laurent Debomy**: thèse en informatique non soutenue pour raison personnelle *2000–2003*
    Titre: *Application de la N-écriture à la comparaison multiple de séquences*
    Directeur: Jean-Paul Delahaye, Pr en informatique
    Co-encadrante: Maude Pupin

- **Ingénieurs**

- **Mohcen Benmounah**
  Développements pour la ressource publique Norine *2013*
  pendant 4 mois
- **Louise Ott**
  Développements d'une base de données privée sur les synthétases *2010*
  pendant 6 mois
- **Laurie Tonon**
  Développements pour la ressource publique Norine *2009–2010*
  pendant 6 mois

Étudiants en stage de master recherche

- **Yoann Dufresne**
  Outils pour les peptides non-ribosomiques *2013*
  co-encadré avec Laurent Noé
- **Zorha Saci**
  Contribution à une plate-formed'analyse des peptides non-ribosomiques *2010*
  co-encadrée avec Valérie Leclère
- **Uciel Pablo Chorostecki**
  Bioinformatics annotation of specific enzymes *2010*
  étudiant argentin
- **Estefania Mancini**
  Ancestral sequence reconstruction with applications to homology search *2009*
  étudiante argentine, co-encadrée avec Laurent Noé
- **Cédric Molendi**
  Identification de regroupements de fragments répétés dans une ou plusieurs séquences *2008*
  co-encadré avec Laurent Noé et Grégory Kuchérov
- **Ségolène Caboche**
  Base de données et comparaison de peptides non-ribosomiques *2006*
  co-encadré avec Valérie Leclère
- **Bertrand Drache**
  Alignement multiple de séquences sans alignement *2005*
- **Gwénaël Monot**
  Identification de séquences reporters pour la conception de biopuces *2004*
  co-encadré avec Hélène Touzet
- **Antony Wojcik**
  Découverte de sites promoteurs dans le génome humain *2003*
  co-encadré avec Hélène Touzet
- **Mickaël Delautre**
  Alignement de séquences génétiques et compression de données *2000*
  co-encadré avec Jean-Stéphane Varré et Jean-Paul Delahaye

Collectivité

Responsablilités

- **Responsable du pôle régional ReNaBi Nord-Est**
  Pôle du réseau national des plates-formes de bio-informatique (ReNaBi) *depuis 2010*
  regroupe les plates-formes de Lille, Nancy, Reims et Strasbourg

- **Responsable de Bilille**
  Plate-forme de services en Bio-informatique de Lille *depuis 2010*
- **Co-responsable du DESS/Master professionnel Bio-informatique de Lille**
  Demande de création, définition du contenu, gestion des enseignants et des étudiants *2001–2006*
  effectué avec Hélène Touzet

- **Jurys de recrutement**
  - **membre du comité de sélection pour le LIFL**
    Recrutement des maîtres de conférences en informatique *2010–2013*
    Université Lille 1
  - **membre du comité Cordi-Postdoc**
    Recrutement des doctorants et postdoctorants en informatique *2010–2011*
    Centre Inria Lille-Nord Europe
  - **membre du comité de sélection pour le laboratoire I3S**
    Recrutement d'un maître de conférences en informatique, pour une chaire avec le CNRS *2010*
    Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis, Université de Nice
  - **membre du comité de sélection pour le laboratoire SADV**
    Recrutement d'un maître de conférences en biologie *2010*
    Laboratoire des Stress Abiotiques et Différenciation des Végétaux cultivés, Université Lille 1
  - **membre d'un jury de recrutement administratif Inria**
    Recrutement du responsable administratif et financier et d'un juriste *2007*
    Centre Inria Lille-Nord Europe
  - **membre expert d'un jury de recrutement technicien CNRS**
    Recrutement de deux techniciens d'exploitation informatique *2005*

- **Jurys de thèse**
  - **Sidahmed Benabderrahmane**: thèse en informatique de l'Université Nancy 1 *dec 2011*
    Titre: *Prise en compte des connaissances du domaine dans l'analyse transcriptomique : Similarité sémantique, classification fonctionnelle et profils flous. Application au cancer colorectal.*
    Directrice: Marie-Dominique Devignes , CR HdR CNRS
    Co-encadrante: Malika Smaïl-Tabonne, MCF en informatique
  - **Walla Hussein**: thèse en ingénierie des fonctions biologiques l'Université Lille 1 *sept 2011*
    Titre: *Study on the regulation and biosynthesis of fengycin and plipastatin produced by bacillus subtilis*
    Directeurs: Philippe Jacques, Pr en microbiologie et Iordan Nikov, Pr en génie des procédés
    Co-encadrante: Frédérique Gancel, MCF en biologie
  - **Arthur Tapi**: thèse en ingénierie des fonctions biologiques l'Université Lille 1 *janv 2010*
    Titre: *Stratégie moléculaire de mise en évidence de peptides actifs d'origine non ribosomiale chez Bacillus sp et Lactobacillus sp*
    Directeur: Philippe Jacques, Pr en microbiologie
    Co-encadrante: Marlène Chollet, MCF en microbiologie
  - **Olivia Jardin-Mathé**: thèse en Biologie-Santé de l'Université Lille 1 *jun 2008*

Titre: *Développement d'un logiciel universel d'imagerie par spectrométrie de masse et application au modèle sangsue et aux maladies neurodégénératives*

Directeur: Micèl Salzet, Pr en biologie

Co-encadrante: Isabelle Fournier, MCF en chimie

- **Céline Meslin**: thèse en informatique de l'Université de Rouen *janv 2007*

  Titre: *Utilisation de la table des suffixes pour la détection des répétitions*

  Directeur: Thierry Lecrocq, Pr en informatique

  Co-encadrant: Laurent Mouchard, MCF en informatique

- **Comités d'organisation**
  - **workshop NRPS 2013**: Bioinformatics tools for NRPS discovery *10–12 juil 2013*
  - **AMP 2012**: Third international symposium on antimicrobial peptides *13–15 juin 2012*
  - **CPM'09**: 20th Annual Symposium on Combinatorial Pattern Matching *22–24 juin 2009*
  - **JOBIM 2008**: La conférence francophone de bio-informatique *30 juin–2 juil 2008*
- **Relecture d'articles**
  - Journal of Biotechnology
  - JOBIM 2006 et 2008
  - Food Technology and Biotechnology
- **Divers**
  - **membre du comité de pilotage de l'IFB**
    Institut Français de Bio-informatique *depuis sept 2013*
  - **membre nommé au CNU 27**
    Conseil National des Universités en informatique *depuis mai 2013*
  - **membre élu au conseil de l'UFR IEEA**
    UFR Informatique, Électronique, Électrotechnique et Automatique *depuis mars 2011*
  - **membre du groupe de travail égalité Femme-Homme**
    Université Lille 1 *depuis sept 2009*

Publications

Articles dans des journaux internationaux avec comité de lecture

1. F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S.-K. Choi, J.-J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Düsterhöft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S.-Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Hénaut, H. Hilbert, S. Holsappel, S. Hosono, M.-F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, **Klaerr-Blanchard, M.**, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S.-M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauël, C. Médigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback,

D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S.-H. Park, V. Parro, T. M. Pohl, D. Portetelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Roche, M. Rose, Y. Sadaie, T. Sato, E. Scanlan, S. Schleich, R. Schroeter, F. Scoffone, J. Sekiguchi, A. Sekowska, S. J. Seror, P. Serror, B.-S. Shin, B. Soldo, A. Sorokin, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Tognoni, V. Tosato, S. Uchiyama, M. Vandenbol, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, E. Wedler, H. Wedler, T. Weitzenegger, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H.-F. Yoshikawa, E. Zumstein, H. Yoshikawa, and A. Danchin. The complete genome sequence of the gram-positive bacterium bacillus subtilis. *Nature*, 390(6657):249–256, November 1997.

2. **Klaerr-Blanchard, Maude**, Hélène Chiapello, and Eivind Coward. Detecting localized repeats in genomic sequences: a new strategy and its application to bacillus subtilis and arabidopsis thaliana sequences. *Computers & Chemistry*, 24(1):57–70, January 2000.

3. Ivan Laprevotte, **Pupin, Maude**, Eivind Coward, Gilles Didier, Christophe Terzian, Claudine Devauchelle, and Alain Hénaut. HIV-1 and HIV-2 LTR nucleotide sequences: Assessment of the alignment by n-block presentation, "Retroviral signatures" of overrepeated oligonucleotides, and a probable important role of scrambled stepwise Duplications/Deletions in molecular evolution. *Molecular Biology and Evolution*, 18(7):1231–1245, January 2001.

4. Ivo G. Boneca, Hilde de Reuse, Jean-Charles Epinat, **Pupin, Maude**, Agnès Labigne, and Ivan Moszer. A revised annotation and comparative analysis of helicobacter pylori genomes. *Nucleic Acids Research*, 31(6):1704–1714, March 2003.

5. Gilles Didier, Ivan Laprevotte, **Pupin, Maude**, and Alain Hénaut. Local decoding of sequences and alignment-free comparison. *Journal of Computational Biology*, 13(8):1465–1476, October 2006.

6. Gilles Didier, Laurent Debomy, **Pupin, Maude**, Ming Zhang, Alexander Grossmann, Claudine Devauchelle, and Ivan Laprevotte. Comparing sequences without using alignments: application to HIV/SIV subtyping. *BMC Bioinformatics*, 8(1):1, January 2007. PMID: 17199892.

7. Ségolène Caboche, **Pupin, Maude**, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE: a database of nonribosomal peptides. *Nucleic Acids Research*, 36(suppl 1):D326–D331, January 2008.

8. Ségolène Caboche, **Pupin, Maude**, Valérie Leclère, Phillipe Jacques, and Gregory Kucherov. Structural pattern matching of nonribosomal peptides. *BMC Structural Biology*, 9(1):15, March 2009. PMID: 19296847.

9. Ségolène Caboche, Valérie Leclère, **Pupin, Maude**, Gregory Kucherov, and Philippe Jacques. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *Journal of Bacteriology*, 192(19):5143–5150, January 2010.

10. Ammar Abdo, Ségolène Caboche, Valérie Leclère, Philippe Jacques, and **Pupin, Maude**. A new fingerprint to predict nonribosomal peptides activity. *Journal of Computer-Aided Molecular Design*, 26(10):1187–1194, October 2012.

11. Max Béchet, Thibault Caradec, Walaa Hussein, Ahmed Abderrahmani, Marlène Chollet, Valérie Leclère, Thomas Dubois, Didier Lereclus, **Pupin, Maude**, and Philippe Jacques. Structure,

biosynthesis, and properties of kurstakins, nonribosomal lipopeptides from bacillus spp. *Applied Microbiology and Biotechnology*, 95(3):593–600, August 2012.

Chapitre de livre

12. Eduardo Rocha, Ivan Moszer, **Klaerr-Blanchard, Maude**, Agnieszka Sekowska, Alain Viari, and Antoine Danchin. In silico genome analysis. In Wolfgang Schumann, Stanislav Dusko Ehrlich, and Naotake Ogasawara, editors, *Functional analysis of bacterial genes: a practical manual*. Chichester, Royaume-Uni, 2001.

Article de vulgarisation

13. **Pupin, Maude**. Des peptides à explorer. *Interstices*, March 2010.

D326–D331 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 2 October 2007*
doi:10.1093/nar/gkm792

NORINE: a database of nonribosomal peptides

Ségolène Caboche[1,2,*], Maude Pupin[1], Valérie Leclère[2], Arnaud Fontaine[1], Philippe Jacques[2] and Gregory Kucherov[1]

[1]Computer Science Laboratory of Lille (UMR USTL/CNRS 8022) and INRIA, and [2]ProBioGEM (UPRES EA 1026), University of Sciences and Technologies of Lille, 59655 Villeneuve d'Ascq, France

Received August 14, 2007; Revised and Accepted September 17, 2007

ABSTRACT

Norine is the first database entirely dedicated to nonribosomal peptides (NRPs). In bacteria and fungi, in addition to the traditional ribosomal proteic biosynthesis, an alternative ribosome-independent pathway called NRP synthesis allows peptide production. It is performed by huge protein complexes called nonribosomal peptide synthetases (NRPSs). The molecules synthesized by NRPS contain a high proportion of nonproteogenic amino acids. The primary structure of these peptides is not always linear but often more complex and may contain cycles and branchings. In recent years, NRPs attracted a lot of attention because of their biological activities and pharmacological properties (antibiotic, immunosuppressor, antitumor, etc.). However, few computational resources and tools dedicated to those peptides have been available so far. Norine is focused on NRPs and contains more than 700 entries. The database is freely accessible at http://bioinfo.lifl.fr/norine/. It provides a complete computational tool for systematic study of NRPs in numerous species, and as such, should permit to obtain a better knowledge of these metabolic products and underlying biological mechanisms, and ultimately to contribute to the redesigning of natural products in order to obtain new bioactive compounds for drug discovery.

INTRODUCTION

Nonribosomal peptides (NRPs) show various particularities and a large structural diversity. They are short (2 to about 50 amino acids) and contain nonproteogenic amino acids. Indeed, amino acids other than the classical 20 ones found in proteins can be incorporated into those peptides. In addition, their particular way of synthesis can lead to chemical modifications of incorporated residues such as epimerization or methylation. Products from other biosynthesis pathways such as lipids or carbohydrates can also be introduced. The NRPs thus show a great diversity of their monomer composition. The primary structure of the NRPs is not always linear but often more complex: they can be linear like classical ribosomal peptides, but also branched, cyclic (partially or totally) or even poly-cyclic. Structural and compositional variety of these peptides allows them to have a broad range of important biological and pharmacological activities. For example, the ACV-tripeptide, the famous penicillin and cephalosporin precursor, is synthesized by this way. NRPs show immunosuppressive (cyclosporine), antitumor (bleomycin) or antibiotic (vancomycine) activities. Other examples include siderophores (pyoverdin), toxins (HC-toxin) or surfactants (surfactin).

NRPs are synthesized by large enzymatic complexes called nonribosomal peptide synthetases (NRPSs). This mechanism has been described for the first time in 1971, during the study of two antibiotics: gramicidin S and tyrocidin (1). A NRPS represents at the same time a template and biosynthetic machinery (2). Genes coding for NRPS are organized in operons or in clusters. NRPSs are modularly organized. Each module is responsible for the incorporation of a specific monomer. Modules are subdivided into domains, each domain catalyzing a specific reaction in the incorporation of a monomer. Four main domains are necessary for a complete synthesis. The first one, the adenylation domain, selects and activates the monomer transforming it into adenylate form. The thiolation or peptidyl carrier protein domain covalently binds the activated monomer to the synthetase. The condensation domain catalyses the peptide bond formation between the residues linked onto two adjacent modules. Finally, the thioesterase domain, only present in the final module, releases the peptide from the synthetase. The product can either be released as a linear compound or get transformed into a cyclic peptide through an intramolecular reaction. In NRPS of iterative type, the thioesterase domain can allow the enzyme to iterate the collinear biosynthesis several times.

Secondary domains that allow residue modifications are present in many NRPSs. For example, an epimerization domain leading to obtaining the D isomer of an amino

*To whom correspondence should be addressed. Tel: +33 3 59 57 79 17; Fax: +33 3 28 77 85 37; Email: caboche@lifl.fr

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

acid can be encountered. Methylation, oxidation or cyclization domains can also be found in some NRPSs.

The NRPS mechanism can produce different variants of a peptide that have the same structure but have different monomers at certain positions. It has been shown that variations of fermentation broths can lead to production of more than 30 cyclosporine variants (3). In other cases, variant synthesis is due to a diversity of genomic sequences. For example, it has been shown that the DNA sequences of NRPSs that produce the bacillomycin D (4) and bacillomycin L (5) variants are different. A better knowledge of the biosynthetic mechanism opens a way to redesign natural products and to obtain new bioactive peptides for drug discovery (6).

When the NRPS mechanism has been discovered, it seemed to be of little significance. It appeared to be more and more important in the literature due to the discovery of numerous genes coding for NRPSs and important biological activities of their products. Currently, there are still few research groups that develop methods or computational tools for manipulating NRPs. Among existing resources, the NRPS-PKS database (7) is focused on the synthetases and contains only 20 or so peptides. Other resources like PubChem (8) contain some NRPS peptides as well as other small biological molecules but only few variants are presented. The Peptaibol Database (9) is focused only on a specific family of non-ribosomal products. A comprehensive resource compiling all known NRPs has been missing so far.

To fill this lack, we developed the Norine database containing a large amount of NRPS peptides with all types of structure and activity. The name Norine stands for *NOnR*ibosomal peptides, with *ine* as a typical ending of NRP names. The database currently contains more than 700 peptides and this number is still growing. Norine is freely available at http://bioinfo.lifl.fr/norine/.

CONTENTS

Several reviews describing the NRPS mechanism have been published (2,10–13). These publications contain some examples of peptides produced by this way but no resource including an exhaustive up-to-date list of NRPs has been available so far. We explored relevant papers published since 1970s to compile an exhaustive list of known NRPs. The Norine database currently features more than 700 peptides extracted from about 350 publications. All data of Norine comes from the scientific literature and has been manually curated (predicted data are not included), which insures the reliability of the annotations. Various types of annotations are stored in the database.

Figure 1 provides a representative screenshot showing the description of a peptide. The web page is organized into several parts. The first part, entitled 'peptide' (Figure 1a), presents general annotations of the peptide such as the peptide name and its synonyms. This is followed by fields presenting known biological activities of the peptide molecule, its molecular formula with the associated molecular weight and a possible comment presenting additional information on the molecule. Finally, the 'entry information' field contains the peptide status (curated or putative nonribosomal product), and creation and last modification dates that allow the user to follow the history of the entry.

The next 'structure' section (Figure 1b) contains the most original data stored in Norine: structural features of the peptide. We chose to represent the peptide structure at the monomeric level rather than use a classical chemical atomic representation. This choice is justified by the fact that NRPs are synthesized by successive addition of monomers and not by atomic reactions, and therefore representing a peptide by its monomeric structure is an adequate way of specification. The first information found in this part is the peptide structural type. In Norine, the NRPs are classified in several groups according to their structural type: linear, branched, cyclic, partial cyclic, double cyclic and other. The group 'other' contains peptides that show a complex structure with several overlapping cycles and branches. The number of monomers composing the peptide is also given. The peptide structure is then presented using two representations. The first one is the 'linear representation' (Figure 2b). We developed this representation as a quick and easy way to represent a (possibly nonlinear) monomeric structure of the peptide by a linear string. In this representation, monomers are encoded according to a set of simple rules. The 20 proteogenic amino acids are encoded by the classical three-letter code (for example, Ala for alanine). When a functional group (like methyl) is added, its symbol is also added of the three-letter code (for example, NMe-Ala for *N*-methyl-alanine). By default, the amino acid is in L-form, the D-symbol is added when it is in D-form. To represent the structure, chained monomers are separated by an underscore sign. Cycles and branchings are represented, respectively, by brackets and braces. Note that this representation does not specify whether the bound involves only the backbone atoms of the monomer or side chain atoms. An example of linear representation of a double cyclic structure is given in Figure 2b. The linear representation provides a fast way of specifying a large class of structures using a set of simple rules. This class contains structures with no overlapping cycles covering a broad range of practical cases. However, structures with overlapping cycles cannot be specified unambiguously by a linear representation.

Another representation is called the 'graph representation' (Figure 2c). In this case, a peptide is represented as an undirected graph with nodes labeled by monomers and edges corresponding to the bonds between monomers. However, a standard computer representation of graphs (such as adjacency lists) does not allow the user to quickly figure out its 2D image. We thus developed a Java applet that draws the peptide structure in two dimensions. This applet is based on the Fruchterman–Reingold graph layout algorithm (14) that avoids edge crossing and keeps uniform edge lengths. The users can save the peptide structure in an image or text format, redraw the structure or still switch the node representation.

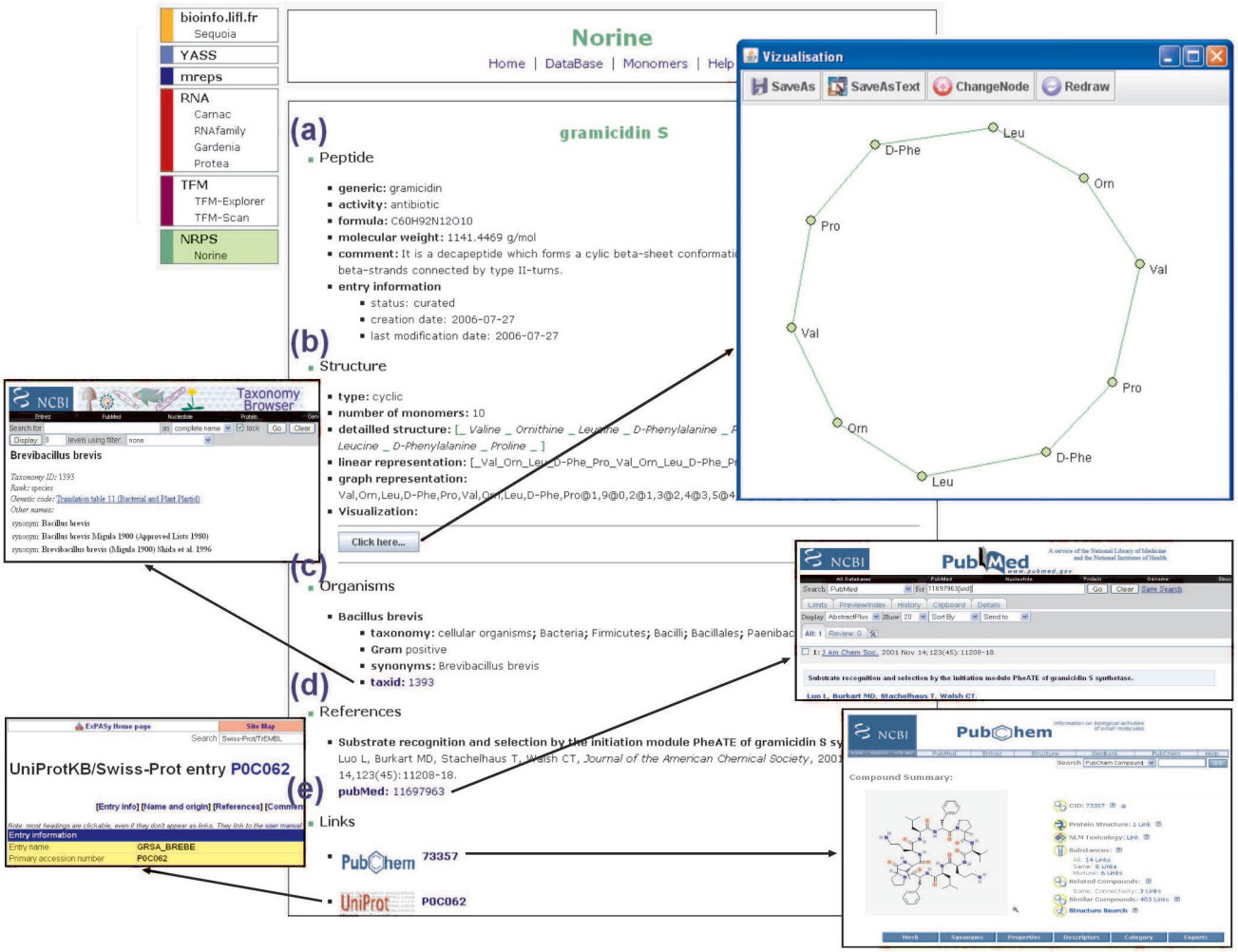


Figure 1. Description page of a peptide. The page is obtained when a peptide is selected after a search in the database. (**a**) Part 'Peptide' contains general information on the peptide. (**b**) Part 'Structure' compiles all the information on the peptide structure. A Java applet allows the peptide visualization in two-dimensions. (**c**) Part 'Organisms' concerns the organisms known to produce the peptide. Via a link, the user can access to NCBI taxonomy. (**d**) Part 'References' contains bibliographical references associated to the peptide. The user can access the corresponding entry of NCBI PubMed through a link. (**e**) Part 'Links' contains links to NCBI PubChem and UniProt.

The annotations in the 'organisms' section (Figure 1c) concern the producing organisms. In general, the same peptide can be produced by several organisms and one organism can produce several NRPS peptides. Different data about these organisms are included: the organism name and its synonyms, taxonomic information, the Gram type for bacteria, a link to the NCBI taxonomy via the NCBI taxonomy identifier (taxid) (8).

The 'references' section (Figure 1d) includes bibliographical references associated to the peptide. For each reference, its title, authors, year of publication, journal name and NCBI PubMed link (8) are given.

The last part, the 'links' section (Figure 1e), includes links to other databases. A link to UniProt (15) allows the user to obtain information on synthetase such as its amino acid sequence. A link to NCBI PubChem compound (8) gives a direct access to the peptide chemical information.

WEB INTERFACE

The data are stored in a relational database using PostgreSQL database system. It can be queried via a friendly web interface in order to select NRPs corresponding to various search criteria. The web interface has been implemented with Java Server Page (JSP) technology using the Apache tomcat server (http://tomcat.apache.org/).

Two types of search are available: a general search and a structure search.

General search

General search can be done in different ways. The first one, called 'basic search', allows the user to search peptides according to several fields. The user can search by general or specific peptide names. A specific name represents the exact name of an NRP (e.g. cyclosporine A). If the peptide

is present in the database, only one result will be obtained in this case. A general name represents a generic name for NRP (e.g. cyclosporine). This kind of search can result in several peptide variants belonging to the same group. The two names can be combined by any one of the Boolean operators AND, OR, AND NOT. Other search criteria include the biological activity and the structural type. When several fields are selected (such as name and activity) the results must match each of them. Peptides can also be searched by their molecular weight. To do this, the user has to specify an interval of possible molecular weights. The whole list of peptides is accessible by clicking a button.

'Reference search' allows a search for peptides by their bibliographical references. One can search by author, title, year of publication, journal or by pmid (NCBI PubMed identifier). Once again, it is possible to combine two reference fields with a Boolean operator.

In 'organism search', the search is done by organisms known to produce the corresponding peptides. The user can specify a desired taxonomic level. For example, querying 'bacteria' will output all the peptides produced by bacteria, and querying '*Bacillus subtilis*' will output only those produced by this organism. Boolean operators can be used to combine two search fields. The complete list of organisms included in the database can be obtained by clicking a button.

Structure search

As structural variability is an important characteristic of NRPs, particular attention was given to 'structure search' that allows the user to search for peptides that verify certain structural properties. Two types of structure search have been implemented.

The first one is 'composition-based search', which looks up for peptides according to monomer composition features. As a simplest example, one can obtain all the peptides which contain a specific monomer, by specifying the corresponding monomer identifier. For example, querying Thr (for threonine) yields the list of all the peptides that contain at least one threonine. One can also obtain all the peptides containing a given monomer and its derivates. For example, querying Thr in this case gives all the peptides containing threonine amino acid, but also those that contain allo-threonine, D-threonine or choro-threonine. The user can search for peptides containing a given number of monomers, or less than or greater than a given number. Finally, the user can search for peptides containing a given list of monomers, either all of them or with a given maximum number of 'errors'. This search does not take into account the peptide structure but only its monomeric composition. For example, querying <Ala,Pro,Val,Gly,Pro> with two possible errors returns all the peptides containing all the five monomers but also the peptides that contain only four (one error) or three (two errors) of the five monomers given in entry. Note that the query list can contain the same monomer several times.

The user can also search for a peptide by specifying its structure in the 'structure-based search' (Figure 3). The peptide structure can be specified using either a

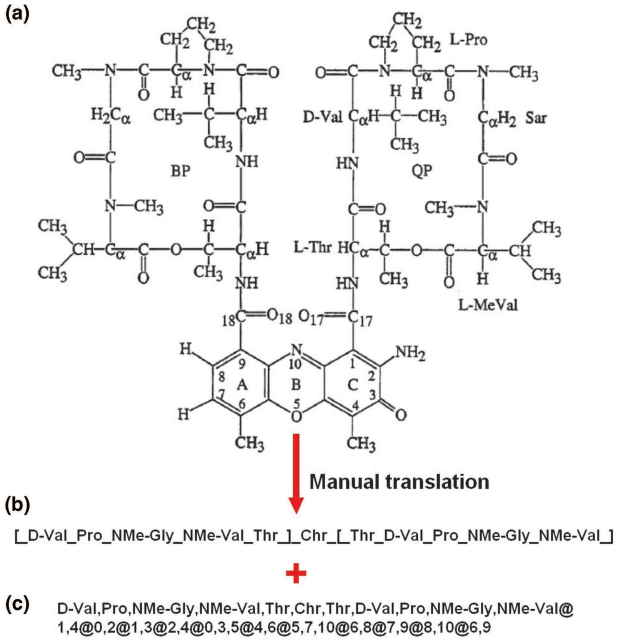


Figure 2. Representation of NRP structures in Norine. (**a**) Chemical representation of actinomycin D (16). (**b**) Linear representation of actinomycin D. Monomers are encoded and separated by an underscore. Cycles are represented by brackets. (**c**) Graph representation of actinomycin D. It starts with a list of monomers each of which is associated with its rank in the list (numbered from zero) and corresponds to a node of the graph. Then, the adjacency list represents the edges incident to each node.

linear or a graph representation. A graph representation can be created using a dedicated structure editor integrated to Norine. The Norine editor is a Java applet that allows one to build quickly and easily a graph representation of a peptide, i.e. to specify monomers and links between them. Complex monomeric structures can be easily drawn with a friendly graphical interface: first, monomers are selected and corresponding graph nodes are created, which are then connected by drawing edges between the nodes. It is also possible to delete some monomers or the whole structure. The user can also open a text file generated by the visualization applet in order to modify the created graph by hand. Once the structure is completed, clicking the 'go' button returns the peptide structure to the appropriate field of the Norine search page and the search for it in the database can be launched right away. Both with the linear and graph representations, the user can specify either the entire peptide structure to look for, or a structural pattern that the peptide must contain. The containment is defined as the usual subgraph relation: a pattern occurs in a peptide if each node of the pattern labeled by a monomer can be associated to a node of the peptide labeled by the same monomer so that the linked (unlinked) nodes of the pattern are linked (respectively unlinked) in the peptide.

A structural pattern is specified in the same way as a regular peptide, i.e. using a linear or graph representation, where the latter is specified with the Norine structure editor. However, these representations are enriched by the possibility for a structural pattern to contain nodes

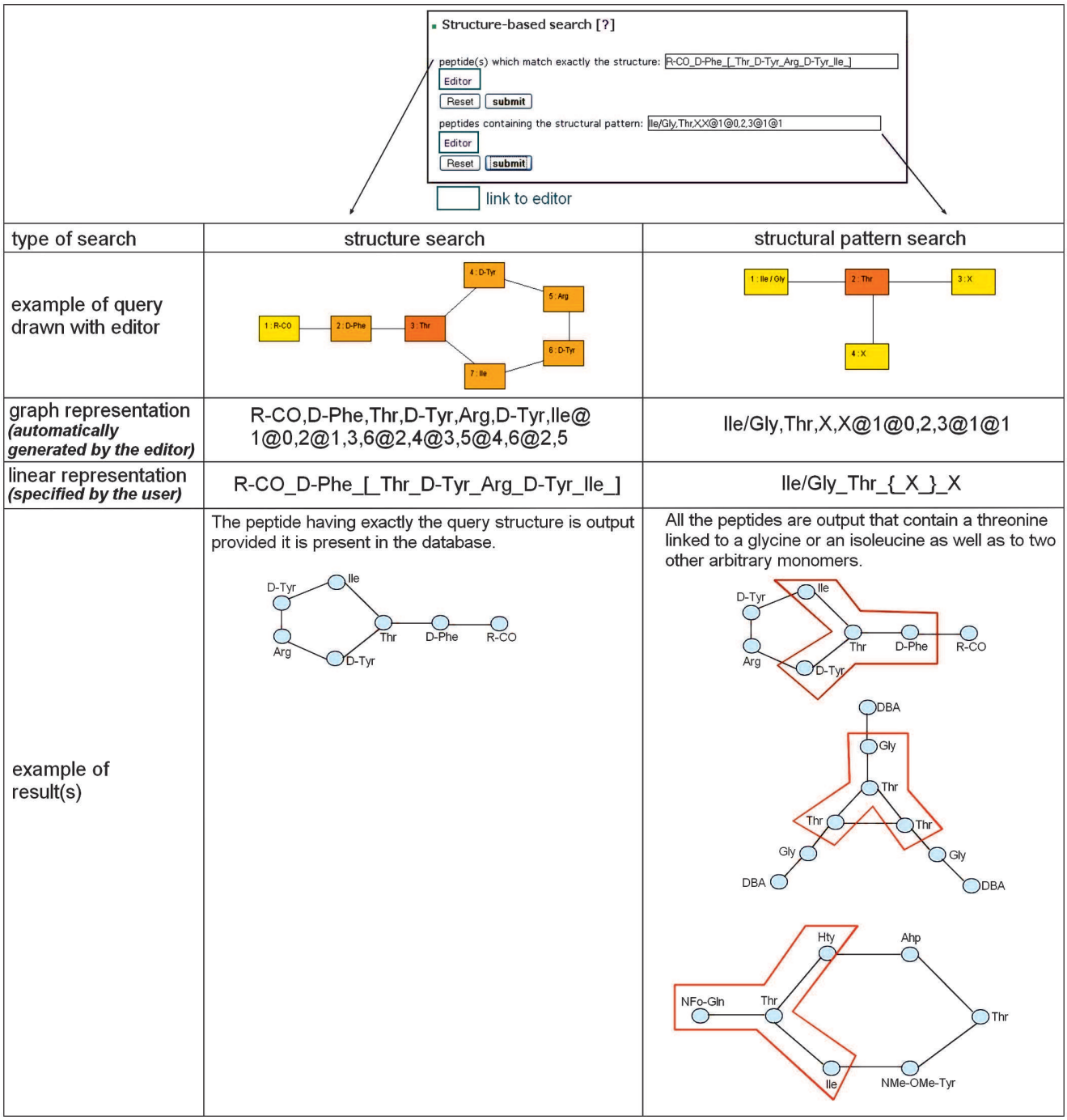


Figure 3. Example of structure-based search. Two search features are provided. The structure search looks up for a peptide having exactly the query structure. The structural pattern search looks up for the peptides containing the query pattern as a subgraph. The query pattern can contain joker or alternatively-labeled nodes (X and/). In both types of search, the query can be specified using either linear or graph representation. A link to the dedicated peptide structure editor (in green) allows the user to automatically obtain the graph representation. Alternatively, the user can specify the query through the linear representation. In the last row, examples of resulting peptides are given.

labeled by several alternative monomers. These can be specified by a list of monomers separated by a special '/' symbol. A special 'X' symbol stands for any monomer. The occurrence of patterns having alternatively labeled nodes is defined in the natural way.

CONCLUSION AND PERSPECTIVES

Nonribosomal synthesis is an original biosynthesis pathway that leads to a great diversity of products. A huge structural diversity of the NRPs allows them to have a broad range of important biological activities.

This work resulted in compiling the first database entirely dedicated to the peptides produced by NRPS. Norine already contains more than 700 peptides and will continue to be completed and regularly updated.

Different features of Norine, and in particular different types of queries the user can make to the database, lead to different possibilities of its usage. In general, the user can easily extract different types of information about known NRPs. For example, Norine can be used to identify a peptide predicted by some other means from an NRPS amino acid sequence. One can then determine if the predicted peptide has already been identified by using the structure search features of Norine. Various other types of information can be extracted from the Norine database.

We expect that structure comparison tools can help to better understand the structure/function relationships of NRPs. More generally, the possibilities of study of different properties of peptides offered by Norine can bring new insights on their impact to their biological activity. We also believe that possibilities provided by Norine, in association with other NRPS enzyme dedicated tools, can lead to facilitate the redesign of natural products in order to develop new bioactive compounds for drug discovery. Indeed, combinatorial biosynthesis, the process of genetic manipulations of natural product biosynthetic machinery, depends, in particular, on the detailed knowledge of the involved metabolic processes. The product data contained in Norine should permit to better identify the specificity of different domains and to facilitate the search for domains incorporating a given residue. Note that few NRPS sequences are currently available in comparison with the number of peptide products.

In near future, we plan to enrich Norine with new computational tools such as the search for similarities between a new or unknown peptide and those already present in the database.

ACKNOWLEDGMENTS

This work was supported by PPF bioinformatique of Lille. S.C. was supported by an INRIA/Région Nord-Pas-de-Calais fellowship. ProBioGEM lab is supported by the *region Nord-Pas-de-Calais*, the *Ministere de l'Enseignement et de la Recherche (ANR)* and the European Funds for the Regional Development. Funding to pay the Open Access publication charges for this article was provided by INRIA.

Conflict of interest statement. None declared.

REFERENCES

1. Lipmann,F., Gevers,W., Kleinkauf,H. and Roskoski,R. (1971) Polypeptide synthesis on protein templates: the enzymatic synthesis of gramicidin S and tyrocidine. *Adv. Enzymol. Relat. Areas Mol. Biol.*, **35**, 1–34.
2. Sieber,S.A. and Marahiel,M.A. (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chem. Rev.*, **105**, 715–738.
3. von Döhren,H. (2004) Biochemistry and general genetics of nonribosomal peptide synthetases in fungi. *Adv. Biochem. Eng. Biotechnol.*, **88**, 217–264.
4. Moyne,A.L., Cleveland,T.E. and Tuzun,S. (2004) Molecular characterization and analysis of the operon encoding the antifungal lipopeptide bacillomycin D. *FEMS Microbiol. Lett.*, **234**, 43–49.
5. Hofemeister,J., Conrad,B., Adler,B., Hofemeister,B., Feesche,J., Kucheryava,N., Steinborn,G., Franke,P., Grammel,N. *et al.* (2004) Genetic analysis of the biosynthesis of non-ribosomal peptide and polyketide-like antibiotics, iron uptake and biofilm formation by Bacillus subtilis. *A1/3. Mol. Genet. Genomics*, **272**, 363–378.
6. Van Lanen,S.G. and Shen,B. (2006) Progress in combinatorial biosynthesis for drug discovery. *Drug Discov. Today*, **3**, 285–292.
7. Ansari,M.Z., Yadav,G., Gokhale,R.S. and Mohanty,D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**(Web Server issue), W405–W413.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**(Database issue), D5–D12.
9. Whitmore,L. and Wallace,B.A. (2004) The peptaibol database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.*, **32**(Database issue), D593–D594.
10. Grünewald,J. and Marahiel,M.A. (2006) Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiol. Mol. Biol. Rev.*, **70**, 121–146.
11. Challis,G.L. and Naismith,J.H. (2004) Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Struct. Biol.*, **14**, 748–756.
12. Keller,U. and Schauwecker,F. (2003) Combinatorial biosynthesis of non-ribosomal peptides. *Comb. Chem. High Throughput Screen.*, **6**, 527–540.
13. Schwarzer,D., Finking,R. and Marahiel,M.A. (2003) Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.*, **20**, 275–287.
14. Fruchterman,T.M.J. and Reingold,E.M. (1991) Graph drawing by force-directed placement. *Software Pract Exper*, **21**, 1129–1164.
15. niProt Consortium1. (2007) The universal protein resource (UniProt). *Nucleic Acids Res.*, **35**(Database issue), D193–D197.
16. Chen,H., Liu,X. and Patel,D.J. (1996) DNA bending and unwinding associated with actinomycin D antibiotics bound to partially overlapping sites on DNA. *J. Mol. Biol.*, **258**, 457–479.

BMC Structural Biology

Methodology article

Open Access

Structural pattern matching of nonribosomal peptides

Ségolène Caboche*[1,2], Maude Pupin[1], Valérie Leclère[2], Phillipe Jacques[2] and Gregory Kucherov[1]

Address: [1]Computer Science Laboratory of Lille, UMR USTL/CNRS 8022, INRIA, F59655, Villeneuve d'Ascq, France and [2]ProBioGEM (UPRES EA 1026), University of Sciences and Technologies of Lille, F59655, Villeneuve d'Ascq, France

Email: Ségolène Caboche* - segolene.caboche@lifl.fr; Maude Pupin - maude.pupin@lifl.fr; Valérie Leclère - valerie.leclere@univ-lille1.fr; Phillipe Jacques - Philippe.Jacques@polytech-lille.fr; Gregory Kucherov - gregory.kucherov@lifl.fr

* Corresponding author

Published: 18 March 2009

BMC Structural Biology 2009, **9**:15 doi:10.1186/1472-6807-9-15

Received: 24 July 2008
Accepted: 18 March 2009

This article is available from: http://www.biomedcentral.com/1472-6807/9/15

© 2009 Caboche et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Nonribosomal peptides (NRPs), bioactive secondary metabolites produced by many microorganisms, show a broad range of important biological activities (e.g. antibiotics, immunosuppressants, antitumor agents). NRPs are mainly composed of amino acids but their primary structure is not always linear and can contain cycles or branchings. Furthermore, there are several hundred different monomers that can be incorporated into NRPs. The NORINE database, the first resource entirely dedicated to NRPs, currently stores more than 700 NRPs annotated with their monomeric peptide structure encoded by undirected labeled graphs. This opens a way to a systematic analysis of structural patterns occurring in NRPs. Such studies can investigate the functional role of some monomeric chains, or analyse NRPs that have been computationally predicted from the synthetase protein sequence. A basic operation in such analyses is the search for a given structural pattern in the database.

Results: We developed an efficient method that allows for a quick search for a structural pattern in the NORINE database. The method identifies all peptides containing a pattern substructure of a given size. This amounts to solving a variant of the maximum common subgraph problem on pattern and peptide graphs, which is done by computing cliques in an appropriate compatibility graph.

Conclusion: The method has been incorporated into the NORINE database, available at http://bioinfo.lifl.fr/norine. Less than one second is needed to search for a pattern in the entire database.

Background

Nonribosomal Peptides (NRPs) are bioactive compounds having various important biological functions (e.g. as antibiotics, siderophores, antitumor agents, immunosuppressants). NRPs are synthesized by large multi-enzymatic complexes called Nonribosomal Peptide Synthetases (NRPSs) that are modularly organized [1]. Each module is responsible for the incorporation of a specific monomer and is itself subdivided into domains catalysing specific enzymatic reactions.

Until about fifteen years ago, the number of known NRPs remained relatively low. However, many new molecules have been reported in the literature during the last years, associated with different biological activities and having a broad range of potential applications. This triggered a

considerable interest among the research community in the nonribosomal synthesis pathway.

Among potential applications of such studies, redesigning natural products by genetic engineering of NRPSs opens an interesting new way in drug discovery [2]. Indeed, modifying the nucleotide sequence of a natural NRPS or combining modules of different NRPSs could potentially yield a more efficient compound or a product with a new biological activity. However, generating a new peptide with a specific function from a modified NRPS nucleic sequence requires a deep understanding of both the assembly line and the resulting products.

NRPS enzymes have been well studied for several years. Stachelhaus *et al*. [3] discovered a specificity-conferring code of adenylation domains. With this discovery, several software programs have been developed [4-6] to predict a produced peptide from the NRPS protein sequence. With the increasing number of sequenced genomes, the number of hypothetical NRPSs increases too. Therefore, this raises the problem of verifying whether a peptide predicted to be produced by a NRPS is already known or even corresponds to a part of a known peptide.

NRP molecules show several important particularities. The first one is related to the incorporation of non-proteinogenic amino acids. Indeed, in addition to the twenty standard amino acids found in proteins, several hundreds of other residues can be encountered in final NRPS products. Incorporated residues can further undergo chemical modifications such as epimerisation or methylation. Products of other biosynthesis pathways, like lipids or carbohydrates, can also be introduced. Because of this composition diversity of NRPs, we will use the term 'monomer' rather than 'amino acid' for NRP structural units. Another interesting property of NRPs is their structure. Unlike regular proteins, the primary structure of NRPs is not always linear but can also be cyclic (partially or totally), branched or even poly-cyclic. A computational treatment of these molecules appears therefore to be very different from standard proteins and requires a development of specific computational methods and resources.

There exist, however, very few computational resources specifically devoted to NRPs and, until recently, there was no one providing a complete inventory of those. To fill this lack, we have developed the NORINE database [7] which is the first resource entirely dedicated to NRPs. It contains various annotations of each peptide such as the producing organism, bibliographic references, activities and, most importantly, its monomeric structure. The choice of representing NRP molecules by their monomeric rather than atomic structure reflects the way they are synthesized by successive addition of monomers. This structure is encoded by an undirected labeled graph representing its (possibly nonlinear) structure. Using undirected (rather than directed) edges is justified by the existence of nonpeptide bonds, appearing e.g. in cyclic or branched peptides, for which the orientation can not be naturally defined. Furthermore, using directed edges could be restrictive for the analysis of peptide families: for example, lipopeptides containing an asparagine-serine dipeptide include the iturin family (produced by different *Bacillus* species). Tsuge *et al*. [8] proposed a model in which iturin or mycosubtilin swapped nucleotide sequences encoding adenylation domains after a common ancestor became established. In this case, looking for a directed asparagine-serine dipeptide would miss mycosubtilin that has a serine-asparagine dipeptide.

Similar to the search for sequence patterns in genomic and protein databases, NORINE raises the need to efficiently search for *structural* patterns. In the simplest case, one needs to identify if a given peptide is already present in the database. An even more important motivation is provided by the close relation between the structure and the function of the peptide. For example, Minowa *et al*. [9] identified motifs that are significantly related to some biological activities. Therefore, a search for a structural pattern can help to assign a biological function to a peptide under study.

In some analyses, one needs to identify a part of the pattern, rather than the whole pattern, occurring in a given peptide. For example, the order of monomers in the resulting peptide can be changed with respect to the order of modules in the synthetase (so-called nonlinear biosynthesis [10]). For instance, in the biosynthesis of syringomycin [11], the *Syr*B1 gene responsible for the incorporation of the threonine monomer is located upstream of the *Syr*E gene in the genome, whereas threonine is the final monomer of the peptide. Therefore, a search for the entire pattern predicted from the synthetase does not produce an output, while a search for a common substructure allows one to identify the peptide.

In this paper, we present an efficient method to identify a substructure of a given structural pattern that occurs in a given NRP, where both the pattern and the peptide are represented by undirected labeled graphs. From the computational viewpoint, this can be expressed as a variant of the Maximum Common Subgraph (MCS) problem, which is NP-complete [12] i.e. is very unlikely to be solvable by an algorithm with a running time polynomially bounded on the graph size. Another related NP-complete problem, called Graph Motif problem [13], is to look for a connected subgraph with the given (multi-)set of labels. Despite of the formal NP-completeness of the underlying computation, our method works very efficiently on NRP

graphs, taking advantage of their relatively small size and specific structural properties.

Our method is based on the commonly used construction of a Compatibility Graph (CG), also called association or product graph, in which the largest clique represents a solution to the MCS problem (see [14] for a review). Here we adapt the method of CG to the structural search for nonribosomal peptides. We propose several ways to reduce the size of the CG, both in terms of number of nodes and edges. Note that the size of the CG is a crucial factor for the efficiency of the whole method, as the clique search in the CG is the computationally most demanding step. We follow the idea of filtration by trying to detect, as early as possible, pairs of nodes that cannot be mapped one to the other by a graph morphism. This considerably reduces the size of the CG and leads to an efficient practical structural search for nonribosomal peptides. The presented algorithm has been implemented in NORINE. Here we present some experimental results showing the efficiency of the method. We also provide some examples of using structural search for nonribosomal peptides in biological studies.

Results and discussion

Theory and algorithms

Graph representation of NRP structure

We encode the monomeric structure of NRPs by an undirected labeled graph. A *peptide graph* is a graph $G = (V, E, M, f)$, where V is a set of nodes, $E \subseteq V \times V$ is a set of undirected edges i.e. pairs (u, v) with u, v in V, and $f: V \rightarrow M$ a labeling mapping of nodes. Here nodes represent monomers, edges correspond to chemical bonds between monomers and labels are monomer names. Monomer names are encoded by a set of simple rules inspired by the IUPAC nomenclature [15]: monomers are denoted by the classical three-letter code, possibly preceded by a symbol of a chemical modification. For example, *NMe-Ala* represents the N-methyl-alanine monomer. Each node in a graph has a unique number in order to distinguish two nodes with the same label. Figure 1(a) shows examples of peptide graphs.

A structural pattern is also represented by a graph. Let $P = (V_P, E_P, L, f)$ be a pattern graph, where V_P is a set of nodes, $E_P \subseteq V_P \times V_P$ a set of undirected edges and $f: V_P \rightarrow L$ is the labeling of nodes. The main difference between graphs G and P is in the set of possible labels: $M \subset L$ but L contains some additional labels. One of them is the "joker label", denoted 'X', that stands for any monomer. L also includes alternative labels denoted by lists of several monomers separated by the '/' symbol. The intended meaning is that any monomer of the list can occur at the corresponding position. Finally, L also includes labels formed by the '*' symbol followed by a monomer. This means that any derivative of the monomer can be found at this position. Figure 1(b) shows some examples of structural patterns. For example, in pattern P4, **Orn* means that at this position, ornithine (Orn) or any of its derivatives, such OH-Orn or Fo-OH-Orn, can be found.

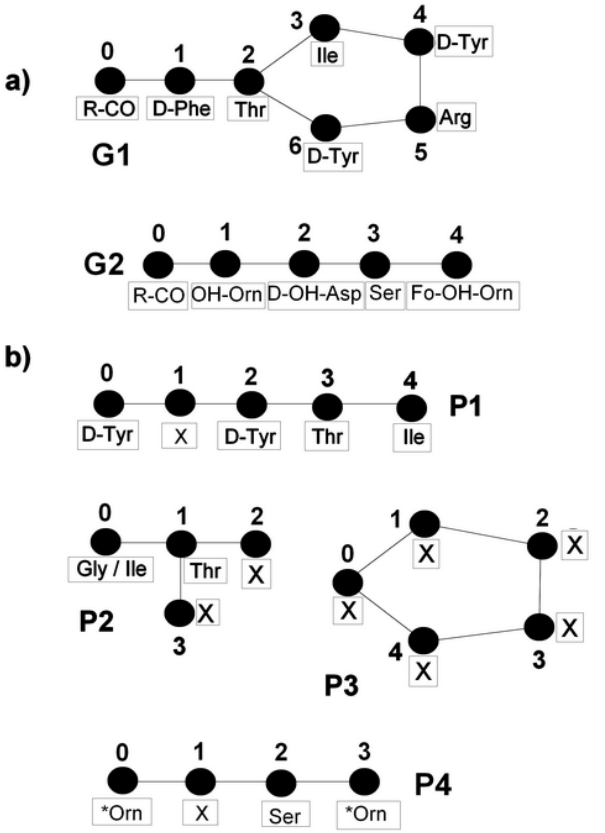


Figure 1
Examples of peptide and pattern graphs. This figure shows examples of (a) peptide graphs and (b) pattern graphs. Nodes and edges represent monomers and chemical bonds respectively. Labels are the monomer names.

Computing a maximal common substructure using the compatibility graph

The construction of the compatibility graph (CG) is often used in chemoinformatics to establish a structure mapping between two molecule graphs [14]. The CG encodes potential mappings between the two graphs. Then, a search for the largest clique in the CG allows one to obtain the maximum common subgraph. First, we describe the classical CG construction.

Compatibility graph

The classical definition of the CG of two graphs P and G is as follows:

• the set of nodes of CG is the cartesian product $V_P \times V$, i.e. a node U (u, u') of CG corresponds to the association of a pattern node u and a peptide node u'; in the case of (unambigously) labeled nodes, only nodes with the same label get associated to form a node of the CG,

• nodes U (u, u') and V (v, v') are adjacent in the CG if and only if $u \neq v$ and $u' \neq v'$ and one of the following conditions holds:

- u is adjacent to v in P and u' is adjacent to v' in G (1)

- u is not adjacent to v in P and u' is not adjacent to v' in G (2)

For our purposes, we modify the classical CG definition to only require that associated nodes have compatible labels. If $f(u) \in M$ (i.e. the label of u is not 'X' nor a "*-label" monomer), then any peptide node u' with $f(u') = f(u)$ gets associated with u. If $f(u)$ = 'X', then u gets associated with any node u' of G. Finally, if $f(u)$ is a "*-label", then u naturally gets associated with any node u' labeled by a derivative of the corresponding monomer.

Figure 2 shows a simple example of CG of a pattern and a peptide graph. Edges between nodes 'a' and 'b' and nodes 'b' and 'c' correspond to condition (1). The edge between nodes 'a' and 'c' corresponds to condition (2).

Clique computation

The CG represents all potential mappings between graphs P and G. Recall that a clique in an undirected graph is a subset of nodes such that every two nodes of this subset are connected by an edge. Each clique in the CG corresponds to a common substructure of graphs P and G, whose size (number of nodes) is equal to that of the corresponding clique. Consequently, searching for a clique of a given size k (k-clique) is equivalent to searching for a common subgraph of size k. In Figure 2, nodes a, b and c form a 3-clique, which corresponds to the occurrence of the whole pattern in the peptide. The general clique detection problem i.e. finding whether there is a k-clique in a graph is NP-complete [12].

Refining CG building rules

Our goal is to detect efficiently and exactly whether a part (connected subgraph) of a size k of a pattern graph P is a substructure of a peptide graph G. We assume that parameter k is specified by the user. If k is equal to the size of the pattern graph, the problem amounts to checking if P is a substructure of G. In other words, the searched pattern P occurs in the tested peptide G. The notion of "substructure" needs to be made precise. In the above construction of CG, a clique corresponds to a common *induced* subgraph of both P and G (see [14]). In our case, we want to allow a node of G to have more incident edges than the associated node of P. For example, we want pattern P1 from Figure 1 to match the peptide graph G1, although there is no edge between the first and the last node in P1 while there is one between the corresponding nodes in G1. In mathematical terms, we are looking for a subset of k nodes in P such that the corresponding induced subgraph of P is connected and occurs as a (not necessarily induced) subgraph of G. This asymmetry between P and G prevents using standard solutions for computing common substructures (see [14]) and raises the need to develop an efficient method appropriate for our setting. For this purpose, we modify the above solution based on clique search in the compatibility graph.

We first modify the definition of compatibility graph, taking into account that if two nodes in G are connected by an edge, the associated nodes in P may or may not be connected. Since the size of the CG (both in terms of the number of nodes and edges) is the crucial factor for efficiency, we need to make sure to keep this size reduced and filter out those node associations which cannot participate in the mapping. Even prior to constructing the CG, we verify simple properties that prevent a common substructure of size k to exist. First, the size of G must be greater than or equal to k. Furthermore, at least k nodes of the pattern must be associated to some nodes of the peptide graph. Only if these two simple tests are verified, we proceed to the construction of the CG and searching for a k-clique.

CG nodes

In order to decrease the number of nodes in the CG, we associate a node u' of G and a node u of P only if the degree of u' is greater than or equal to the degree of u. This is justified by the above definition of common substructure of P in G.

CG edges

According to our definition of common substructure, we have to modify the above definition of an edge in the CG. Conditions (1) and (2) are replaced by the following:

• nodes U (u, u') and V (v, v') are adjacent in the CG if and only if $u \neq v$ and $u' \neq v'$ and u' is adjacent to v' in G *provided that* u is adjacent to v in P (3)

In other words, if two nodes in the pattern graph are connected, then the corresponding nodes in the peptide graph must be connected too, but the opposite is not necessarily true. With this definition we achieve that if two nodes u and v are not connected in the pattern graph P, their corresponding nodes u' and v' in the peptide graph G may or may not be connected, in both cases the corresponding nodes U (u, u') and V (v, v') in the CG are connected. However, this rule leads to an increase of the

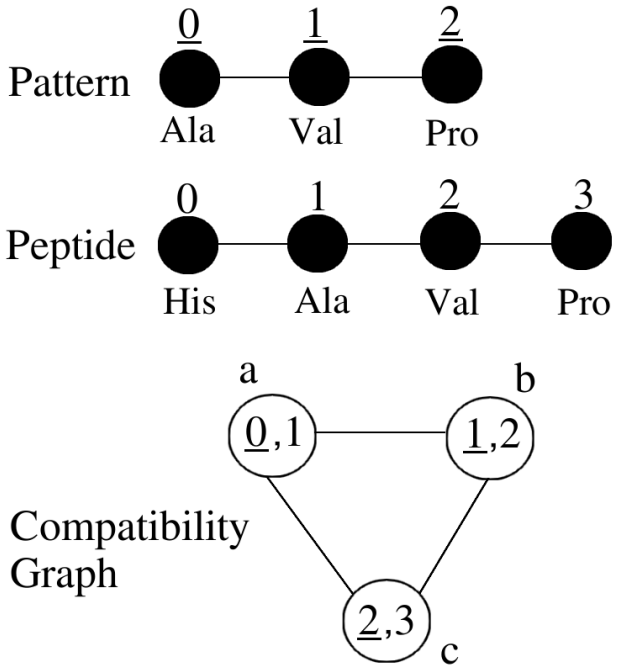


Figure 2
A simple example of compatibility graph. The Figure shows a pattern graph, a peptide graph and the corresponding CG. Each node of pattern and peptide graphs has a label (for example 'Ala') and a number that is a unique identifier of this node. Identifiers of pattern nodes are underlined in order to distinguish them from peptide nodes. A node of the CG corresponds to the association of a node of the pattern graph (underlined number) and a node of the peptide graph with the same label. Nodes of the CG are named by letters. For example, node 'a' corresponds to the association of pattern node 0 and peptide node 1 that have both label 'Ala'. Edges between nodes 'a' and 'b' and between nodes 'b' and 'c' in the CG correspond to condition (1). Edge between nodes 'a' and 'c' corresponds to condition (2) of the definition of CG. A clique of size 3 exists in the CG and corresponds to the occurrence of the pattern in the peptide.

number of edges in the CG. Indeed, according to condition (3), the CG has an egde between U (u, u') and V (v, v') even if u and v are not connected in P while u' and v' are connected in G. The classical CG constructed according to conditions (1) and (2) would not include an edge in this case. We then introduce a stronger rule in order to reduce the number of edges and to make the search for a k-clique efficient. The rule is based on the computation of elementary paths.

An elementary path (EP) in a graph is a path without loops. For each node in P and G, we compute the size of all EPs from this node to all the others. Since we are interested in connected subgraphs of size k, the EP size in such subgraphs is limited to k - 1 (which is the maximal number of nodes that can be visited along an EP in a graph of size k). For a graph G, we store the EP sizes in a matrix EPS_G, where the $EPS_G[i, j]$ contains the list of all EP sizes between the nodes i and j.

Figure 3 shows the matrices for pattern graph P1 and peptide graph G1 from the Figure 1 with k equal to the pattern size. For example, there are two EPs between nodes 1 and 4 in G1, one of size 3 (path 1 - 2 - 3 - 4) and another of size 4 (path 1 - 2 - 6 - 5 - 4). Nodes 0 and 4 are connected by two EPs of size 4 and 5, but the second one is not considered as it is greater than k - 1 (P1 has 5 nodes).

We then define an edge between U (u, u') and V (v, v') in the CG if and only if the EP size list of (u, v) in P (considered as a multiset) is included in the EP size list of (u', v') in G. This means that the distances between two nodes in P must be included in the respective distances in G in order for an edge in the CG to exist. In other words, the monomers of the EPs beteeen u and v in P and between u' and v' in G are not directly compared but the distances of possible paths in P must be also distances of possible paths in G. This new rule decreases the number of edges in the CG without losing any information on a possible occurrence of the pattern.

Figure 4 shows the resulting CG of P1 and G1 built with the classical and the new CG building rules, with k = 5 (size of pattern P1). Observe that there is no edge between nodes a and l in the CG constructed with classical building rules (nodes 0 and 4 are not adjacent in P whereas nodes 4 and 3 are adjacent in G) whereas this edge exists in the CG constructed with the new building rules and implies a clique of size 5 that shows that P1 occurs in G1. In addition, the number of nodes and edges in the two CGs are different: the CG obtained with the classical building rules has 13 nodes and 22 edges whereas the CG obtained with the new building rules has 12 nodes and 19 edges. For example, in the CG obtained with the classical building rules there is an edge between nodes b(0, 6) and l(4, 3) that does not exist in the CG obtained with the new building rules. This is because the EP sizes between nodes 0 and 4 in P1 (here {4}) are not included in the EP sizes between nodes 6 and 3 in G1 (here {2, 3}). The new CG building rules exclude this kind of edges and thus decrease the overall number of edges in the CG.

New CG building rules: summary
We conclude this section by summarizing the CG building rules for a pattern graph P and a peptide graph G:

• each CG node U (u, u') corresponds to the association of a node u of P and a node u' of G such that $deg(u) \leq deg(u')$ and f (u) is compatible with f (u'). (4)

• two nodes U (u, u') and V (v, v') are adjacent in the CG if and only if $u \neq v$ and $u' \neq v'$ and EPS_P [u, v] $\subseteq EPS_G$ [u', v']. (5)

Search for a k-clique

The presence of a k-clique in the CG implies that there is an induced subgraph of P that is a subgraph of G. In the case when k is smaller than the size of P, we have to verify, in addition, that the corresponding subgraph is connected in P (and consequently in G).

To search for a k-clique, we use a branch and bound algorithm (see Chapter 6 in [16]). It is essentially an exhaustive algorithm that explores the depth-search tree of the graph. For a node of depth h in the tree, we try to extend the current clique of size h with a new node in order to obtain a clique of size h + 1. The tree is pruned by not exploring the branches with the length smaller than k. Once a k-clique is found, the search terminates and the pattern occurrence is output.

Another heuristic we use to speed up the clique search is based on the fact that once we identified more than ($|V_P|$ - k) nodes of pattern that do not participate in the clique, the search for a k-clique can be stopped. In the case of search for the entire pattern (k = $|V_P|$), each pattern node has to contribute to the clique. For example, in Figure 4(b) with k = 5 (pattern size), node 0 of the pattern participates in two nodes a and b of the graph, which implies that one of these two nodes must belong to the clique. If a k-clique containing one of these two nodes is not found, the search is stopped. Finally, another speeding heuristics is to start the search with CG nodes that correspond to pattern nodes of maximal degree and have therefore most chances to lead to a fast detection of non-occurrence of the pattern. Applying all these heuristics leads to a practically fast and exact clique search, as confirmed by experimental results provided in the next section.

a)

	0	1	2	3	4
0	{0}	{1}	{2}	{3}	{4}
1	{1}	{0}	{1}	{2}	{3}
2	{2}	{1}	{0}	{1}	{2}
3	{3}	{2}	{1}	{0}	{1}
4	{4}	{3}	{2}	{1}	{0}

b)

	0	1	2	3	4	5	6
0	{0}	{1}	{2}	{3}	{4}	{4}	{3}
1	{1}	{0}	{1}	{2}	{3,4}	{3,4}	{2}
2	{2}	{1}	{0}	{1,4}	{2,3}	{2,3}	{1,4}
3	{3}	{2}	{1,4}	{0}	{1,4}	{2,3}	{2,3}
4	{4}	{3,4}	{2,3}	{1,4}	{0}	{1,4}	{2,3}
5	{4}	{3,4}	{2,3}	{2,3}	{1,4}	{0}	{1,4}
6	{3}	{2}	{1,4}	{2,3}	{2,3}	{1,4}	{0}

Figure 3
Matrix of elementary path sizes. This figure shows matrix of elementary path sizes for (a) P1 of Figure 1 and (b) G1 of Figure 1.

Testing

Case study of structural properties of NRPs

We studied the distribution of patterns of size 4 in all peptides of the database. The results are shown in Figure 5(a). They show that the most frequent 4-pattern is the linear pattern. We also computed the distribution of the number of peptide graphs depending on their size. The results are shown in Figure 5(b). More than 70% of peptides have at least seven monomers. This means that a search for a pat-

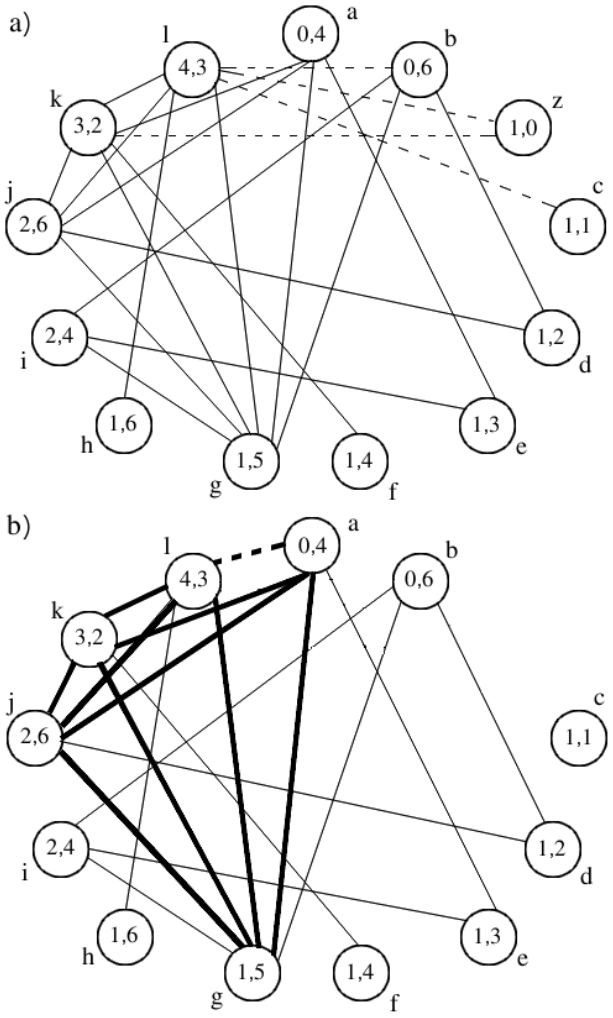


Figure 4
Example of compatibility graph constructed with classical and new methods. The CG of pattern graph *P*1 and peptide graph *G*1 of Figure 1 constructed with (a) classical and (b) new CG building rules. Each CG node is identified by a letter. It represents an association between a node of *P*1 and a node of *G*1 with compatible labels. For example, node 'a' associates node 0 of *P*1 and node 4 of *G*1 that both carry the 'D-Tyr' label. Dashed edges correspond to the edges that differ between the two CGs and the bold edges correspond to a clique of size 5 (size of *P*1).

tern containing seven 'X' labels triggers the construction of the CG for more than 70% of peptides of the NORINE database.

Efficiency of the method

In order to test the efficiency of our method, we compared the number of nodes and edges in the CGs obtained with the classical and the new building rules on different examples in the case of search for an entire pattern. The results are shown in Table 1. In the case of matching P3 against G2, the CG constructed with modified rules has no edges because the EP sizes of the pattern are not included in the corresponding EP sizes of the peptide graph. Indeed, P3 is cyclic and each pair of nodes is connected by two EPs whereas G2 is linear and there is only one EP for each pair of nodes. Therefore, our method outputs the answer without looking for a clique. In the last example corresponding to the linear pattern of 19 'X' against alamethicin F50, the difference in the number of nodes (346 against 380) is due to the additional condition on the degree of associated nodes. Moreover, in this example, our version of CG has about 13 times less edges than the CG constructed with the classical rules. These examples illustrate that our method produces a compact compatibility graph, suitable for an efficient clique search.

In order to validate this speed-up in running time, we measured the search time for different complete patterns in the NORINE database. The results are shown in Figure 6. The first observation is that the number of results is often smaller with the classical rules. This is due to the case when the pattern graph has a number of edges different from the peptide graph. In example 6, we search for peptides containing any pair of monomers which is the case for all 711 peptides of NORINE. However, only 698 peptides are output if the classical building rules are used. There are 13 cyclic dipeptides in Norine. This is due to a special case where two nodes of a peptide are connected by two edges, which corresponds to a cyclisation between the two monomers. This special case cannot be detected with the classical building rules but is taken into account by our method.

For the linear pattern of size 7 (example 7), which is contained in more than 70% of the database peptides, the classical rules show a 8-fold slow-down of the running time compared to the new rules. For a linear pattern composed of 14 'X' (example 9), the classical method required 7 hours to produce the result whereas our method took less than 300 ms. Example 12 is the search for a pattern composed of 49 'X', the size of the largest peptide of the database. About 5 minutes were needed for the classical method to obtain the result whereas our method took only about 600 ms. Example 14 represents a negative test as this pattern does not occur in NORINE. In this example, the classical method did not terminate in 8 hours, whereas our method output the result in 280 ms.

These experiments illustrate the effficiency and adequacy of our method for the search for structural patterns in NRPs.

Examples of practical applications

In this part, we give some examples of using structural pattern matching of NRPs in biological studies.

Structural features

Structural search can allow one to identify a structural motif common to peptides of a given family. As an example, a search for a cyclic 8-node pattern composed of seven 'X' and a fatty acid moiety (represented by the monomer code '*R-'), with $k = 8$, outputs the peptides of the iturin family (iturins, bacillomycins and mycosubtilin), surfactins and lichenysins. Therefore, this pattern represents a common structural feature of this family.

Another example is the search for a pattern associated with a biological activity. For example, pattern P4 occurs in G2 that represents ornibactin. Ornibactin is a siderophore, an iron-chelating molecule. This type of molecule needs bidendate functions that can ensure a six-fold coordination of the ferric iron. Ornithine and its derivatives can harbour this function. A search for complete pattern P4 returns a list of six siderophore peptides

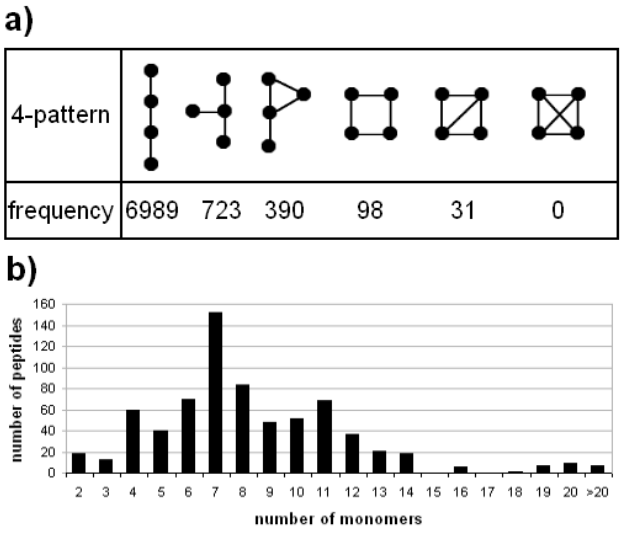


Figure 5
Structural properties of NRPs contained in NORINE. Distribution of (a) 4-patterns and (b) peptide sizes in the NORINE database.

Table 1: Number of nodes and edges of the CG constructed with classical and new building rules

pattern	peptide	# CG nodes	# CG edges
P1	G1	13/**12**	22/**19**
P2	G1	16/**16**	43/**29**
P3	G1	35/**30**	210/**100**
P3	G2	25/**15**	100/**0**
P4	G2	10/**8**	14/**9**
Ala-1[(a)]	Ala[(b)]	73/**73**	1918/**286**
(X)19[(c)]	Ala[(b)]	380/**346**	53010/**3948**

Patterns P1–P4 and peptides G1–G2 refer to Figure 1. In all examples, *k* is equal to the pattern size. In columns 3 and 4, data shown in regular and bold font concern respectively the classical and modified CG building rules.
[(a)] linear pattern of size 19 corresponding to alamethicin F50 without the last monomer
[(b)] alamethicin F50 [NORINE:NOR00007]
[(c)] linear pattern of 19 'X' monomers

such as ornibactin, pyoverdin or foroxymithine. A search for the pattern R-CO_*OH-Orn_*Asp_*Ser_*Orn derived from ornibactin, with $k = 2$, returns a list of 51 peptides. Among them 46 peptides are annotated as siderophores. This example illustrates that structure-function relationships can also be elucidated by searching for a common substructure between a pattern derived from a peptide characterizing the function and the other peptides of the database.

Product identification

Another application of structural pattern matching is the search for a predicted peptide. Several studies (see [17,18] for recent examples) start with searching for putative NRPS genes within a genome and predicting the produced peptide out of this sequence. Such a prediction can result in an undetermined monomer or several possible monomers at some positions. The possibility of using 'X' or '/' labels in the structural pattern allows for a look-up for the predicted peptide in the database in order to find out whether the peptide has been discovered before, possibly in another species. To give a concrete example, we submitted six NRPS proteins found in UniProtKB database [19] ([UniProtKB:Q2VQ12–Q2VQ17]) to the NRPSPREDICTOR[6] and obtained predicted peptides. By combining them, we obtained the structural pattern X_NP_X_NP_NP_NP_X_NP_*Leu_X_*Phe/*Trp/*Tyr_*Leu_NP, where NP stands for non-polar amino acids and corresponds to *Val/*Ile/*Leu/*Abu/*Iva in NORINE notation. A search for this complete pattern in NORINE resulted in only one peptide, BT1583, that is consistent with the bibliographic data of UniProtKB. This example illustrates that the structural search can help associating a product with a set of nonribosomal synthetases.

	pattern	#matches	time
1	P1	0 / 1	152 ms / 147 ms
2	P2	10 / 11	2,3 s / 186 ms
3	P3	105 / 105	7,7 s / 309 ms
4	P4	4 / 6	271 ms / 219 ms
5	Gln/Glu_X_D-Leu_X_Asp_D-Leu_X	12 / 12	178 ms / 175 ms
6	X_X	698 / 711	180 ms / 179 ms
7	X_(X)5_X	332 / 511	3,1 s / 383 ms
8	X_(X)9_X	113 / 175	7,1 min / 387 ms
9	X_(X)12_X	33 / 48	7 h / 267 ms
10	X_(X)16_X	ND / 24	ND / 265 ms
11	X_(X)18_X	ND / 15	ND / 377 ms
12	X_(X)47_X	1 / 1	4,7 min / 598 ms
13	X_(X)14_X(X,X)	ND / 7	ND / 394 ms
14	X_(X)14_X(X,X,X)	ND / 0	ND / 260 ms

Figure 6
Search time for different complete patterns in the NORINE database. Here, *k* is equal to the size of the pattern. In the 2nd and 3rd columns, the first and second value corresponds respectively to the classical and new building rules. 'ND' means that the result has not been obtained as the running time exceeded 8 hours.

Analysis of a putative peptide

From the analysis of protein sequence similarity, some proteins can be predicted as putative NRPSs. Examples of such predictions can be found in the UniProtKB database. Even though the produced peptide has not been identified, one can infer some properties of a putative NRPS product using the structural search. An example can be provided by the putative NRPS [UniProtKB:Q1I964] from UniProtKB found in *Pseudomonas entomophila*. By studying the sequence of this synthetase, four modules can be predicted. Pattern Val_Leu_Ser_Ile is obtained using the NRPSPREDICTOR. This pattern occurs in the lipopeptide putisolvin I stored in NORINE. The search for a more general pattern NP_NP_Ser_NP gives three results, putisolvin I, II and PFL2145, that are all lipopeptides. One can observe that putisolvin I is produced by *Pseudomonas putida*, the same genus of bacteria than [UniProtKB:Q1I694]. This bacteria genus is known to produce various cyclic lipopeptides [20]. By analysing the gene environment of [UniProtKB:Q1I964], we found another gene coding for a putative NRPS [UniProtKB:Q1I963], which probably produces the beginning of the peptide. A condensation domain characteristic of lipopeptide production can be

predicted at the begining of the protein. This domain binds the lipid moiety to the peptide part. This is another clue for lipopeptide production. NRPSpredictor outputs the octapeptide X_NP_NP_X_NP_NP_X_Ser which matches no peptide of NORINE. However, the final product would be composed of 12 monomers and a lipid moiety like putisolvin I. In addition, the composition of the predicted peptide does not match any peptide in Norine but is close to the composition of lipopeptides. Indeed, if we compare the monomeric composition of the predicted peptide with putisolvin I, both compositions are similar. Thus, all data converge to a lipopeptide production. This example illustrates that structural pattern search can assist the biological identification of a predicted peptide by inferring its properties.

Conclusion

Nonribosomal peptides are important bioactive compounds that have various important biological activities and are increasingly studied. With this motivation, we developed the NORINE database [21] that is the first computational resource entirely devoted to NRPs. All peptides stored in NORINE are annotated with their monomeric (possibly non-linear) structure encoded by undirected labeled graphs.

In this paper, we presented an efficient dedicated method to search for a structural pattern in the database. We refined the CG building rules previously used in the literature and improved them to adapt to our problem. The main idea of refinement is to use the information on elementary path sizes and on the node degrees in order to decrease the number of nodes and especially the number of edges in the resulting CG. This, in turn, leads to a considerable speed-up in the search for a clique in the CG, which is the final step in the identification of a pattern occurrence.

As a result, a search for a pattern in the NORINE database currently containing 711 peptide structures takes typically less than one second. Note that the proposed method is exact, i.e. outputs precisely all the peptides that contain the pattern, without any error allowed. Note also that the efficiency of the method can be further increased by precomputing the matrices of EP sizes for all stored peptides. This would obviously improve the performance of querying the database with different patterns.

Searching for a structural pattern in the database can be used in different biological studies. For example, it can help to identify members of a peptide family that share common structural properties. It can also help to identify a predicted peptide by searching for it in the NORINE database in case the peptide has been discovered before in other species. Furthermore, a search for a structural pattern can provide new insights into peptide features and help to isolate this peptide experimentally. Finally, it can help to elucidate the relationship between structure and function by searching for patterns occurring in peptides that share a common biological activity.

An obvious weakness of the method is that in general it might be unable to identify a common structure if the correspondence is not exact, i.e. some monomers "get replaced" by others (not specified explicitely with the joker or alternative labels), or do not have their counterparts at all. Therefore, an interesting direction for future research would be to extend the method to an "error-tolerant" pattern matching dealing with possible deletions, insertions or substitutions of monomers.

Methods

NORINE

The method presented in this paper is included in NORINE. NORINE http://bioinfo.lifl.fr/norine is a public Web resource entirely dedicated to NRPs. As of today, NORINE stores 711 peptides, each annotated with different information such as producing organisms, biological activities or its monomeric structure. Monomeric structures are encoded by undirected labeled graphs with nodes representing monomers. Peptides currently stored in NORINE contain overall more than 400 different monomers. Those include all standard amino acids but also many non-standard amino acids incorporated in NRPs during the biosynthesis. Lipids, carbohydrates and polyketides also occur in NRPs and are considered by NORINE as monomers. More details on the NORINE system can be found in [7].

Implementation

The method has been implemented in Java within the NORINE system. The program looks up for a structural pattern or a common substructure in all the peptides of the NORINE database ([NORINE:NOR00001] to [NORINE:NOR00711] were considered in this publication). All peptides containing the pattern or a common substructure are output. Time measures reported below have been obtained on a PC with a 1.73 GHz processor, 512 MB of RAM and 265 Mflops. The Java code implementing the method can be provided by request to the first author.

Authors' contributions

SC carried out most of the work on the algorithm design, implementation in NORINE and experiments. MP and GK participated in the algorithmic setup and the design of computational experiments. VL and PJ contributed to various biological issues related to this study. GK and PJ organized the project. SC, MP and GK wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the *PPF bioinformatique* program of the Lille 1 University. S.C. was supported by an INRIA/Région Nord-Pas-de-Calais fellowship. ProBioGEM lab is supported by the region Nord-Pas-de-Calais, the *Ministère de l'Enseignement et de la Recherche*, French ANR agency, and the European Funds for the Regional Development. Funding to pay the Open Access publication charges for this article was provided by INRIA. The authors thank Jesper Jansson for reading and commenting the manuscript.

References

1. Sieber SA, Marahiel MA: **Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics.** *Chem Rev* 2005, **105(2):**715-738.
2. Menzella HG, Reeves CD: **Combinatorial biosynthesis for drug development.** *Curr Opin Microbiol* 2007, **10(3):**238-245.
3. Stachelhaus T, Mootz HD, Marahiel MA: **The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases.** *Chem Biol* 1999, **6(8):**493-505.
4. Challis GL, Ravel J, Townsend CA: **Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains.** *Chem Biol* 2000, **7(3):**211-224.
5. Ansari MZ, Yadav G, Gokhale RS, Mohanty D: **NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases.** *Nucleic Acids Res* 2004:W405-413.
6. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH: **Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs).** *Nucleic Acids Res* 2005, **33(18):**5799-5808.
7. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G: **NORINE: a database of nonribosomal peptides.** *Nucleic Acids Res* 2008:D326-331.
8. Tsuge K, Akiyama T, Shoda M: **Cloning, sequencing, and characterization of the iturin A operon.** *J Bacteriol* 2001, **183(21):**6265-6273.
9. Minowa Y, Araki M, Kanehisa M: **Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes.** *J Mol Biol* 2007, **368(5):**1500-1517.
10. Mootz HD, Schwarzer D, Marahiel MA: **Ways of assembling complex natural products on modular nonribosomal peptide synthetases.** *Chembiochem* 2002, **3(6):**490-504.
11. Guenzi E, Galli G, Grgurina I, Gross DC, Grandi G: **Characterization of the syringomycin synthetase gene cluster. A link between prokaryotic and eukaryotic peptide synthetases.** *J Biol Chem* 1998, **273(49):**32857-32863.
12. Garey MR, Johnson DS: *Computers and Intractability: A Guide to the Theory of NP-Completeness* New York, NY, USA: W. H. Freeman & Co; 1979.
13. Fellows M, Fertin G, Hermelin D, Vialette S: **Sharp tractability borderlines for finding connected motifs in vertex-colored graphs.** *Proceedings of the 34th International Colloquium on Automata, Languages and Programming (ICALP), June 9–13, 2007, Wroclaw (Poland)* 2007, **4596:**340-351 [http://www.springerlink.com/content/978-3-540-73419-2]. Lecture Notes in Computer Science, Springer Verlag
14. Raymond JW, Willett P: **Maximum common subgraph isomorphism algorithms for the matching of chemical structures.** *J Comput Aided Mol Des* 2002, **16(7):**521-533.
15. IUPAC, IUB: **IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983.** *Biochem J* 1984, **219(2):**345-373.
16. Mehlhorn K: *Graph algorithms and NP-completeness* New York, NY, USA: Springer-Verlag New York, Inc; 1984.
17. Tobiasen C, Aahman J, Ravnholt KS, Bjerrum MJ, Grell MN, Giese H: **Nonribosomal peptide synthetase (NPS) genes in Fusarium graminearum, F. culmorum and F. pseudograminearium and identification of NPS2 as the producer of ferricrocin.** *Curr Genet* 2007, **51(1):**43-58.
18. de Bruijn I, de Kock MJ, Yang M, de Waard P, van Beek TA, Raaijmakers JM: **Genome-based discovery, structure prediction and functional analysis of cyclic lipopeptide antibiotics in Pseudomonas species.** *Mol Microbiol* 2007, **63(2):**417-428.
19. **UniProtKB** [http://www.uniprot.org/]
20. Raaijmakers JM, de Bruijn I, de Kock MJ: **Cyclic lipopeptide production by plant-associated Pseudomonas spp.: diversity, activity, biosynthesis, and regulation.** *Mol Plant Microbe Interact* 2006, **19(7):**699-710.
21. **Norine** [http://bioinfo.lifl.fr/norine/]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

JOURNAL OF BACTERIOLOGY, Oct. 2010, p. 5143–5150 Vol. 192, No. 19
0021-9193/10/$12.00 doi:10.1128/JB.00315-10
Copyright © 2010, American Society for Microbiology. All Rights Reserved.

Diversity of Monomers in Nonribosomal Peptides: towards the Prediction of Origin and Biological Activity▽†

Ségolène Caboche,[1,2]* Valérie Leclère,[1] Maude Pupin,[2] Gregory Kucherov,[2] and Philippe Jacques[1]

ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, Polytech-Lille/IUTA, F59655 Villeneuve d'Ascq, France,[1] *and LIFL, UMR USTL/CNRS 8022, INRIA Lille-Nord Europe, F59655 Villeneuve d'Ascq, France*[2]

Received 22 March 2010/Accepted 26 July 2010

Nonribosomal peptides (NRPs) are molecules produced by microorganisms that have a broad spectrum of biological activities and pharmaceutical applications (e.g., antibiotic, immunomodulating, and antitumor activities). One particularity of the NRPs is the biodiversity of their monomers, extending far beyond the 20 proteogenic amino acid residues. Norine, a comprehensive database of NRPs, allowed us to review for the first time the main characteristics of the NRPs and especially their monomer biodiversity. Our analysis highlighted a significant similarity relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that some NRPs isolated from sponges are actually synthesized by symbiotic bacteria rather than by the sponges themselves. A comparison of peptide monomeric compositions as a function of biological activity showed that some monomers are specific to a class of activities. An analysis of the monomer compositions of peptide products predicted from genomic information (metagenomics and high-throughput genome sequencing) or of new peptides detected by mass spectrometry analysis applied to a culture supernatant can provide indications of the origin of a peptide and/or its biological activity.

Nonribosomal peptides (NRPs) are molecules produced by microorganisms and synthesized by huge multienzymatic complexes (38, 41), called *n*on*r*ibosomal *p*eptide *s*ynthetases (NRPSs). These megaenzymes are organized into modules, one for each amino acid to be built into the peptide product. This is accomplished by division of each catalytic step into specialized semiautonomous domains. The basic set of domains (adenylation, thiolation, and condensation) within a module can be extended by substrate-modifying domains, including domains for substrate epimerization, β hydroxylation, N methylation, and heterocyclic ring formation. The peptide release is catalyzed by a thioesterase domain which can also, in many cases, be involved in an intramolecular reaction leading to a cyclic or partially cyclic peptide or, in fewer cases, in the oligomerization of peptide units (iterative biosynthesis). NRPs show a broad spectrum of biological activities and pharmaceutical applications. They can harbor antimicrobial, immunomodulator, or antitumor activities. Cyclosporine (5), an immunosuppressant drug widely used in organ transplantation, daptomycin (60) (marketed in the United States under the trade name Cubicin), used in the treatment of certain infections caused by Gram-positive bacteria, aminoadipyl-cysteinyl-valine (ACV)-tripeptide, which is the precursor of cephalosporin and penicillin (29), the most famous antibiotic, and also bleomycin (57), used in the treatment of several cancers, are some common examples of NRPs of high therapeutic importance. Two main structural traits distinguish these peptides from ribosomally synthesized peptides: first, their primary structure is more frequently cyclic (partially or totally) branched or polycyclic rather than linear and, second, the biodiversity of monomers incorporated in NRPs goes far beyond the 20 proteogenic amino acids residues. NRP monomers include modified versions of the proteogenic amino acids (e.g., methylated, hydroxylated, and D-forms) but also other monomers, such as, for example, 2-aminoisobutyric acid (Aib), hydroxyphenylglycine (Hpg), and 2,3-dihydroxybenzoic acid (diOH-Bz). However, essential characteristics of this diversity and its relationship with biological functions and producing organisms have been poorly understood until now.

The development of the Norine database, the first resource entirely dedicated to NRPs (8, 9), filled this gap. Based on Norine data, we performed the first large-scale analysis of about a thousand peptides which represent a total coverage of more than 10,000 monomer occurrences, revealing the presence of as many as 500 different monomer types. A data-mining analysis of the monomeric compositions of NRPs allowed us to reveal a strong relationship between certain monomeric characteristics of NRPs and their biological function and producing organism. In addition to providing a comprehensive overview of monomeric biodiversity in NRPs, this work demonstrated (i) a dissimilarity of structural properties between bacterial and fungal NRPs; (ii) a significant relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that the peptides isolated from sponges are in reality synthesized by symbiotic bacteria rather than by the sponges themselves; and (iii) a certain monomer specificity to a class of biological activities. Those observations are supported by successful statistical predictions of biological activities of NRPs based on their monomeric compositions.

* Corresponding author. Mailing address: ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, F59655 Villeneuve d'Ascq, France. Phone: 33(0) 328 76 7440. Fax: 33(0) 328 76 7356. E-mail: segolene.caboche@lifl.fr.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 6 August 2010.

TABLE 1. Repartition of NRPs in groups showing great diversity

Group	All NRPs ($n = 9$)	Only curated NRPs ($n = 4$)
Dolastatines	4	4
Kahalalides	16	16
Pyoverdins	57	57
Serrawettins	2	2
Guineamides	6	
Hymenamides	10	
Kapakahines	5	
Phakellistatins	14	
Stylopeptides	2	

MATERIALS AND METHODS

NRP set. Here, we define the data sets we used in our analyses. We started with the whole set of the first 1,071 peptides stored in the Norine database (http://bioinfo.lifl.fr/norine). Based on the annotations of the Norine database, we selected several training sets as described below.

First, we distinguished between curated and putative NRPs as annotated in the Norine database. For curated peptides, either corresponding synthetase genes have been identified or their nonribosomal origin has been universally accepted by the scientific community, as is the case for polytheonamide, for example (32). For putative peptides, there is no experimental evidence of their synthesis pathway, and other characteristics, such as their nonlinear structure and/or found nonproteogenic amino acids, suggest their nonribosomal origin. Examples are oscillarin (27) and aurilide (62). Of the 1,071 Norine peptides, 790 (that is, nearly three-quarters) are curated.

Second, some Norine peptides are considered variants belonging to the same group. Until now, no universal definition of the NRP variants has been proposed. Most of the time, peptides are called variants if they show similar compositions. For example, surfactin (3) and [Val7]surfactin (44) differ by only one monomer at position 7. More rarely, variants may have different structures and/or sizes, such as cyclic gramicidin S (11), composed of 10 monomers, and linear gramicidin A (31), composed of 16 monomers. However, some peptides can also be considered variants when they share a specific function or have features in common. For example, pyoverdins (67) form a large group of siderophores (7) produced by species of *Pseudomonas* or a related genus; they have a large diversity in structure and monomeric composition but have features in common, such as the presence of a fluorescent chromophore.

To reduce a bias in learning structural or compositional properties of NRPs belonging to different groups, we eliminated close variants and kept only one variant per group. For example, in some groups, only few monomers vary, and therefore invariable monomers of these groups may be given an overestimated significance. In order to identify groups for which we need to keep only one representative variant, we computed an average distance between all variants of the group (see the supplemental material). If the peptides of a given group were similar, i.e., the average phylogenetic distance in the group was low, we kept only one random member of this group; otherwise, we kept all the peptides of the group. The 1,071 NRPs (790 curated) were divided into 183 groups (100 curated), of which 62 groups (24 curated) contain only one peptide. Among the 121 groups (76 curated) containing at least two variants, 9 (4 curated) showed a high diversity (Table 1): dolastatines (47), guineamides (64), hymenamides (58), kahalalides (26), kapakahines (68), phakellistatins (42), pyoverdins (67), serrawettins (35), and stylopeptides (6). For those groups, we kept all the variants in all analyses. We want to mention that 130 peptaibols (16) are stored in the Norine database, divided into 20 groups of variants. All the peptaibols are curated NRPs. This means that there are 20 representative peptaibols among 290 (175 curated) representative NRPs. When we consider a data set containing only one variant per group (except for the 9 groups mentioned above), we use the term "excluding variants." The NRP set excluding variants contains 290 peptides, and the curated NRP set excluding variants contains 175 peptides.

Correlation coefficient. The correlation coefficient (CC) is computed in order to evaluate the relationship between two data sets. It is comprised between -1 and 1; the closer the CC is to these extreme values, the more the data are correlated. In two series, X ($x1, \ldots, xn$) and Y ($y1, \ldots, yn$), the CC is computed as follow:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

In our case, X and Y will be either two monomeric distributions, i.e., the number of monomer occurrences in two sets of NRPs, or two distributions of peptide sizes.

RESULTS

The arrival of the Norine database provided us with the possibility of obtaining a general overview of NRPs. Below we present three analyses that we performed on Norine data. First, we studied some general statistical characteristics of NRPs, and then we focused on properties related to producing organisms; finally, we analyzed the monomeric distribution depending on biological activities. The observations that we drew from these analyses should be used for predicting the activities and origins of newly identified products.

General statistics. To avoid a bias in our analysis, we chose to restrict the data set for general statistics to the curated peptides excluding variants, which refers to 175 peptides (see Materials and Methods). However, the results obtained with the total set of NRPs lead to the same conclusions (data not shown).

(i) How variable are nonribosomal peptide structures? It can be deduced from Norine data that nonlinear structures represent nearly three-quarters of the NRP structures (Fig. 1). The majority (64%) of NRPs contain at least one cycle, but only a few peptides (1%) possess only branchings, and up to 8% present complex structures with overlapping cycles and branching. Those complex structures are found mainly in glycopeptides (15), a large group of antibiotics, the most famous of which is vancomycin (Fig. 2), used in treatment of infections caused by Gram-positive bacteria (28).

(ii) What is the size of a nonribosomal peptide? We have defined the size of an NRP to be the total number of monomers, with fatty acids in lipopeptides being considered individual monomers, as they are most frequently added as a single block by a condensation domain (as in arthrofactin synthesis [51]), or by using a single module in PKS-NRPS hybrid synthetases (as in the mycosubtilin synthetase [17]). As the synthetases are generally constituted of as many modules as monomers incorporated into the peptide (except in the iterative mode of synthesis), the sizes of NRPs are limited (Fig. 3). The most frequent sizes are 7 to 9 monomers, sizes shared by about one-third of the set. The sizes vary between 2 and 23 monomers, except for polytheonamide B (not shown in Fig. 3), which has 49 monomers. Polytheonamide B has been extracted

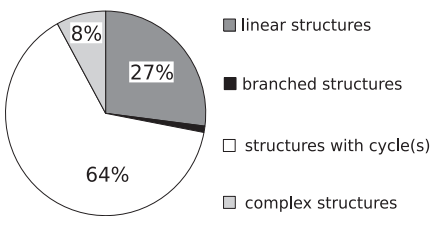


FIG. 1. Distribution of primary structures in curated peptides excluding variants (175 peptides).

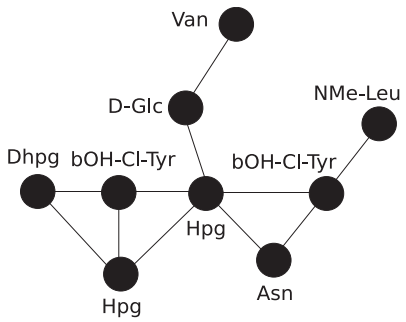


FIG. 2. Monomeric structure of vancomycin. Hpg, hydroxyphenylglycine; NMe-Leu, *N*-methylleucine; bOH-Cl-Tyr, beta-hydroxy-chloro-tyrosine; Asn, asparagine; Dhpg, 3,5-dihydroxyphenylglycine; D-Glc, D-glucose; Van, vancosamine.

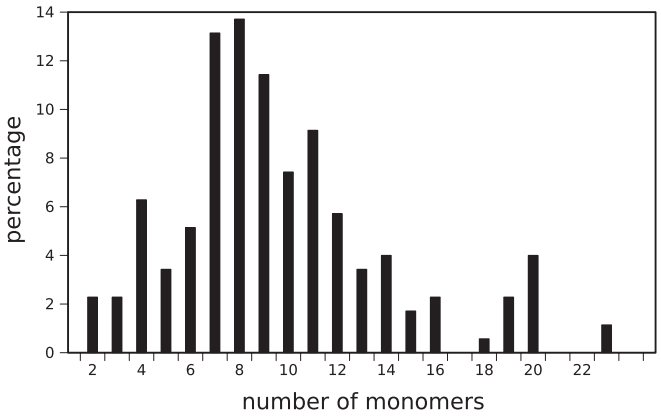


FIG. 3. Size distribution of curated peptides excluding variants (175 peptides). Polytheonamide B, composed of 49 monomers, does not appear in the figure, as it is the only peptide known with more than 23 monomers.

from the marine sponge *Theonella swinhoei* (25), but its synthesis process is not known. However, it contains several nonproteogenic and D-amino acids that suggest that its synthesis is nonribosomal (32).

(iii) What are the most frequently found monomers in nonribosomal peptides? In addition to having specific primary structures, NRPs contain nonproteogenic amino acids and other monomers. Among 1,725 monomers contained in 175 curated peptides excluding variants, 1,614 are incorporated and/or modified directly by nonribosomal peptide synthetases, 44 are lipids, 11 are carbohydrates, 5 are polyketides, and 51 are of other or unknown origins. Proteogenic L-amino acids represent 40% of monomers found in NRPs. The most frequent monomer in curated peptides excluding variants is 2-aminoisobutyric acid (Aib) (Fig. 4). The monomer Aib is characteristic of peptaibols (2), which are linear antibiotics produced only by fungi. The Norine database contains 130 peptaibols, forming 20 groups of variants (see Materials and Methods), all of which contain at least one Aib residue. Aib can occur several times in the same peptaibol, on average 6 times per peptide, which explains why Aib is the most frequent monomer in Norine NRPs. Note that serine (Ser), threonine (Thr), and their derivatives are very frequent in NRPs. These amino acids present a hydroxyl function that allows the formation of an additional chemical bond in order to obtain nonlinear primary structures. For example, in syringomycins (56), the hydroxyl group of serine is responsible for the branching, and the hydroxyl group of two threonines is used to form two cycles in actinomycin D (45). Many amino acids appear in their D-form in NRPs. D-Monomers are epimerized mainly by a specific domain of the synthetase or, in some cases, can be directly incorporated into their D-form, as was shown for arthrofactin synthesis (51). The monomer isomery can play an important role in a peptide's structure and properties, also by limiting protease degradation. Nonproteogenic amino acids such as 2,4-diaminobutyric acid (Dab) or ornithine (Orn) and derivates are often found in NRPs, for example, in marinobactins (39) or pyoverdins (67), respectively.

Finally, it is interesting to note that the two proteogenic amino acids containing the thiol function are underrepresented in NRPs. Methionine (Met) is present in only one curated peptide, oscillamide B (55), synthesized by bacteria, and in four putative peptides extracted from sponges (haligramides A and B [48], hymenamide F [33], and phakellistatin 5 [43]). Cysteine (Cys) occurs in the famous ACV (the penicillin precursor) synthesized by both bacteria and fungi but also occurs in 9 bacitracins (40) and in different siderophores (curated or not), such as pyochelin and related compounds (wataseymycin, thiazostatin, aeruginoic acid, desferriferrithiocin, micacocidin, yersiniabactin, and anguibactin) (7), synthesized by bacteria where Cys forms a cycle with another monomer. The high reactivity of the sulfhydryl group may explain the low representativeness of free cysteine in NRPs.

The Norine database provides an interface to query the monomers composing nonribosomal peptides (see http://bioinfo.lifl.fr/norine/search_amino.jsp). The entire list of monomers can be browsed, and information on each monomer can be consulted.

Study of producing organisms. Most of the peptides stored in the Norine database are isolated from inoculated media or natural environments, but only a few are inferred from a synthetase protein sequence (154 database peptides are linked to synthetases). Until now, synthetase genes have been identified only in bacteria and fungi; none have been identified in archaea or nonfungal eukaryotes. The major part of curated Norine peptides (61%) are synthesized by bacteria, and 34% of them are synthesized by fungi (Fig. 5). However, some peptides are extracted from other organisms (5% of data), such as sponges (like cyclotheonamides [20] and polytheonamides [25, 32] from *Theonella*), tunicates (like didemnins [66] isolated from the *Didemnidae*), gastropoda (such as antitumor peptide dolastatines [7] extracted from species of *Dolabella*), or even plants (e.g., a putative cyclolinopeptide [46] from linseed oil). In these cases, the NRPS genes have not been identified, but based on the presence of unusual monomers and/or structural features in these peptides, a nonribosomal origin can be hypothesized. For example, didemnins form a group of cyclic depsipeptides isolated from two groups of tunicates, *Didemnidae* and *Polyclinidae*, and show antibiotic, antitumor, and immunomodulating activities. Didemnins contain several unusual monomers, such as isostatine (Ist) and hydroxyisovalerylpropionyl (Hip), suggesting their nonribosomal origin (24). However, in the case of organisms other than bacteria and fungi,

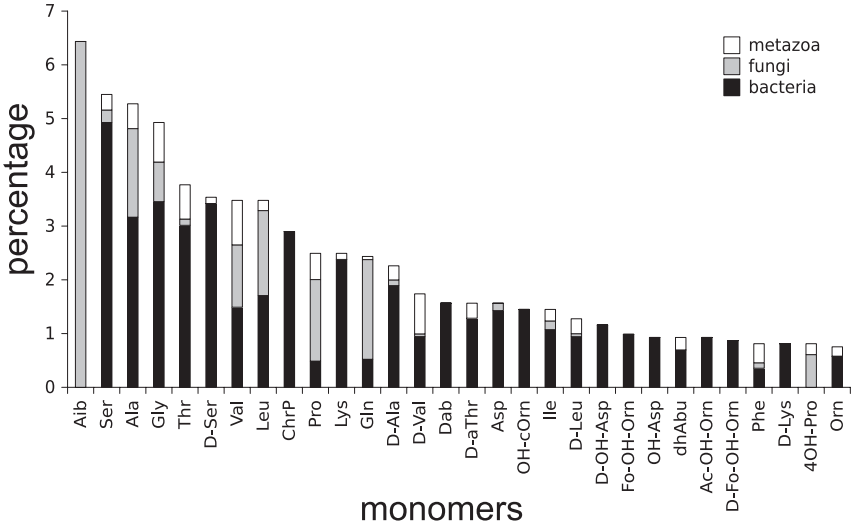


FIG. 4. Thirty of the most frequently found monomers among curated peptides excluding variants (175 peptides). The monomers' proportions occurring in bacteria, fungi, and metazoa are represented. dhAbu, 2,3-dehydro-2-aminobutyric acid.

the hypothesis that NRPs are actually synthesized by symbiotic bacteria cannot be excluded. Indeed, several recent studies have shown that NRPs extracted from sponges are in fact synthesized by symbiotic bacteria (for recent studies, see references 30, 36, 53, and 70).

We compared properties of NRPs isolated from bacteria, fungi, and metazoa, the last being mainly sponges. In this study, we considered curated NRPs for bacteria and fungi and all NRPs for metazoa. We did not use sets excluding variants or curated metazoan NRPs in order to keep a significant number of peptides in each set. To begin, we studied the size distribution of NRPs in the three groups of producing organisms (Fig. 6). We observed that the NRP sizes are different depending on the group (bacteria, fungi, or metazoa). For example, both bacteria and metazoa display a peak for sizes 7 and 8, while those sizes are nearly absent in fungal peptides. Numerous fungal peptides have a size between 14 and 20, while few bacterial and no metazoan peptides have those sizes. A peak for size 4 is shared by fungi and metazoa. However, the high proportion of size 4 metazoan peptides comes from geodiamolides (10), a family of 19 variants extracted only from sponges. We have computed the correlation coefficients (CCs) (Materials and Methods) between size distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa (Table 2). The closer the CC is to 0, the less the monomeric distributions are related. This experiment confirms the previous observations that fungal NRPs are unrelated to both bacterial and metazoan NRPs (CC close to 0), while metazoan NRPs can be related to bacterial NRPs (CC close to 1).

Furthermore, we have computed the CC between peptide monomeric distributions depending on the producing organism. The results are shown in Table 3.

The CC between the monomeric distribution of NRPs synthesized by bacteria and those synthesized by fungi is the lowest observed, pointing out differences between the monomers used by both (super)kingdoms. On the other hand, the CC between the distributions of bacteria and metazoa is the highest, highlighting their similarity. Analyzing the monomers contained in peptides of different (super)kingdoms confirms the correlation coefficient tendency. For example, Aib (2-aminoisobutyric acid), the characteristic monomer of peptaibols, occurs in 131 fungal peptides, from which only one putative cyclic antitumor peptide is not a peptaibol: chlamydocin (4). Only one bacterial peptide of the Norine database, microcystin L-Aib (21), contains an L-Aib and no metazoan peptide (see Fig. 4). Other monomers are specific to fungal NRPs and, more

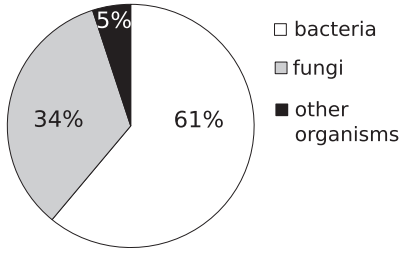


FIG. 5. Distribution of producing organisms for curated peptides (790 peptides).

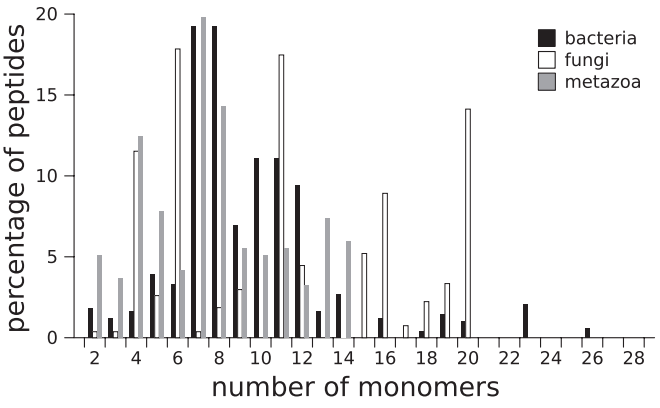


FIG. 6. Size distribution of curated peptides synthesized by bacteria (n = 488), curated peptides synthesized by fungi (n = 269), and peptides extracted from metazoa (n = 218).

TABLE 2. Correlation coefficients between size distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa

Organism 1	Organism 2	CC
Bacteria (488 peptides)	Fungi (269 peptides)	0.236
Bacteria (488 peptides)	Metazoa (218 peptides)	0.817
Fungi (269 peptides)	Metazoa (218 peptides)	0.254

precisely, to peptaibols, such as isovaline (Ival), occurring in 23 peptaibols, phenylalaninol (Pheol), occurring in 68 peptaibols, Leucinol (Leuol), occurring in 43 peptaibols, and valinol (Valol), occurring in 29 peptaibols. The C-terminal amino alcohol in the peptaibols plays an important role in liposome permeabilization and ion channel formation (18).

Other monomers seem to be found only in bacterial NRPs (Fig. 4), which have specific properties. The monomer hydroxyphenylglycine (Hpg) is present in 56 NRPs, all of them produced by bacteria. This monomer is the only amino acid able to form 5 bonds with other monomers by oxidative ring closure. Hpg is usually found in peptides with complex primary structures, forming overlapping cycles and branching, such as vancomycin (Fig. 2) (28), balhimycin (49), decaplanin (54), eremomycin (22), and galacardin (63). Other monomers specific to bacteria are chromophores (Chr), which occur mainly in siderophores, except for actinomycins.

Study of biological activity. (i) Statistical results. In this section, we present a statistical analysis of some peptide characteristics depending on the peptide's biological activity. Here we consider all the variants because, in spite of their close compositions and structures, two variants can exhibit different activities. For example, actinomycin D is known to have antibiotic and antitumoral activities, but for the majority of other actinomycin variants, an antitumoral activity has not been reported. Figure 7 shows the distribution of six main activities found in curated peptides from the Norine database. Note that some peptides can show more than one activity. For example, there are 9 curated peptides in the Norine database that present both antibiotic and immunomodulator activities (such as edeines [12]) and 14 didemnins (66) that present three activities, antibiotic, antitumor, and immunomodulator.

We analyzed the monomeric distribution of NRPs depending on their activities. The distribution of the 30 most frequent monomers found in the Norine database-curated siderophores are presented in Fig. 8.

At neutral and alkaline pHs, ferric ions form insoluble polymeric hydroxide complexes that cannot be assimilated by microorganisms. To tackle this low iron bioavailability, many bacteria and fungi biosynthesize and excrete high-affinity iron

TABLE 3. Correlation coefficients between monomeric distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa

Organism 1	Organism 2	CC
Bacteria (488 peptides)	Fungi (269 peptides)	0.252
Bacteria (488 peptides)	Metazoa (218 peptides)	0.534
Fungi (269 peptides)	Metazoa (218 peptides)	0.359

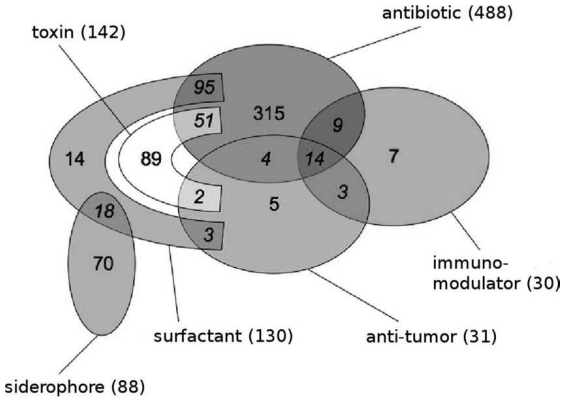


FIG. 7. Repartition of six main biological activities displayed by curated peptides in the Norine database (790 NRPs).

chelators known as siderophores. A common structural trait of most of these molecules is the presence of three bidentate groups which can ensure the 6-fold coordination of the ferric iron. Catecholate, hydroxamate, and hydroxycarboxylate groups are the three found most frequently that are able to play this role in the coordination of iron. Different monomers of NRPS products harbor such groups. For example, 2,3-dihydroxybenzoate (diOH-Bz, often denoted Dhb, which may result in confusion with another monomer, dehydrobutyrin, present in ribosomal bacteriocins like nisin or subtilin [19]), found in enterobactin or bacillibactin (1), or the chromophore of pyoverdins (ChrP, a dihydroxyquinoline group which results from the condensation and subsequent modification of diaminobutyric acid, tyrosine, and a dicarboxylic acid or its monoamide [7]) contains a catecholate group. *N*-Formyl-*N*-OH-ornithine (Fo-OH-Orn), found in ornibactins (61), *N*-acetyl-*N*-OH-ornithine (Ac-OH-Orn), found in aquachelins (69), *N*-OH-cyclo-ornithine (OH-cOrn), found in pseudobactins (65), OH-histidine (OH-His), found in corrugatin (50), and OH-lysine (OH-Lys), found in mycobactins (59) all contain an hydroxamate group. OH-aspartate (OH-Asp) of the azotobactin D (13) contains a hydroxycarboxylate group. Note that these monomers appear in the 30 most frequent siderophore monomers (Fig. 8). We also analyzed the monomeric distribution of NRPs in the other five classes of main activities (data not shown). The results suggest that the monomeric composition of a peptide can be used as a determiner of its biological activity. From this observation, we developed a method helping to predict the biological activity of a peptide from its monomeric composition.

(ii) Toward the prediction of biological activities. For each monomer of the database, we precomputed its frequency in each of the six activity classes. Then, for a given peptide, we computed the average frequency of its monomers to appear within each activity class and then deduced *P* values reflecting the probability of the peptide belonging to each class. The lower the *P* value for a given activity class, the more likely the peptide is to present this given activity.

We tested this method on several NRPs which have not yet been annotated in the Norine database. Table 4 gives some examples of resulting predictions.

Orfamide is a surfactant showing antibiotic activity (23). A

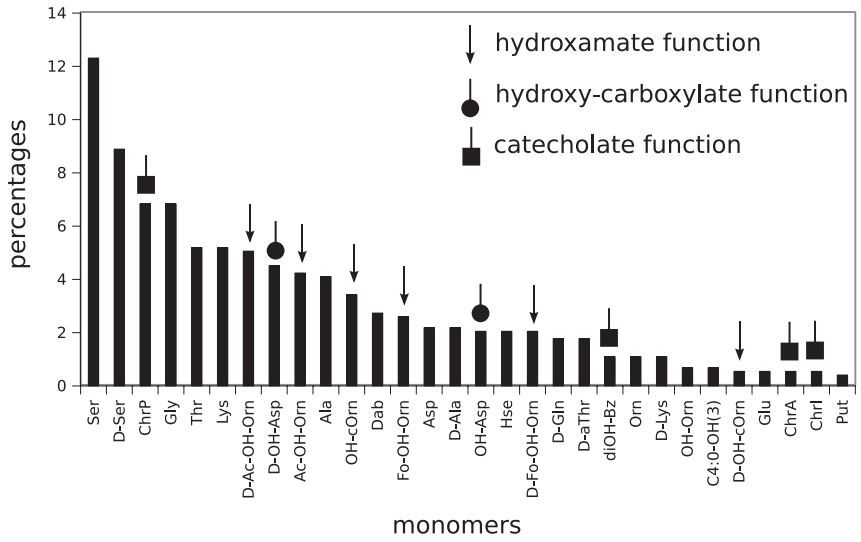


FIG. 8. Repartition of 30 of the most frequently found monomers in Norine database-curated siderophores (88 peptides). Hse, homoserine; ChrA, azotobactin chromophore; ChrI, isopyoverdin chromophore; Put, putrescine.

significant *P* value of 3.4E−4 is obtained for this peptide for the surfactant class, which suggests its surfactant properties. Fuscachelin is a nonribosomal siderophore (14), for which our method provided a strong indication. Aibellin is an antibiotic NRP (34). The *P* value obtained for the antibiotic class is close to 0, accurately predicting the biological activity of aibellin. These results suggest that the monomeric compositions of peptides can be of great help in predicting their biological activities.

DISCUSSION

General statistics. NRPs have numerous particularities. First, more than 500 different monomers are found in those peptides. Fatty acids, carbohydrates, and nonproteogenic amino acids are often incorporated in NRPs. We showed that in NRPs, a large part of monomers appear in their D-form. The D-residues play an important role in the structural conformation of the peptide, crucial for its biological activity and resistance against degradation process. NRPs can have a linear primary structure (less than 30% of peptides in the Norine database) but often present a more complex primary structure, including branchings and cycles. These nonlinear structures are essential for NRPs to have important biological activities, such as antibiotic and antitumor activities, preserving molecules from being degraded by proteases, which is essential for their biological functions.

Study of producing organisms. NRPS genes are experimentally identified only for bacteria and fungi. In this paper, we showed that NRPs produced by bacteria and those produced by fungi present different characteristics. The monomers used by bacteria are very different from those used by fungi. In addition, the NRP sizes are also very different for bacterial and fungal NRPs. These results showed that monomeric and other features of NRPs provide strong indications of their bacterial or fungal origin.

In addition to NRPs from fungi and bacteria, various peptides extracted from other organisms, such as metazoa, are believed to have a nonribosomal origin with regard to the presence of nonproteogenic amino acids (as in apramides [37], containing *N*-methyl-glycine-thiazole, or the cyclic barangamides [52], containing D-alloisoleucine) or their nonlinear structure. The current hypothesis is that the peptides extracted from metazoa are in fact produced by symbiotic bacteria rather than the metazoa themselves. Many recent papers tend to confirm this hypothesis by experimental validation of special cases of symbiosis (30, 36, 53, 70). In this paper, we have shown that monomers and structural features of NRPs produced by bacteria and those isolated from metazoa are very similar, which provides a supplementary argument supporting this hypothesis or the idea of a potential transfer of bacterial genes to metazoan genomes.

Study of biological activities. NRPs harbor a large spectrum of important biological activities. In this paper, we demonstrated that each activity class is associated with some specific monomers. By using frequencies of monomers in each activity class, we showed that the monomeric composition of a peptide

TABLE 4. Examples of biological activity predictions for some NRPs not annotated in the Norine database

Tested NRP	Known activity(ies)	*P* value obtained for the indicated class of activity					
		Antibiotic	Antitumor	Immunomodulator	Siderophore	Surfactant	Toxin
Orfamide	Surfactant, antibiotic	0.46	0.52	0.86	0.68	3.40E−04	0.29
Fuscachelin	Siderophore	0.65	0.82	0.96	4.90E−05	0.99	0.99
Aibellin	Antibiotic	<1.0E−30	0.98	0.97	0.97	0.97	0.93

can be an indicator of its biological activities. This can be of great interest for the study of new NRPs and can provide guidance for experimental studies.

Conclusion. For the last decade, the interest in NRPs and their biosynthetic enzymes has been considerably increased, as witnessed by an exponentially growing number of publications in this field. These peptides, indeed, are or can be used in many existing or potential biotechnological and pharmaceutical applications. A major part of published reviews focus on synthesis enzymes or on a genomic analysis, and much less work has been devoted to the global diversity of NRPs themselves. This situation led us to develop the Norine database, which is the only public resource devoted to NRPs and currently contains more than a thousand peptides. This tool allowed us to review, for the first time, the remarkable monomer diversity of these compounds. To our knowledge, no extensive study of monomers incorporated into NRPs has previously been done. In this paper, we presented the first large-scale analysis of monomers incorporated in NRPs. In addition to providing an overview of this biodiversity, this work demonstrated (i) a dissimilarity of structural properties between bacterial and fungal NRPs, (ii) a significant relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that the peptides isolated from sponges are in reality synthesized by symbiotic bacteria rather than by the sponges themselves, and (iii) a certain monomer specificity to classes of biological activities. An analysis of the monomer compositions of peptide products predicted from genomic information (metagenomics and high-throughput genome sequencing) can provide an indication of the origin of a peptide and/or its biological activity. Furthermore, new peptides detected by mass spectrometry analysis applied to a culture supernatant or carried out directly on colonies could be studied in the same way, leading to predictions concerning their origin or activities. Finally, those observations can be of great interest for developing combinatorial biosynthesis of NRPS and for the design of more-active antibiotic, immunomodulator, or anticancer drugs.

ACKNOWLEDGMENTS

This work was supported by the PPF bioinformatique program of Lille 1 University. S.C. was supported by an INRIA/Région Nord-Pas-de-Calais fellowship. The ProBioGEM lab is supported by the Région Nord-Pas-de-Calais, the Ministère de l'Enseignement Supérieur et de la Recherche, the French ANR Agency, and European Funds for Regional Development.

REFERENCES

1. **Abergel, R., A. Zawadzka, T. Hoette, and K. Raymond.** 2009. Enzymatic hydrolysis of trilactone siderophores: where chiral recognition occurs in enterobactin and bacillibactin iron transport. J. Am. Chem. Soc. **131:**12682–12692.
2. **Aravinda, S., N. Shamala, and P. Balaram.** 2008. Aib residues in peptaibiotics and synthetic sequences: analysis of non helical conformations. Chem. Biodivers. **5:**1238–1262.
3. **Arima, K., A. Kakinuma, and G. Tamura.** 1968. Surfactin, a crystalline peptidelipid surfactant produced by *Bacillus subtilis*: isolation, characterization and its inhibition of fibrin clot formation. Biochem. Biophys. Res. Commun. **31:**488–494.
4. **Bernardi, E., J. Fauchere, G. Atassi, P. Viallefont, and R. Lazaro.** 1993. Antitumoral cyclic peptide analogues of chlamydocin. Peptides **14:**1091–1093.
5. **Borel, J.** 2002. History of the discovery of cyclosporine and of its early pharmacological development. Wien. Klin. Wochenschr. **114:**433–437.
6. **Brennan, M., C. Costello, S. Maleknia, G. Pettit, and K. Erickson.** 2008. Stylopeptide 2, a proline-rich cyclodecapeptide from the sponge *Stylotella* sp. J. Nat. Prod. **71:**453–456.
7. **Budzikiewicz, H.** 2004. Siderophores of the *Pseudomonadaceae* sensu strict (fluorescent and non-fluorescent *Pseudomonas* spp.). Fortschr. Chem. Org. Naturst. **87:**81–237.
8. **Caboche, S., M. Pupin, V. Leclère, A. Fontaine, P. Jacques, and G. Kucherov.** 2008. Norine: a database of nonribosomal peptides. Nucleic Acids Res. **36:**D326–D331.
9. **Caboche, S., M. Pupin, V. Leclère, P. Jacques, and G. Kucherov.** 2009. Structural pattern matching of nonribosomal peptides. BMC Struct. Biol. **18:**9–15.
10. **Coleman, J., R. V. Soest, and R. Andersen.** 1999. New geodiamolides from the sponge *Cymbastela* sp. collected in Papua New Guinea. J. Nat. Prod. **62:**1137–1141.
11. **Consden, R., A. Gordon, and A. Martin.** 1947. Gramicidin S; the sequence of the amino-acid residues. Biochem. J. **41:**596–602.
12. **Czajgucki, Z., R. Andruszkiewicz, and W. Kamysz.** 2006. Structure activity relationship studies on the antimicrobial activity of novel edeine A and D analogues. J. Pept. Sci. **12:**653–662.
13. **Demange, P., A. Bateman, A. Dell, and M. Abdallah.** 1988. Structure of azotobactin D, a siderophore of *Azotobacter vinelandii* strain D (ccm289). Biochemistry **27:**2745.
14. **Dimise, E., P. Widboom, and S. Bruner.** 2008. Structure elucidation and biosynthesis of fuscachelins, peptide siderophores from the moderate thermophile *Thermobifida fusca*. Proc. Natl. Acad. Sci. U. S. A. **105:**15311–15316.
15. **Donadio, S., and M. Sosio.** 2008. Biosynthesis of glycopeptides: prospects for improved antibacterials. Curr. Top. Med. Chem. **8:**654–666.
16. **Duclohier, H.** 2007. Peptaibiotics and peptaibols: an alternative to classical antibiotics? Chem. Biodivers. **4:**1023–1026.
17. **Duitman, E., L. Hamoen, M. Rembold, G. Venema, H. Seitz, W. Saenger, F. Bernhard, R. Reinhardt, M. Schmidt, C. Ullrich, T. Stein, F. Leenders, and J. Vater.** 1999. The mycosubtilin synthetase of *Bacillus subtilis* atcc6633: a multifunctional hybrid between a peptide synthetase, an aminotransferase, and a fatty acid synthase. Proc. Natl. Acad. Sci. U. S. A. **96:**13294–13299.
18. **Duval, D., P. Cosette, S. Rebuffat, H. Duclohier, B. Bodo, and G. Molle.** 1998. Alamethicin-like behaviour of new 18-residue peptaibols, trichorzins PA. Role of the C-terminal amino-alcohol in the ion channel forming activity. Biochim. Biophys. Acta **1369:**309–319.
19. **Entian, K., and W. de Vos.** 1996. Genetics of subtilin and nisin biosyntheses: biosynthesis of lantibiotics. Antonie Van Leeuwenhoek **69:**109–117.
20. **Fusetani, N., and S. Matsunaga.** 1990. Cyclotheonamides, potent thrombin inhibitors, from a marine sponge *Theonella* sp. J. Am. Chem. Soc. **112:**7053–7054.
21. **Gathercole, P., and P. Thiel.** 1987. Liquid chromatographic determination of the cyanoginosins, toxins produced by the cyanobacterium *Microcystis aeruginosa*. J. Chromatogr. **408:**435–440.
22. **Gause, G., M. Brazhnikova, N. Lomakina, T. Berdnikova, G. Fedorova, N. Tokareva, V. Borisova, and G. Batta.** 1989. Eremomycin: new glycopeptide antibiotic: chemical properties and structure. J. Antibiot. (Tokyo) **42:**1790–1797.
23. **Gross, H., V. Stockwell, M. Henkels, B. Nowak-Thompson, J. Loper, and W. Gerwick.** 2007. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem. Biol. **14:**53–63.
24. **Grubb, D., E. Wolvetang, and A. Lawen.** 1995. Didemnin B induces cell death by apoptosis: the fastest induction of apoptosis ever described. Biochem. Biophys. Res. Commun. **215:**1130–1136.
25. **Hamada, T., S. Matsunaga, G. Yano, and N. Fusetani.** 2005. Polytheonamides A and B, highly cytotoxic, linear polypeptides with unprecedented structural features, from the marine sponge, *Theonella swinhoei*. J. Am. Chem. Soc. **127:**110–118.
26. **Hamann, M., C. Otto, P. Scheuer, and D. Dunbar.** 1996. Kahalalides: bioactive peptides from a marine mollusk *Elysia rufescens* and its algal diet *Bryopsis* sp. J. Org. Chem. **61:**6594–6600.
27. **Hanessian, S., M. Tremblay, and J. Petersen.** 2004. The N-acyloxyiminium ion aza-prins route to octahydroindoles: total synthesis and structural confirmation of the antithrombotic marine natural product oscillarin. J. Am. Chem. Soc. **126:**6064–6071.
28. **Hubbard, B., and C. Walsh.** 2003. Vancomycin assembly: nature's way. Angew. Chem. Int. Ed. Engl. **42:**730–765.
29. **Kallow, W., T. Neuhof, B. Arezi, P. Jungblut, and H. von Döhren.** 1997. Penicillin biosynthesis: intermediates of biosynthesis of delta-alpha-aminoadipyl-cysteinyl-valine formed by ACV synthetase from *Acremonium chrysogenum*. FEBS Lett. **414:**74–78.
30. **Kennedy, J., P. Baker, C. Piper, P. Cotter, M. Walsh, M. Mooij, M. Bourke, M. Rea, P. O'Connor, R. Ross, C. Hill, F. O'Gara, J. Marchesi, and A. Dobson.** 2009. Isolation and analysis of bacteria with antimicrobial activities from the marine sponge *Haliclona simulans* collected from Irish waters. Mar. Biotechnol. **11:**384–396.
31. **Kessler, N., H. Schuhmann, S. Morneweg, U. Linne, and M. Marahiel.** 2004. The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. J. Biol. Chem. **279:**7413–7419.

32. **Kleinkauf, H., and H. V. Döhren.** 1996. Polytheonamide, the longest peptide reported of presumably enzymatic origin. J. Biochem. **236:**335–351.
33. **Kobayashi, J., T. Nakamura, and M. Tsuda.** 1996. Hymenamide F, new cyclic heptapeptide from marine sponge *Hymeniacidon* sp. Tetrahedron **52:** 6355–6360.
34. **Kumazawa, S., M. Kanda, H. Aoyama, M. Utagawa, J. Kondo, S. Sakamoto, H. Ohtani, T. Mikawa, I. Chiga, and T. Hayase.** 1994. Structural elucidation of aibellin, a new peptide antibiotic with efficiency enhancing activity on rumen fermentation. J. Antibiot. (Tokyo) **47:**1136–1144.
35. **Li, H., T. Tanikawa, Y. Sato, Y. Nakagawa, and T. Matsuyama.** 2005. Serratia marcescens gene required for surfactant serrawettin W1 production encodes putative aminolipid synthetase belonging to nonribosomal peptide synthetase family. Microbiol. Immunol. **49:**303–310.
36. **Luesch, H., G. Harrigan, G. Goetz, and F. Horgen.** 2002. The cyanobacterial origin of potent anticancer agents originally isolated from sea hares. Curr. Med. Chem. **9:**1791–1806.
37. **Luesch, H., W. Yoshida, R. Moore, and V. Paul.** 2000. Apramides A–G, novel lipopeptides from the marine cyanobacterium *Lyngbya majuscula*. J. Nat. Prod. **63:**1106–1112.
38. **Marahiel, M., and L. Essen.** 2009. Nonribosomal peptide synthetases mechanistic and structural aspects of essential domains. Methods Enzymol. **458:** 337–351.
39. **Martinez, J., and A. Butler.** 2007. Marine amphiphilic siderophores: marinobactin structure, uptake, and microbial partitioning. J. Inorg. Biochem. **101:**1692–1698.
40. **Ming, L., and J. Epperson.** 2002. Metal binding and structure-activity relationship of the metalloantibiotic peptide bacitracin. J. Inorg. Biochem. **91:** 46–58.
41. **Mootz, H., D. Schwarzer, and M. Marahiel.** 2002. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. Chembiochem **3:**490–504.
42. **Pettit, G., Z. Cichacz, J. Barkoczy, A. Dorsaz, D. Herald, M. Williams, D. Doubek, J. Schmidt, L. Tackett, and D. Brune.** 1993. Isolation and structure of the marine sponge cell growth inhibitory cyclicpeptide phakellistatin 1. J. Nat. Prod. **56:**260–267.
43. **Pettit, G., B. Toki, J. Xu, and D. Brune.** 2000. Synthesis of the marine sponge cycloheptapeptide phakellistatin 5. J. Nat. Prod. **63:**22–28.
44. **Peypoux, F., J. Bonmatin, H. Labbé, B. Das, M. Ptak, and G. Michel.** 1991. Isolation and characterization of a new variant of surfactin, the [Val7]surfactin. Eur. J. Biochem. **202:**101–106.
45. **Pfennig, F., F. Schauwecker, and U. Keller.** 1999. Molecular characterization of the genes of actinomycin synthetase I and of a 4-methyl-3-hydroxyanthranilic acid carrier protein involved in the assembly of the acylpeptide chain of actinomycin in *Streptomyces*. J. Biol. Chem. **274:**12508–12516.
46. **Picur, B., M. Cebrat, J. Zabrocki, and I. Siemion.** 2006. Cyclopeptides of *Linum usitatissimum*. J. Pept. Sci. **12:**569–574.
47. **Poncet, J.** 1999. The dolastatins, a family of promising antineoplastic agents. Curr. Pharm. Des. **5:**139–162.
48. **Rashid, M., K. Gustafson, J. Boswell, and M. Boyd.** 2000. Haligramides A and B, two new cytotoxic hexapeptides from the marine sponge *Haliclona nigra*. J. Nat. Prod. **63:**956–959.
49. **Recktenwald, J., R. Shawky, O. Puk, F. Pfennig, U. Keller, W. Wohlleben, and S. Pelzer.** 2002. Nonribosomal biosynthesis of vancomycin-type antibiotics: a heptapeptide backbone and eight peptide synthetase modules. Microbiology **148:**1105–1118.
50. **Risse, D., H. Beiderbeck, K. Taraz, H. Budzikiewicz, and D. Gustine.** 1998. Corrugatin, a lipopeptide siderophore from *Pseudomonas corrugata*. Z. Naturforsch. C **53:**295–304.
51. **Roongsawang, N., K. Hase, M. Haruki, T. Imanaka, M. Morikawa, and S. Kanaya.** 2003. Cloning and characterization of the gene cluster encoding arthrofactin synthetase from *Pseudomonas* sp. MIS38. Chem. Biol. **10:**869–880.
52. **Roy, M., I. Ohtani, J. Tanaka, T. Higa, and R. Satari.** 1999. Barangamide A, a new cyclic peptide from the Indonesian sponge *Theonella swinhoei*. Tetrahedron Lett. **40:**5373–5376.
53. **Salomon, C., N. Magarvey, and D. Sherman.** 2004. Merging the potential of microbial genetics with biological and chemical diversity: an even brighter future for marine natural product drug discovery. Nat. Prod. Rep. **21:**105–121.
54. **Sanchez, M., R. Wenzel, and R. Jones.** 1992. In vitro activity of decaplanin (M86-1410), a new glycopeptide antibiotic. Antimicrob. Agents Chemother. **36:**873–875.
55. **Sano, T., T. Usui, K. Ueda, H. Osada, and K. Kaya.** 2001. Isolation of new protein phosphatase inhibitors from two cyanobacteria species, *Planktothrix* spp. J. Nat. Prod. **64:**1052–1055.
56. **Segre, A., R. Bachmann, A. Ballio, F. Bossa, I. Grgurina, N. Iacobellis, G. Marino, P. Pucci, M. Simmaco, and J. Takemoto.** 1989. The structure of syringomycins A1, E and G. FEBS Lett. **255:**27–31.
57. **Shen, B., L. Du, C. Sanchez, D. Edwards, M. Chen, and J. Murrell.** 2001. The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillus* ATCC 15003 as a model for hybrid peptide-polyketide natural product biosynthesis. J. Ind. Microbiol. Biotechnol. **27:**378–385.
58. **Shiki, Y., M. Onai, D. Sugiyama, S. Osada, I. Fujita, and H. Kodama.** 2009. Synthesis and biological activities of cyclic peptide, hymenamide analogs. Adv. Exp. Med. Biol. **611:**323–324.
59. **Snow, G.** 1970. Mycobactins: iron-chelating growth factors from mycobacteria. Bacteriol. Rev. **34:**99–125.
60. **Steenbergen, J., J. Alder, G. Thorne, and F. Tally.** 2005. Daptomycin: a lipopeptide antibiotic for the treatment of serious gram-positive infections. J. Antimicrob. Chemother. **55:**283–288.
61. **Stephan, H., S. Freund, W. Beck, G. Jung, J. Meyer, and G. Winkelmann.** 1993. Ornibactins—a new family of siderophores from *Pseudomonas*. Biometals **6:**93–100.
62. **Suenaga, K., S. Kajiwara, S. Kuribayashi, T. Handa, and H. Kigoshi.** 2008. Synthesis and cytotoxicity of aurilide analogs. Bioorg. Med. Chem. Lett. **18:**3902–3905.
63. **Takeuchi, M., S. Takahashi, R. Enokita, Y. Sakaida, H. Haruyama, T. Nakamura, T. Katayama, and M. Inukai.** 1992. Galacardins A and B, new glycopeptide antibiotics. J. Antibiot. (Tokyo) **45:**297–305.
64. **Tan, L., N. Sitachitta, and W. Gerwick.** 2003. The guineamides, novel cyclic depsipeptides from a Papua New Guinea collection of the marine cyanobacterium *Lyngbya majuscula*. J. Nat. Prod. **66:**764–771.
65. **Teintze, M., and J. Leong.** 1981. Structure of pseudobactin A, a second siderophore from plant growth promoting *Pseudomonas* B10. Biochemistry **20:**6457–6462.
66. **Vera, M., and M. Joullié.** 2002. Natural products as probes of cell biology: 20 years of didemnin research. Med. Res. Rev. **22:**102–145.
67. **Visca, P., F. Imperi, and I. Lamont.** 2007. Pyoverdine siderophores: from biogenesis to biosignificance. Trends Microbiol. **15:**22–30.
68. **Yeung, B., Y. Nakao, R. Kinnel, J. Carney, W. Yoshida, P. Scheuer, and M. Kelly-Borges.** 1996. The kapakahines, cyclic peptides from the marine sponge *Cribrochalina olemda*. J. Org. Chem. **61:**7168–7173.
69. **Zhang, G., S. Amin, F. Küpper, P. Holt, C. Carrano, and A. Butler.** 2009. Ferric stability constants of representative marine siderophores: marinobactins, aquachelins, and petrobactin. Inorg. Chem. **48:**11466–11473.
70. **Zhang, W., Z. Li, X. Miao, and F. Zhang.** 2009. The screening of antimicrobial bacteria with diverse novel nonribosomal peptide synthetase (NRPS) genes from South China Sea sponges. Mar. Biotechnol. **11:**346–355.

J Comput Aided Mol Des (2012) 26:1187–1194
DOI 10.1007/s10822-012-9608-4

A new fingerprint to predict nonribosomal peptides activity

Ammar Abdo · Ségolène Caboche · Valérie Leclère · Philippe Jacques · Maude Pupin

Received: 18 July 2012 / Accepted: 20 September 2012 / Published online: 29 September 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract Bacteria and fungi use a set of enzymes called nonribosomal peptide synthetases to provide a wide range of natural peptides displaying structural and biological diversity. So, nonribosomal peptides (NRPs) are the basis for some efficient drugs. While discovering new NRPs is very desirable, the process of identifying their biological activity to be used as drugs is a challenge. In this paper, we present a novel peptide fingerprint based on monomer composition (MCFP) of NRPs. MCFP is a novel method for obtaining a representative description of NRP structures from their monomer composition in fingerprint form. Experiments with Norine NRPs database and MCFP show high prediction accuracy (>93 %). Also a high recall rate (>82 %) is obtained when MCFP is used for screening NRPs database. From this study it appears that our fingerprint, built from monomer composition, allows an effective screening and prediction of biological activities of NRPs database.

Keywords Nonribosomal peptides · Target Prediction · Similarity searching · Drug discovery

A. Abdo (✉) · S. Caboche · M. Pupin
LIFL UMR CNRS 8022 Université Lille 1 and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq cedex, France
e-mail: ammar_utm@yahoo.com

A. Abdo
Computer Science Department, Hodeidah University, Al Hudaydah, Yemen

S. Caboche · V. Leclère · P. Jacques
ProBioGEM, UPRES EA 1026, Polytech'Lille, Av P. Langevin, Univ Lille 1-Sciences et Technologies, 59655 Villeneuve d'Ascq cedex, France

Introduction

For thousands years, natural products are an important source of drugs [1]. They are produced by marine or terrestrial organisms (plants, vertebrates, invertebrates…) and microorganisms (fungi, bacteria, algae). Many studies in the literature discuss the importance of natural products in drug discovery [2–5]. They are still important sources for many drugs in the market (e.g. morphine, cocaine, penicillin, taxols…) and are also good lead compounds suitable for further modification during drug development. Introducing a new compound on the market is time consuming and cost-intensive process [6, 7], in particular for natural products, so that strategies allowing time saving are welcomed.

The discovery of natural products requires specific steps as they are synthesized by living organisms. For example, scientists need to determine which organisms produce interesting compounds and define the conditions of production. The produced compounds have to be extracted from cultured media or from natural environments. Finally, chemical structures are determined. Those structures can, finally, be mimicked leading to artificial compounds. To reduce the time and cost of the specific steps, the optimal process is to predict the compounds produced by an organism directly from its genome sequence. This strategy can be particularly performed with nonribosomal peptides.

Those peptides are synthesized by a ribosome-independent cell machinery. This alternative pathway produces peptides using large multi-enzymatic complexes called nonribosomal synthetases (NRPSs) [8]. Those synthetases are composed of proteins organized in modules, each one being responsible for the incorporation of one specific amino acid in the final peptide. A relationship between specific signatures and a given incorporated amino acid

have been determined from protein sequences of NRPSs [9–12]. So, from a genome sequence, bioinformatics analysis allows to extract genes coding for NRPSs, to deduce their protein sequences and to predict the amino acids incorporated in the produced peptide [13]. This predicted peptide can then be analyzed by bioinformatics tools to infer its putative activity.

We have collected nonribosomal peptides in Norine (http://bioinfo.lifl.fr/norine/) [14], the first and still unique computational resource dedicated to nonribosomal peptides (NRPs). Each peptide has a unique Norine identifier in the form NOR followed by a number of 5 digits. The database contains more than 1,100 nonribosomal peptides extracted from scientific literature with manually curated annotations such as biological activity, producing organisms or bibliographic references and, most importantly, their monomeric structure. We used the universal term monomer instead of amino acid because the entities encountered into those peptides do not only include the 21 proteogenic amino acids, but also derivates or unusual ones; other compounds such as carbohydrates or lipids can also be incorporated. Norine currently references 526 different monomers occurring in the listed peptides. The monomeric structures are encoded by undirected labelled graphs, with nodes representing monomers and edges corresponding to chemical bonds between them. One monomer can display more than two peptidic bonds, and non peptidic bonds are also observed in NRPs leading to peptides with cycles and/or branches. The database can be queried for peptide search through their annotations as well as through their monomeric structures. It also contains a section dedicated to the monomers incorporated into the peptides stored in Norine.

Due to the particular way of synthesis, nonribosomal peptides are a valuable source of a wide range of structural and biological activities, produced by microbial cells (typically bacteria and fungi). The NRPs may represent novel drugs for several pharmaceutical areas including antibiotics (penicillin and cephalosporin the precursor of which is ACV, NOR00006), antitumors (actinomycin D, NOR00228), and immunosuppressive agents (cyclosporin A, NOR00033). They can also be exploited in biotechnological applications such as biosurfactants. Their various and interesting biological activities almost comes from their original mode of synthesis that offers huge flexibility by including non proteogenic monomers and cycles and branching.

As they are small and exploited in pharmacology and biotechnology, nonribosomal peptides are usually represented by atomic structures and stored in chemical compounds databases. Classical chemo-informatics tools are applied to them as part of generalist chemical databases to predict their activity or do some structure search or comparison. Norine contains few links to structural conformation databases such as PDB (25 NRPs). However, the length of this data set is too low to be exploited for NRP comparison or activity prediction.

Due to the similar property principle, structurally similar compounds are expected to exhibit similar properties and similar biological activities. This principle is exploited for in silico drug discovery. The chemical compounds are virtually screened either by docking into the active site of interest or by virtue of their similarity to a known active. Many studies suggest that knowledge about a target obtained from known bioactive ligand is as valuable as knowledge of the target structures for identifying novel bioactive scaffolds through virtual screening [15, 16].

But, NRPs exhibit specificities in comparison to typical synthetic compounds (synthesis pathway, complex structures). So, published numerical representations for chemical compounds, such as fingerprints, may not be the optimal choice to represent NPRs. Our monomeric approach opens new ways to analyze them. As first observations showing that some monomers are specific to a given activity [17] were promising, we decided to further investigate the relationship between the NRP monomer structures and their activity.

In this paper a new fingerprint based on monomeric composition of NRPs is introduced. Monomer composition fingerprint (MCFP) is a new method for obtaining a representative description of NRP structures from their monomer composition in fingerprint form. In this work, we present experiments that show the usefulness of monomer composition fingerprint when used for similarity searching and activity prediction of NRPs.

Materials and methods

Monomer composition fingerprint (MCFP)

MCFP is represented as an integer vector, in which each element represents the presence (number of occurrences) or absence (“0” value) of a specific monomer. The process of generating the MCFP for each peptide starts by extracting the monomer compositions from Norine and then filling the corresponding positions in the MCFP vector. We use the 526 monomers referenced in Norine as individual elements (see Fig. 1). For example, the peptaibolin (NOR01028) is composed of the monomers NAc-Leu (N-acetyl-leucine), Aib (2-aminoisobutyric acid), Leu (leucine), Aib (2-Aminoisobutyric acid) and Pheol (phenylalaninol) and generates a fingerprint with three elements set to “1”, one to “2” and the rest (522) set to “0”. Four elements are “on” for this peptide of length five because the monomer Aib is repeated twice.

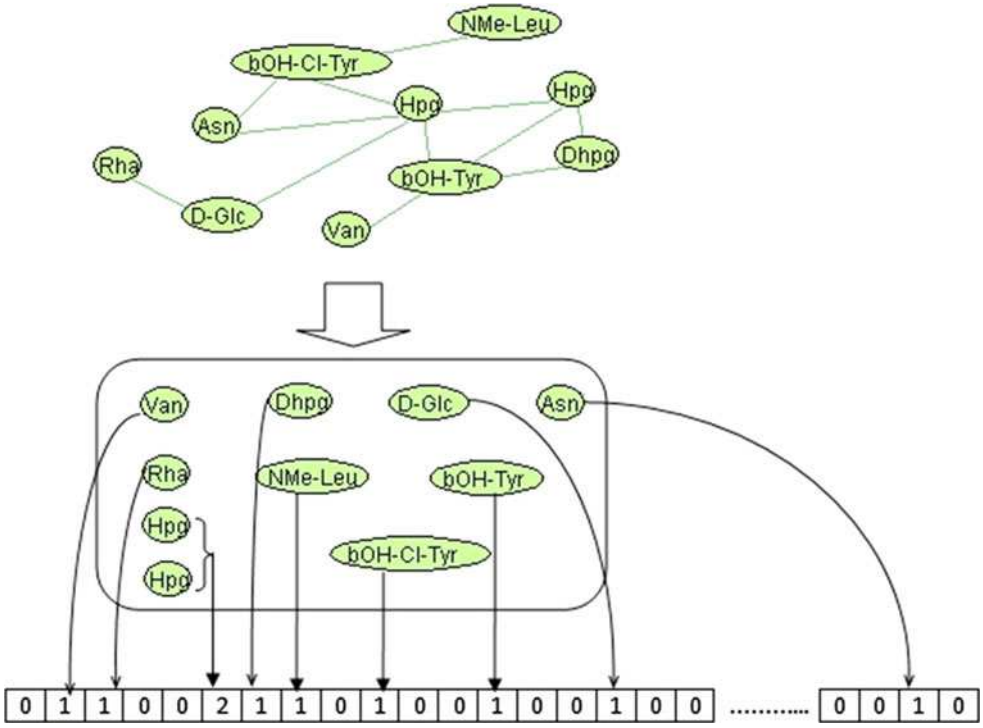


Fig. 1 MCFP generation process

Similarity search system

We use Tanimoto-based similarity search system (TAN). This system is based on the Tanimoto coefficient that is a well established method in similarity-based virtual screening and was therefore used as reference. In particular, the continuous form of the Tanimoto coefficient was used. If A_i and B_i represent the ith monomer occurrence in the peptides A and B, respectively, the similarity score $S_{A,B}$ between peptides A and B was calculated by the following equation.

$$S_{A,B} = \sum_{i=1}^{n} A_i B_i \Big/ \sum_{i=1}^{n} (A_i)^2 + \sum_{i=1}^{n} (B_i)^2 - \sum_{i=1}^{n} A_i B_i \qquad (1)$$

The advantage of this score is the direct use of monomer occurrences in the equation and the neutrality of empty elements. This equation has been widely used for chemical similarity searching. However, a detailed study of fragment weighting schemes has recently suggested that superior screening performance is obtained if the square roots of the element occurrence frequencies are used rather than the unmodified frequencies [18–21]. We have hence carried out experiments in which the raw monomer occurrences in the TAN similarity measures are replaced by the square roots of those occurrences. The TAN coefficient varies between 0 (totally different monomer compositions) and 1 (identical monomer compositions).

Activity prediction system

We use in our experiments three machine learning algorithms available in WEKA-Workbench [22, 23]. The naive Bayesian classifier [24], the linear (LibLinear) classifier [25], and the SMO classifier [26]. Details on these algorithms can be found in their references. The machine learning algorithms are used with their default settings in the WEKA-Workbench.

Data sets

The data set for this study is taken from the Norine database (version of April 2012), which contains 1122 peptides with 11 distinct activities. We don't consider the surfactant activity as it is more a physico-chemical property (being a lipopeptide or not) than a biological activity. The database is first filtered so that, activity classes containing less than 20 peptides are removed. Then, peptides with same monomer lists, even with different number of occurrences (same elements "on" in the MCFP), within an activity class are removed. Finally, we only consider the peptides with only one known activity. A total of 605 peptides were available for forming our test set, belonging to 5 different activity classes.

(1) The **antibiotics class** (319 NRPs) includes different NRPs categories, which are peptaibols (linear peptides

produced by fungi), glycopeptides (vancomycin-like with several cycles in their monomer structure), lipopeptides, pure peptides and even chromopeptides. It is to notice that, in Norine, 210 peptides share antibiotic with other activities (antitumor, toxin, surfactant or immuno-modulator). Those 210 peptides are included in the evaluation data set (see discussion section).

(2) **Toxins** (157 NRPs) harbor different modes of action to kill cells. They are pure peptides or lipopeptides. In Norine, 103 NRPs that are toxins are also antibiotics, antitumors or surfactants. They are also in the evaluation data set.

(3) **Siderophores** (82 NRPs) chelate (bind) iron molecules with specific monomers, including chromophores. They are mainly chromopeptides, but can also be lipopeptides or pure peptides. Among the 82 siderophore peptides, 18 are also known as surfactants.

(4) **Antitumors** (25 NRPs) operate with different modes of action, being mainly pure peptides. In Norine, 71 NRPs that are antitumors are also antibiotics, toxins or immuno-modulator. They are also in the evaluation data set.

(5) **Protease inhibitors** (22 NRPs) are all pure peptides. This class never crosses with other classes, as far as we know.

Performance of machine learning algorithms depends on the training data set (peptides with or without a given activity). The negative set, peptides without the studied activity, for any single activity class derives from the positive sets, that are peptides having any other activity.

Validation

The similarity searching experiments were performed with 20 peptides selected randomly (as queries) from each activity class. The recall results were averaged over each such set of active peptides. The recall is the percentage of active peptides retrieved in the top-1 % or the top-5 % of the ranked list resulting from a similarity search.

For activity prediction experiments, 10-fold cross-validation was used to validate the results of different machine learning algorithms. In this cross-validation, the data set is split into 10 parts; one part is used for testing, the remaining 9 parts for training. This is repeated 10 times, so all the data have been used as test data once. Each activity class is tested against all the others, grouped. As in the case of many prediction methods, we used the F-measure as quality criterion to quantify the performance of MCFP with different classification algorithms. F-measure is defined as the harmonic mean of precision and recall. The precision is defined by $prec = tp/(tp + fp)$ and the recall (or sensitivity) is defined by $rec = tp/(tp + fn)$, where tp, fp and fn are the number of true positives, false positives, and false negatives, respectively. We also used accuracy (ac) and area under the Receiver Operating Characteristic (ROC) curve (AUC) measures to quantify the performance of MCFP with different classification models. Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications $ac = (tp + tn)/(tp + tn + fp + fn)$.

Further metrics of statistical performance analysis involved the ROC curve, which has been used in various fields (medicine, meteorology, etc.) [27] and also in drug discovery field [28]. A ROC curve describes the tradeoff between sensitivity and specificity, where the sensitivity is defined as the ability of the model to avoid false negatives, and the specificity relates to its ability to avoid false positives. The area under the ROC curve (AUC) is a measure of the model performance: the closer to 1, the better is the performance of the prediction.

Results

Similarity-based results

Details of the pairwise similarities among the activity classes are given in Table 1. A rough guide to the diversity of each of the chosen sets of NRPs is provided by matching each peptide with every other in its activity class (intra-class) or with all the 605 used in this study (inter-class), calculating the Tanimoto coefficient applied to MCFP. The class diversity is measured by computing the mean and the number of comparisons having a coefficient greater than or equal to 0.7 for these intra-class similarities. The histogram of Fig. 2 gives an overview of the pairwise distances obtained among intra- and inter-classes. The number of pairwise comparisons with a high score is low for all the classes, confirming a high diversity.

Table 1 Pairwise similarity and retrieval results for Tanimoto coefficient

Activity class	NRPs number	Pairwise TAN		TAN recall	
		Mean	% ≥0.7	Top 1 %	Top 5 %
Antibiotics	319	0.09	3.69	88.33	81.50
Toxin	157	0.09	1.65	75.00	59.33
Siderophore	82	0.18	2.11	100.00	90.83
Antitumor	25	0.27	8.67	67.50	45.21
Protease inhibitors	22	0.26	9.52	80.83	56.90
All against all	605	0.05	1.21		
Mean				82.33	66.75

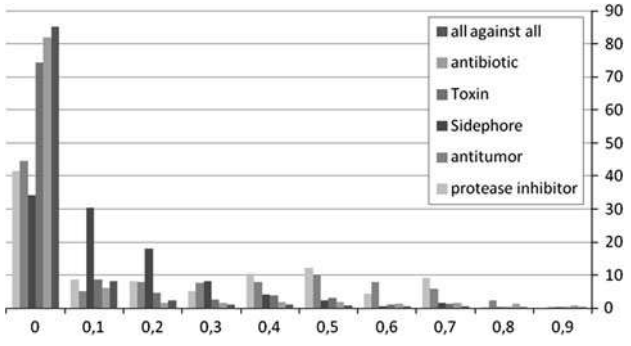


Fig. 2 Histogram for pairwise similarity using Tanimoto coefficient

The results for the searches in the data set are shown in Table 1. Each row corresponds to one activity class and lists the recall for the top 1 and 5 % of a sorted ranking when averaged over the ten searches for this activity class.

Results reported in Table 1 show that TAN system with MCFP obtains overall average recall rates of 82 and 67 % for top 1 and 5 %, respectively. It has the best performance for siderophore, antibiotics, and protease inhibitors activity classes while performing least well for antitumor and toxin. We observe a diminution of the recall between top-1 % and top-5 %.

Biological activity prediction results

Visual inspection of the precision, recall, F-measure and accuracy rates in Table 2 enables one to make comparisons between the effectiveness of using MCFP with various prediction models. The MCFP with LibLinear approaches produce the best performance across the five activity classes, with SMO and NaiveB also performing well. In only one class (antitumor activity), the performance of the MCFP with different prediction models was low. In terms of the overall correctness of the prediction, MCFP fingerprint with different prediction approaches produced high accuracy rates, especially with LibLinear model (>93 %).

In this study we used the ROC curve to study the performance of MCFP with different prediction models. Table 2 shows that the AUC value is always close to 1 (>0.93).

Discussion

The main aim of this study is to introduce the monomer composition fingerprint as a useful representation for NRPs and then identify the effectiveness of using such representation in similarity-based and prediction of the activity for those peptides displaying many different biological activities. The best selection of descriptors/fingerprints is based on their accuracy in predicting the property/activity of a peptide from another peptide that is considered similar to it, by using either a similarity method, or a clustering or its k-nearest neighbors. For those descriptors, and for predicting the activity class of peptides, the best descriptors are those yielding the highest number of correct predictions (peptides with similar activity class), taking into account the total number of peptides having this activity in the database used. To achieve this aim, the Tanimoto similarity system (TAN, see Eq. 1) and three different machine learning approaches (NaiveB, LibLinear, and SMO) have been applied.

The TAN calculated on monomer composition fingerprint demonstrates good results for the recall computed on the top-1 %, except for the toxin and antitumor classes. The toxin class has only 14 % of specific monomers and shares up to 81 % of its monomers with the antibiotic class (see Table 4). So, they can match with antibiotics or other peptides because of their common monomers. This is not surprising as those activities are biologically closed and can even be both harbored by a single peptide (72 peptides of Norine are known to be antibiotics and toxins, we tested them as an evaluation data set). This is even worse for antitumors that have no specific monomers and share 96 % of their monomers with antibiotics and toxins. Their TAN recall is lower than the one of toxin. At the opposite, protease inhibitors have also no specific monomers and share 88 % of their monomers with antibiotics and toxins, but show the third best recall of the set. This is certainly

Table 2 Precision, recall, F-measure, accuracy and AUC rates for the prediction models

Activity class	Naïve Bayesian				LibLinear				SMO			
	Prec	Rec	F	AUC	Prec	Rec	F	AUC	Prec	Rec	F	AUC
Antibiotics	0.971	0.737	0.838	0.961	0.950	0.962	0.956	0.953	0.947	0.953	0.950	0.942
Toxin	0.656	0.898	0.758	0.946	0.899	0.904	0.902	0.934	0.889	0.917	0.902	0.937
Siderophore	0.890	0.988	0.936	0.998	0.988	0.963	0.975	0.981	1	0.951	0.975	0.994
Antitumor	0.471	0.640	0.542	0.935	0.696	0.64	0.667	0.814	0.696	0.640	0.667	0.868
Protease inhibitors	0.870	0.909	0.889	0.996	0.952	0.909	0.930	0.954	0.952	0.909	0.930	0.975
Accuracy	81.49				93.22				92.89			

Table 3 Confusion matrix for different prediction models

Activity class	Naïve Bayesian					LibLinear					SMO				
	a	b	c	d	e	a	b	c	d	e	a	b	c	d	e
a	235	64	9	11	0	307	9	0	3	0	304	11	0	4	0
b	5	141	1	7	3	9	142	1	4	1	9	144	0	3	1
c	0	1	81	0	0	1	2	79	0	0	2	2	78	0	0
d	2	7	0	16	0	6	3	0	16	0	4	5	0	16	0
e	0	2	0	0	20	0	2	0	0	20	2	0	0	0	20

a antibiotics, b Toxin, c Siderophore, d Antitumor, e Protease inhibitors

Table 4 Percentages of common and specific monomers

	Antibiotics (%)	Toxin (%)	Siderophore (%)	Protease inhibitors (%)	Antitumor (%)
Antibiotics	**38**	55	22	26	13
Toxin	81	**14**	24	20	39
Siderophore	74	55	**26**	32	13
Protease inhibitors	96	96	36	**0**	29
Antitumor	88	88	25	50	**0**

You should read the table by row. For example, antibiotics share 55 % of their monomers with toxins; antibiotics have 38 % of specific monomers. The sum of the rows is not equal to 100 % because some monomers are shared between several classes

The numbers in bold are the percentages of monomers specific of each activity

due to the fact that they are small peptides (3, 4, 5 or 7 monomers) in comparison to the other peptides (mean number of monomers is around 10) and that their composition is specific of their activity. It is to notice that no peptide of Norine share protease inhibitor activity with another activity. Finally, antibiotics have the second best recall (88 %, in top 1 %), but it is not so good (as siderophore) because, as mentioned before, antibiotic class is constituted by several sub-groups that differ in monomer composition, structure and mode of action (they are peptaibols, glycopeptides, lipopeptides, pure peptides or chromopeptides). But the number of NRPs in each subclass is sufficiently high to find similar peptides in top-1 % and top-5 % lists. Generally, the recall results presented here are highly interesting and promising. That is because this data set comprises heterogeneous activity classes which are normally considered as very challengeable in similarity-based searching. We plan to study more deeply the intra-class similarities to distinguish sub-classes among the actual activity classes, if some can be designed.

The prediction accuracy rates obtained with the three machine learning approaches are promising because they are higher than 90 %. Again, and for the same reasons, antitumor class gives lower rates. However, the mispredicted cases in antitumor class (see Table 3) are not really incorrect. This is because these cases are predicted as antibiotic and toxin classes and as we mentioned above, in Norine, NRPs that are antitumors can also be antibiotics and toxins. This finding is also supported by the number of common monomers between antitumor, antibiotic and toxin classes (Table 4). We plan to study the data sets within each class and across classes to improve the predictions. For example, isolated peptides can be removed from the classes.

In order to assess the true predictivity of any model it is necessary to have an independent data set (evaluation data set) against which the model predictions can be compared. The evaluation data sets are different from the training data sets used to build the model. This approach makes it possible for users to judge the robustness and predictivity of the model when making predictions. Therefore, we predict the activity of 5 peptides that are not yet included in Norine

Table 5 Description and results for peptides that are not in Norine

Name	Ref.	Known activities	Predicted activity	Monomer composition
Coelichelin	[29]	Siderophore	**Siderophore**	D-Fo-OH-Orn, D-aThr, OH-Orn, D-Fo-OH-Orn
Hypomurocin A1	[30]	Antibiotic	**Antibiotic**	Ac-Aib, Gln, Val, Val, Aib, Pro, Leu, Leu, Aib, Pro, Leuol
Orfamide A	[31]	Antibiotic	Toxin	C14:0-OH(3); Leu,D-Glu, D-aThr, D-aIle, Leu,D-Ser, Leu, Leu, D-Ser, Val
Pyoverdin PSEN	[32]	Siderophore	**Siderophore**	ChrP, D-Ala, Asn,Dab, OH-His, Gly, Gly, Ser, Thr, D-Ser, OH-cOrn
TVB I	[33]	Antibiotic	**Antibiotic**	Ac-Aib, Gly, Ala, Val, Aib, Gln, Aib, Ala, Aib, Ser, Leu, Aib, Pro, Leu, Aib, Aib, Gln, Valol

The good predictions are in bold

Table 6 Results for evaluation data set extracted from Norine

NRPs number	Known activities	Predicted activity
62	Antibiotic, toxin	**32 antibiotic**
		29 toxin
		1 protease inhibitor
7	Antibiotic, toxin, surfactant	**7 antibiotic**
5	Antibiotic, antitumor, toxin	**5 antibiotic**
17	Antibiotic, antitumor	**10 antibiotic**
		6 toxin
		1 antitumor
14	Antibiotic, antitumor, immunomodulating	8 toxin
		6 antitumor
95	Antibiotic, surfactant	**74 antibiotic**
		5 toxin
		4 siderophore
		12 antitumor
29	Antitumor, toxin	2 antibiotic
		22 toxin
		5 antitumor
3	Antitumor, immunomodulating	2 toxin
		1 antitumor

The good predictions are in bold

(see Table 5) and built an exhaustive evaluation set with 232 peptides that are in Norine but not in the initial data set as they have at least two known activities. The data sets and predictions obtained with LibLinear method are presented in Tables 5 and 6. The correct activity, described in source papers, is predicted for 4 out of the 5 new peptides. Orfamide A is an antibiotic predicted as toxin, but crosses between antibiotic and toxin predictions are also observed in our initial data set. The results obtained for the evaluation set are promising as we predict correctly one of the activities for 83 % among the 237 tested peptides. This rate is similar to the one found with the cross-validation done with the initial data set, even if the activities represented in this set are challenging because they are the ones with the higher rate of crossing (antibiotic, antitumor and toxin). The prediction results for the evaluation data set clearly show the usefulness and robustness of our approach.

To improve the results in both similarity search and activity prediction, we will work on the fingerprints. On one hand, determining clusters of monomers will reduce the numbers of elements in the fingerprints and increase the common elements between peptides. On the other hand, adding of structure information such as monomer neighborhood will increase the number of elements in the fingerprints and improve the discrimination between two NRPs with similar monomer compositions but different structures.

The results obtained show that monomer composition fingerprint provides an interesting alternative to the widely used atomic fingerprints for similarity-based searching and biological activity prediction of nonribosomal peptides. However, beside the good performance of MCFP, it is efficient compared to any other representation approach, since dealing with fingerprint calculation is faster and conduct at minimal computational cost.

Conclusion

In this paper, we present a new peptide fingerprint (MCFP) based on monomer composition of NRPs. Experiments with the Norine NRPs database, clearly show the usefulness and effectiveness of MCFP for similarity-based searching and biological activity prediction of nonribosomal peptides.

Acknowledgments This work was supported by PPF Bioinformatique of Lille 1 University and INRIA.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

1. Newman DJ, Cragg GM, Snader KM (2000) The influence of natural products upon drug discovery. Nat Prod Rep 17:215–234. doi:10.1039/A902202C
2. Balunas MJ, Kinghorn AD (2005) Drug discovery from medicinal plants. Life Sci 78:431–441. doi:10.1016/j.lfs.2005.09.012
3. Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. Nat Rev Drug Discov 4:206–220. doi:10.1038/nrd1657
4. Newman DJ, Cragg GM, Snader KM (2003) Natural products as sources of new drugs over the period 1981–2002. J Nat Prod 66:1022–1037. doi:10.1021/np030096l
5. Paterson I, Anderson EA (2005) The renaissance of natural products as drug candidates. Science 310:451–453. doi:10.1126/science.1116364
6. Entzeroth M, Chapelain B, Guilbert J, Hamon V (2000) High throughput drug profiling. J Assoc Lab Autom 5:69–71. doi:10.1016/s1535-5535(04)00085-1
7. Merino A, Bronowska AK, Jackson DB, Cahill DJ (2010) Drug profiling: knowing where it hits. Drug Discov Today 15:749–756. doi:10.1016/j.drudis.2010.06.006
8. Marahiel MA (2009) Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. J Pept Sci 15:799–807. doi:10.1002/psc.1183
9. Stachelhaus T, Mootz HD, Marahiel MA (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol 6:493–505. doi:10.1016/S1074-5521(99)80082-9
10. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O (2011) NRPSpredictor2—a web server for predicting NRPS

adenylation domain specificity. Nucleic Acids Res 39:362–367. doi:10.1093/nar/gkr323
11. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res 33:5799–5808. doi:10.1093/nar/gki885
12. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol 7:211–224. doi:10.1016/S1074-5521(00)00091-0
13. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39(2):W339–W346. doi:10.1093/nar/gkr466
14. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G (2008) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326–D331. doi:10.1093/nar/gkm792
15. Bajorath J (2008) Computational analysis of ligand relationships within target families. Curr Opin Chem Biol 12:352–358. doi:10.1016/j.cbpa.2008.01.044
16. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152:9–20. doi:10.1038/sj.bjp.0707305
17. Caboche S, Leclère V, Pupin M, Kucherov G, Jacques P (2010) Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J Bacteriol 192:5143–5150. doi:10.1128/jb.00315-10
18. Abdo A, Chen B, Mueller C, Salim N, Willett P (2010) Ligand-based virtual screening using bayesian networks. J Chem Inf Model 50:1012–1020. doi:10.1021/ci100090p
19. Abdo A, Salim N (2011) New fragment weighting scheme for the bayesian inference network in ligand-based virtual screening. J Chem Inf Model 51:25–32. doi:10.1021/ci100232h
20. Abdo A, Salim N, Ahmed A (2011) Implementing relevance feedback in ligand-based virtual screening using bayesian inference network. J Biomol Screen 16:1081–1088. doi:10.1177/1087057111416658
21. Arif S, Holliday J, Willett P (2009) Analysis and use of fragment-occurrence data in similarity-based virtual screening. J Comput Aided Mol Des 23:655–668. doi:10.1007/s10822-009-9285-0
22. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl 11:10–18. doi:10.1145/1656274.1656278
23. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
24. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Paper presented at the proceedings of the eleventh conference on uncertainty in artificial intelligence, Montréal, Qué, Canada
25. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: a library for large linear classification. J Mach Learn Res 9:1871–1874
26. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput 13:637–649. doi:10.1162/089976601300014493
27. Swets J (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293. doi:10.1126/science.3287615
28. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the "Receiver Operating Characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. J Med Chem 48(7):2534–2547. doi:10.1021/jm049092j
29. Challis GL, Ravel J (2000) Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS Microbiol Lett 187:111–114. doi:10.1111/j.1574-6968.2000.tb09145.x
30. Pradeille N, Zerbe O, Möhle K, Linden A, Heimgartner H (2005) The first total synthesis of the Peptaibol Hypomurocin A1 and its conformation analysis: an application of the 'Azirine/Oxazolone Method'. Chem Biodivers 2:1127–1152. doi:10.1002/cbdv.200590084
31. Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH (2007) The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Biol 14:53–63. doi:10.1016/j.chembiol.2006.11.007
32. Matthijs S, Laus G, Meyer J-M, Abbaspour-Tehrani K, Schäfer M, Budzikiewicz H, Cornelis P (2009) Siderophore-mediated iron acquisition in the entomopathogenic bacterium Pseudomonas entomophila L48 and its close relative Pseudomonas putida KT2440. Biometals 22:951–964. doi:10.1007/s10534-009-9247-y
33. Wiest A, Grzegorski D, Xu B-W, Goulard C, Rebuffat S, Ebbole DJ, Bodo B, Kenerley C (2002) Identification of peptaibols from Trichoderma virens and cloning of a peptaibol synthetase. J Biol Chem 277:20862–20868. doi:10.1074/jbc.M201654200

Appl Microbiol Biotechnol (2012) 95:593–600
DOI 10.1007/s00253-012-4181-2

MINI-REVIEW

Structure, biosynthesis, and properties of kurstakins, nonribosomal lipopeptides from *Bacillus* spp.

Max Béchet · Thibault Caradec · Walaa Hussein · Ahmed Abderrahmani · Marlène Chollet · Valérie Leclère · Thomas Dubois · Didier Lereclus · Maude Pupin · Philippe Jacques

Received: 2 March 2012 / Revised: 15 May 2012 / Accepted: 15 May 2012 / Published online: 9 June 2012
© Springer-Verlag 2012

Abstract A new family of lipopeptides produced by *Bacillus thuringiensis*, the kurstakins, was discovered in 2000 and considered as a biomarker of this species. Kurstakins are lipoheptapeptides displaying antifungal activities against *Stachybotrys charatum*. Recently, the biosynthesis mechanism, the regulation of this biosynthesis and the potential new properties of kurstakins were described in the literature. In addition, kurstakins were also detected in other species belonging to *Bacillus* genus such as *Bacillus cereus*. This mini-review gathers all the information about these promising bioactive molecules.

Keywords Kurstakins · Lipopeptides · *Bacillus cereus* · *Bacillus thuringiensis* · NRPS · Spreading

M. Béchet · T. Caradec · W. Hussein · M. Chollet · V. Leclère · P. Jacques
Laboratoire des Procédés Biologiques, Génie Enzymatique et Microbien (ProBioGEM), UPRES-EA 1026, Polytech'Lille/IUT A, Université Lille Nord de France-Sciences et Technologies, USTL, Avenue Paul Langevin,
59655 Villeneuve d'Ascq Cedex, France

A. Abderrahmani
Laboratoire de Biologie Cellulaire et Moléculaire,
Faculté des Sciences Biologiques,
Université des Sciences et de la Technologie Houari Boumediene,
B.P. 32, El Alia,
Alger, Algérie

T. Dubois · D. Lereclus
INRA,
UMR1319 Micalis, La Minière,
78280 Guyancourt, France

M. Pupin
LIFL,
UMR USTL/CNRS 8022, INRIA Lille-Nord Europe,
59655 Villeneuve d'Ascq Cedex, France

P. Jacques (✉)
Laboratoire des Procédés Biologiques, Génie Enzymatique et Microbie (ProBioGEM), UPRES-EA 1026, Polytech'Lille/IUT A, Université Lille Nord de France—Sciences et Technologies, USTL, Avenue Paul Langevin,
59655 Villeneuve d'Ascq Cedex, France
e-mail: philippe.jacques@polytech-lille.fr

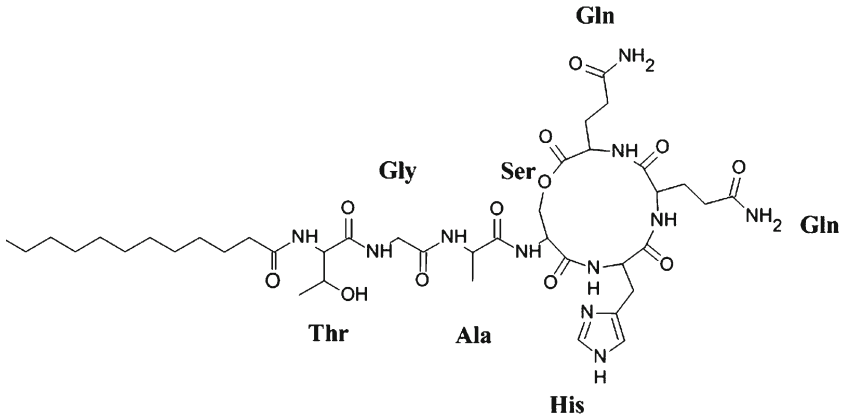
Introduction

Between 1949 and 1986, three different families of nonribosomal lipopeptides were identified in *Bacillus* spp.: surfactins, iturins, and fengycins (Jacques 2011). A new family produced by *Bacillus thuringiensis* and named the kurstakins, was discovered in 2000. The recent characterization of the biosynthesis mechanism of these compounds (in 2009 and 2011), their main properties (in 2011 and 2012) and the regulation of their biosynthesis in 2012, open new perspectives for these lipopeptidic compounds. This mini-review is the first one dedicated to them.

Discovery, structure, and mass spectrometry detection

Among the spores from six ATCC *B. thuringiensis* strains, the presence of a lipophilic biomarker, named kurstakin, was detected for the first time from the *B. thuringiensis* subsp. *kurstaki* strain HD-1 (Hathout et al. 2000). After a spore washing, the authors identified using LC-MS four $[M + H]^+$ molecular ions of *m/z* 879, 893, 893, and 907 (with a common fragment ion at *m/z* of 609). The molecular masses of these four compounds differed by 14 Da (–CH_2–), suggesting that these molecules were homologous lipopeptides. Further acid hydrolysis led to the identification of the corresponding free fatty acids from the four compounds which are 9-

methyldecanoic, dodecanoic, 10-methylundecanoic, and 11-methyldodecanoic acids, generating kurstakins C_{11} (*iso*-), C_{12} (n- and *iso*-), and C_{13} (*iso*-), respectively. Amino acid analyses revealed the same residues for the four molecules: Thr, His, Ala, Gly, Ser, and Glx (Gln or Glu) with molar ratios of 1:1:1:1:1:2. The ascertainment of the chemical structures of the kurstakins was completed by the determinations of (1) the sequence of the heptapeptide: Thr-Gly-Ala-Ser-His-Gln-Gln; (2) the presence of an amide bond between the fatty acid chain and the first threonine residue, and (3) the presence of a lactone linkage between the serine at position 4 and the C terminus of glutamine at position 7 (Fig. 1). Such a peculiar lactone ring between the fourth and seventh amino acids of the peptidic part was recently reported for a new biosurfactant (licheniformin) produced by *Bacillus licheniformis* MS3 (Biria et al. 2010). However, neither the nonribosomal peptide synthesis status of kurstakins in *B. thuringiensis* subsp. *kurstaki* strain HD-1 nor the possible occurrence of D- forms among the seven amino acids of their peptidic moiety (compared to other known nonribosomal lipopeptides from *Bacillus* spp.) were demonstrated. Three years later, a homologous series of three ions at *m/z* 892, 906, and 920 similar to those of kurstakins was detected by Madonna et al. (2003) in *Bacillus subtilis* ATCC 6051 but these results were not confirmed by genetic analyses (see below). Another study dealing with numerous *Bacillus* sp. strains isolated worldwide further revealed the presence of kurstakins in 20 from 54 strains tested, using MALDI-TOF-MS fingerprinting of whole bacterial cells (Price et al. 2007). These were typically identified by the molecular ions of *m/z* 889, 905, 917, and 933 but their primary structures or those of other putative kurstakins of *m/z* about 942 and 958 were not elucidated. The authors confirmed that these secondary metabolites were retained by the spores or cells and not secreted because they predominantly found them in the bacterial colonies on agar plates.



I.

II.

	FA_Thr-Gly-Ala-Ser-His-Gln-Gln	**FA**_Thr-Gly-Ala-Ser-His-Gln-Gln
	O—	
iC_{11}	877.465	n.r.
C_{12}	891.481	n.r.
C_{12} [-OH(3)]	907.476	925.486
iC_{12}	891.481	n.r.
iC_{13}	905.496	n.r.
iC_{13} [-OH(3)]	921.491	939.502
C_{14} [-OH(3)]	935.507	953.51
	A	**B**

Fig. 1 **I** Chemical structure of kurstakin with a C_{12} fatty acid chain and a cycle. **II** Structures and calculated molecular masses (Da) of the different isoforms of kurstakins characterized until now by MALDI-TOF-MS. **A** Partially cyclic molecules (Hathout et al. 2000; Bumpus et al. 2009); the *square brackets* **[-OH(3)]** in *bold* correspond to the presence of a β-hydroxylated fatty acid; **B** linear molecules (Bumpus et al. 2009). The amino acids Thr and Gln (in position 6) are expected to be under the D-form (Abderrahmani et al. 2011) (It is worthy of note that L- and D- forms have never been chemically determined). The first kurstakin homologues, chemically characterized by Hathout et al. (2000), did not contain a β-hydroxylated fatty acid, and comprised the sole iC_{11}-, iC_{12}-, and iC_{13} isoforms isolated so far. **FA** = fatty acid; n.r. = not reported to date

New information about the structural diversity of the kurstakins was recently reported by MALDI-TOF-MS analyses of the *Bacillus* sp. strain NK2018 (Bumpus et al. 2009) and six *B. thuringiensis* strains (Abderrahmani et al. 2011). From culture supernatants of strain NK2018 grown in an M9 minimal medium, the first authors showed the presence of six kurstakin variants in differing amounts, with molecular masses ranging from 907.4765 to 953.5192 Da (formulae from $C_{40}H_{65}O_{13}N_{11}$ to $C_{42}H_{71}O_{14}N_{11}$), corresponding to kurstakins with a β-hydroxylated fatty acid chain with 12, 13, and 14 atoms of carbon. The occurrence of three molecules differing by exactly 18.0103 Da from three other ones was attributed to the presence or lack of a lactone ring inside the peptidic part of the molecules, suggesting that kurstakins might be found currently in culture supernatant as variants with linear peptidic parts. After growth on either AK or LB medium, Abderrahmani et al. (2011) detected the presence of the three C_{11}, C_{12}, and C_{13} kurstakin isoforms in six *B. thuringiensis* strains from the 11 tested. Some other molecular ions with *m/z* of 920, 942, and 958 were also sometimes detected and could correspond to the kurstakin with a C_{14} fatty acid chain (Abderrahmani 2011). To summarize, the kurstakins synthesized by *Bacillus* spp. consist of lipoheptapeptides which (1) are linked to between C_{11} and C_{14} fatty acids, β-hydroxylated or not, with two isoforms (n-, *iso*-); (2) are partially cyclic (lactone bond between Ser/4 and C-terminal Gln/7) and might even be linear; and (3) are expected to contain two D-configured amino acids. In MALDI-TOF-MS experiments, the values of the $[M + H]^+$, $[M + Na]^+$, and $[M + K]^+$ molecular ions detected should range from 878.473 (cyclic C_{11} isoform $[M + H]^+$) to 992.481 Da (linear β-hydroxylated C_{14} isoform $[M + K]^+$), under conditions allowing these biomarkers to be detected (Figs. 1 and 2).

Biosynthesis

The operons potentially encoding kurstakin synthetases in *Bacillus cereus* and *B. thuringiensis* were identified by bioinformatics analyses using two new approaches. In the first one, Bumpus et al. (2009) took advantage of the size of the NRPS enzymes and the presence of unique marker ions derived from the common phosphopantetheinyl cofactor to adapt mass spectrometry-based proteomics to detect selectively NRPS and PKS gene clusters in microbial proteomes without requiring genome sequence information. In these conditions, the authors highlighted in strains *Bacillus* sp. NK2018 and *B. cereus* AH1134 the genes involved in the biosynthesis of the kurstakins. The second approach used PCR with degenerate primers based on the intraoperon DNA sequence alignment of adenylation and thiolation domains of all enzymes implicated in the biosynthesis of the lipopeptide family (Tapi et al. 2010). Two sets of primers elaborated from first bacillomycin genes, then from kurstakin genes led to the discovery of genes implied in kurstakin biosynthesis (Tapi 2010; Tapi et al. 2010; Abderrahmani et al. 2011). From these two studies the organization of the kurstakin cluster could be predicted. This cluster (Fig. 3) contains three genes (*krsA*, *krsB*, and *krsC*) which encode three large multifunctional proteins (KrsA, KrsB, and KrsC) constituting the complete synthetase. This latter is organized as follows: KrsA comprises one module (m1), KrsB is constituted of two modules (m2 and m3), and KrsC includes four modules (m4 to m7) and for each module a condensation–adenylation–thiolation motif can be found. In addition, m1 and m6 harbour a supplementary epimerization domain. The module 7 includes a final thioesterase domain enabling

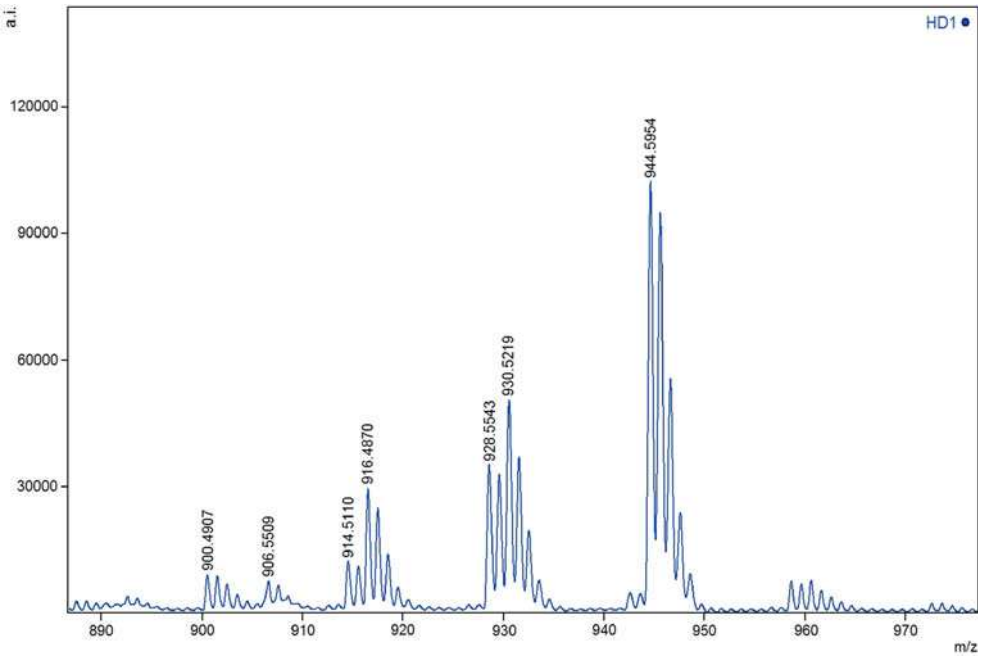


Fig. 2 Typical pattern of whole cell MALDI-TOF MS analysis of the kurstakin producing strain *B. thuringiensis* subsp. *kurstaki* HD-1 (Caradec et al., unpublished data). Isoform with C_{11} fatty acid chain and a cycle: 900.4907 $[M+Na]^+$; 916.4870 $[M+K]^+$. Isoform with C_{12} fatty acid chain and a cycle: 914.5110 $[M+Na]^+$; 930.5219 $[M+K]^+$. Isoform with C_{13} fatty acid chain and a cycle: 906.5509 $[M+H]^+$; 928.5543 $[M+Na]^+$; 944.5944 $[M+K]^+$

the unhooking of the neo-formed peptide from the NRPS and its possible cyclization. The combination of the predictions obtained from different bioinformatics tools (Ansari et al. 2004; Rausch et al. 2005; Bachmann and Ravel 2009) led to a peptide with the primary structure: D-Thr_Gly/Ala/Gln/Glu_Ala_Ser_X_D-Gln_Gln.

While the precise mechanism of biosynthesis was not yet experimentally analyzed, it could easily be deduced from the bioinformatics analysis and the high quantity of available information on the biosynthesis of other lipopeptides from *Bacillus* or *Pseudomonas* species (Sieber and Marahiel 2005; Raaijamakers et al. 2010; Roongsawang et al. 2010) (Fig. 3). The first synthetase, KrsA, contains one module with a starting condensation domain (Rausch et al. 2007; Kraas et al. 2010) which catalyses the link between a fatty acid chain of different length and isomery and a threonine residue activated by the adenylation domain and covalently fixed to the thiolation domain of this first module. The last domain of this first module is an epimerization domain that will transform the L-Thr to D-Thr. The second synthetase, KrsB, contains two modules responsible for the activation and the incorporation of two other amino acid residues, Gly and Ala. In the third synthetase, KrsC, four modules are involved in the incorporation of the four other amino acid residues: Ser, His, Gln, and Gln. The presence of an epimerization domain in the third module probably modifies the incorporated L-Gln to form the D-configuration. The last module contains a thioesterase which will catalyse the liberation of the peptide and probably its partial cyclization (Kopp and Marahiel 2007). A gene encoding a phosphopantetheinyl transferase was identified downstream from the NRPS complex. This enzyme could be involved in the transformation of the apo-form of the enzymes to the holoform by the addition of a phosphopantetheinyl group from the coenzymeA to the different thiolation domains (Mofid et al. 2004). A second thioesterase (TEII) is encoded by the next gene (*krsD*). The results obtained by Schwarzer et al. (2002) on the biochemical characterization of two second similar thioesterases named TEII and involved in the surfactin and bacitracin peptide antibiotics biosynthesis showed that these enzymes play a role in the regeneration of the misacylated peptidyl carrier proteins (thiolation domains). The presence of linear peptide in strain NK2018 could result from the action of this second thioesterase or an insufficient expression or efficacy of the first one.

A sixth gene (*krsE*) situated upstream from the *krsA-C* genes and belonging to the kurstakin cluster was identified by Dubois et al. (2012) in *B. thuringiensis* Bt407. The

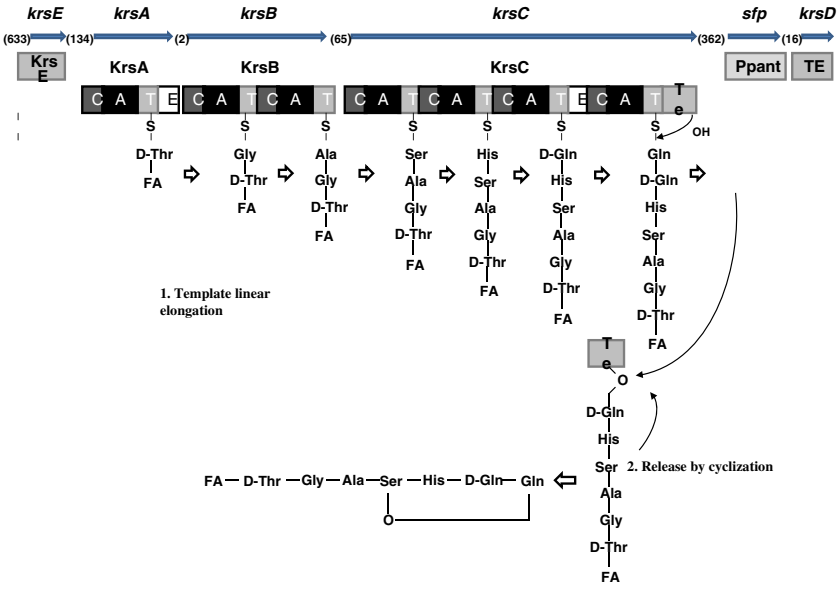


Fig. 3 Hypothetical kurstakin biosynthetic assembly line based on bioinformatics analysis. Situated downstream from the putative *krsE* gene which potentially encodes for an efflux protein which could be involved in the kurstakin secretion (bthur0010_59510), the biosynthetic complex consists of three kurstakin synthetases: KrsA, KrsB, and KrsC (encoded by three genes: bthur0010_59520, bthur0010_59530, and bthur0010_59540 in *B. thuringiensis* serovar pondicheriensis BSCG 4BA1), divided into seven distinct modules. Each module is responsible for recognition, activation and loading of a single amino acid substrate. In the first module, a starter condensation domain links a fatty acid chain to the amino acid residue (Thr) activated and fixed in this module. Two epimerization domains are found in modules 1 and 6, converting the corresponding amino acid in the D-stereoisomer. The cyclization and release of the final heptapeptide is catalyzed by the first adjacent Te domain. Immediately downstream are situated two genes, coding for a phosphopantetheinyl transferase (bthur0010_59550) and a closely adjacent type II thioesterase (bthur0010_59560), which are expected to belong to the kurstakin cluster. The *numbers in brackets* correspond to the spaces (in nucleotides) between these different open reading frames borne by strain BSCG 4BA1. **C** Condensation domain; **A** adenylation domain; **T** thiolation domain; **E** epimerization domain; **Te** thioesterase domain; **Ppant** phosphopantetheinyl transferase (Sfp); **TE** type II thioesterase; **FA** fatty acid

product of this gene, the protein KrsE, is a presumed efflux protein and could be involved in the secretion of the lipopeptide.

Overview on kurstakin potentially producing microorganisms

In order to identify the genetic potential for kurstakin production among microorganisms, we performed a BLASTp method using KrsA, KrsB, KrsC, KrsD, KrsE, and Ppant sequences from *B. thuringiensis* serovar pondicheriensis BGSC 4BA1 as queries. The same cluster leading to kurstakin synthesis was retrieved in the genomes of 32 strains for which genomes are sequenced, assembled, and either finished or still as drafts (V. Leclère, M. Pupin, W. Hussein, and P. Jacques, unpublished data). Without exception, all the strains pointed out belong to the *Bacillus* genus, more especially to the *B. cereus* species group, indicating that kurstakin production could be considered as the marker for this group. However, no sequences of kurstakin synthetases were present in *Bacillus anthracis* and *Bacillus cytotoxicus* although they belong to the same *B. cereus* group. In addition, only one genome is available for *B. cytotoxicus* and five genomes are completely sequenced and assembled for *B. anthracis*. So the question of lack of kurstakin synthetase in the *B. anthracis* species remains open and should be related later to the high virulence of the strains. In addition, kurstakins were also detected in *B. subtilis* ATCC 6051 strain (Madonna et al. 2003). This strain was an ancestor of the reference strain 168 (Zeigler et al. 2008), the genome of which was completely sequenced and, surprisingly, no traces of kurstakin genes were found in this genome. The kurstakin cluster is present in most of the genomes of *B. cereus* and *B. thuringiensis* for which genomic data are available (Table 1). Kurstakin genes were also detected in the partially sequenced genome of *B. cereus* BDRD-Cer4 and *B. cereus* AH1134. When the sequences are present in the genomes, they are highly conserved and the organization of the cluster (KrsE-KrsA-KrsB-KrsC-Ppant-TE) is also conserved. However, the strains *B. cereus* AH603, BDRD-ST196, *Bacillus mycoides* DSM 2048, and *Bacillus weihenstephanensis* KBAB4 might produce a variant form with a Glu or Asp instead of Gln at the last position as predicted by NRPSpredictor2 (Röttig et al. 2011). As no amino acid residue can be predicted for module 6, the strain *B. cereus* Rock4-2 can be supposed to produce another member of the kurstakin family varying by the residue at this position.

Table 1 Presence of kurstakin genes in sequenced genomes of *B. cereus* and *B. thuringiensis*

Strains	Sequenced genome status	Genes					
		krsE	*krsA*	*krsB*	*krsC*	*sfp*	*krsD*
Reference strain: *Bacillus thuringiensis* serovar pondicheriensis BGSC 4BA1	C	+	+	+	+	+	+
B. thuringiensis							
BMB171; sv chinensis CT-43	C	+	+	+	+	+	+
sv finitimus; sv konkukian str. 97-27; str. Al Hakam	C	−	−	−	−	−	−
Bt407; IBL 200; sv berliner ATCC 10792; sv huazhongensis BGSC 4BD1; sv kurstaki str. T03a001; sv thuringiensis str. T01001; sv pakistani str. T13001; sv pulsiensis BGSC 4CC1; sv sotto str. T04001	U	+	+	+	+	+	+
IBL 4222	U	+	+	+	TCT?	+	+
B. cereus							
ATCC 14579; B4264; G9842	C	+	+	+	+	+	+
03BB102; AH187; AH820; ATCC 10987; E33L; Q1; bv anthracis str. CI	C	−	−	−	−	−	−
172560W; ATCC 10876; BDRD-ST24; BGSC 6E1; F65185; m1550; Rock1-15; Rock1-3	U	+	+	+	+	+	+
BDRD-Cer4	P	+	+	+	+	+	+
AH1134	P	+	+	NF	+	NF	NF
AH603; BDRD-ST196	U	+	+	+	7-D/E	+	+
AH621	U	+	NF	NF	TCT?	+	NF
AH676	U	+	+	+	+	+	NF
Rock4-2	U	+	+	+	6-X	+	+
Others species							
B. mycoides DSM 2048	C	+	+	+	7-D/E	+	+
B. weihenstephanensis KBAB4	C	+	+	+	7-D/E	+	+

C complete, U unfinished, P partial, NF not found, TCT? truncated ?, 7-D/E prediction of amino acid residue incorporated by module 7 is D/E instead of Q, 6-X no possible prediction for amino acid residue incorporated by module 6, + protein present with at least 90 % identity with the reference one, −: not present

Regulation

The regulation system of kurstakin production has been partially described in *B. thuringiensis* Bt407 (Fig. 4). A transcriptomic analysis indicates that the four genes *krsEABC* form a cluster whose transcription is activated by the NprR–NprX quorum-sensing system during late stationary phase (Dubois et al. 2012). NprR is a quorum sensor activated by its cognate signaling peptide, NprX. NprR–NprX functions as a typical Gram-positive quorum-sensing system: the pro-signaling peptide NprX is exported from the bacterial cell and after being processed to its active form (presumably a heptapeptide), the peptide is reimported into the bacteria, where it binds to NprR allowing the recognition of its DNA target (Perchat et al. 2011).

The NrpX–NrpR system regulates 41 different genes, divided into four different groups. The first group is composed of genes coding for stress resistance proteins, including cytochrome P450, cysteine dioxygenase, and several metabolite exporters. The second group is composed of four genes encoding the Opp permease system, involved in the import of small peptides into the cells. The third group is composed of the NRPS *krs* genes. The last group codes for degradative enzymes and proteins able to bind organic material (Dubois et al. 2012).

Properties

Kurstakins are not recovered in the culture supernatant of producing strains but are found in association with the bacterial cells and particularly on spores (Hathout et al. 2000; Price et al. 2007; Abderrahmani et al. 2011). However, co-infection experiments carried out with a producing strain and a non-producing one in the insect larvae *Galleria mellonella* suggest that this molecule is secreted (Dubois et al. 2012). This apparent discrepancy between these results suggests that kurstakin is a secreted molecule with a high affinity for membranes. This high affinity could be due to the presence of the basic amino acid histidine which confers a cationic charge to the lipopeptide and thus facilitates its electrostatic interaction with phospholipids of the cell membrane.

Purified kurstakins displayed an antifungal activity against *Stachybotrys charatum*, showing a halo of inhibition identical to the one obtained with polymyxin B used as positive control (Hathout et al. 2000). Nevertheless, Abderrahmani et al. (2011) showed that no correlation exists between the antifungal activities of the strains and the presence of kurstakins. Indeed, some producing strains did not show any antifungal activity whereas some other ones did not produce kurstakin and showed antifungal properties. However, their evaluation was made with the fungi *Mucor rouxii* DSM 1191, *Rhizopus orizae* DSM 907, *Penicillium roqueforti* DSM 1080, *Aspergillus niger* DSM 737, and *Fusarium oxysporum* DSM 62297, different from those used by Hathout et al. These data indicate that kurstakin might be a pore-forming molecule with a limited spectrum of activity.

The fact that significant colonization of solid media was detected neither for non-producing kurstakin strains nor for a kurstakin-deficient mutant indicates that kurstakins are responsible for the invasive growth (Abderrahmani et al. 2011). Moreover, a strain where the genes *krs*A, *krs*B, and *krs*C (Δ*krs*ABC) were deleted was unable both to swarm and to form a biofilm at the air/liquid interface (Dubois et al. 2012). A very interesting property of kurstakin is its ability to enhance the survival of *B. thuringiensis* in the insect

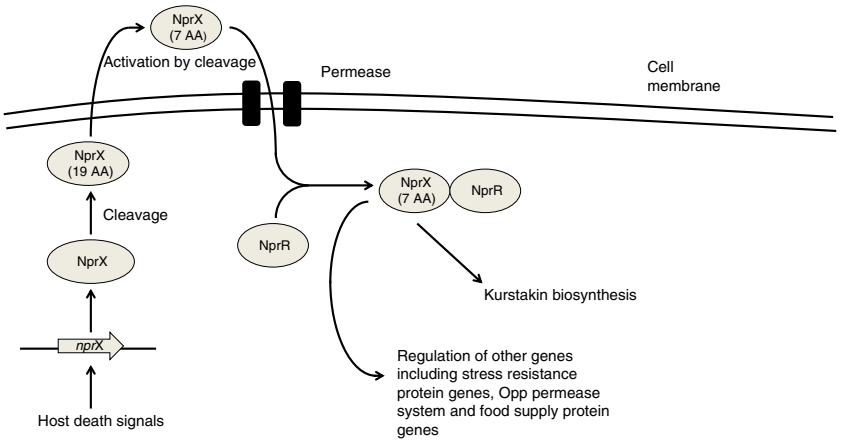


Fig. 4 Regulation of kurstakin biosynthesis. The NprX peptide is produced as an inactive form and is then activated by two successive cleavages. The first cleavage happens in the cell cytoplasm, and leads to a 19 amino acid peptide formation. After this first cleavage, the peptide is exported out of the cell, and is cleaved a second time in a seven amino acids peptide, leading to its active form NprX. This peptide is then reimported within the bacterial cell through the Opp permease system, and is bound with the NprR regulator. NprR forms a complex with the heptapeptide NprX whose production is regulated by host-death signals. The NprX–NprR complex activates the kurstakin production, (Perchat et al. 2011)

cadaver (Dubois et al. 2012). In view of these various properties, kurstakin might allow *B. thuringiensis* to spread across the cadaver, thereby facilitating access to new substrates and increasing its ability to disseminate in the environment.

Perspectives

The research on this fourth family of lipopeptides produced by *Bacillus* spp. has yet to be developed, and several perspectives are worth considering. The precise structure of the different variants should be confirmed by chemical analysis: the presence of the D-amino acid residues should be validated by analysis of amino acid residues after acid hydrolysis and derivatization, e.g., by GC using chirasyl-L-Val column. Confirmation of the presence of linear structure or C14 fatty acid chain should be done by LC-MS-MS analysis and NMR. The predicted biosynthetic pathway proposed in this review should be confirmed by biochemical analysis of the different domains of the synthetase. Particular attention will have to be paid to the thioesterase domains and their role in the concomitant presence of cyclized and linear forms of the lipopeptides.

Hathout et al. (2000) have evaluated the amount of kurstakin produced at about 15–20 μg/mg of spore. Overproducing mutant cells could be constructed using similar strategies developed by Leclère et al. (2005), Fickers et al. (2009), and Coutte et al. (2010), for the overproduction of other families of lipopeptides from *B. subtilis*. Purification techniques need to be developed to extract the lipopeptides or collect them in the supernatant.

Lipopeptides from *Bacillus* spp. are well-known for their potential applications in several fields (Jacques 2011) including biocontrol of plant pathogens (Ongena and Jacques 2008). Purified compounds could be thus used in different physico-chemical or biological tests in order to characterize their physico-chemical properties and biological activities and their potential applications.

Acknowledgments This work received financial support from the Université Lille 1, Sciences et Technologies and ARCIR funds from Nord-Pas-de-Calais and the European Funds of INTERREG IV PhytoBio Project. We acknowledge William Everett for his kind English proofreading.

References

Abderrahmani A (2011) Identification du mécanisme de synthèse non ribosomique d'un nouveau lipopeptide, la kurstakine et étude de son influence sur le phénotype de souches de *Bacillus thuringiensis* isolées en Algérie. Thèse de Doctorat d'Etat, Université des Sciences et de la Technologie Houari Boumediene, Alger, Algérie

Abderrahmani A, Tapi A, Nateche F, Chollet M, Leclère V, Wathelet B, Hacene H, Jacques P (2011) Bioinformatics and molecular approaches to detect NRPS genes involved in the biosynthesis of kurstakin from *Bacillus thuringiensis*. Appl Microbiol Biotechnol 92:571–581

Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32:405–413

Bachmann BO, Ravel J (2009) Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 458:181–217

Biria D, Maghsoudi E, Roostaazad R, Dadafarin H, Sahebghadam Lotfi A, Amoozegar MA (2010) Purification and characterization of a novel biosurfactant produced by *Bacillus licheniformis* MS3. World J Microbiol Biotechnol 26:871–878

Bumpus SB, Evans BS, Thomas PM, Ntai I, Kelleher NL (2009) A proteomics approach to discovering natural products and their biosynthetic pathways. Nat Biotechnol 27:951–956

Coutte F, Leclère V, Béchet M, Guez JS, Lecouturier D, Chollet-Imbert M, Dhulster P, Jacques P (2010) Effect of *pps* disruption and constitutive expression of *srfA* on surfactin productivity, spreading and antagonistic properties of *Bacillus subtilis* 168 derivatives. J Appl Microbiol 109:480–491

Dubois T, Faegri K, Perchat S, Lemy C, Buisson C, Nielsen-LeRoux C, Gohar M, Jacques P, Ramarao N, Kolsto AB, Lereclus D (2012) Necrotrophism is a quorum-regulated lifestyle in *Bacillus thuringiensis*. PloS Pathogens 8:e1002629

Fickers P, Guez JS, Damblon C, Leclère V, Béchet M, Jacques P, Joris B (2009) High-level biosynthesis of the anteiso-C(17) isoform of the antibiotic mycosubtilin in *Bacillus subtilis* and characterization of its candidacidal activity. Appl Environ Microbiol 75:4636–4640

Hathout Y, Ho YP, Ryzhov V, Demirev P, Fenselau C (2000) Kurstakins: a new class of lipopeptides isolated from *Bacillus thuringiensis*. J Nat Prod 63:1492–1496

Jacques P (2011) Surfactin and other lipopeptides from *Bacillus* spp. In: Soberon-Chavez G (ed) Biosurfactants microbiology monographs, vol 20. Springer, Berlin, pp 57–91

Kopp F, Marahiel MA (2007) Macrocyclization strategies in polyketide and nonribosomal peptide biosynthesis. Nat Prod Rep 24:735–749

Kraas FI, Helmetag V, Wittman M, Strieker M, Marahiel MA (2010) Functional dissection of surfactin synthetase initiation module reveals insights into the mechanism of lipoinitiation. Chem Biol 17:872–880

Leclère V, Béchet M, Adam A, Guez JS, Wathelet B, Ongena M, Thonart P, Gancel F, Chollet-Imbert M, Jacques P (2005) Mycosubtilin overproduction by *Bacillus subtilis* BBG100 enhances the organism's antagonistic and biocontrol activities. Appl Environ Microbiol 71:4577–4584

Madonna AJ, Voorhees KJ, Taranenko NI, Laiko VV, Doroshenko VM (2003) Detection of cyclic lipopeptide biomarkers from *Bacillus* species using atmospheric pressure matrix-assisted laser desorption/ionization mass spectrometry. Anal Chem 75:1628–1637

Mofid MR, Finking R, Essen L, Marahiel MA (2004) Structure-based mutational analysis of the 4′-phosphopantetheinyl transferases Sfp from *Bacillus subtilis*: carrier protein recognition and reaction mechanism. Biochemistry 14:4128–4136

Ongena M, Jacques P (2008) *Bacillus* lipopeptides: versatile weapons for plant diseases biocontrol. Trends Microbiol 16:115–125

Perchat S, Dubois T, Zouhir S, Gominet M, Poncet S, Lemy C, Aumont-Nicaise M, Deutscher J, Gohar M, Nessler S, Lereclus D (2011) A cell–cell communication system regulates protease production during sporulation in bacteria of the *Bacillus cereus* group. Mol Microbiol 82:619–633

Price NP, Rooney AP, Swezey JL, Perry E, Cohan FM (2007) Mass spectrometric analyses of lipopeptides from *Bacillus* strains isolated from diverse geographical locations. FEMS Microbiol Lett 271:83–89

Raaijamakers JM, De Bruijn I, Nybroe O, Ongena M (2010) Natural functions of lipopeptides from *Bacillus* and *Pseudomonas*: more than surfactants and antibiotics. FEMS Microbiol Rev 34:1037–1062

Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using Transductive Support Vector Machines (TSVM). Nucl Acids Res 33:5799–5808

Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol Biol 7:78

Roongsawang N, Washio K, Morikawa M (2010) Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. Int J Mol Sci 12:141–172

Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O (2011) NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucl Acids Res 39(Web Server issue):W362–W367

Schwarzer D, Mootz HD, Linne U, Marahiel MA (2002) Regeneration of misprimed nonribosomal peptide synthetases by type II thioesterases. Proc Natl Acad Sci USA 99:14083–14088

Sieber SA, Marahiel MA (2005) Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem Rev 105:715–738

Tapi A (2010) Stratégie moléculaire de mise en évidence de peptides actifs d'origine non ribosomiale chez *Bacillus* sp. et *Lactobacillus* sp. PhD thesis, Université Lille1 Sciences et Technologies, France

Tapi A, Chollet-Imbert M, Scherens B, Jacques P (2010) New approach for the detection of non ribosomal peptide synthetase genes in *Bacillus* strains by polymerase chain reaction. Appl Microbiol Biotechnol 85:1521–1531

Zeigler DR, Pragai Z, Rodriguez S, Chevreux B, Muffler A, Albert T, Bai R, Wyss M, Perkins JB (2008) The origins of 168, W23, and other *Bacillus subtilis* legacy strains. J Bacteriol 190:6983–6995