Mise en place d'une plate-forme logicielle pour l'analyse des peptides non-ribosomiaux

THÈSE

présentée et soutenue publiquement le 8 septembre 2009

pour l'obtention du

Doctorat de l'Université de Lille 1 – Sciences et Technologies

(spécialité informatique)

par

Ségolène Caboche

Composition du jury

Rapporteurs :	Pierre Cornelis, Professeur	VIB, Université de Bruxelles
	Marie-Dominique Devignes, CR CNRS	LORIA, Nancy
Examinateurs :	Daslav Hranueli, Professeur	PBF, Université de Zagreb
	Philippe Jacques, Professeur	ProBioGEM, Université de Lille 1
	Valérie Leclere, Maître de conférences	ProBioGEM, Université de Lille 1
	Maude Pupin, Maître de conférences	LIFL-INRIA, Université de Lille 1
	Bernard Wathelet, Professeur	UCBI, Faculté Univ. de Gembloux
Directeur :	Gregory Kucherov, DR CNRS	LIFL-INRIA, Université de Lille 1

UNIVERSITÉ DE LILLE 1 – SCIENCES ET TECHNOLOGIES
ÉCOLE DOCTORALE SCIENCES POUR L'INGÉNIEUR
Laboratoire d'Informatique Fondamentale de Lille — UMR 8022
U.F.R. d'I.E.E.A. – Bât. M3 – 59655 VILLENEUVE D'ASCQ CEDEX
Tél. : +33 (0)3 28 77 85 41 – Télécopie : +33 (0)3 28 77 85 37 – email : `direction@lifl.fr`

Remerciements

Je tiens tout d'abord à remercier Marie-Dominique Devignes et Pierre Cornelis pour avoir accepté d'être rapporteurs, ainsi que Bernard Whatelet et Daslav Hranueli d'être examinateurs.

Un grand merci à Gregory, mon directeur de thèse, pour la confiance et la liberté accordées ainsi que pour son esprit critique et son expérience scientifitique (et son anglais). Je remercie sincèrement Maude pour sa disponibilité, sa patience, sa gentillesse et ses conseils. Je tiens à remercier Valérie pour ses conseils, sa bonne humeur, sa disponibilité et son enthousiasme. Enfin, je remercie Philippe pour son expérience scientifique, son enthousiasme, sa confiance et la liberté accordée. En bref, un grand merci à tous mes encadrants qui m'ont permis de mener à bien mon travail et qui m'ont supportée, soutenue et beaucoup appris durant ces trois ans.

Je tiens à remercier Hélène, sans qui je n'aurais pas découvert les NRPS, ainsi que pour son aide apportée notamment pour mes premiers enseignements. Merci également à Jean-Stéphane, Maude et Laurent pour l'aide et leur expérience lors des enseignements. Je remercie aussi Stéphane, Mathieux, Azadé, Marta, Benjamin pour leur bonne humeur et leur gentillesse. Un merci particulier à Antoine pour son aide préciseuse sur NORINE. Merci à Aude L. pour sa joie de vivre et ses petites histoires. Merci aussi à Aude D. pour son humour, sa compagnie et son soutien permanent. Je remercie Arnaud pour toute l'aide apportée mais aussi pour avoir fait de moi une geekette. L'équipe SEQUOIA m'a permis d'effectuer ma thèse dans un environnement idéal.

Je remercie également Lucie, François et tous les membres du laboratoire ProBioGEM, pour m'avoir acceuillie parmi eux.

Enfin, je dédicace ma thèse à ma famille, sans qui je ne serais pas qui je suis aujourd'hui et qui m'a toujours soutenue durant toutes mes études. Merci à mon père, qui m'a transmis son perfectionisme, qui m'a toujours soutenue dans mes choix et qui n'a pas eu peur de se mouiller pour moi. Merci à ma mère pour son soutien et sa confiance en moi. Merci à ma grand-mère qui m'a toujours soutenue et pour son aide permanente. Merci à ma soeur et mon frère pour leur confiance. Un merci particulier à Julien, qui m'a supportée tout ce temps et qui m'a toujours soutenue dans mes choix, même quand ceux-ci m'obligeaient à partir. Je les remercie sincérement pour leur amour et leur présence qui m'ont aidée à tenir dans les moments difficiles. Je terminerais en remerciant mes amis de Bauvin, en particulier ma cobel et Alex, pour leur présence et les bons monment partagés, hors du laboratoire, durant ces dernières années.

Remerciements

Table des matières

Table des matières

Introduction

Le dogme central de la biologie moléculaire présente la synthèse protéique comme un transfert de l'information portée par l'ADN vers les protéines, via l'ARN messager. Il existe cependant chez les bactéries et les champignons une voie alternative synthétisant des peptides actifs, par le biais de grands complexes multi-enzymatiques, les synthétases ou NRPS (*NonRibosomal Peptide Synthetases*). Les peptides non-ribosomiaux sont des molécules d'une grande importance pharmaceutique car elles ont été optimisées durant des millions d'années d'évolution pour jouer des rôles importants dans la défense et la communication pour les organismes producteurs. Les peptides non-ribosomiaux couvrent un large spectre d'activités biologiques et d'applications pharmaceutiques. En effet, ils peuvent être des antibiotiques, des immuno-suppresseurs, des anti-tumoraux, des toxines ou encore des sidérophores. La cyclosporine, un immuno-suppresseur utilisé après une greffe, la daptomycine (commercialisée sous le nom de Cubicin) utilisée dans le traitement de certaines infections par des bactéries à Gram positif, le tripeptide ACV, précurseur de la pénicilline, le plus célèbre antibiotique ou encore la bléomycine, utilisée dans le traitement de certains cancers sont des exemples de peptides non-ribosomiaux présentant des propriétés pharmaceutiques importantes.

Les peptides non-ribosomiaux présentent des caractéristiques spécifiques par rapport aux peptides classiques, notamment dans leur composition et leur structure primaire. En plus des vingt acides aminés protéogéniques, d'autres acides aminés sont incorporés, comme par exemple la kynurénine, ou encore des acides aminés protéogéniques modifiés, comme l'isomère D d'un acide aminé. Ils peuvent également être modifiés par la synthétase après leur sélection, par exemple, avec l'ajout d'un groupement méthyl. Enfin, des composés provenant d'autres voies de synthèse sont également incorporés pour former des lipopeptides ou glycopeptides.

La structure primaire des peptides non-ribosomiaux peut être linéaire, comme dans le cas des peptides ribosomiaux, mais est souvent plus complexe. En effet, elle peut être cyclique (partiellement ou totalement), branchée et même poly-cyclique.

Les activités biologiques importantes et variées font de ces peptides des composés de grand intérêt et de plus en plus étudiés. L'enjeu actuel est le développement d'approches génétiques ou enzymatiques pour l'obtention de nouveaux peptides. En modifiant les synthétases, le but est d'obtenir des nouveaux composés actifs ou de modifier des composés naturels afin d'améliorer leurs activités. Pour ce faire, une bonne connaissance du système de biosynthèse, mais aussi des peptides produits, est nécessaire.

Une équipe du laboratoire ProBioGEM (Laboratoire des Procédés Biologiques Génie Enzymatique et Microbien) travaille sur les peptides non-ribosomiaux. Durant leurs recherches, ils ont eu besoin d'outils informatiques permettant la comparaison des peptides non-ribosomiaux. En effet, lors de l'identification d'un peptide, la première étape est de savoir si ce peptide a déjà été décrit, ensuite, dans le cas où ce peptide n'a jamais été décrit, la seconde étape est d'identifier des peptides connus présentant des similarités avec le peptide d'intérêt afin de pouvoir mettre en

évidence des propriétés biologiques potentielles et orienter les validations expérimentales. Cependant, aucun outil de ce type n'était disponible. Ils se sont donc rapprochés de l'équipe SEQUOIA du LIFL (Laboratoire d'Informatique Fondamentale de Lille), spécialisée dans le développement d'outils informatiques pour la biologie. La collaboration entre ces deux équipes a commencé par le co-encadrement de cette thèse dont le but est de centraliser les données concernant les peptides non-ribosomiaux et leurs monomères, ainsi que de développer des outils informatiques facilitant leur analyse.

La première étape a été de réaliser une étude bibliographique sur les peptides non-ribosomiaux et leur voie de biosynthèse afin de comprendre et de mettre en évidence leurs spécificités par rapport aux peptides ribosomiaux beaucoup mieux connus (chapitre 1).

Nous avons ensuite fait le bilan des outils bioinformatiques existants (chapitre 2). Plusieurs banques de données contiennent des peptides non-ribosomiaux mais aucune ressource centralisant les données sur ces peptides n'était disponible au commencement de ce travail. De même, des outils d'analyse des synthétases et de prédiction de peptides sont disponibles, mais aucun outil permettant l'analyse de ces peptides n'existait.

Les peptides non-ribosomiaux présentent des caractéristiques spécifiques au niveau de leur composition et de leur structure, ce qui rend impossible l'utilisation des outils et modélisations utilisés pour les peptides ribosomiaux. Nous avons donc développé une modélisation spécifique, basée sur des graphes étiquetés non-orientés, pour représenter ces peptides particuliers, ainsi que des outils informatiques permettant leur analyse (chapitre 3). Nous avons développé des outils de comparaison de compositions monomériques, de comparaison stricte de structures, de recherche de motifs structuraux et aussi de recherche de similarités.

Dans le but d'évaluer toute la diversité des peptides non-ribosomiaux et de centraliser les données sur ces peptides, nous avons développé Norine, la première ressource publique entièrement dédiée aux peptides non-ribosomiaux (`http://bioinfo.lifl.fr/norine`) (chapitre 4). Norine contient une base de données regroupant actuellement plus de 1 000 peptides ainsi qu'une interface web permettant d'interroger la base en fonction des différentes annotations, mais aussi à partir de données structurales, grâce aux outils informatiques présentés dans le chapitre 3.

Nous avons réalisé des analyses statistiques sur les données contenues dans Norine, pour étudier les spécificités des peptides non-ribosomiaux et de leurs monomères (chapitre 5). Ces expériences sont les premières à traiter un nombre aussi important de monomères et de peptides non-ribosomiaux et ont permis la mise en évidence de propriétés très intéressantes. Par exemple, nous avons montré qu'il existe une spécificité des monomères en fonction de l'activité biologique. Cette observation nous a conduit au développement d'un outil d'aide à la prédiction de l'activité biologique à partir de la composition en monomères d'un peptide.

Nous avons ensuite procédé à l'analyse de peptides non-ribosomiaux putatifs à partir de la séquence de génomes (chapitre 6). Nous avons prédit des peptides produits par des synthétases putatives identifiées dans un génome, puis à l'aide de Norine nous avons déduit des propriétés biologiques pour ces peptides. En étudiant le génome de *Pseudomonas entomophila*, nous avons mis en évidence la synthèse potentielle d'une nouvelle pyoverdine ; nous avons ensuite validé cette prédiction expérimentalement.

Chapitre 1

Contexte Biologique

Dans cette section, nous introduisons les différentes notions biologiques nécessaires à la compréhension du travail réalisé. Dans un premier temps, nous rappellerons les éléments intervenant dans la synthèse peptidique. Nous présenterons ensuite la voie non-ribosomiale qui est une voie alternative à la voie classique pour la production de peptides. Nous verrons ensuite les caractéristiques des peptides synthétisés par la voie non-ribosomiale. Pour finir, nous nous intéresserons aux perspectives ouvertes par cette voie de synthèse originale pour l'obtention de nouvelles molécules.

1.1 Synthèse peptidique

1.1.1 Acides aminés

Les acides aminés sont les unités structurales de base des protéines et des peptides. Les acides aminés naturels sont essentiellement des acides α-aminés. Un acide α-aminé est une molécule organique possédant une fonction amine (-NH2) et un acide carboxylique (-COOH) tous deux liés à un atome de carbone appelé carbone α [Voet et al., 2007a]. Le terme « acide α-aminé » est souvent généralisé par le terme « acide aminé ». Tous les acides aminés présentent cette structure commune appelée chaîne principale. Le carbone α porte également la chaîne latérale ou radical, symbolisé par R, qui est spécifique à un acide aminé donné. La figure 1.1 montre la structure commune à tous les acides aminés.

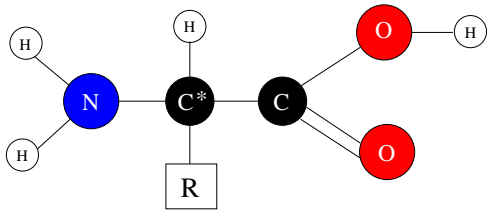


FIG. 1.1 – Structure commune à tous les acides aminés. Le carbone α est marqué par *.

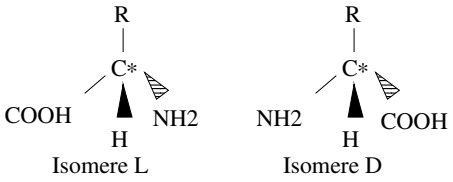
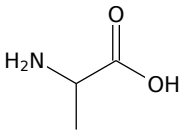
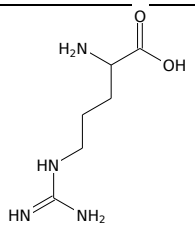
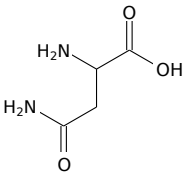
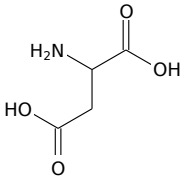
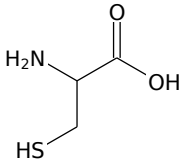
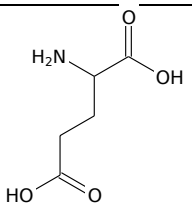
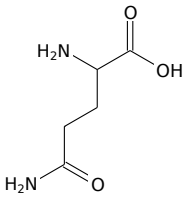
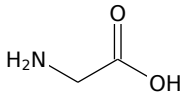
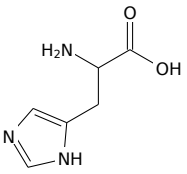
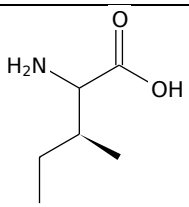
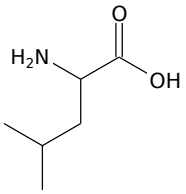
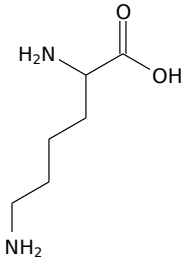
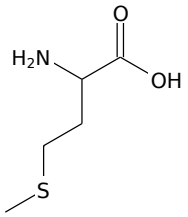
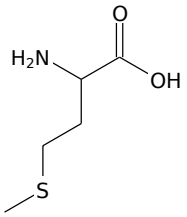
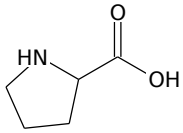
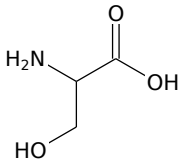
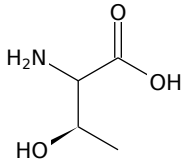
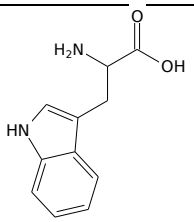
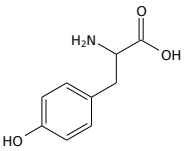
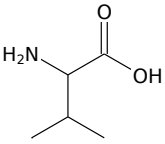


FIG. 1.2 – Isomérie L et isomérie D pour les acides aminés. Le carbone α est marqué par *.

Le carbone α est asymétrique, c'est-à-dire qu'il porte quatre groupements différents, sauf pour la glycine où $R = H$. L'asymétrie du carbone α fait donc des acides aminés des molécules chirales. Deux isomères sont observés pour un acide aminé donné, l'un de la série D et l'autre de

la série L [Voet et al., 2007a]. L'isomérie est donnée par l'ordre d'apparition dans le sens horaire ou anti-horaire des groupements du carbone α. La figure 1.2 montre les deux isomères D et L pour les acides aminés. Les acides aminés L représentent la majorité des acides aminés trouvés au sein des molécules du vivant.

 alanine A - Ala	 arginine R - Arg	 asparagine N - Asn	 acide aspartique D - Asp	 cystéine C - Cys
 acide glutamique E - Glu	 glutamine Q - Gln	 glycine G - Gly	 histidine H - His	 isoleucine I - Ile
 leucine L - Leu	 lysine K - Lys	 méthionine M - Met	 phénylalanine F - Phe	 proline P - Pro
 sérine S - Ser	 thréonine T - Thr	 tryptophane W - Trp	 tyrosine Y - Tyr	 valine V - Val

TAB. 1.1 – Les 20 acides aminés protéogéniques.

Les acides aminés se différencient par la nature de leur radical. Chez l'homme et la plupart des espèces, 20 acides aminés sont incorporés dans les protéines et peptides [Voet et al., 2007a]. Ces vingt acides aminés sont appelés acides aminés protéogéniques. Par raison de commodité, une nomenclature universelle a été adoptée, représentant les vingt acides aminés par un code à trois lettres et un code à une lettre. La table 1.1 montre la structure chimique des vingt acides aminés protéogéniques, avec leur nom et les codes à une et trois lettres correspondants.

Chaque acide aminé possède un radical spécifique présentant des propriétés physico-chimiques particulières. Ainsi, les acides aminés sont classés en groupes présentant des propriétés similaires. Par exemple, ils peuvent être classés en fonction de la nature de leur radical. Dans ce cas, les acides aminés présentant une chaîne hydrocarbonée (contenant uniquement des atomes de carbone et d'hydrogène) sont regroupés (Gly, Ala, Ile, Leu et Val), ainsi que ceux présentant une fonction hydroxyle (Ser et Thr), une fonction carboxylique supplémentaire (Asp et Glu), un groupement soufré (Cys et Met), une fonction amide (Asn et Gln), une fonction amine supplémentaire (Lys, Arg et His), un hétérocycle (Pro) ou un cycle aromatique (Phe, Trp et Tyr). Les acides aminés peuvent également être classés en fonction des propriétés physico-chimiques données par le radical [Voet et al., 2007a]. En général, trois propriétés sont distinguées : les acides aminés apolaires, les acides aminés polaires chargés et les acides aminés polaires non-chargés .

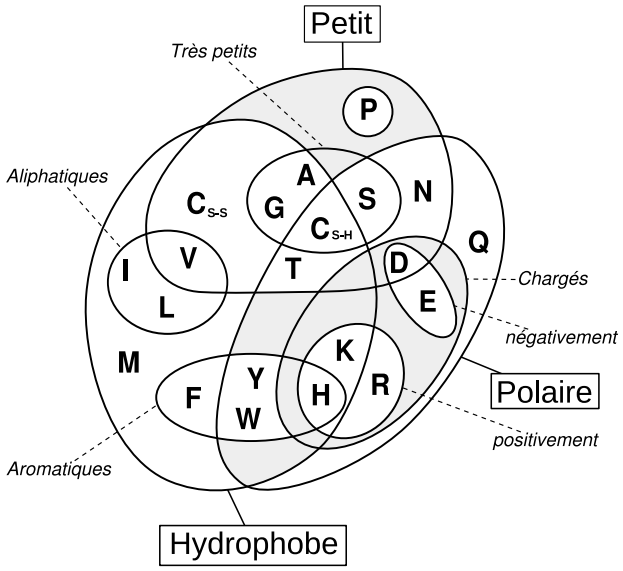


FIG. 1.3 – Diagramme de Venn des propriétés des acides aminés

La figure 1.3 présente une classification possible des acides aminés, basée sur les données issues d'une étude de leurs caractéristiques [Livingstone and Barton, 1993].

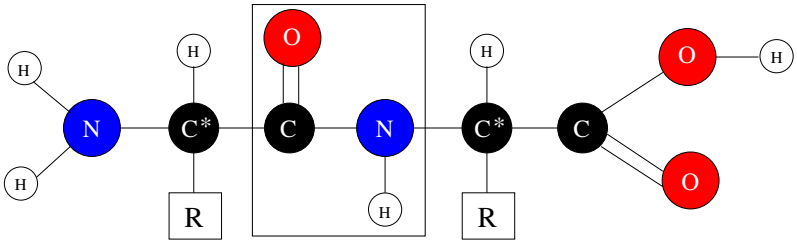


FIG. 1.4 – Formation d'une liaison peptidique entre deux acides aminés

Deux acides aminés sont liés entre eux au sein d'une protéine par une liaison peptidique formée entre l'acide carboxylique du premier acide aminé et la fonction amine du second, avec élimination d'une molécule d'eau [Voet et al., 2007a]. La figure 1.4 montre la formation d'une liaison peptidique entre deux acides aminés.

1.1.2 Peptides et protéines

Une protéine est une macromolécule composée d'acides aminés liés entre eux par des liaisons peptidiques. Les protéines remplissent des rôles importants et variés au sein des organismes vivants et sont essentielles à la vie. En général, le terme protéine représente une chaîne de plus de 40 acides aminés et celui de peptide une chaîne plus petite [Voet et al., 2007b]. Cependant, cette règle n'est pas absolue et dans certains cas, des composés de 50 acides aminés sont encore appelés peptides. La différence entre peptide et protéine se situe simplement au niveau du nombre d'acides aminés, par conséquent, toutes les propriétés données par la suite pour les protéines sont les mêmes pour les peptides, sauf pour la structure car les peptides sont souvent trop petits pour présenter des structures autres que primaires.

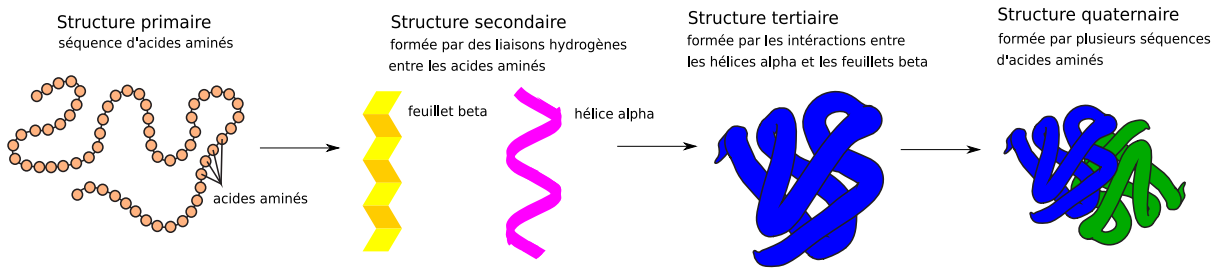


FIG. 1.5 – Structure des protéines : de la structure primaire à la structure quaternaire

La figure 1.5 montre la structure des protéines. La structure primaire d'une protéine, également appelée séquence protéique, correspond à l'enchaînement des acides aminés. Ces polymères sont linéaires et non ramifiés [Voet et al., 2007a]. La structure primaire d'une protéine a un sens défini. Le premier acide aminé de la séquence protéique est par convention, et aussi selon l'ordre de la synthèse, celui qui a la fonction amine libre. Il constitue l'extrémité N-terminale. Le dernier acide aminé est celui qui présente l'acide carboxylique libre. Il constitue alors l'extrémité C-terminale. La structure secondaire des protéines décrit le repliement local de la chaîne principale. L'existence des structures secondaires est due aux repliements énergétiques favorables de la chaîne peptidique. Les structures secondaires comprennent principalement les hélices α, les feuillets β et les coudes [Voet et al., 2007c]. La structure tertiaire correspond au repliement de la protéine dans l'espace, provoqué par les interactions entre les différentes structures secondaires [Voet et al., 2007c]. Enfin, la structure quaternaire correspond à l'agencement spatial de plusieurs chaînes peptidiques qui s'associent de manière non covalente [Voet et al., 2007c].

Les protéines sont synthétisées par ajouts successifs d'acides aminés. L'ordre des acides aminés incorporés est donné par l'information portée par le génome. Cette synthèse s'inscrit dans le dogme central de la biologie moléculaire.

1.1.3 Synthèse peptidique : le dogme central de la biologie moléculaire

Le dogme central de la biologie moléculaire fut introduit par Francis Crick en 1958. Il fait intervenir trois entités biologiques. Tout d'abord, l'Acide DésoxyriboNucléique (ADN) est le support de l'information génétique, information permettant le développement et le fonctionnement d'un organisme. Il constitue le génome des êtres vivants. L'ADN est formé par l'assemblage de nucléotides [Voet et al., 2007d]. Quatre nucléotides forment l'ADN : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). Ces nucléotides sont complémentaires deux à deux, le A est complémentaire du T et le C du G. L'ADN est en fait composé de deux brins se faisant

face, et formant une double hélice grâce à la complémentarité des nucléotides se trouvant face à face sur chacun des brins. La seconde entité biologique est l'Acide RiboNucléique (ARN). Il existe différents ARN. L'ARN messager (ARNm) est le support temporaire de l'information génétique. En effet, c'est lui qui sert ensuite de matrice pour la synthèse protéique. L'ARNm est simple brin et est composé comme l'ADN de quatre nucléotides : les nucléotides A, C et G mais dans l'ARNm le T est remplacé par l'uracile (U) [Voet et al., 2007d]. Enfin, la dernière entité biologique intervenant dans le dogme central est la protéine, présentée précédemment.

Le dogme central de la biologie moléculaire présente la synthèse protéique comme un transfert de l'information portée par l'ADN vers les protéines, via l'ARNm. La figure 1.6 montre de manière schématique le principe du dogme central de la biologie moléculaire.

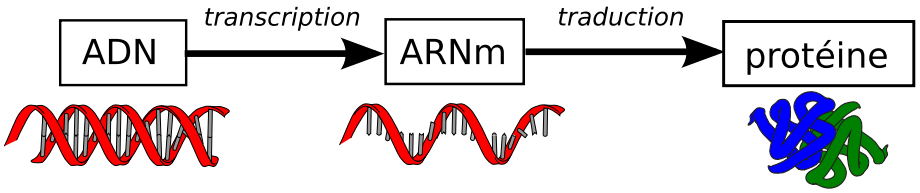


FIG. 1.6 – Le dogme central de la biologie moléculaire

Lors de la transcription, les régions codantes, de l'ADN, c'est-à-dire les gènes, sont copiées en molécules d'ARNm grâce aux ARN-polymérases [Voet et al., 2007e]. Lors de la traduction, l'ARNm est traduit en protéine. La traduction est l'interprétation des codons de l'ARNm en acides aminés qui se fait grâce aux ribosomes [Voet et al., 2007f] et aux ARN de transfert (ARNt). Un ARNt transfère un acide aminé donné à la chaine protéique en cours d'élongation au sein du ribosome. Les ribosomes sont constitués d'ARN ribosomiques et de protéines ribosomiques. Les ribosomes sont constitués de deux sous-unités, une plus petite qui « lit » l'ARNm et une plus grosse qui se charge de la synthèse de la protéine correspondante. Un codon est un triplet de nucléotides. Chaque codon correspond à un acide aminé donné. A l'inverse, un acide aminé peut être codé par plusieurs codons. L'ensemble des correspondances codon/acide aminé forme le code génétique. Il existe 64 codons pour 20 acides aminés. C'est pour cette raison que le code génétique est dit « dégénéré ».

Selon le dogme central, toute protéine (ou peptide) est synthétisée à partir de l'information portée par l'ADN, via l'ARNm. Cependant, il existe chez les bactéries et les champignons, une voie alternative qui ne suit pas ce schéma. Cette synthèse peptidique non-ribosomiale, comme son nom l'indique, ne s'effectue pas au sein des ribosomes mais par le biais de grandes protéines multi-fonctionnelles.

1.2 Synthèse peptidique non-ribosomiale

1.2.1 Généralités

Cette voie originale fut décrite pour la premières fois en 1971 dans le cadre de l'étude du mécanisme de biosynthèse de deux antibiotiques : la gramicidine S et la tyrocidine [Lipmann et al., 1971].

Considéré comme relativement anecdotique dans les années 70, le mécanisme non-ribosomial a pris de plus en plus d'importance dans la littérature scientifique. La figure 1.7 montre l'évolution du nombre de publications, au sein de PubMed, traitant de la synthèse non-ribosomiale en fonction de l'année de parution.

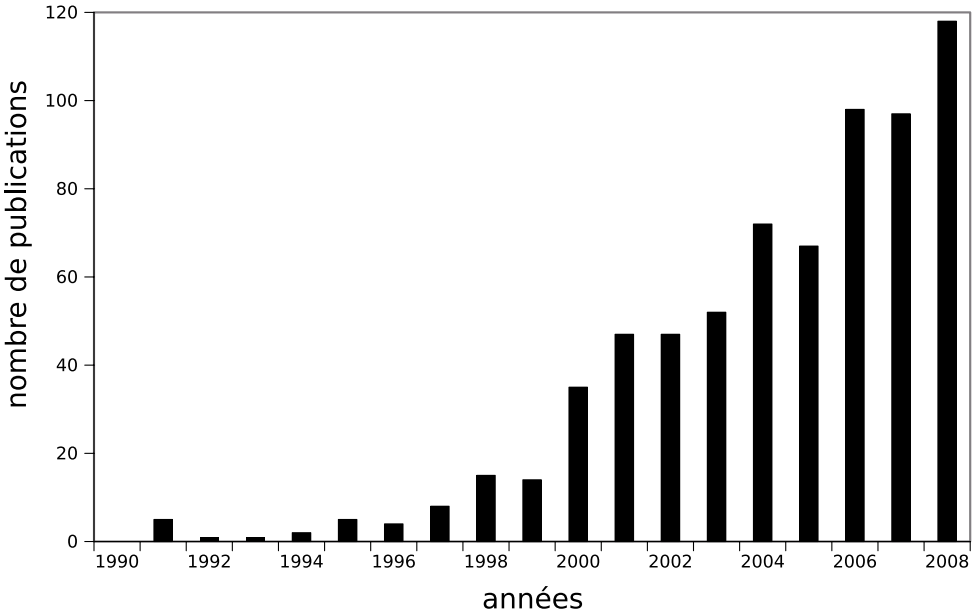


FIG. 1.7 – Nombre de publications par année, de 1990 à 2008, au sein de PubMed, traitant de la synthèse non-ribosomiale (requête utilisée : "nonribosomal peptide" OR "non-ribosomal peptide")

Le nombre de publications concernant cette synthèse subit une croissance exponentielle, ce qui montre l'intérêt grandissant des chercheurs pour cette voie originale. Une première raison est le fait que les peptides produits par la voie non-ribosomiale présentent des activités variées et ayant des débouchés industriels potentiels. Par exemple, le précurseur de la pénicilline, célèbre antibiotique découvert par Alexander Fleming en 1928, est synthétisé par la voie non-ribosomiale [Kallow et al., 1997]. D'autre part, le séquençage de nombreux génomes a permis la mise en évidence de nombreux gènes impliqués dans cette voie de synthèse.

Les synthétases ou NRPS (*non-ribosomal peptide synthetases*) sont les enzymes responsables de la synthèse peptidique non-ribosomiale. Ce sont de grands complexes qui représentent à la fois la matrice, rôle comparable à celui de l'ARNm dans la voie classique, mais aussi la machinerie biosynthétique, fonction effectuée par les ribosomes et les ARNt au sein de la synthèse protéique classique.

1.2.2 Synthétases

Organisation des synthétases

Les synthétases sont de grands complexes multi-enzymatiques. Elles mêmes sont synthétisées par la voie classique du dogme central. Plusieurs synthétases peuvent être nécessaires à la synthèse d'un peptide. Les gènes codant des synthétases intervenant dans la synthèse d'un même peptide sont généralement organisés en opérons ou en clusters [Schwarzer et al., 2003]. La figure 1.8 montre l'organisation des synthétases conduisant à la production de la surfactine. La surfactine présente différentes activités biologiques (antibiotique, surfactant). Elle est produite par *Bacillus subtilis* [Peypoux et al., 1999]. La figure 1.9 montre la structure de la surfactine qui est un lipopeptide cyclique composé d'un heptapeptide et d'un acide gras.

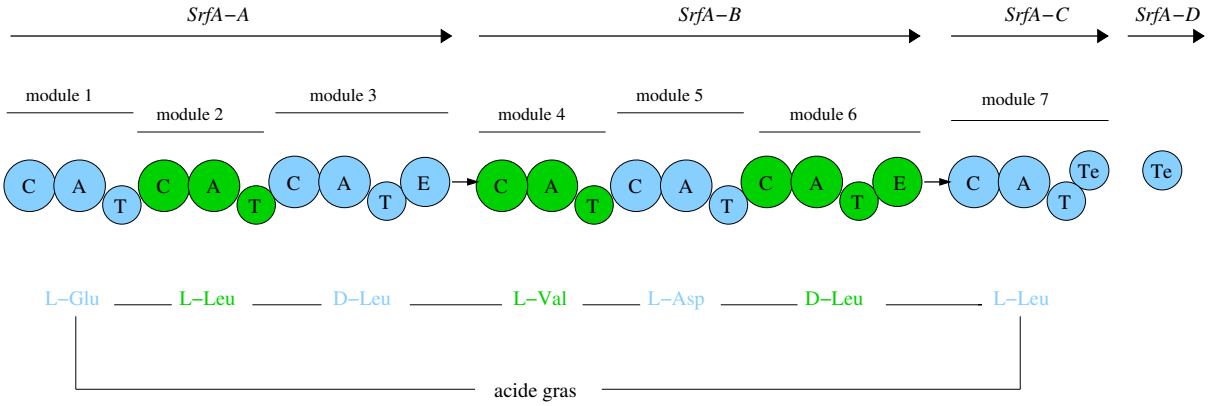


FIG. 1.8 – Organisation des synthétases : exemple de la surfactine
A : domaine d'adénylation ; T : domaine de thiolation ; C : domaine de condensation ; Te : domaine de la thioestérase ; E : domaine d'épimérisation

Quatre protéines sont nécessaires pour la production de la surfactine (cf. figure 1.8). Elles sont codées par les gènes *SrfA-A*, *SrfA-B*, *SrfA-C* et *SrfA-D* [Peypoux et al., 1999]. Les synthétases sont organisées en modules. Chaque module est responsable de l'incorporation d'un acide aminé spécifique dans la chaîne peptidique en formation. Sept modules sont nécessaires à la production de la surfactine (cf. figure 1.8). Chaque module incorpore un acide aminé donné. Par exemple, le *module 1* incorpore l'acide glutamique et le *module 5* incorpore l'acide aspartique.

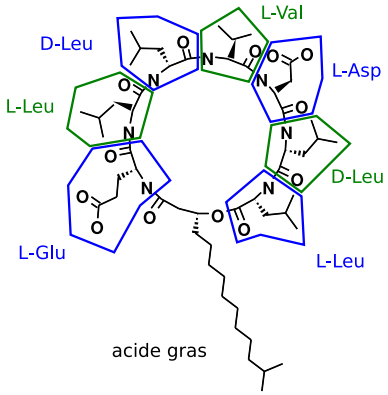


FIG. 1.9 – Structure de la surfactine

Chaque module est lui même organisé en domaines. Chaque domaine présente une fonction enzymatique particulière dans l'incorporation ou la modification des acides aminés. Les différents domaines sont classés en deux catégories : les domaines principaux et les domaines secondaires.

Domaines principaux

Les domaines principaux sont les domaines obligatoires pour le fonctionnement d'une synthétase. Ils sont au nombre de quatre [Schwarzer et al., 2003] :

– le domaine d'adénylation (A) : il sélectionne et active un acide aminé spécifique. L'acide aminé est transformé en aminoacyl adénylate.

– le domaine de thiolation (T), aussi appelé domaine PCP (peptidyl carrier protein) : il fixe l'acide aminé sur la synthétase de façon covalente par l'intermédiaire d'une liaison thioester.
– le domaine de condensation (C) : il forme la liaison peptidique entre les deux acides aminés de modules adjacents.
– le domaine de la thioestérase (Te) : il libère le peptide néoformé. Ce domaine permet également la cyclisation de certains peptides [Trauger et al., 2000].

Ces domaines forment trois catégories de modules. Tout d'abord, le module d'initiation est composé d'un domaine A et d'un domaine T. Il initie la biosynthèse peptidique. Le second, appelé module d'élongation, comprend les domaines C, A et T. Il accomplit l'élongation de la chaîne peptidique en formation. Enfin, le dernier, appelé module de terminaison, comprend en plus des domaines C, A et T un domaine Te. Il permet l'incorporation d'un dernier acide aminé et la libération du peptide.

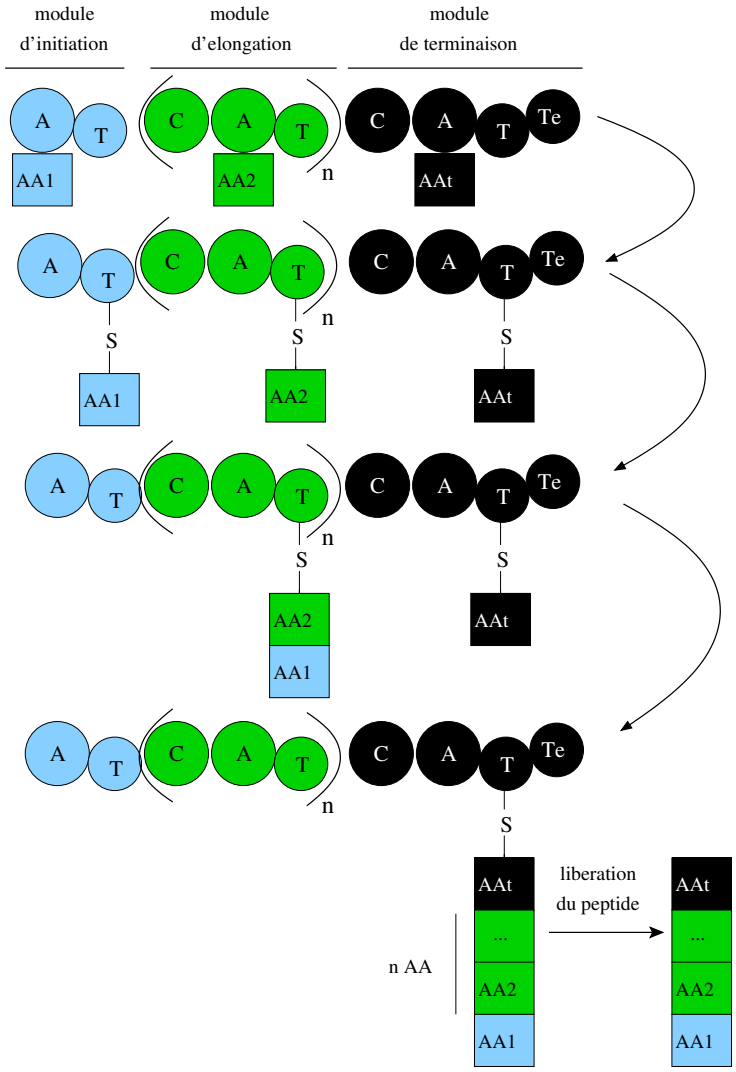


FIG. 1.10 – Biosynthèse d'un peptide

La figure 1.10 montre de façon schématique la biosynthèse d'un peptide par une synthétase. Dans un premier temps, le domaine A du module d'initiation sélectionne et active l'acide aminé 1. Ensuite, le domaine T du premier module fixe de façon covalente l'acide aminé 1 sur la synthétase. Les mêmes réactions se déroulent au sein du module d'élongation conduisant à la fixation de l'acide aminé 2 sur la synthétase. L'étape suivante est réalisée par le domaine C du second module. Le domaine C forme la liaison peptidique entre les acides aminés 1 et 2. Les n domaines d'élongation ajoutent n acides aminés à la chaîne peptidique en formation. La dernière étape est la libération du peptide grâce au domaine Te du module de terminaison.

Les domaines principaux suffisent pour la biosynthèse de peptides. Cependant, il existe une autre catégorie de domaines, les domaines secondaires qui modifient les acides aminés incorporés.

Domaines secondaires

Les domaines secondaires sont des domaines facultatifs qui modifient les acides aminés incorporés lors de la biosynthèse. Il existe différents domaines secondaires. Le plus répandu est le domaine d'épimérisation (E) qui transforme un acide aminé L en son isomère D (figure 1.11). La synthétase de la surfactine contient deux domaines E (figure 1.8). Ces deux domaines E forment les deux D-leucines présentes dans la surfactine. De nombreux isomères D sont observés au sein des peptides NRPS. Ils sont obtenus soit par la présence d'un domaine E, soit par la sélection d'un acide aminé D par le domaine A. Par exemple, la synthétase de l'arthrofactine, un biosurfactant produit par *Pseudomonas sp. MIS38*, ne possède pas de domaine E alors que des acides aminés D sont présents au sein du peptide [Roongsawang et al., 2003]. Il existe également des domaines de condensation présentant une activité d'épimérisation [Balibar et al., 2005]. Ce domaine, appelé domaine dual de condensation/épimérisation, épimérise l'acide aminé avant la formation de la liaison peptidique avec l'acide aminé du module adjacent.

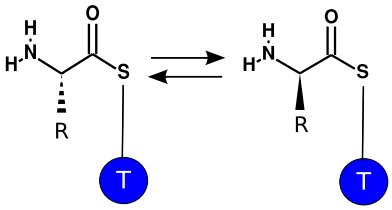


FIG. 1.11 – Réaction d'épimérisation sur l'acide aminé fixé sur le domaine de thiolation

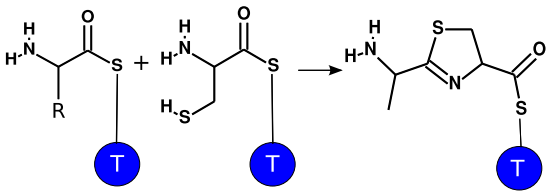


FIG. 1.12 – Réaction d'hétérocyclisation entre deux acides aminés fixés sur les domaines de thiolation de deux modules adjacents

Un autre domaine secondaire est le domaine d'hétérocyclisation (Cy) (figure 1.12). Ce domaine peut parfois remplacer le domaine C. Il forme de petits hétérocycles au sein des peptides. Ces cyclisations sont obtenues par la formation d'une liaison entre la chaîne latérale de la sérine, la thréonine ou la cystéine avec la chaîne peptidique principale. Les produits de ces cyclisations sont des thiazolines (cystéine, sérine) ou des oxazolines (thréonine). Un exemple de thiazoline est observé dans la bacitracine A, un antibiotique produit par *Bacillus subtilis* [Konz et al., 1997]. La vibriobactine, un sidérophore produit par *Vibrio cholerae*, contient une oxazoline [Keating et al., 2000]. Les cycles thiazoles ou oxazoles requièrent la présence d'un domaine additionnel. Le domaine d'oxydation (Ox) forme des cycles thiazoles et oxazoles, respectivement à partir des thiazolines et des oxazolines (figure 1.13). L'épothilone, agent anti-tumoral produit par *Sorangium cellulosum*, contient un cycle thiazole [Molnár et al., 2000].

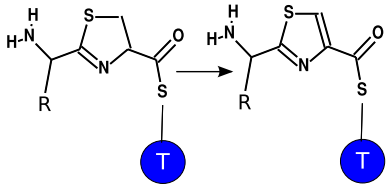


FIG. 1.13 – Réaction d'oxydation sur l'acide aminé fixé sur le domaine de thiolation

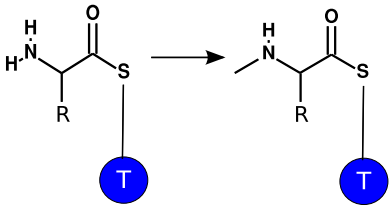


FIG. 1.14 – Réaction de méthylation sur l'acide aminé fixé sur le domaine de thiolation

Le domaine de méthylation (M) ajoute un groupement méthyl sur l'atome d'azote (N) de l'acide aminé (figure 1.14). Les acides aminés N-méthylés sont fréquents au sein des peptides NRPS. Par exemple, la cyclosporine A possède 7 acides aminés N-méthylés sur les 11 qu'elle contient [Weber et al., 1994].

Le domaine de formylation (F) ajoute un groupement formyle sur l'atome d'azote de l'acide aminé (figure 1.15).

Une glutamine N-formylée est présente au sein de l'anabaenopeptilide 90-A [Rouhiainen et al., 2000]. Dans la myxocheline A le groupement carboxyl C-terminal est réduit en un groupement aldéhyde [Li et al., 2008]. Cette réaction est réalisée par le domaine de la réductase (R) qui remplace le domaine Te dans ce cas particulier (figure 1.16).

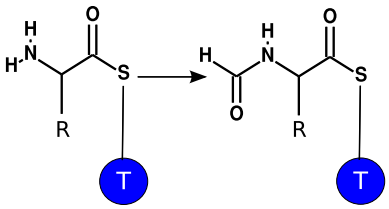


FIG. 1.15 – Réaction de formylation sur l'acide aminé fixé sur le domaine de thiolation

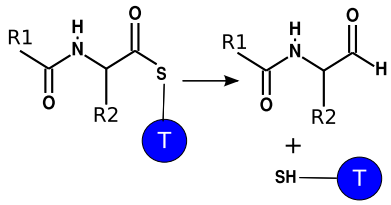


FIG. 1.16 – Réaction de réduction sur l'acide aminé fixé sur le domaine de thiolation

Autres modifications

Certains peptides produits par la voie non-ribosomiale peuvent contenir un acide gras lié à l'extrémité N-terminale. Ce sont les lipopeptides. Dans ce cas, le module d'initiation commence souvent par un domaine de condensation supplémentaire. Ce domaine effectue la liaison entre l'acide gras et le premier acide aminé.

Après la biosynthèse, il est possible que le peptide subisse des modifications comme une glycosylation ou une halogénation. Ces modifications sont réalisées par des enzymes associées aux synthétases. Par exemple, la vancomycine subit des glycosylations et des halogénations [van Wageningen et al., 1998].

De nombreux peptides mixtes NRPS-PKS ont été identifiés. Les polykétides (ou polycétides) sont synthétisés par la condensation itérative d'acides carboxyliques par des enzymes spécialisées, les polykétides synthétases (PKS) [Schwarzer and Marahiel, 2001]. Cette synthèse montre une grande ressemblance avec celle des acides gras. Il existe 3 types de PKS. Les PKS de type II

forment un ensemble de protéines mono-fonctionnelles. Les PKS de type III n'ont pas de domaines ACP (*acyl carrier protein*). Les PKS de type I sont divisées en deux sous-types : les PKS de type I itératives, qui réutilisent des domaines, et les PKS de type I modulaires. Ces dernières possèdent une organisation modulaire comparable aux NRPS. Chaque module est responsable de l'incorporation d'une unité propionate ou acétate au sein du polykétide en formation. Chaque module est lui même organisé en domaines. Comme dans la synthèse non-ribosomiale, certains domaines sont essentiels, d'autres facultatifs. Les quatre domaines principaux sont :

- le domaine de la kétosynthase (KS) : il catalyse la réaction d'élongation lors de la synthèse.
- le domaine de l'acétyltransférase (AT) : il transfert l'unité activée vers le domaine ACP. Ce domaine est comparable au domaine A des NRPS.
- le domaine *acyl carrier protein* (ACP) : il a la même fonction que le domaine T des NRPS, c'est-à-dire la fixation de l'unité carbonée sur l'enzyme via une liaison thioester.
- le domaine de la thioestérase (Te) : il libère le polykétide.

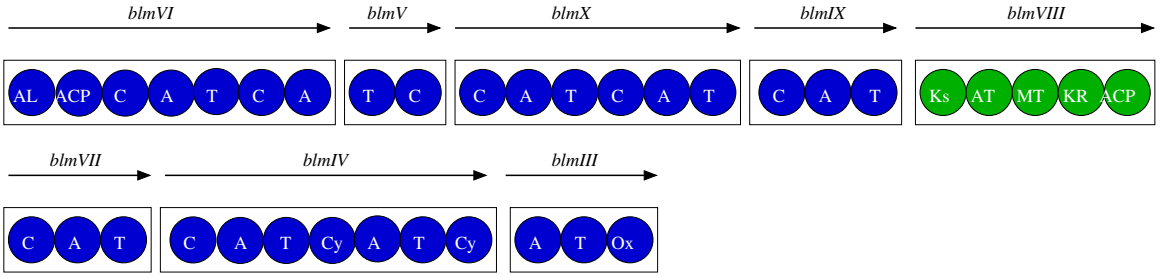


FIG. 1.17 – Organisation des enzymes impliquées dans la synthèse de la bléomycine
A : adénylation ; ACP : acyl carrier protein ; AL : acylCoA ligase ; AT : acyltransférase ; C : condensation ; Cy : cyclisation ; KR : kétoréductase ; KS : kétoacyl synthase ; MT : méthyltransférase ; Ox : oxidation ; PCP : peptidyl carrier protein

Les domaines secondaires sont la kétoréductase (KR), la déhydratase (DH) et l'énoylréductase (ER). Le domaine KR forme un groupement hydroxyl à partir de la partie cétone de l'intermédiaire. Le domaine DH déhydrate le groupement hydroxyle menant à la formation d'une double liaison entre deux atomes de carbones. Enfin, le domaine ER réduit le groupement énoyl en groupement alkyl.

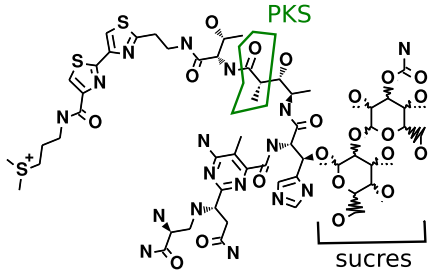


FIG. 1.18 – Structure de la bléomycine. La partie synthétisée par les domaines PKS est encadrée.

La figure 1.17 montre l'organisation des enzymes impliquées dans la biosynthèse d'un peptide mixte NRPS-PKS avec l'exemple de la bléomycine produite par *Streptomyces verticillus* [Shen et al., 2002, Shen et al., 2001]. La bléomycine (figure 1.18) est un agent anti-cancéreux utilisé dans le traitement des lymphomes ou du cancer des testicules.

1.2.3 Motifs caractéristiques et spécificité des domaines

Motifs caractéristiques des domaines

Les différents domaines jouent un rôle spécifique dans la biosynthèse peptidique. En effet, chaque domaine présente une fonction bien définie dans l'incorporation d'un acide-aminé, par conséquent, il doit exister au sein des séquences protéiques de chaque domaine, des acides aminés très conservés impliqués directement dans la fonction du domaine. En comparant plusieurs séquences d'un domaine donné, des motifs très conservés et caractéristiques du domaine étudié ont pu être mis en évidence [Marahiel et al., 1997, Schwarzer et al., 2003]. Un ensemble de motifs est caractéristique d'un domaine.

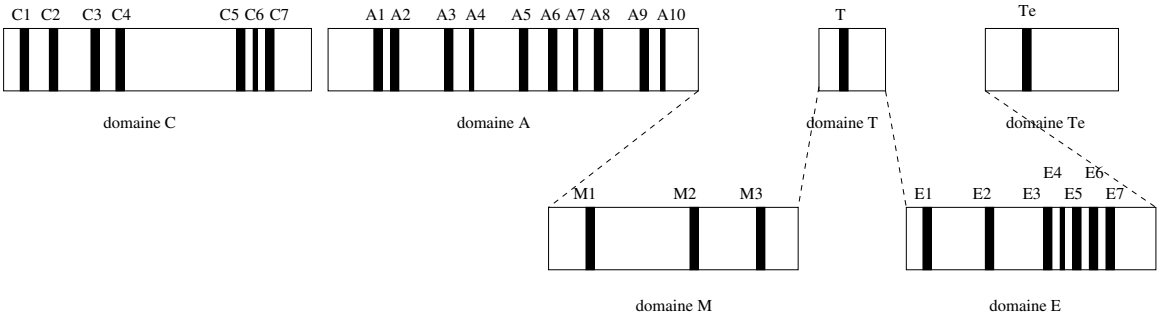


FIG. 1.19 – Motifs caractéristiques de quelques domaines [Marahiel et al., 1997]

	Domaine de condensation (C)		**Domaine de méthylation (M)**
C1	SxAQxR(LM)(WY)xL	M1	VL(DE)xGxGxG
C2	RHExLRTxF	M2	NELSxYRYxAV
C3	MHHxISDG(WV)S	M3	VExSxARQxGxLD
C4	YxD(FY)AVW		**Domaine de thiolation (T)**
C5	(IV)GxFVNT(QL)(CA)xR	T	LGG(DH)SL
C6	(HN)QD(YD)PFE		**Domaine d'épimérisation (E)**
C7	RDxSRNPL	E1	PIQxWF
	Domaine d'adénylation (A)	E2	HHxISDG(WV)S
A1	L(TS)YxEL	E3	DxLLxAxG
A2	LKAGxAYL(VL)P(LI)D	E4	EGHGRE
A3	LAYxxYTSG(ST)TGxPKG	E5	RTVGWFTxxYP(YV)PFE
A4	FDxS	E6	PxxGxGYG
A5	NxYGPTE	E7	FNYLG(QR)
A6	GELxJGx(VL)ARGYL		**Domaine de la thioestérase (Te)**
A7	Y(RK)TGDL	Te	GSxG
A8	GRxPXQVQIRGxIRIELGEIE		
A9	LPxYM(IV)P		
A10	NGK(VL)DR		

TAB. 1.2 – Motifs conservés caractéristiques de domaines impliqués dans la synthèse non-ribosomiale. Dans un motif, la liste des acides aminés possibles à une position donnée est indiquée entre parenthèses. Le X symbolise n'importe quel acide aminé.

La position des différents motifs est également conservée. En effet, le repliement de la protéine

lui confère sa fonction, par conséquent les motifs caractéristiques ont une localisation bien précise au sein de la séquence protéique. La figure 1.19 montre les motifs caractéristiques au sein de quelques domaines. La séquence de ces motifs est donnée dans la table 1.2.

Les acides aminés composant ces motifs jouent un rôle important dans la fonction du domaine. Les domaines A ont la fonction très importante de la sélection de l'acide-aminé. Une région comprise entre les motifs A4 et A5 confère la spécificité du substrat aux domaines A.

Spécificité des domaines d'adénylation

La structure 3D du domaine d'adénylation qui active la phénylalanine (PheA) dans la synthèse de la gramicidine S a été obtenue par cristallographie [Conti et al., 1997]. Elle a permis de mettre en évidence dix acides aminés situés dans le site de liaison au substrat et qui interviennent directement dans la sélection et l'activation de ce dernier. La figure 1.20 montre de manière schématique ces dix acides aminés au sein du site de fixation. Le substrat apparaît en rouge au centre. Les dix acides aminés sont situés dans un rayon de $5,5\mathring{A}$ autour du substrat. Parmi ces dix acides aminés, la lysine en position 517 est invariable car elle interagit avec l'AMP de l'adénylate.

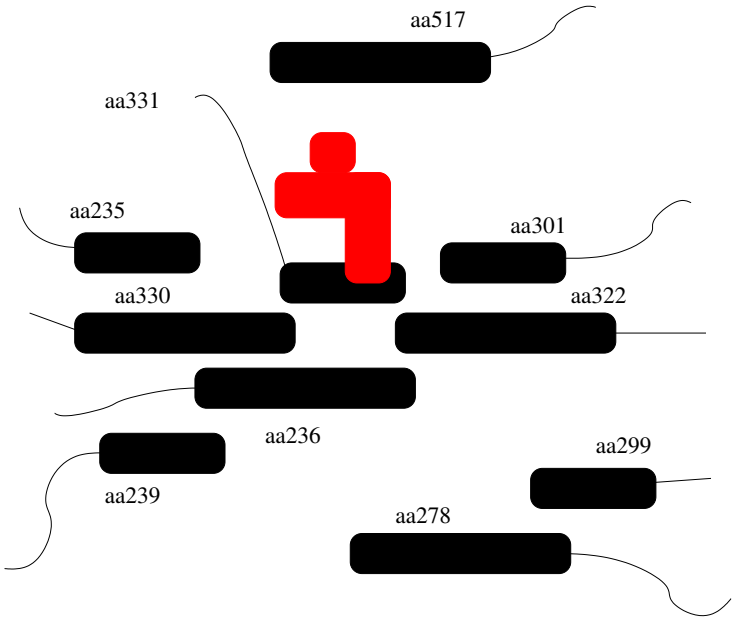


FIG. 1.20 – Représentation schématique des dix acides aminés impliqués dans le site de fixation du substrat au sein d'un domaine d'adénylation [Stachelhaus et al., 1999]

La structure cristallographique de ce domaine a permis d'établir un code conférant la spécificité des domaines A [Stachelhaus et al., 1999]. Ce code, appelé 'code NRPS' ou 'code Stachelhaus', peut rappeler le code génétique de la synthèse peptidique classique. Mais, il propose un acide aminé probablement incorporé dans le peptide final contrairement au code génétique qui n'est pas ambigu (sauf dans certains cas rares). La spécificité d'un domaine A peut être prédite à partir des dix acides aminés caractéristiques. Cependant, la prédiction du substrat avec cette méthode n'est applicable que dans 80% des cas. En effet, pour certains domaines A, ces dix acides aminés ne suffisent pas pour prédire le substrat. De plus, la prédiction de l'acide

aminé capté par un domaine A ne suffit pas à prédire le peptide final car différentes biosynthèses peptidiques non-ribosomiales existent.

1.2.4 Biosynthèse peptidique non-ribosomiale

Différents modes de biosynthèse

Il existe trois modes de biosynthèse [Mootz et al., 2002b]. Le premier est la **biosynthèse linéaire** (cf. figure 1.21 avec l'exemple de l'ACV).

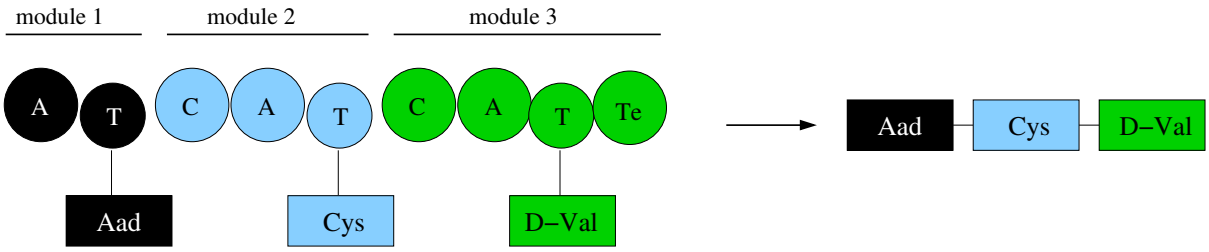


FIG. 1.21 – La biosynthèse linéaire : exemple de l'ACV

Dans cette biosynthèse, il y a colinéarité entre la synthétase et le peptide produit : l'enchaînement des acides aminés incorporés dans le produit correspond à celui des modules de la synthétase. La séquence du peptide est entièrement déterminée par le nombre et l'ordre des modules au sein de la synthétase. Dans certains cas, le domaine Te peut être remplacé par un domaine spécial menant à la cyclisation du peptide produit. Cette synthèse forme des peptides linéaires ou cycliques (totalement ou partiellement). Elle est la plus fréquente. Par exemple, la cyclosporine A, immuno-suppresseur produit par *Beauveria nivea* [Weber et al., 1994], l'ACV [Smith et al., 1990] ou encore la fengycine, antibiotique produit par *Bacillus subtilis* [Lin et al., 1999], sont produits par une biosynthèse linéaire.

Le deuxième mode rencontré est la **biosynthèse itérative** (cf. figure 1.22 avec l'exemple de l'entérobactine).

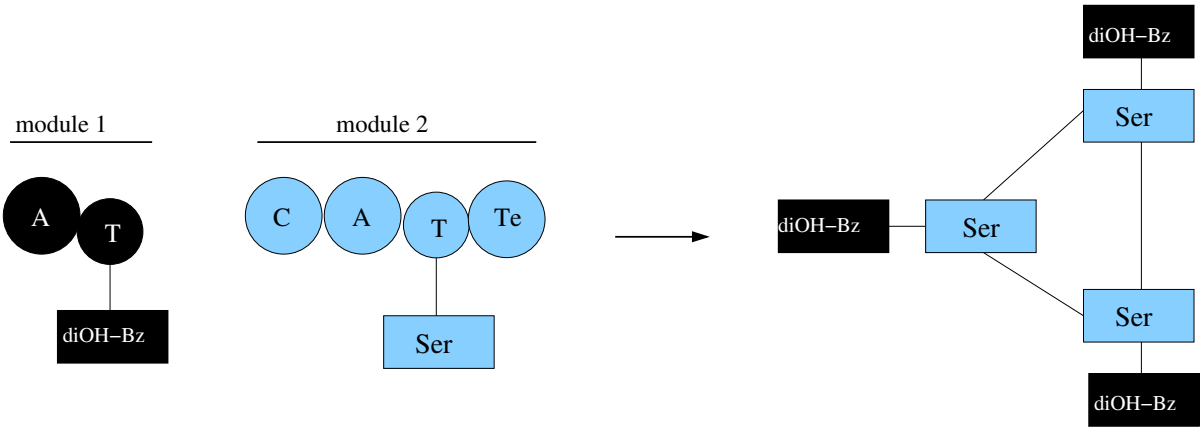


FIG. 1.22 – La biosynthèse itérative : exemple de l'entérobactine

Dans ce cas, certains modules de la synthétase sont utilisés plus d'une fois lors de la synthèse d'un même peptide. Le domaine Te est souvent à l'origine de l'itération. Les peptides pro-

duits contiennent alors un motif peptidique de taille variable répété au moins deux fois. Cette biosynthèse mène à des peptides linéaires, cycliques ou branchés. Par exemple la gramicidine S est un dimère d'un pentapeptide [Kohli et al., 2001]. L'enniatine est un trimère cyclique d'un dipeptide [Haese et al., 1993].

Enfin, le dernier mode est la **biosynthèse non-linéaire** (cf. figure 1.23 avec l'exemple de la vibriobactine).

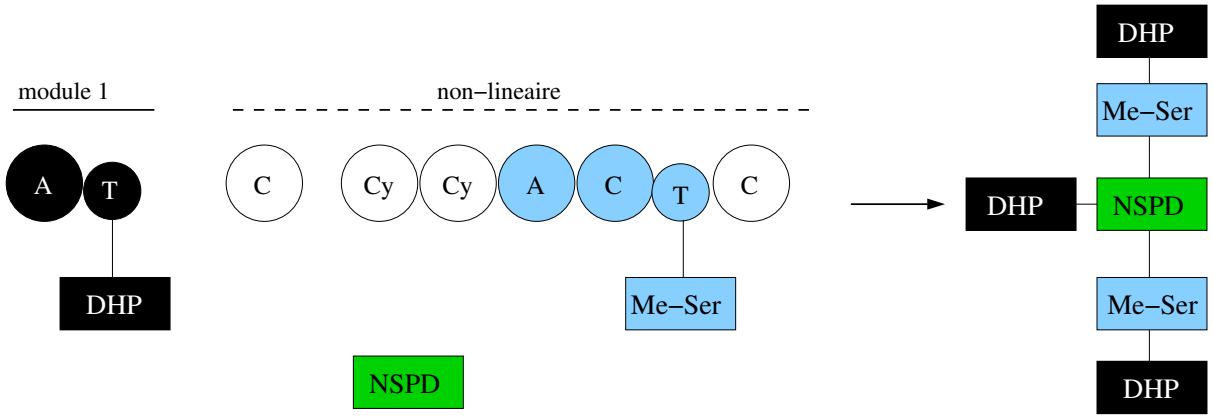


FIG. 1.23 – La biosynthèse non-linéaire : exemple de la vibriobactine

Dans cette biosynthèse, l'agencement des modules de la synthétase est différent de celui des monomères du peptide. L'organisation des domaines à l'intérieur des modules est différente de l'organisation classique $(CAT)_n$. De plus, des résidus provenant d'une autre voie de synthèse peuvent également être incorporés, comme c'est le cas de la norspermidine (NSPD) de la figure 1.23. La biosynthèse non-linéaire peut mener à des hétérocyclisations (cyclisations internes) ou à des branchements. Par exemple, la syringomycine, biosurfactant antibiotique produit par *Pseudomonas syringae* [Guenzi et al., 1998], ou encore la yersiniabactine, sidérophore produit par différentes espèces de *Yersinia* [Pelludat et al., 1998], sont synthétisées par cette voie.

Production de variants

Une autre caractéristique de la synthèse peptidique non-ribosomiale est que cette voie de synthèse originale peut conduire à la production de différents variants. Les variants sont des peptides ayant une structure et une composition proches. Par exemple, le terme de variant est utilisé pour les peptides présentant une séquence consensus avec des acides aminés variables à des positions données. Des variants environnementaux ou génétiques sont observés.

Les **variants environnementaux** sont obtenus à partir de changements du milieu dans lequel évolue l'organisme. En effet, selon les acides aminés présents dans le milieu, la composition du peptide produit peut varier. Par exemple, il a été montré que différents variants de la cyclosporine peuvent être obtenus en fonction du milieu de culture [von Döhren et al., 1997]. La table 1.3 montre quelques exemples de variants de la cyclosporine.

Une autre étude portant sur les microcystines synthétisées au sein d'une population de cyanobactéries [Kurmayer et al., 2002] a montré que la diversité des variants de la microcystine résulte de l'activation d'acides aminés variés durant la biosynthèse. Les domaines d'adénylation de la synthétase peuvent activer non pas un seul acide aminé spécifique, mais un ensemble d'acides aminés possibles. En effet, même si le domaine A sélectionne un acide aminé spécifique, en cas

variant	position de l'acide aminé dans la chaîne peptidique										
	1	2	3	4	5	6	7	8	9	10	11
A	D-Ala	MeLeu	MeLeu	MeVal	MeBmt	Abu	Sar	MeLeu	Val	MeLeu	Ala
B	D-Ala	MeLeu	MeLeu	MeVal	MeBmt	Ala	Sar	MeLeu	Val	MeLeu	Ala
C	D-Ala	MeLeu	MeLeu	MeVal	MeBmt	Thr	Sar	MeLeu	Val	MeLeu	Ala
D	D-Ala	MeLeu	MeLeu	MeVal	MeBmt	Val	Sar	MeLeu	Val	MeLeu	Ala
E	D-Ala	MeLeu	MeLeu	Val	MeBmt	Abu	Sar	MeLeu	Val	MeLeu	Ala
I	D-Ala	Leu	MeLeu	MeVal	MeBmt	Val	Sar	MeLeu	Val	MeLeu	Ala
V	D-Ala	MeLeu	MeLeu	MeVal	MeBmt	Abu	Sar	MeLeu	Val	MeLeu	Abu

TAB. 1.3 – Quelques variants de la cyclosporine produits par *Tolypocladium niveum* dans différents milieux de culture [von Döhren et al., 1997]. La cyclosporine est un peptide cyclique composé de 11 acides aminés. MeLeu :N-méthyl-leucine, MeVal :N-méthyl-valine, MeBmt :N-méthyl-4-buténynl-4-méthylthréonine, Abu :acide 2-Aminobutyrique, Sar :sarcosine

d'absence de ce dernier, un acide-aminé présentant des propriétés physico-chimiques proches, peut être incorporé. Ainsi, selon les acides aminés disponibles dans le milieu, l'acide aminé incorporé est différent et un nouveau variant est obtenu.

Les variants peuvent également avoir une **origine génétique**. En effet, une séquence nucléique différente au niveau du gène de la synthétase conduit à la production d'un peptide différent. Par exemple, la séquence de la synthétase produisant la bacillomycine D [Moyne et al., 2004] est différente de celle produisant la bacillomycine L [Hofemeister et al., 2004] dans des souches différentes. Un autre exemple est celui des pyoverdines. Les pyoverdines sont des sidérophores produits par différentes espèces de *Pseudomonas*. Selon l'espèce et la souche, les gènes codant les synthétases sont différents et conduisent à des pyoverdines très diverses [Meyer, 2000].

En général, les variants environnementaux présentent des structures très proches. En effet, seuls quelques acides aminés, souvent à des positions conservées, varient. Au contraire, les variants génétiques peuvent être très différents les uns des autres comme c'est le cas pour les pyoverdines qui présentent des structures très diverses.

La synthèse peptidique non-ribosomiale est une voie de synthèse originale conduisant à la synthèse de peptides très divers et présentant de nombreuses particularités. Ces peptides présentent également un grand intérêt de par leurs fonctions biologiques importantes et variées.

1.3 Peptides non-ribosomiaux

Les peptides non-ribosomiaux présentent de nombreuses particularités par rapport aux peptides synthétisés par la voie classique. D'un côté, les plus petits peptides non-ribosomiaux sont des dipeptides tels que la bacilysine [Tabata et al., 2005]. D'un autre, les polythéonamides contiennent 48 acides aminés [Hamada et al., 2005] et sont à ce jour les plus grands peptides non-ribosomiaux identifiés. Les peptides NRPS montrent une grande diversité au niveau de leur composition, de leur structure primaire et de leurs activités biologiques.

1.3.1 Diversité de composition

La biosynthèse non-ribosomiale utilise, en plus des 20 acides aminés protéogéniques (section 1.1.1), beaucoup d'autres acides aminés. En effet, plusieurs centaines d'acides aminés différents

peuvent être sélectionnés et incorporés par les domaines A au sein des peptides durant la biosynthèse. Par exemple, l'acide 3-méthyl-glutamique et la kynurénine sont deux acides aminés non-protéogéniques sélectionnés par les domaines A et incorporés au sein de la daptomycine (figure 1.24) [Miao et al., 2005]. La daptomycine est un antibiotique utilisé en cas d'infection par des bactéries à gram positif.

De plus, les synthétases peuvent contenir des domaines secondaires capables de modifier les acides aminés incorporés, ce qui augmente encore la diversité des acides aminés identifiés au sein des peptides produits. La daptomycine (figure 1.24) possède trois domaines d'épimérisation permettant d'obtenir les formes D des acides aminés incorporés par les modules contenant ces domaines secondaires. La bacitracine A (figure 1.24) contient un hétérocycle formé grâce à un domaine Cy. Elle contient également des acides aminés sous forme D obtenus grâce à des domaines E [Konz et al., 1997]. La cyclosporine A (figure 1.24) contient des acides aminés N-méthylés obtenus par des domaines de méthylation [Lawen and Zocher, 1990]. La 4-butényl-4-methyl-thréonine est reconnue et activée par un domaine A puis N-méthylée par un domaine de méthylation.

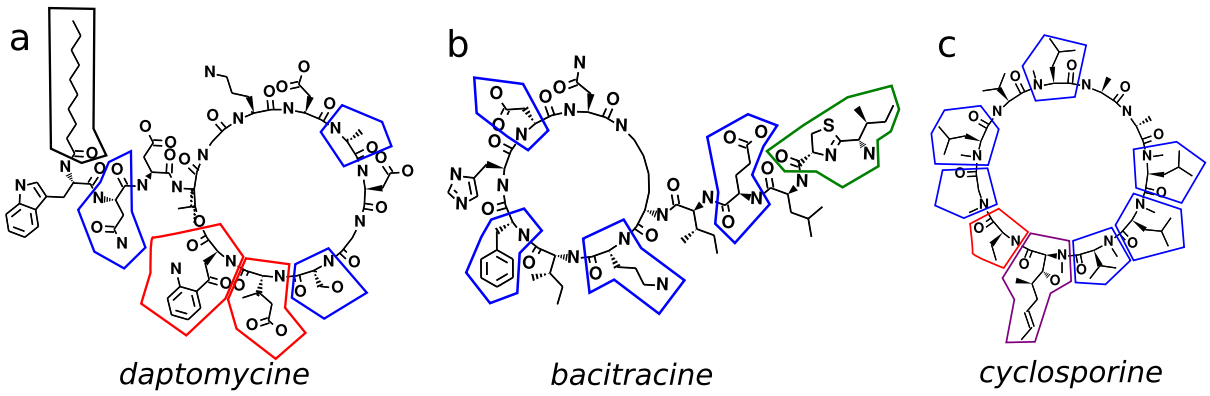


FIG. 1.24 – Structure de a) la daptomycine, b) la bacitracine A et c) la cyclosporine. Les acides aminés modifiés par des domaines secondaires sont encadrés en bleu, les acides aminés non-protéogéniques en rouge, l'hétérocycle (obtenu par un domaine secondaire) en vert, l'acide aminé non-protéogénique et modifié par un domaine secondaire en violet et l'acide gras en noir.

Une autre particularité des peptides non-ribosomiaux est l'incorporation de composés provenant d'autres voies de biosynthèse. Comme nous l'avons vu précédemment, un acide gras peut être greffé à la partie peptidique. Il existe beaucoup de lipopeptides provenant de la voie NRPS [Strieker and Marahiel, 2009]. Par exemple, la daptomycine (figure 1.24) contient un acide gras. Des sucres ou des acides carboxyliques peuvent également être ajoutés lors de la synthèse peptidique. Enfin, il existe des peptides hybrides NRPS/PKS qui contiennent à la fois des unités peptidiques et polykétides. La bléomycine (figure 1.18) contient à la fois des sucres et des unités provenant de la synthèse PKS.

L'utilisation d'acides aminés non-protéogéniques, la modification des acides aminés incorporés par les domaines secondaires, l'incorporation d'acide gras, de sucres, d'acides carboxyliques ou encore de polykétides au sein des peptides non-ribosomiaux mènent à une très grande diversité de composition de ces peptides. C'est pourquoi nous utiliserons par la suite le terme de « **monomère** » plutôt que celui d'acide aminé dans le cas des peptides non-ribosomiaux.

En plus de la diversité des monomères incorporés, une diversité au niveau de la structure primaire de ces peptides est également observée.

1.3.2 Diversité des structures primaires

Comme nous l'avons vu dans la section 1.1.2, les peptides et protéines classiques ont une structure primaire linéaire. Ce n'est pas toujours le cas pour les peptides non-ribosomiaux. En effet, ces peptides présentent souvent des structures primaires plus complexes. Les différentes biosynthèses (section 1.2.4) ainsi que certains domaines (section 1.2.2) forment des structures primaires complexes. La figure 1.25 montre la diversité des structures primaires observées pour les peptides non-ribosomiaux.

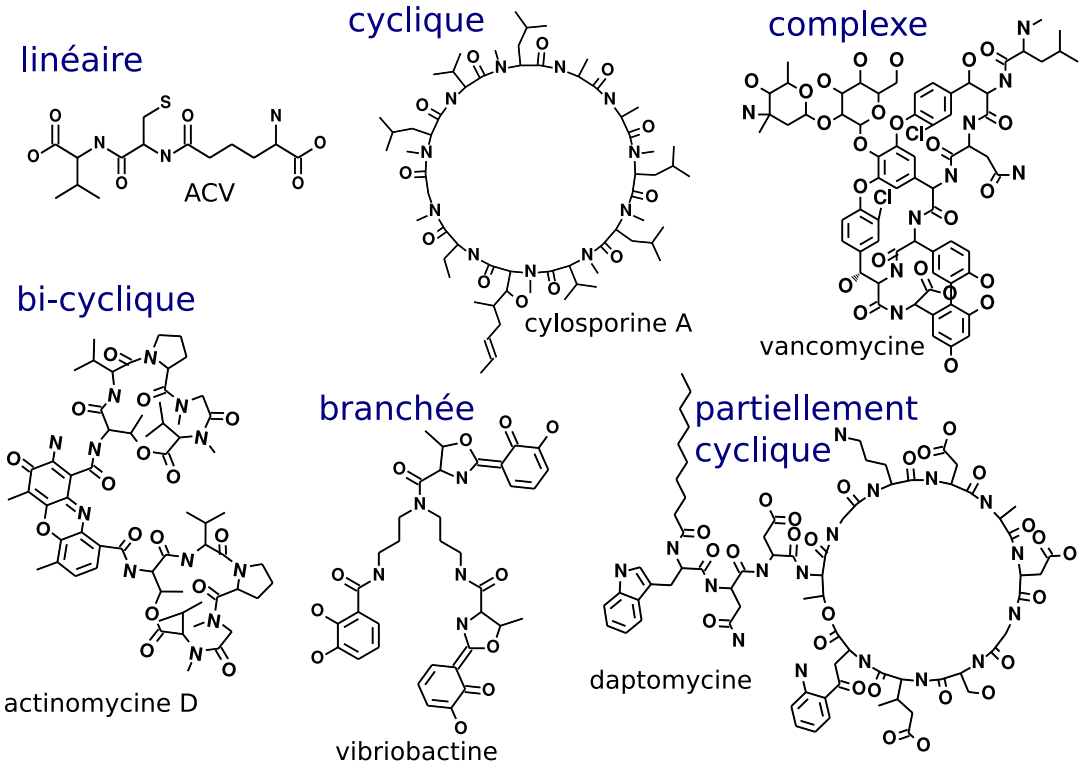


FIG. 1.25 – Diversité des structures primaires

Comme les peptides classiques, certains peptides non-ribosomiaux ont une structure primaire linéaire. Par exemple, les peptaibols sont un grand groupe de peptides antibiotiques synthétisés par voie non-ribosomiale [Duclohier, 2007]. Ces peptides partagent des caractéristiques communes telles que leur richesse en acide 2-amino-isobutyrique (Aib) et leur structure primaire linéaire. Un autre exemple est l'ACV qui est un tripeptide linéaire.

Le domaine de la thioestérase (Te) libère le peptide néoformé. De plus, certains domaines Te permettent également la macrocyclisation du peptide produit [Grünewald and Marahiel, 2006]. La macrocyclisation mène à l'obtention de peptides totalement ou partiellement cycliques. Par exemple, la cyclosporine est un peptide totalement cyclique alors que la daptomycine contient un macrocycle et un branchement sur celui-ci menant à une structure partiellement cyclique.

La biosynthèse itérative peut produire des peptides contenant deux macrocycles. Par exemple, l'actinomycine D présente une structure bi-cyclique obtenue par la condensation de deux macrocycles [Pfennig et al., 1999].

La biosynthèse itérative et la biosynthèse non-linéaire produisent, entre autres, des structures

branchées. Par exemple, la vibriobactine, synthétisée par une biosynthèse non-linéaire, présente une structure branchée.

La voie non-ribosomiale est souvent associée à d'autres voies de biosynthèse, comme la voie PKS par exemple, et à des enzymes accessoires. Ces associations produisent des structures encore plus complexes que celles citées précédemment. La vancomycine présente une structure contenant des branchements et des cycles chevauchants. Elle est issue de la combinaison de différentes voies de biosynthèse dont la voie NRPS et la voie PKS auxquelles viennent s'ajouter des réactions de glycosylation, d'oxydation et d'halogénation [Geib et al., 2008].

La structure primaire d'un peptide est essentielle à sa fonction. Les peptides non-ribosomiaux montrant une grande diversité de structures primaires présentent également une grande diversité au niveau de leurs activités biologiques.

1.3.3 Diversité des activités biologiques

Les peptides non-ribosomiaux forment un ensemble très varié de produits naturels présentant un très large éventail d'activités biologiques et pharmacologiques [Schwarzer et al., 2003, Felnagle et al., 2008]. En effet, de nombreux peptides non-ribosomiaux sont actuellement utilisés pour le traitement de diverses maladies.

Antibiotiques

Certains peptides non-ribosomiaux sont des antibiotiques antibactériens. Ces substances tuent les bactéries ou inhibent leur croissance. Nous avons déjà cité l'exemple de la pénicilline synthétisée à partir du tripeptide-ACV par divers champignons et bactéries. Elle est utilisée dans le traitement d'infections bactériennes en inhibant la croissance des bactéries [Erlanger and Goode, 1967]. La daptomycine, commercialisée sous le nom de Cubicin®, est utilisée dans le traitement des infections de la peau par des bactéries à gram positif [Weis et al., 2008]. Elle est synthétisée par *Streptomyces roseosporus* comme un composant mineur du complexe A21978C. La vancomycine, synthétisée par *Nocardia orientalis*, est utilisée dans le traitement des infections hospitalières, contre certaines bactéries à gram positif [Levine, 2008]. Son spectre d'activité inclut, par exemple, les Staphylocoques, les Streptocoques ou encore les Pneumocoques.

Certains peptides non-ribosomiaux présentent une activité anti-fongique. Ces molécules sont capables de tuer les champignons ou d'inhiber leur croissance. Par exemple, la mycosubtiline, produite par *Bacillus subtilis*, inhibe la croissance de *Saccharomyces cerevisiae* [Besson and Michel, 1989] ou de *Fusarium oxysporum* [Leclère et al., 2005].

D'autres peptides non-ribosomiaux font partie de la classe des anti-viraux, substances permettant de lutter contre les virus. Par exemple, la surfactine, produite par *Bacillus subtilis*, possède une activité anti-virale [Kracht et al., 1999].

Par la suite, nous utiliserons le terme générique « antibiotique » pour désigner à la fois les anti-bactériens, les anti-fongiques et les anti-viraux sans aucune distinction.

Immuno-modulateurs

Il existe des peptides non-ribosomiaux présentant une activité immuno-modulatrice. Ces molécules inhibent ou activent le système immunitaire. Les molécules qui inhibent le système immunitaire sont appelés immuno-suppresseurs. L'exemple le plus connu est celui de la cyclosporine

A produite par *Tolypocladium inflatum*. Elle est utilisée après une greffe pour réduire les risques de rejet ou dans le traitement des maladies auto-immunes [Green, 1981].

Antitumoraux

Certains peptides non-ribosomiaux sont utilisés comme antitumoraux dans le traitement de certains cancers. La bléomycine A2 et la bléomycine B2, produites par *Streptomyces verticillus*, sont les principaux constituants de la Blenoxane®, utilisée dans le traitement du cancer de l'oesophage, des lymphomes, des carcinomes à cellules squameuses ou du cancer des testicules [Lazo, 1999]. L'actinomycine D, produite par diverses bactéries du genre *Streptomyces*, est un antibiotique présentant également une activité antitumorale. Elle est l'un des plus anciens traitements utilisés pour traiter certains cancers des os, de la peau, des testicules, ou encore de l'utérus [Koba and Konopa, 2005].

Sidérophores

La grande majorité des sidérophores bactériens et fongiques sont produits par la voie non-ribosomiale. Ces substances permettent la chélation du fer, c'est-à-dire la capture, dans le milieu extérieur, du fer nécessaire à la survie des microorganismes. Les sidérophores les plus connus sont les pyoverdines, produites par différentes espèces de *Pseudomonas* [Budzikiewicz, 2004], plus de 60 ont été identifiées [Budzikiewicz, 2004]. Ces molécules sont indispensables à la survie de l'organisme. Il existe d'autres sidérophores non-ribosomiaux tels que la yersiniabactine, produite par *Yersinia pestis* [Miller et al., 2002] ou encore l'entérobactine produite par différentes bactéries [Raymond et al., 2003].

Toxines

Certains peptides non-ribosomiaux sont des toxines. Une toxine est une substance toxique qui confère un pouvoir pathogène à son producteur. Par exemple, l'HC-toxine, produite par *Cochliobolus carbonum*, est déterminante pour la spécificité et la virulence du champignon envers son hôte, le maïs [Walton, 2006]. Les microcystines sont produites par les cyanobactéries [Lam et al., 2000] et plus de 80 variants ont été identifiés. Elles sont toxiques pour les plantes et les animaux, dont l'homme. Leur hépatotoxicité peut conduire à des dommages importants du foie.

Surfactants

Il existe également des surfactants. Ce sont des substances améliorant la miscibilité entre des phases hydrophiles et hydrophobes. Les surfactants incluent les détergents, les agents moussants, les agents dispersants ou encore les émulsifiants. Les biosurfactants présentent un grand intérêt car contrairement aux surfactants synthétiques, ils sont dégradés facilement. Par exemple, *Bacillus subtilis* produit la surfactine [Abdel-Mawgoud et al., 2008]. La lichenysine, produite par *Bacillus licheniformis*, a une structure proche de celle de la surfactine (une glutamine remplace l'acide glutamique de la surfactine) mais montre un pouvoir surfactant plus élevé que celui de la surfactine [Grangemard et al., 2001].

Autres

Certains peptides non-ribosomiaux montrent des activités biologiques plus rares. Par exemple, la cyclosporine A possède une activité anti-inflammatoire en plus de son activité immuno-modulatrice. L'indigoidine, produite par *Erwinia chrysanthemi*, est un pigment [Reverchon et al., 2002]. Enfin, quelques peptides non-ribosomiaux sont de fonction inconnue tels que l'anabaenopeptilide 90-A produit par la souche 90 d'*Anabaena* [Rouhiainen et al., 2000].

La grande diversité et l'importance des activités biologiques présentées par les peptides non-ribosomiaux font de ces peptides des molécules de grand intérêt. Modifier la voie de synthèse non-ribosomiale pourrait conduire à la production de composés plus actifs ou de composés nouveaux présentant des activités biologiques intéressantes.

1.4 Approches pour l'obtention de nouveaux peptides

L'un des objectifs actuels est la modification de produits naturels dans le but d'améliorer ou de changer leurs activités biologiques. Pour ce faire, deux approches existent. Les approches génétiques consistent à modifier la machinerie biosynthétique au niveau du génome. Les approches chemo-enzymatiques consistent à combiner des méthodes chimiques et enzymatiques pour créer des collections de nouvelles molécules [Sieber and Marahiel, 2005].

1.4.1 Approches génétiques

L'ordre et la composition des modules d'une synthétase naturelle sont le résultat d'une sélection pendant l'évolution pour produire un peptide présentant une bonne adéquation entre structure et activité. L'idée est de reprogrammer génétiquement la machinerie non-ribosomiale dans le but d'obtenir de nouveaux peptides. En effet, d'un point de vue théorique, il est possible de créer des synthétases en fusionnant différents modules ou domaines provenant de diverses NRPS connues, dans le but d'obtenir la synthèse d'un peptide sur mesure. Cependant, en réalité, il n'est pas facile de construire des NRPS sur mesure. Différentes stratégies ont été développées dans le but d'obtenir des peptides modifiés à partir de modifications génétiques au niveau des synthétases. Une première expérience de modification génétique a été menée sur le dernier module de la synthétase de la surfactine [Stachelhaus et al., 1995]. Le dernier module de la synthétase de la surfactine (position 7) contient les domaines C-A-T-Te et est chargé de l'incorporation de la leucine. Les domaines A-T ont été remplacés par diverses unités A-T provenant de bactéries et de champignons et présentant des spécificités pour des acides aminés différents. Des variants de la surfactine avec une valine, une ornithine et une phénylalanine en position 7 ont été obtenus. Cependant, le taux de production de ces variants est faible, surement à cause de la spécificité du domaine C. En effet, les domaines de condensation (C) montrent également une spécificité par rapport au substrat. Cette étude a également montré que la même expérience menée cette fois avec le second domaine de la synthétase de la surfactine ne produit pas de variants, surement à cause de l'existence de régions de liaison entre les différents domaines. En effet, des séquences d'environ 15 acides aminés, situées entre les différents domaines, semblent importantes pour l'activité enzymatique. Ces régions, appelées régions de liaison, présentent des séquences variables et une grande flexibilité. De plus, leur localisation en font des cibles intéressantes pour des fusions artificielles sans perte de l'activité enzymatique.

La stratégie par fusion artificielle a été testée pour la première fois avec la synthétase de la tyrocidine [Mootz et al., 2000]. Le module 2 qui incorpore la proline a été fusionné soit avec le

module 9 qui incorpore l'ornithine soit avec le module 10 qui incorpore la leucine. Ces deux enzymes hybrides ont été incubées avec le module 1, qui incorpore la D-phénylalanine, et ont produit des tripeptides artificiels D-Phe-Pro-Orn et D-Phe-Pro-Leu. La libération du peptide a été observée seulement en présence d'un domaine Te à l'extrémité C-terminale. Une autre expérience a été réalisée sur le gène *dptD* intervenant dans la synthèse de la daptomycine [Doekel et al., 2008]. Dans cette étude, les régions de liaison ont été utilisées pour fusionner le module 12 de la daptomycine, aux modules terminaux du peptide CDA (calcium-dependent antiobiotics), incorporant le tryptophane (Trp), et du peptide A54145 incorporant l'isoleucine (Ile). Ainsi, de nouveaux dérivés de la daptomycine présentant un Trp ou une Ile en position 13 ont été obtenus, sans diminution des taux de production.

Beaucoup de peptides naturels contiennent des hétérocycles directement impliqués dans la fonction biologique. Une synthétase hybride a été construite en utilisant un domaine de cyclisation (Cy) [Duerfahrt et al., 2004]. Dans cette étude le module 1 A-T de la synthétase de la bacitracine, incorporant l'isoleucine a été fusionné avec soit le module Cy-A-T activant la thréonine de la synthétase de la myxobactine soit avec le module Cy-A-(Ox)-T activant la cystéine de la synthétase du myxothiazole. Le domaine d'oxydation (Ox) forme un cycle thiazole à partir de la thiazoline. Pour assurer la libération du peptide, le domaine Te de la tyrocidine a été ajouté aux deux hybrides. Les deux peptides attendus ont été obtenus, à savoir Ile-Thr-oxazoline et Ile-Cys-thiazole. Cependant, le produit majoritaire est Ile-Ser-oxazole, ce qui montre la tolérance des domaines A pour différents acides aminés. Cette étude a montré qu'il est possible d'obtenir des hétérocycles en reprogrammant la machinerie NRPS. Cependant, le taux de production est faible et les peptides produits ne sont pas toujours ceux attendus.

En plus de l'échange de modules ou de domaines, une autre stratégie peut être utilisée. La délétion ou l'insertion de modules peut modifier le peptide produit. Lors d'une étude, le module 2 de la synthétase de la surfactine a été délété [Mootz et al., 2002a]. Cette manipulation a conduit à la production d'un peptide non-naturel contenant une leucine en moins par rapport à la surfactine. La recombinaison de modules entiers représente une intervention drastique qui peut mener à une efficacité enzymatique réduite et à des taux de production faibles. Une stratégie plus conservatrice consiste à modifier la spécificité des domaines d'adénylation (A).

Une seule mutation ponctuelle peut changer la spécificité d'un domaine A. Par exemple, la substitution d'un seul acide aminé parmi les 10 résidus critiques du code Stachelhaus peut changer la spécificité d'un domaine A sélectionnant l'acide glutamique (Glu) en un domaine sélectionnant la glutamine (Gln). En effet, le code Stachelhaus du domaine A sélectionnant Glu et celui du domaine A sélectionnant Gln ne diffèrent que par un seul acide aminé. Cette expérience a été réalisée sur le module 1 (activant Glu) de la surfactine [Eppelmann et al., 2002]. Comme attendu, cette mutation a bien mené à l'obtention d'un peptide dans lequel Glu est remplacé par Gln, sans réduction de l'efficacité enzymatique.

Il existe également une autre stratégie basée sur la délétion d'un gène intervenant dans la synthèse d'un acide aminé non-protéogénique. Cette méthode est appelée « mutasynthèse ». Cette stratégie a été utilisée pour la production d'analogues du peptide CDA (calcium-dependent antibiotics) [Hojati et al., 2002]. Le gène codant une enzyme intervenant dans la synthèse de l'HPG (hydroxy-phénylglycine) a été délété. En présence de dérivés de l'HPG, ces derniers sont incorporés dans le peptide à la place de l'HPG et de nouveaux analogues du CDA sont obtenus. Cette méthode produit des peptides non-naturels. Cependant, la mutasynthèse ne s'applique qu'aux acides aminés non-protéogéniques et seuls des dérivés de l'acide aminé d'origine peuvent être incorporés.

L'amélioration des connaissances sur la machinerie non-ribosomiale a permis des progrès

considérables dans les approches génétiques, menant à la production de produits non-naturels ou semi-naturels. Cependant, nous sommes encore loin de pouvoir construire des synthétases sur mesure. De plus, certains organismes produisant des peptides d'intérêt sont difficiles à cultiver en laboratoire. D'autres techniques produisent des peptides modifiés, les approches chemo-enzymatiques.

1.4.2 Approches chemo-enzymatiques

L'un des objectifs actuels est l'identification rapide de pharmacophores par synthèse puis screening des activités. Les approches génétiques représentent une voie intéressante pour la production de nouveaux composés, mais elle demande beaucoup de travail et d'efforts, ainsi que des connaissances supplémentaires sur la voie non-ribosomiale. Les approches chemo-enzymatiques allient des méthodes chimiques et des méthodes enzymatiques. Il est possible d'obtenir rapidement des librairies de peptides par synthèse chimique sur phase solide. Cependant, les réactions de cyclisation posent souvent problème. En effet, il est difficile d'obtenir une cyclisation car elle est défavorable d'un point de vue énergétique. Les méthodes enzymatiques permettent de pallier ces problèmes de cyclisation.

Des études combinent la force de la synthèse peptidique chimique à la force enzymatique des domaines Te [Sieber and Marahiel, 2003, Kohli and Walsh, 2003]. En effet, il est possible d'obtenir un peptide linéaire par synthèse chimique sur phase solide puis de former un peptide cyclique à l'aide d'un domaine Te. Cependant, ces domaines sont sélectifs. Chacun est spécialisé dans la cyclisation d'un peptide donné. Des études ont été menées pour évaluer la spécificité du substrat et les restrictions enzymatiques des domaines Te. Le domaine Te de la tyrocidine accepte des changements d'acides aminés sauf pour les acides aminés terminaux, ainsi que des changements de longueur du peptide [Trauger et al., 2001]. Les mêmes études ont été menées avec le domaine Te de la surfactine [Tseng et al., 2002]. Comme pour la tyrocidine, les acides aminés terminaux ne peuvent pas être changés, par contre le domaine Te de la surfactine ne tolère pas les variations de longueur du peptide. Le domaine Te de la tyrocidine semble être un bon candidat en tant que catalyseur de la cyclisation.

La tyrocidine agit en déstabilisant la membrane plasmique des bactéries mais aussi celle des cellules humaines et ne peut donc pas être utilisée comme antibiotique. Des variants de la tyrocidine ont été produits dans le but d'être plus sélectifs contre les bactéries [Kohli et al., 2002]. Une collection de peptides a été obtenue par synthèse chimique avec cyclisation enzymatique. Parmi ces variants non-naturels, deux ont montré une meilleure sélectivité contre les membranes bactériennes.

La synthèse chimique est souvent longue et très onéreuse. De plus, la plupart des antibiotiques utilisés en médecine sont produits par fermentation, c'est-à-dire par des micro-organismes. L'idée ici est de combiner les différentes approches. Tout d'abord, les approches chemo-enzymatiques produisent des collections de peptides non-naturels. Ensuite, les peptides présentant des activités intéressantes sont recherchés parmi cette collection. Enfin, des approches génétiques sont développées dans le but de produire les peptides d'intérêt par fermentation. Les approches génétiques et chemo-enzymatiques sont donc complémentaires et non-concurrentes.

Dans cette section, nous avons vu que la synthèse peptidique non-ribosomiale prend de plus en plus d'importance dans la littérature. Cela s'explique tout d'abord par les activités biologiques importantes présentées par les peptides produits par cette voie de biosynthèse encore trop méconnue. En effet, de nombreux peptides utilisés aujourd'hui dans le traitement de diverses maladies sont produits par la voie NRPS. De plus, le séquençage de nombreux génomes de micro-organismes a permis l'identification de nombreux gènes codant des synthétases. Les connaissances acquises et le grand intérêt pour la voie non-ribosomiale ont conduit au développement d'outils bioinformatiques dédiés à cette voie de biosynthèse.

Chapitre 2

Outils bioinformatiques existants

Du fait du grand intérêt de la voie non-ribosomiale, ainsi que des peptides produits par cette voie, des outils bioinformatiques dédiés ont été développés. Ces outils peuvent être divisés en deux catégories. Tout d'abord, il existe des banques et bases de données contenant des informations sur les synthétases et les produits. Par ailleurs, des outils d'analyse des synthétases ont été développés.

2.1 Banques et bases de données

Il existe à l'heure actuelle de nombreuses banques et bases de données publiques accessibles sur le Web. Les deux termes « banque de données » et « base de données » sont souvent confondus. La différence réside dans le fait qu'une base de données s'appuie sur un schéma et un logiciel de gestion de bases de données alors qu'une banque de données mémorise les données dans des "fichiers texte" formatés et des logiciels sont développés pour accéder aux données. En pratique, les deux termes sont utilisés pour définir un ensemble d'informations autour d'un domaine donné. Deux catégories de bases de données existent : les bases de données généralistes, contenant un grand nombre d'informations sur des domaines vastes, et les bases de données spécialisées, contenant des informations sur des domaines très ciblés.

2.1.1 Banques et bases de données généralistes

Certaines banques de données sont développées par de grands organismes dans le but de regrouper et rendre accessibles les informations provenant du séquençage et de l'analyse des nombreux génomes et autres séquences. Trois grands organismes de recherche développent des outils et gèrent des banques de données centrées sur les informations en biologie :

- Le NCBI (National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov/`) des Etats-Unis d'Amérique est une ressource nationale pour la biologie moléculaire fondée en 1988. Il s'implique dans la création de banques publiques et la recherche en bioinformatique. Son but est de développer des outils informatiques permettant l'analyse des données du génome et de diffuser l'information médicale pour mieux comprendre les processus moléculaires touchant la santé humaine.
- L'EBI (European Bioinformatics Institut, `http://www.ebi.ac.uk/`) est une organisation académique à but non lucratif formée en 1992. C'est un centre de recherche et de services en bioinformatique qui gère des bases de données biologiques (ADN-ARN, protéines, struc-

tures 3D). Il a pour but de rendre accessible gratuitement les données issues de la recherche en biologie moléculaire et génomique afin de promouvoir le progrès scientifique.
- Le CIB-DDBJ (Center for Information Biology and DNA DataBank of Japan, `http://www.cib.nig.ac.jp/Welcome.html`) est le centre japonais créé en 1995. Il a pour but de stocker les informations issues de la biologie. Il est constitué de cinq laboratoires.

Tous ces organismes centralisent et rendent accessible l'énorme quantité d'informations obtenue aujourd'hui grâce aux nouvelles technologies de plus en plus efficaces.

Séquences nucléiques

Il existe trois banques de données principales regroupant des séquences nucléiques, développées par les trois organismes cités précédemment :
- GenBank (`http://www.ncbi.nlm.nih.gov/Genbank`) au NCBI
- EMBL bank (European Molecular Biology Laboratory, `http://www.ebi.ac.uk/embl/`) à l'EBI
- DDBJ (DNA Data Bank of Japan, `http://www.ddbj.nig.ac.jp/`) au CIB

Ces banques échangent des données quotidiennement et chacune est chargée de la collecte des données de son continent. Une entrée contient une séquence nucléique, que ce soit un génome complet, un fragment ou encore un ARN ; ainsi que les annotations de cette séquence. De nombreuses séquences sont disponibles dont des gènes codant des synthétases.

UniProt

La banque de données de référence pour les séquences protéiques est Uniprot. En 2002, trois instituts, l'EBI, le SIB (Swiss Institut of Bioinformatics) et la PIR (Protein Information Ressource) ont décidé de créer le consortium UniProt (Universal Protein Ressource) afin de regrouper l'ensemble des séquences protéiques et leurs annotations à un même endroit. UniProt est formée de plusieurs banques de données dont UniProtKB (UniProt knowledgeBase, `http://www.uniprot.org/`) composée de Swiss-Prot et TrEMBL (*translated EMBL*). Swiss-Prot (`http://www.expasy.ch/sprot/`) contient des données corrigées et validées par des experts. Elle propose un haut niveau d'annotation avec une redondance minimale. TrEMBL contient les séquences protéiques qui ne sont pas dans SwissProt. Dans TrEMBL, les séquences sont annotées automatiquement, c'est-à-dire par informatique, à partir de la traduction automatique des séquences nucléiques codantes contenues dans EMBL. TrEMBL comprend également des entrées provenant de soumissions spontanées. UniProtKB contient un grand nombre de séquences protéiques diverses, parmi lesquelles des séquences protéiques de synthétases. Les entrées d'UniProt contiennent bien entendu la séquence protéique, mais également différentes annotations. Dans le cas des synthétases, les différents domaines peuvent être localisés sur la séquence. Ces annotations sont obtenues à partir de l'existence de motifs spécifiques pour les différents domaines.

wwPDB

wwPDB (world wide Protein Data Bank, `http://www.wwpdb.org/`) est la seule banque de données de structures 3D expérimentales de protéines et grosses molécules biologiques. En 2003, les banques de données de structures 3D de protéines RSCB (Research Collaboratory for Stuctural Bioinformatics), MSD (Macromolecular Structure Database) et PDBj (Protein Data Bank of Japan) se regroupent pour former une seule banque de données : wwPDB. Au sein de wwPDB,

beaucoup d'entrées sont disponibles parmi lesquelles la structure 3D de différents domaines de synthétases tels que le domaine d'adénylation de la gramicidine S ou la structure 3D de modules avec par exemple le module terminal de la synthétase de la surfactine. Elle contient également la structure 3D de petites molécules telles que trois actinomycines.

Pfam

Pfam est une grande collection d'alignements multiples de séquences et de modèles de Markov cachés (HMM, *Hidden Markov Model*) couvrant beaucoup de familles et de domaines protéiques. Un HMM est un modèle statistique permettant de représenter des systèmes dont le processus est markovien, c'est-à-dire un processus stochastique possédant la propriété de Markov (la prédiction du futur, sachant le présent, n'est pas rendue plus précise par des éléments d'information supplémentaires concernant le passé). Grâce aux HMM, il est possible de générer des profils. Un profil est une description statistique caractéristique d'une portion de séquence. Par exemple, il existe des profils caractérisant une famille donnée de protéines. Les HMM peuvent être utilisés pour rechercher des domaines protéiques dans des bases de données de séquence à l'aide de HMMER. HMMER est un logiciel permettant de rechercher un profil HMM dans une séquence [Eddy, 1998]. Pfam contient des profils caractéristiques de différents domaines présents dans les NRPS : les domaines A, T, C, ou encore Te.

PubChem

PubChem est une ressource publique sur les activités biologiques de petites molécules. Elle est organisée en trois bases de données dont une, PubChem compound (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound`), qui contient des informations chimiques validées sur des molécules bioactives. PubChem Compound contient des millions de molécules dont certaines sont synthétisées par la voie non-ribosomiale. Cependant, aucune différence n'est faite entre les molécules naturelles et les molécules synthétiques et la voie de synthèse des molécules n'est pas mentionnée. De plus, tous les peptides non-ribosomiaux connus ne sont pas présents dans cette base de données. Enfin, lorsqu'un peptide est présent dans PubChem, ses variants naturels ne le sont pas forcément. Par exemple, les pyoverdines représentent un grand groupe comprenant plus de 60 variants, alors que seulement 3 pyoverdines sont répertoriées dans PubChem.

2.1.2 Banques et bases de données spécialisées

Contrairement aux bases de données dites généralistes citées précédemment, les bases de données spécialisées contiennent des informations centrées sur un domaine très spécifique. Nous présentons ici celles qui sont spécifiques des peptides non-ribosomiaux et de leurs synthétases.

Peptaibol database

Une base de données dédiée aux peptaibols naturels a été développée [Whitmore and Wallace, 2004b], accessible en ligne (`http://www.cryst.bbk.ac.uk/peptaibol/introduction.htm`). Cette base de données regroupe plus de 300 peptaibols. Les peptaibols sont des peptides linéaires antibiotiques synthétisés par voie NRPS. La séquence, l'organisme producteur et la référence bibliographique sont donnés pour chaque peptaibol de la base.

NRPS-PKS

NRPS-PKS est une ressource publique axée sur les NRPS et les PKS [Ansari et al., 2004], accessible en ligne (`http://www.nii.res.in/nrps-pks.html`). Elle contient un outil d'analyse des NRPS/PKS (section 2.2.3) ainsi que 4 bases de données :

- NRPSDB, contenant des NRPS
- PKSDB, contenant des PKS modulaires
- ITERDB, contenant des PKS itératives de type I
- CHSDB, contenant des PKS de type III

La base de données NRPSDB contient 17 clusters NRPS et 5 clusters hybrides NRPS/PKS. Pour chaque entrée, l'organisation des domaines, la séquence protéique des différents domaines et celle des 10 acides aminés impliqués dans la liaison au substrat dans le site de fixation des domaines d'adénylation sont disponibles. La structure du peptide produit par le cluster est également donnée.

ClustScan database

ClustScan DataBase (CSDB, `http://csdb.bioserv.pbf.hr/csdb/ClustScanWeb.html`) est une base de données créée à partir de données obtenues avec le logiciel ClustScan, que nous présenterons dans la section suivante (section 2.2.5). Il permet l'annotation semi-automatique des clusters NRPS et PKS. CSDB contient des données génétiques et biochimiques sur les systèmes NRPS et PKS. CSDB contient les séquences nucléiques et protéiques des gènes, modules et domaines formant les clusters. Elle contient également les différents composés incorporés durant les biosynthèses NRPS et PKS, permettant la prédiction des produits synthétisés. Pour le moment, CSDB contient uniquement cinq clusters PKS.

Nous avons vu dans cette partie qu'il existe de nombreuses bases de données disponibles sur le Web. Certaines d'entres elles contiennent des informations sur les synthétases ou sur les peptides synthétisés par la voie non-ribosomiale. Cependant, ces informations sont noyées dans la masse d'information contenue dans ces bases de données. En effet, aucune base de données exhaustive dédiée à la voie NRPS ou à ses produits n'était disponible au commencement de ce travail. A l'inverse, certains outils bioinformatiques spécifiques d'analyse des synthétases et de prédiction des peptides produits ont été développés par diverses équipes et sont disponibles sur le web.

2.2 Outils d'analyse des synthétases et prédiction du peptide produit

La découverte de motifs conservés propres aux différents domaines ainsi que celle du code conférant la spécificité aux domaines d'adénylation (voir section 1.2.3), le code Stachelhaus, a permis le développement de différents outils bioinformatiques d'analyse et de prédiction des peptides à partir de la séquence protéique des synthétases. Nous donnerons ici les différents outils dédiés aux synthétases.

2.2.1 Quelques définitions

Nous donnerons ici quelques notions et définitions utiles pour la compréhension des paragraphes suivants.

Alignement de séquences

Un alignement entre deux séquences protéiques ou nucléiques consiste à identifier les éléments communs (les acides aminés ou les nucléotides) aux deux séquences dans le but de déterminer les zones similaires. Il existe plusieurs algorithmes qui calculent le meilleur alignement possible entre deux séquences, en fonction d'un jeu de paramètres. La figure 2.1 montre un exemple d'alignement entre deux séquences protéiques.

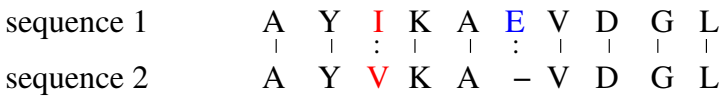


FIG. 2.1 – Exemple d'alignement entre deux séquences protéiques. Un mismatch est représenté par : et un match par |.

Certains outils alignent deux séquences, d'autres en alignent plusieurs. ClustalW [Thompson et al., 2002] est un logiciel d'alignement multiple, c'est-à-dire qu'il aligne plusieurs séquences entre elles dans le but d'identifier la ou les régions communes à toutes ces séquences. Un exemple d'alignement multiple généré par ClustalW est donné dans la figure 2.2.

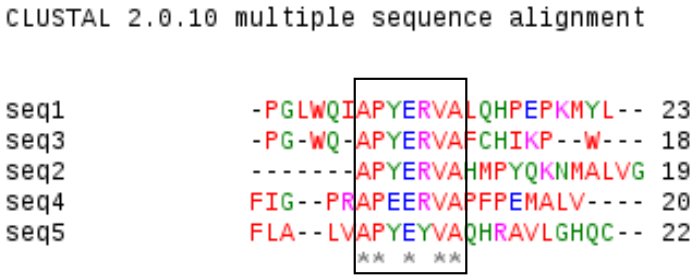


FIG. 2.2 – Exemple d'alignement multiple généré par ClustalW. Les acides aminés présentant des propriétés physico-chimiques similaires apparaissent dans la même couleur. Le symbole * en bas d'une colonne signifie que la colonne est conservée. La région encadrée est une région conservée entre les séquences.

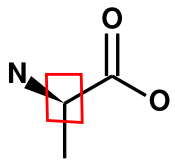
BLAST

BLAST (Basic Local Alignment Search Tool) est une méthode heuristique permettant d'identifier les régions similaires entre deux séquences de nucléotides ou d'acides aminés [Altschul et al., 1990]. Ce programme retrouve rapidement dans les banques de données, des séquences similaires à la séquence entrée par l'utilisateur. Il est utilisé pour trouver des relations fonctionnelles ou évolutives entre les séquences et peut identifier les membres d'une même famille. Depuis sa création en 1990 par Stephen Altschul, Warren Gish et David Lipman au

NCBI, plusieurs versions ont été développées. Parmi elles, BlastN identifie des similarités locales entre séquences nucléiques et BlastP entre séquences protéiques.

SMILES et SMARTS

SMILES (*Simplified Molecular Input Line Entry System*) est une notation linéaire, c'est-à-dire une chaîne de caractères ASCII, qui représente les molécules chimiques et qui a été développée par la société *Daylight Chemical Information Systems*. Les SMILES encodent des structures chimiques comme des graphes, avec des indications éventuelles sur l'isomérie. Ils représentent la structure chimique en 2-D utilisée par les chimistes. Les SMILES génériques décrivent uniquement la structure d'une molécule, c'est-à-dire les atomes et les liaisons, sans information sur l'isomérie. La notation SMILES consiste en une série de caractères sans espace. Les hydrogènes peuvent être omis ou inclus. Les atomes sont représentés par leur symbole atomique. La liaison simple n'est pas représentée, c'est-à-dire que l'on considère comme connecté par une liaison simple les atomes adjacents dans la notation SMILES. La double liaison est représenté par le symbole « = » et la triple liaison par le symbole « : ». Les branchements sont spécifiés par des parenthèses. Par exemple, dans le SMILES « CCN(CC)CC », deux atomes de carbones forment un branchement sur l'atome d'azote (N). Les cycles sont représentés en cassant une liaison dans chaque cycle et en ajoutant un chiffre au premier et au dernier atome formant le cycle. Par exemple, le SMILES « C1CCCCC1 » représente un cycle formé de 6 atomes de carbone. Un exemple de notation SMILES générique est donné dans la figure 2.3 avec l'exemple de la D-alanine. Dans cet exemple, l'isomérie n'est pas prise en compte et par conséquent le SMILES générique encode à la fois la L-alanine et la D-alanine.



SMILES
générique: CC(C(=O)O)N
isomérique: C[C@H](C(=O)O)N

FIG. 2.3 – Exemple de notation SMILES : D-alanine

Les SMILES isomériques prennent en compte l'isomérie. La chiralité d'un carbone est donnée par le symbole « @ » pour le sens anti-horaire et par le symbole « @@ » pour le sens horaire. Un exemple de SMILES isomérique est donné dans la figure 2.3. En règle générale, il existe plusieurs notations SMILES possibles pour une même molécule. Un algorithme de canonisation a été développé dans le but de générer un SMILES générique spécial parmi toutes les possibilités. Ce SMILES est appelé SMILES unique.

SMARTS (*SMiles ARbitrary Target Specification*) est une notation dérivée des SMILES qui définit des motifs et des propriétés chimiques. Ils contiennent un ensemble de symboles permettant d'inclure des incertitudes dans le motif à rechercher, comme par exemple le symbole « * » représentant n'importe quel atome.

2.2.2 PKS/NRPS Analysis Web-site

En 1997, la structure du domaine d'adénylation qui active la phénylalanine (PheA) de la gramicidine S a été obtenue par cristallographie [Conti et al., 1997]. Elle a permis de mettre en évidence 10 acides aminés qui interviennent dans la liaison du substrat. En 1999, à partir des 10 acides aminés intervenant dans la liaison du substrat, le code Stachelhaus a été mis en place, conférant la spécificité aux domaines A [Stachelhaus et al., 1999] (section 1.2.3). En 2000, une autre étude a été réalisée sur la spécificité des domaines A [Challis et al., 2000]. Dans cette étude, plus de 150 domaines A ont été analysés. Parmi les 10 acides aminés du code Stachelhaus, 8 acides aminés ont été retenus comme suffisant pour prédire la spécificité de plus de 80 % de domaines A, avec plus de 30 acides aminés différents activés par ces domaines. La cystéine située en position 331 a été exclue des acides aminés critiques car sa chaîne latérale est en dehors du site de liaison du substrat. De même, la lysine en position 517 n'est pas retenue car elle est conservée au sein de tous les domaines A. 154 domaines A ont été alignés avec le domaine A de la synthétase de la gramicidine incorporant la phénylalanine (PheA) afin de récupérer les 8 acides aminés critiques. Une étude phylogénétique a été menée. Deux arbres ont été construits. Le premier a été construit en utilisant les 180-200 acides aminés situés entre les sites A3 et A6 des domaines A (section 1.2.3). Le second a été construit à partir des 8 acides aminés critiques. L'arbre obtenu avec les séquences de 180-200 acides aminés regroupe les séquences en fonction des espèces. Au contraire, l'arbre phylogénétique construit avec les ensembles des 8 acides aminés critiques regroupe les séquences en fonction de la spécificité du domaine A. Cette étude prouve ainsi que les 8 acides aminés suffisent pour prédire la spécificité des domaines A. A partir de cette étude, un serveur Web a été mis en place afin de prédire la spécificité d'un domaine A inconnu.

La première version de cet outil, développée en 1999 et appelée *the predictive BLAST server*, est accessible en ligne (`http://www.tigr.org/jravel/nrps/blast/index2.html`). Dans cette version, l'utilisateur doit donner les 8 acides aminés critiques. Cet ensemble est comparé à une base de données grâce à BLASTP. Deux bases de données sont disponibles. La première, « assigned database », contient les 8 acides aminés critiques pour 198 domaines A dont le substrat a été validé expérimentalement. La seconde, « Unassigned database », contient les 8 acides aminés critiques pour 88 domaines A dont le substrat n'a pas été validé expérimentalement. L'inconvénient de cet outil est que l'utilisateur doit extraire l'ensemble des 8 acides aminés critiques lui-même, ce qui n'est pas toujours facile.

Une seconde version de cet outil a été développée en 2008. *PKS-NRPS Analysis Web-site* est disponible sur le Web (`http://www.tigr.org/jravel/nrps/`). Dans cette nouvelle version, l'utilisateur donne la séquence protéique d'une synthétase. Les différents domaines sont identifiés, ainsi que la spécificité des domaines A. Des HMM sont utilisés pour identifier les différents domaines présents au sein de la synthétase. Les domaines principaux sont identifiés (C, T , A et Te) mais aussi certains domaines secondaires comme les domaines E. Ensuite, les domaines A sont analysés. Pour ce faire, la sous-séquence comprise entre A3 et A6 est extraite et alignée avec PheA grâce à ClustalW pour obtenir les 8 acides aminés critiques. Cet ensemble de 8 acides aminés est ensuite soumis au *BLAST server* afin de prédire la spécificité du domaine A. Cet outil prédit, à partir de la séquence protéique d'une synthétase, les différents domaines composant la synthétase, ainsi que la spécificité des domaines A présents et n'a pas été publié.

2.2.3 NRPS-PKS

NRPS-PKS est une ressource permettant l'analyse de mégasynthases NRPS et PKS [Ansari et al., 2004]. Cette ressource est accessible en ligne

(`http://www.nii.res.in/nrps-pks.html`). En plus des bases de données vues précédemment, elle contient également une interface offrant la possibilité d'analyser des séquences protéiques de synthétases inconnues. BLAST est utilisé pour rechercher au sein de la base de données, les domaines similaires aux sous-séquences composant la séquence requête. Ainsi, les différents domaines présents au sein de la synthétase sont identifiés. Cet outil est capable d'identifier les domaines C, T, A, Te, Cy, E et M. Une fois un domaine A identifié, les 10 acides aminés du code Stachelhaus sont extraits et comparés à ceux des domaines A présents dans la base de données, dans le but de prédire la spécificité du domaine. Cet outil fut mis en ligne en 2004. Il était le seul permettant la prédiction d'un peptide à partir de la séquence protéique de la synthétase ainsi que l'identification des différents domaines présents au sein de la synthétase.

2.2.4 NRPSpredictor

En 2005, un autre outil de prédiction, basé sur l'apprentissage automatique, a été développé dans le but de prédire la spécificité du substrat des domaines A [Rausch et al., 2005]. Les méthodes d'apprentissage automatique permettent de classer des objets en différentes catégories à partir d'une fonction obtenue par apprentissage sur un échantillon de données. Les SVMs (Support Vector Machines) font partie de la classe des méthodes d'apprentissage supervisé. Le but est de trouver le meilleur classifieur permettant de séparer les données portant des étiquettes différentes. Dans l'étude réalisée ici, les SVMs serviront à classer les domaines A en fonction de leur substrat. Une extension des SVMs existe : les transductive SVMs (TSVMs). Cette méthode permet d'utiliser des données non-étiquetées lors de la phase d'apprentissage. En effet, une étiquette sera attribuée à chaque point en fonction de celle portée par les points les plus proches. Cette méthode est très utile ici car beaucoup de domaines A dont la spécificité est inconnue sont disponibles dans les bases de données et pourront être utilisés lors de la construction des modèles.

Dans cette étude, les domaines A sont caractérisés par 34 acides aminés. Les 10 acides aminés du code Stachelhaus sont situés dans un rayon de $5,5\mathring{A}$ autour du substrat. Or les acides aminés situés dans un rayon de $8\mathring{A}$ autour du site de fixation du substrat peuvent intervenir dans la liaison de ce dernier ou influencer cette liaison. 34 acides aminés sont situés dans un rayon de $8\mathring{A}$ autour du site de fixation du substrat et ce sont ces 34 acides aminés qui sont utilisés dans cette étude. Pour chaque acide aminé, 12 valeurs représentant ses propriétés physico-chimiques sont stockées. Finalement, un domaine A est caractérisé par un vecteur contenant 408 valeurs (34 acides aminés * 12 valeurs).

Les domaines A peuvent activer un ensemble de monomères présentant des propriétés physico-chimiques similaires. Pour cette raison, les domaines A ont été regroupés en fonction de la similarité de leurs substrats. Deux niveaux de clustering sont ainsi utilisés. Les *large clusters* regroupent les domaines A dont les substrats sont similaires. Par exemple, les domaines A qui activent des acides aminés dont la chaîne latérale est non-polaire (Gly, Ala, Val, Leu, Ile, Abu, Iva) sont regroupés au sein d'un même cluster. Les *small clusters* regroupent des domaines A dont les substrats sont très proches. Par exemple, les domaines A activant Gly et Ala forment un cluster car les acides aminés activés sont petits et non-polaires.

Pour chaque cluster de domaines A, un modèle est obtenu avec les TSVMs grâce à un échantillon de domaines A présentant les spécificités désirées. Lorsqu'un domaine A de spécificité inconnue est disponible, il est possible de prédire sa spécificité en le comparant aux différents modèles caractéristiques des différents clusters. L'implémentation de cette méthode a été réalisée. L'outil résultant, *NRPSpredictor*, est disponible sur le

Web (`http://www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor`). L'utilisateur donne une ou plusieurs séquences protéiques de NRPS. Les domaines A sont identifiés et chaque domaine A est classé dans un des clusters, prédisant ainsi sa spécificité. Pour chaque domaine A, trois résultats sont donnés : le *small cluster*, le *large cluster*, ainsi que la prédiction obtenue à partir des 10 acides aminés du code Stachelhaus. La méthode par TSVM prédit la spécificité de 18 % de domaines A en plus de la méthode utilisant le code Stachelhaus. Cependant la méthode Stachelhaus rend un seul substrat par domaine alors que la méthode par TSVM donne un ensemble de substrats possibles afin d'étendre le nombre d'acides aminés prédits. L'idée est de combiner la méthode par TSVM avec la méthode Stachelhaus. Cet outil prédit la spécificité d'un domaine A inconnu, grâce à l'apprentissage automatique. Cependant, il ne précise pas les différents domaines présents dans la synthétase.

2.2.5 ClustScan

ClustScan est un ensemble de programmes permettant l'annotation semi-automatique des séquences nucléiques codant des NRPS, des PKS et des hybrides NRPS/PKS, publié en 2008 [Starcevic et al., 2008]. *ClustScan* propose, une fois les gènes identifiés, la structure chimique potentielle de la molécule produite par un cluster de gènes. Pour commencer, GeneMark et Glimmer sont utilisés pour prédire les gènes présents sur la séquence nucléique données en entrée. Ensuite, HMMER identifie les domaines présents au sein des séquences codantes grâce à des profils issus de PFAM ou fournis par l'utilisateur. D'autres profils spécifiques sont utilisés pour extraire les acides aminés afin de prédire le substrat. Ces acides aminés critiques sont ensuite comparés à ceux connus pour établir la spécificité des différents domaines. La prédiction de la structure chimique est réalisée par l'assemblage des différentes unités prédites données sous forme de SMILES ou de SMARTS si la spécificité n'a pas pu être identifiée. La structure chimique prédite peut être exportée, sous forme de SMILES ou de SMARTS, pour une analyse par d'autres programmes. Le programme comprend une interface graphique conviviale permettant à l'utilisateur de visualiser les différents clusters prédits. *ClustScan* peut être utilisé en s'inscrivant via le site Web : `http://bioserv.pbf.hr/cms/index.php?page=clustscan`.

ClustScan a été réalisé de façon à pouvoir introduire facilement des connaissances supplémentaires, comme par exemple de nouveaux profils. De même, il autorise l'insertion d'autres programmes non contenus dans le logiciel initial. Il permet de réaliser l'annotation de grandes séquences nucléiques. Par exemple, l'identification et l'annotation de tous les gènes NRPS/PKS d'un génome d'actinobactérie nécessite deux à trois heures de travail. Cependant, *ClustScan* est, pour le moment, plus adapté à l'annotation des PKS qu'à celle des NRPS.

2.2.6 Clusean

Clusean (CLUster SEquence ANalyzer) est un outil informatique pour l'analyse automatique des clusters de gènes bactériens codant des métabolites secondaires, publié en 2009 [Weber et al., 2009]. C'est un pipeline de programmes « open source » permettant l'analyse de clusters de gènes codant des NRPS ou des PKS de type I. Il intègre des outils standards d'analyse tels que BLAST et HMMER, ainsi que des outils spécifiques d'analyse des NRPS tel que *NRPSpredictor*. BLAST est utilisé pour annoter les gènes contenus dans la séquence entrée par l'utilisateur. HMMER, utilisé avec différents profils, permet l'identification des domaines protéiques ainsi que les motifs conservés de ces domaines. La spécificité des différents domaines d'adénylation est prédite à l'aide de *NRPSpredictor* et du code Stachelhaus.

Le format d'entrée est le format EMBL. Le format de sortie est également le format EMBL et un format sous forme de tables est également disponible. Le format EMBL permet l'utilisation d'autres outils d'analyses. Les différents programmes composant *Clusean* peuvent être utilisés de façon indépendante et tous les résultats intermédiaires peuvent être stockés. *Clusean* permet ainsi d'annoter de manière automatique un cluster de gènes bactériens et d'obtenir la structure des domaines, les motifs conservés des domaines, ainsi que la prédiction de la spécificité des domaines A. *Clusean* peut être téléchargé gratuitement à partir du site Web. Cependant, l'installation de *Clusean* nécessite l'installation d'un grand nombre de logiciels et reste délicate pour la plupart des utilisateurs.

Nous avons vu dans cette section qu'il existe des bases de données contenant des informations sur les synthétases comme leur séquence nucléique, protéique ou encore leur structure 3D. Il existe également certaines bases de données comprenant des peptides produits par la voie non-ribosomiale. Cependant, ces bases ne répertorient que quelques peptides non-ribosomiaux (PubChem) ou sont centrées sur une classe précise, comme par exemple la base de données sur les peptaibols. Certains outils bioinformatiques permettant l'analyse des synthétases et la prédiction du produit ont été développés. Cependant, tous ces outils sont axés sur les synthétases et aucun n'est réellement dédié aux peptides. Au commencement de notre travail, il n'existait aucune ressource publique regroupant les informations sur les peptides non-ribosomiaux, ni aucun outil bioinformatique dédié à leur analyse.

Chapitre 3

Modélisation et comparaison des peptides non-ribosomiaux

Après avoir présenté les peptides non-ribosomiaux et les outils bioinformatiques existants, nous allons maintenant introduire notre contribution. Comme nous l'avons vu dans le chapitre précédent, au commencement du travail, aucun outil dédié aux peptides non-ribosomiaux n'était disponible. Cependant, les scientifiques du domaine éprouvaient le besoin de centraliser les données sur les peptides non-ribosomiaux, ainsi que de comparer ces peptides entre eux. Nous avons donc développé NORINE, la première ressource publique dédiée aux peptides non-ribosomiaux. NORINE contient une base de données regroupant diverses informations sur un grand nombre de peptides non-ribosomiaux. NORINE sera présentée dans le chapitre suivant. Dans ce chapitre, nous introduisons les méthodes informatiques que nous avons développées pour la comparaison et l'analyse des peptides issus de la voie non-ribosomiale.

Les peptides non-ribosomiaux présentent de nombreuses particularités par rapport aux peptides classiques. La diversité des acides aminés et autres composés incorporés (monomères), ainsi que les structures primaires non-linéaires, empêchent l'utilisation de la modélisation et des outils développés pour les peptides classiques. Dans ce chapitre, nous présentons les modélisations que nous avons développées pour les peptides non-ribosomiaux. Nous introduisons ensuite une méthode adaptée et efficace permettant la recherche de motifs structuraux au sein des peptides non-ribosomiaux. Enfin, nous présentons l'extension de cette méthode à la comparaison des peptides synthétisés par la voie non-ribosomiale.

3.1 Modélisation de la structure des peptides non-ribosomiaux

Dans le but de pouvoir comparer et analyser les peptides non-ribosomiaux, nous devons dans un premier temps modéliser leur structure primaire. Les peptides non-ribosomiaux sont des molécules chimiques. Il est donc possible de les représenter comme telles, c'est-à-dire par la description des atomes et des différentes liaisons entre ces atomes. Cependant, les peptides non-ribosomiaux sont synthétisés par l'assemblage de monomères et non d'atomes. En effet, durant leur biosynthèse, les domaines des synthétases sélectionnent et incorporent des monomères complets qui peuvent éventuellement être modifiés. Dans le but de modéliser leur biosynthèse, nous avons décidé de représenter les peptides non-ribosomiaux comme un assemblage de monomères plutôt qu'à un niveau atomique. De plus, cette modélisation s'apparente à celle utilisée pour les peptides classiques.

La structure primaire des peptides classiques est représentée par une chaîne de caractères orientée de l'extrémité dite N-terminale vers l'extrémité dite C-terminale. Ces chaînes de caractères utilisent un alphabet de 20 lettres représentant les 20 acides aminés protéogéniques. Or, dans le cas des peptides non-ribosomiaux, plusieurs centaines de monomères différents peuvent être incorporés, ce qui rend impossible l'utilisation du code à une lettre défini pour les peptides et protéines classiques. De plus, les structures non-linéaires des peptides non-ribosomiaux ne peuvent pas être représentées par des chaînes de caractères. Nous avons donc défini une modélisation capable d'intégrer les particularités de ces peptides originaux. Dans un premier temps, une représentation linéaire a été développée. Cependant, elle s'est avérée insuffisante pour représenter les structures peptidiques les plus complexes. Nous avons alors opté pour une modélisation à l'aide de graphes.

3.1.1 Codage des monomères

Les acides aminés

Les acides aminés protéogéniques sont généralement représentés par un code à trois lettres ou par un code à une lettre (section 1.1.1). Le grand nombre de monomères pouvant être incorporés rend impossible l'utilisation du code à une lettre et non intuitif le code à trois lettres pour les acides aminés issus du mécanisme non-ribosomial. Nous avons mis au point une nomenclature spéciale pour ces monomères, inspirée de la nomenclature IUPAC pour les acides aminés et les peptides [authors listed, 1984]. IUPAC (*International Union of Pure and Applied Chemistry*) est un organisme dont le but est de mettre en place des recommandations pour les composés chimiques.

Dans notre nomenclature, les acides aminés protéogéniques sont représentés par leur code à trois lettres (par exemple, « Ala » pour l'alanine). Certains acides aminés non-protéogéniques sont également représentés par un code à trois lettres. Par exemple, « Hpg » est l'abréviation de HydroxyPhenylGlycine ou encore « Kyn » celle de la kynurénine. Par défaut, les monomères sont les isomères L, forme rencontrée dans les protéines classiques. Lorsque l'acide aminé est l'isomère D, la lettre D, séparée par un tiret, est ajoutée devant l'acide aminé. Par exemple, la D-alanine est notée « D-Ala ». Lorsqu'un groupe fonctionnel est ajouté, comme par exemple un groupement méthyl ou un groupement hydroxyl, le symbole correspondant est ajouté devant le code de l'acide aminé, séparé par un tiret. Les principaux groupes fonctionnels rencontrés sont le groupement méthyl (Me), le groupement formyl (Fo) et le groupement hydroxyl (OH). Leur position au niveau de l'acide aminé est également notée. Par exemple, la N-méthylalanine est codée par « NMe-Ala ». Lorsque plusieurs modifications apparaissent, elles sont séparées par un tiret. Un ordre, arbitraire, a été défini lorsque plusieurs modifications sont présentes afin de ne pas représenter un même monomère avec deux nomenclatures différentes. Par exemple, la N-méthyl-D-alanine est notée « D-NMe-Ala ».

Les acides gras

Des acides gras peuvent également être incorporés dans les peptides non-ribosomiaux formant ainsi des lipopeptides. Les acides gras ont été introduits plus récemment dans nos modélisations. En effet, au commencement du travail, nous représentions un acide gras par le symbole « R-CO », sans distinction des différents acides gras rencontrés au sein des peptides. Cependant, un changement d'acide gras dans un peptide peut modifier son activité biologique. Nous avons introduit la nomenclature spécifique aux acides gras dans le but de différencier certains variants.

Les acides gras sont représentés par une nomenclature couramment utilisée à savoir la lettre C suivie du nombre de carbones de l'acide gras, puis de « : », puis du nombre d'insaturations (liaisons doubles) de la chaîne et pour finir des positions des insaturations entre parenthèses. Par exemple, l'acide hexadécen-9-oïque est noté « C16 :1(9) », c'est-à-dire qu'il comprend 16 atomes de carbone et une double liaison entre le neuvième et le dixième atome de carbone. Une double liaison peut être *cis*, lorsque les deux hydrogènes sont du même côté ou *trans*, lorsque les deux hydrogènes sont opposés. Si la double liaison est *trans*, la lettre « t » est ajoutée devant la position correspondante. Par exemple, « C14 :2(t4.6) » (figure 3.1) représente l'acide gras composé de 14 carbones avec une double liaison *trans* en position 4 et une double liaison *cis* en position 6.

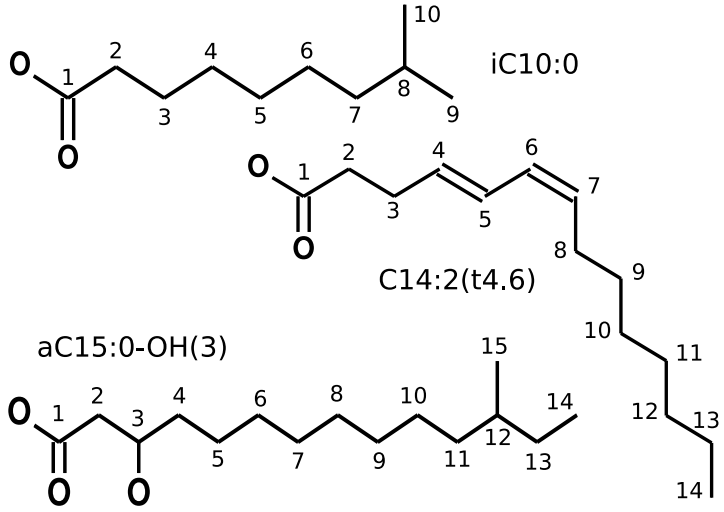


FIG. 3.1 – Quelques exemples de codage d'acides gras

Il existe également des acides gras ramifiés. Lorsque la ramification est portée par le carbone en position $n-1$, c'est-à-dire l'avant dernier carbone de la chaîne, c'est un acide gras *iso*, représenté par l'ajout de la lettre « i » devant le codage de l'acide gras. Lorsque le méthyle est porté par le carbone en position $n-2$ c'est un acide gras *anteiso*, représenté cette fois-ci par la lettre « a ». Par exemple, l'acide isodécanoïque (ou acide 8-méthylnonanoïque) est noté « iC10 :0 » (figure 3.1). Lorsque des groupements supplémentaires sont ajoutés à l'acide gras, le symbole de ces groupements ainsi que leur position sont ajoutés en séparant les différentes modifications par un tiret. Par exemple, l'acide 3-hydroxy-12-méthyl-tétradécanoïque est noté « aC15 :0-OH(3) » (figure 3.1).

Les sucres

Les sucres sont représentés par la nomenclature traditionnelle à trois lettres. Par exemple, le glucose est noté « Glc ». Lorsque l'unité de base est modifiée, les modifications sont ajoutées avant le code correspondant au sucre. Par exemple, la 4-oxo-vancosamine est codée par « 4oxo-Van ».

Autres

D'autres monomères sont rencontrés au sein des peptides non-ribosomiaux tels que des composés issus de la voie PKS. Quelques composés proviennent de voies de synthèses différentes de celles évoquées précédemment ou de voies non identifiées.

Les dérivés d'acide sont représentés avec le code correspondant à leur nom trivial. Par exemple, l'acide valérique est noté « Vaa ». Les monomères correspondants à une chaîne carbonée modifiée sont codés en utilisant une nomenclature dérivée de celle des acides gras. La longueur de la chaîne carbonée est donnée par le C suivi du nombre de carbones de la chaîne. Le nombre d'insaturations est donné après les deux points. Les positions des insaturations sont données entre parenthèses. Les modifications apportées à la chaîne carbonée sont données à la suite, avec la position entre parenthèses, et sont séparées par des tirets. Par exemple, « C6 :0-OH(3.5)-NH2(4) » représente l'acide 3,5-dihydroxy-4-amino-hexanoïque, c'est-à-dire une chaîne carbonée contenant six carbones sans insaturation, présentant deux groupements hydroxyls en position 3 et 5 ainsi qu'un groupement amine en position 4.

3.1.2 Représentation linéaire

Comme nous l'avons vu précédemment, contrairement aux peptides classiques, les peptides non-ribosomiaux présentent des structures primaires complexes. En effet, des structures totalement ou partiellement cycliques sont rencontrées, mais également des structures branchées et même des structures plus complexes. Ces structures primaires particulières ne nous permettent pas de représenter simplement l'enchaînement des monomères, comme dans le cas des peptides classiques. Nous avons conçu une représentation linéaire afin de pouvoir représenter facilement les structures primaires des peptides non-ribosomiaux. Cette représentation linéaire se veut intuitive et compréhensible. Nous nous sommes inspirés des conventions d'écriture utilisées par les SMILES (section 2.2.1).

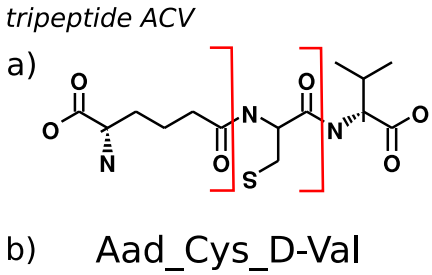


FIG. 3.2 – Représentation d'une structure linéaire : exemple du tripeptide ACV. a) Formule développée du tripeptide ACV et b) représentation linéaire correspondante.

Les structures primaires linéaires sont représentées par l'enchaînement des monomères, séparés par un tiret bas (cf. figure 3.2).

Les branchements sont représentés à l'aide d'accolades. Les monomères situés à l'intérieur des accolades font partie du branchement. Le monomère précédant les accolades est le monomère sur lequel est situé le branchement (cf. figure 3.3).

Les cycles sont représentées à l'aide de crochets. Les monomères à l'intérieur des crochets forment un cycle. Le premier et le dernier monomère situés entre les crochets sont reliés entre eux. Dans le cas de structures partiellement cycliques ou bi-cycliques, les monomères du cycle peuvent également être reliés au monomère se situant juste avant ou juste après le cycle (cf. figure 3.4).

La représentation linéaire que nous avons développée est compréhensible et lisible. Nous l'avons définie au début du travail, à un moment où nous ne pensions pas rencontrer des structures trop "complexes" pour être représentées de façon linéaire. En effet, cette représentation

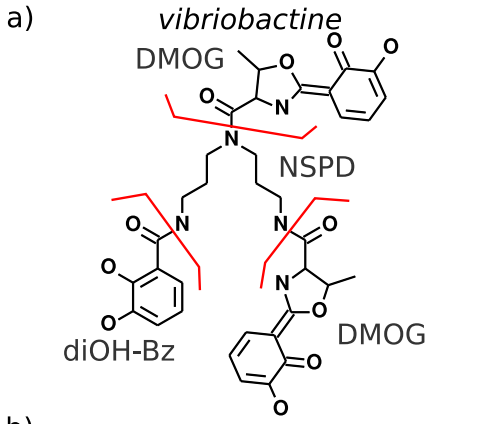


FIG. 3.3 – Représentation d'une structure branchée : exemple de la vibriobactine. a) Formule développée de la vibriobactine et b) représentation linéaire correspondante.

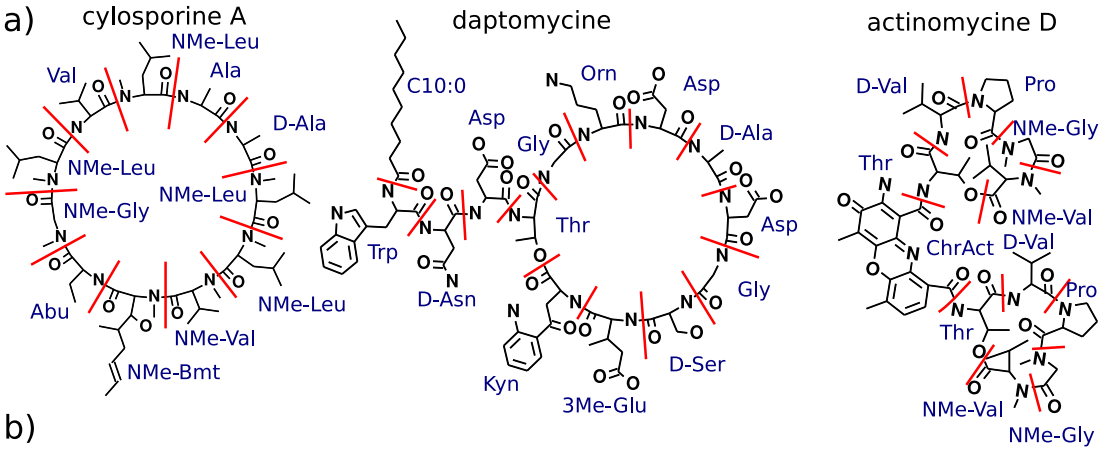


FIG. 3.4 – Représentations de structures contenant des cycles. a) Formules développées et b) représentations linéaires correspondantes.

permet l'encodage des structures linéaires, branchées ou contenant des cycles non-chevauchants. Cependant, il est impossible de représenter des structures contenant des organisations plus complexes comme, par exemple, des cycles chevauchants. De plus, la représentation linéaire n'est pas toujours unique, c'est-à-dire qu'il peut exister plusieurs représentations linéaires possibles pour représenter un même peptide. Par exemple dans le cas d'une structure cyclique, la représentation linéaire peut commencer par n'importe quel monomère du cycle, ce qui amène à

plusieurs représentations linéaires valides pour une même structure cyclique. Afin de représenter de façon unique et non ambigüe toutes les structures des peptides non-ribosomiaux, nous avons développé une autre modélisation de ces structures basée sur les graphes.

3.1.3 Modélisation par les graphes

Un graphe est un ensemble de points, appelés nœuds, pouvant être reliés par des liens appelés arêtes, dans le cas où les liens n'ont pas d'orientation, ou arcs, dans le cas où les liens sont orientés. Dans le but de pouvoir représenter l'ensemble des structures primaires rencontrées au sein des peptides non-ribosomiaux, nous avons décidé de modéliser les structures par des graphes étiquetés non-orientés, c'est-à-dire des graphes dont les nœuds portent un nom (une étiquette) et dont les liens n'ont pas de direction (orientation). Les nœuds représentent les monomères, les arêtes les liaisons entre ces monomères et les étiquettes sont les noms des monomères (cf. figure 3.5). Nous avons choisi les graphes non-orientés, malgré le fait que les liaisons peptidiques sont orientées (N-terminale vers C-terminale), pour plusieurs raisons. Premièrement, au sein des peptides non-ribosomiaux, il existe des liaisons non-peptidiques pour lesquelles l'orientation n'est pas définie. Deuxièmement, les graphes non-orientés sont plus généraux, permettant ainsi d'être moins stringent lors d'une comparaison. Par exemple, dans la mycosubtiline, la sérine est en position 6 et l'asparagine en position 7. L'ordre de ces monomères est inversé dans l'iturine, c'est-à-dire que l'asparagine est en position 6 et la sérine en position 7. Nous voulons être capables de détecter ce type d'inversion, ce qui est possible uniquement en utilisant des graphes non-orientés.

Formellement, la structure d'un peptide est représentée par un graphe $G(V, E, M, f)$ où V est l'ensemble des nœuds, $E \subseteq V \times V$ est l'ensemble des arêtes (liens non-orientés), c'est-à-dire des paires (u, v) où u, v sont dans V, $f : V \to M$ est une fonction qui associe une étiquette à un nœud et M est l'ensemble des étiquettes, c'est-à-dire le nom des monomères. Chaque nœud est identifié par un nombre unique afin de pouvoir différencier des nœuds portant la même étiquette.

La figure 3.5 montre quelques exemples de structures primaires de peptides non-ribosomiaux modélisées par des graphes étiquetés non-orientés.

Les graphes permettent la modélisation de toutes les structures primaires des peptides non-ribosomiaux. La structure de la vancomycine est très complexe car elle contient des cycles chevauchants et des branchements. Cette structure ne peut pas être modélisée avec la représentation linéaire introduite précédemment, mais il est possible de le faire à l'aide d'un graphe étiqueté non-orienté. De plus, cette représentation permet de symboliser le cas particulier des hétérocycles formés entre deux monomères en traçant deux arêtes entre les monomères concernés. Par exemple, la pyoverdine R de la figure 3.5 possède un hétérocycle entre les monomères 1 (D-Ser) et 2 (Dab).

Les graphes sont encodés sous la forme de chaînes de caractères afin de faciliter leur stockage. Cet encodage reprend la définition des listes d'adjacence des graphes. Un graphe peut être défini soit par une matrice d'adjacence soit par une liste d'adjacence. Une matrice d'adjacence consiste en une matrice carrée $n \times n$, avec n le nombre de nœuds du graphe. En position i, j de la matrice un 0 est écrit s'il n'y a pas d'arête entre les nœuds i et j, un 1 lorqu'une arête existe entre ces nœuds ou un 2 dans le cas particulier d'un hétérocycle entre les nœuds i et j. Les listes d'adjacences consistent à définir une liste par nœud. Pour le nœud i, la liste contient tous les nœuds auxquels le nœud i est relié (cf. figure 3.6).

Pour des raisons de programmation, nous utilisons les listes d'adjacence pour encoder les graphes sous forme de chaînes de caractères. Les différentes listes sont séparées par le symbole « @ ». La première liste reprend le nom des différents monomères contenus dans le peptide.

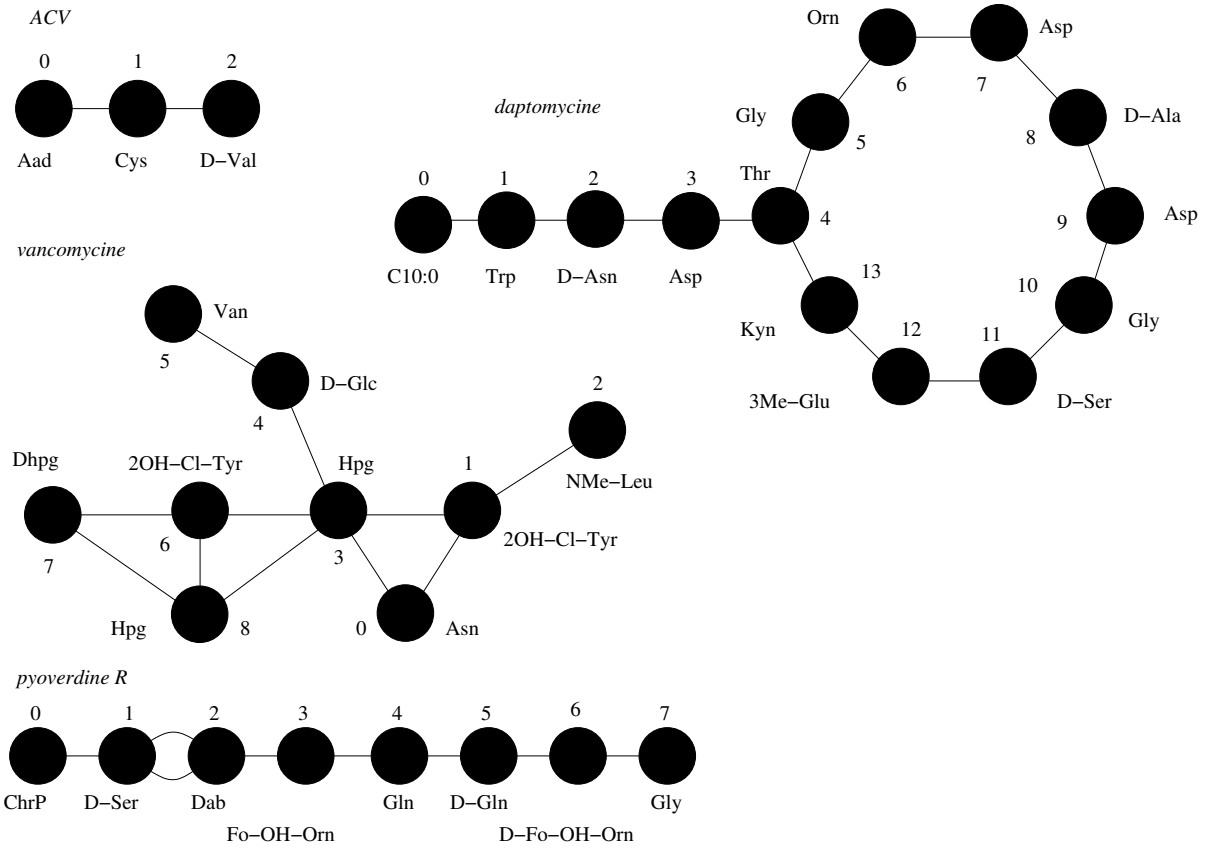


FIG. 3.5 – Quelques exemples de structures modélisées par des graphes

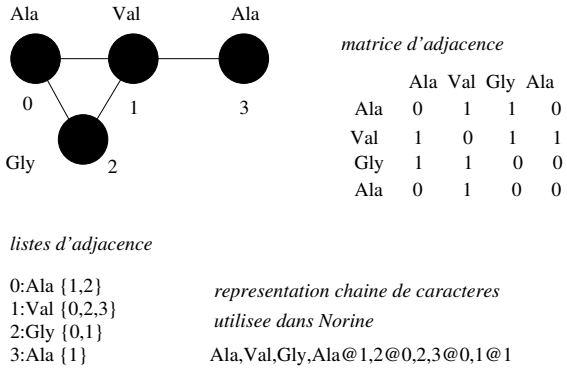


FIG. 3.6 – Représentations des graphes

Ensuite, pour chaque monomère, la liste des monomères auxquels il est relié est donnée. Un exemple est montré dans la figure 3.6. Cette représentation sous forme de chaînes de caractères facilite le stockage informatique de la structure des peptides.

3.2 Recherche d'un peptide selon sa composition en monomères

Il peut être utile de rechercher les peptides contenant certains monomères, sans avoir de contrainte sur la position de ces monomères au sein du peptide. Comme nous l'avons vu dans le chapitre 1, il existe différentes biosynthèses. Par exemple, lors de la biosynthèse non-linéaire, l'agencement des modules de la synthétase est différent de celui des monomères dans le peptide produit. Dans ce cas, lors de la prédiction des monomères incorporés à partir de la séquence protéique de la synthétase, l'information sur la position des monomères est inconnue. Il est donc nécessaire de rechercher les peptides contenant la liste des monomères prédits, sans information sur la structure. De plus, la prédiction n'étant pas certaine, il est intéressant de pouvoir autoriser un nombre maximum d'erreurs entre la composition monomérique recherchée et celle des peptides testés.

La recherche d'une composition monomérique au sein des peptides présents dans la base de données est basée sur l'intersection de deux multi-ensembles. La liste de monomères est transformée en un ensemble, c'est-à-dire une liste d'objets uniques (un élément d'un ensemble n'est pas répété dans cet ensemble). Or dans le cas des peptides non-ribosomiaux, un même monomère peut être présent plusieurs fois au sein d'un peptide. L'idée est donc de transformer la liste des monomères présents dans un peptide en un ensemble, en numérotant les monomères répétés. La liste des monomères recherchés par l'utilisateur est transformée en un ensemble $E1$ et la liste des monomères d'un peptide en un ensemble $E2$. Ensuite, l'intersection entre les deux ensembles $E1$ et $E2$ forme l'ensemble $E3$ qui contient tous les monomères communs aux deux ensembles. Le nombre d'erreurs entre les deux compositions est donc le nombre d'éléments contenus dans l'ensemble de départ $E1$ moins le nombre d'éléments contenus dans l'ensemble intersection $E3$. Les peptides dont la composition correspond à celle recherchée avec le nombre d'erreurs fixé par l'utilisateur sont ainsi retournés. Des exemples d'utilisation de cette fonction seront donnés dans le chapitre suivant.

3.3 Recherche de motifs structuraux

Comme pour les séquences protéiques ou nucléiques, il peut être intéressant de rechercher des motifs structuraux au sein des peptides non-ribosomiaux. Il existe également une forte relation entre la structure et la fonction d'un peptide non-ribosomial. Par exemple, Minowa *et al.* [Minowa et al., 2007] ont identifié des motifs associés, de façon significative, à des activités biologiques. De plus, différents logiciels prédisent un peptide ou une partie d'un peptide à partir de la séquence protéique d'une synthétase (section 2.2). Rechercher le motif obtenu à partir de la prédiction peut être une étape nécessaire dans la recherche de nouvelles molécules ou l'étude de gènes impliqués dans la synthèse d'un peptide donné.

Dans certaines analyses, il est nécessaire de pouvoir identifier une partie d'un motif, plutôt que le motif complet, au sein d'un peptide donné. En effet, dans certains cas, l'ordre des monomères peut être différent de celui des modules de la synthétases (section 1.2.4). Par exemple, lors de la biosynthèse de la syringomycine [Guenzi et al., 1998], le gène *SyrB1*, responsable de l'incorporation de la thréonine est situé en amont du gène *SyrE* dans le génome alors que la thréonine est le dernier monomère du peptide. Dans ce cas, la recherche du motif complet prédit à partir de la synthétase ne permet pas l'identification de la syringomycine. Par contre, la recherche d'une sous-structure commune aboutit à l'identification de la syringomycine.

C'est pourquoi nous avons développé une méthode efficace pour identifier la sous-structure d'un motif donné au sein d'un ensemble de structures de peptides non-ribosomiaux. Cette

méthode a été publiée dans *BMC Structural Biology* [Caboche et al., 2009].

3.3.1 Modélisation des motifs structuraux

Un motif est modélisé, comme la structure des peptides, par un graphe étiqueté non-orienté $P = (V_P, E_P, L, f)$ où V_P est un ensemble de nœuds, $E_P \subseteq V_P \times V_P$ un ensemble d'arêtes (non-orientées) et $f : V_P \rightarrow L$ est une fonction qui associe une étiquette à un nœud. La différence entre un graphe modélisant un peptide et un graphe modélisant un motif est l'ensemble des étiquettes possibles. En effet, $M \subset L$ mais L contient des étiquettes supplémentaires :

- « X », le symbole joker, représente n'importe quel monomère (symbole utilisé pour les protéines classiques)
- une liste de monomères possibles, séparés par le symbole « / », à une position donnée.
- le symbole « * » suivi par le nom d'un monomère symbolise le monomère ou un de ses dérivés

L'introduction de ces symboles permet d'obtenir des motifs contenant des incertitudes ou une alternative entre plusieurs monomères à certaines positions.

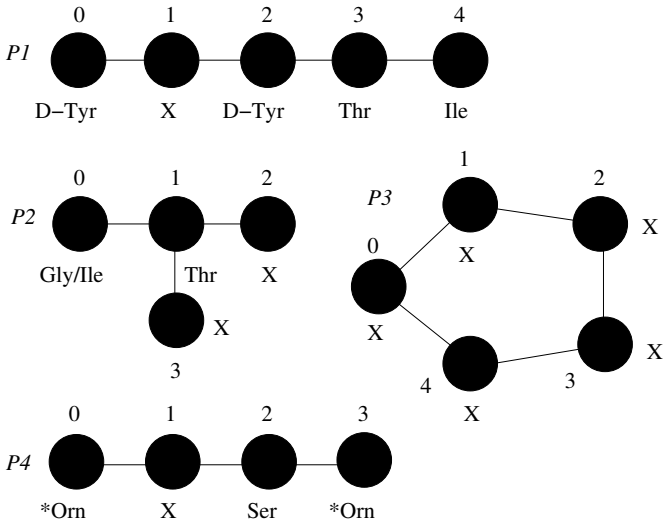


FIG. 3.7 – Quelques exemples de motifs structuraux

Par exemple, le motif $P2$ de la figure 3.7 contient soit une glycine soit une isoleucine en position 0. Le motif $P3$ correspond à un cycle composé de 5 monomères quelconques. Le graphe $P4$ contient l'ornithine ou un dérivé de l'ornithine, comme par exemple OH-Orn (hydroxyornithine), en position 0 et 3.

Nous voulons pouvoir identifier un motif linéaire dans un peptide qui ne l'est pas forcément. La figure 3.8 montre un exemple de motif que l'on souhaite pouvoir identifier dans diverses structures peptidiques.

3.3.2 Méthode classique

La recherche d'un motif au sein d'un peptide correspond à la recherche d'une sous-structure, de taille fixée, commune à deux graphes. Ce problème est une variante du problème de recherche du sous-graphe commun maximum (SCM). Le problème du SCM est un problème NP-complet [Garey and Johnson, 1979], c'est-à-dire un problème qui ne peut, a priori, pas être résolu par

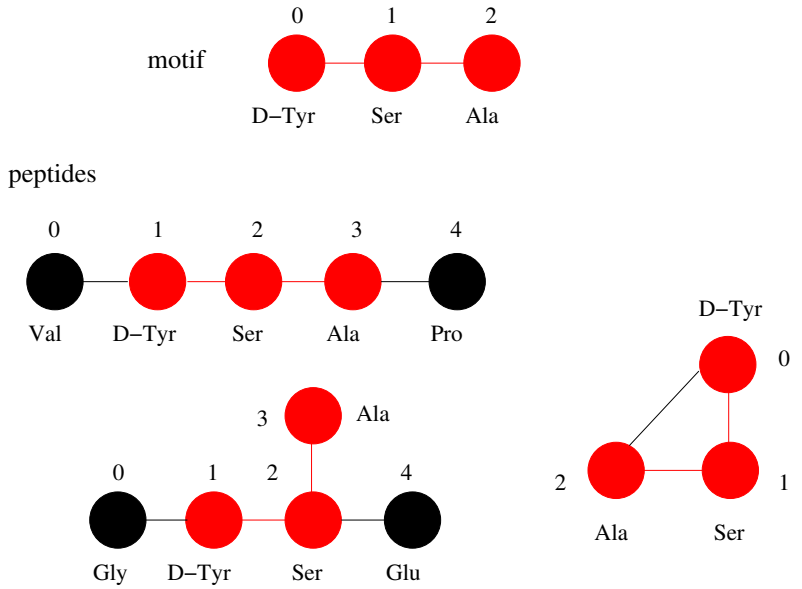


FIG. 3.8 – Exemple d'un motif que l'on veut identifier dans diverses structures peptidiques

un algorithme en temps polynomial en la taille des graphes. Un autre problème similaire à la recherche de sous-structure est le problème de la recherche d'un sous-graphe connexe avec un (multi-)ensemble d'étiquettes qui est également NP-complet [Fellows et al., 2007].

En chémoinformatique, un graphe de compatibilité (GC), aussi appelé graphe d'association ou graphe de produit, est souvent utilisé pour établir la correspondance entre des structures chimiques représentées par des graphes [Raymond and Willett, 2002]. Ce graphe comprend toutes les correspondances potentielles entre les deux graphes, c'est-à-dire toutes les sous-structures communes entre les deux graphes. La recherche de la clique maximale au sein de ce graphe permet d'obtenir le SCM. Une clique dans un graphe non-orienté est un sous-ensemble de nœuds dans lequel chaque paire de nœuds est connectée par une arête. La clique maximale est celle qui contient le plus de nœuds.

Construction du graphe de compatibilité

La définition classique du GC entre deux graphes P et G est la suivante :

- l'ensemble des nœuds du GC est le produit cartésien $V_P \times V$, c'est-à-dire un nœud $U(u, u')$ du GC correspond à l'association du nœud u de P et du nœud u' de G. Dans le cas de nœuds étiquetés, seuls les nœuds avec la même étiquette peuvent être associés et former un nœud dans le GC.
- les nœuds $U(u, u')$ et $V(v, v')$ sont adjacents (reliés par une arête) dans le GC si et seulement si $u \neq v$ et $u' \neq v'$ et si l'une des conditions suivantes est observée :
  - u est adjacent à v dans P et u' est adjacent à v' dans G (1)
  - u n'est pas adjacent à v dans P et u' n'est pas adjacent à v' dans G (2)

Dans notre cas, nous devons modifier les règles de construction du GC car deux nœuds peuvent être associés s'ils ont des étiquettes compatibles mais pas forcément identiques. Si $f(u) \in M$, c'est-à-dire si l'étiquette représente un seul monomère, alors tous les nœuds u' avec $f(u') = f(u)$ sont associés au nœud u. En d'autres termes, un nœud u de P portant une étiquette avec un monomère donné est associé avec tous les nœuds de G portant la même étiquette. Si $f(u) =$'R/S',

tous les nœuds de G portant l'étiquette 'R' ou 'S' sont associés au nœud u de P. Si $f(u)$ ='X', tous les nœuds de G sont associés au nœud u de P. Enfin, si $f(u)$ est une étiquette '*R' alors tous les nœuds de G présentant le monomère 'R' ou un de ses dérivés sont associés avec le nœud u de P.

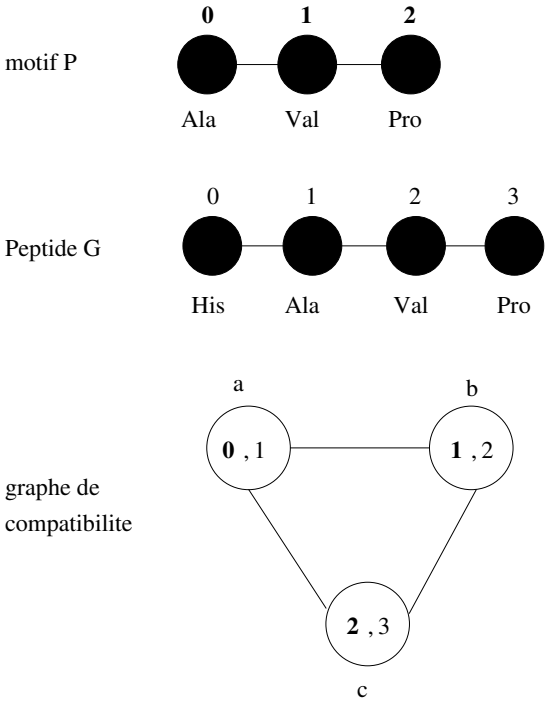


FIG. 3.9 – Exemple de construction d'un graphe de compatibilité entre les graphes P, représentant le motif, et G, représentant le peptide

La figure 3.9 montre un exemple simple de construction d'un GC. Un nœud est identifié par une lettre. Chaque nœud représente l'association entre un nœud de P et un nœud de G. Par exemple, le nœud a représente l'association entre le nœud 0 de P et le nœud 1 de G qui portent tous deux l'étiquette « Ala ». Les arêtes situées entre les nœuds a et b et entre les nœuds b et c sont obtenues à partir de la condition (1). En effet, les nœuds 0 et 1 sont adjacents dans P et les nœuds 1 et 2 sont également adjacents dans G. L'arête entre les nœuds a et c est obtenue à partir de la condition (2). Les nœuds 0 et 2 ne sont pas adjacents dans P et les nœuds 1 et 3 ne le sont pas dans G.

Recherche d'une k-clique

Le GC représente toutes les correspondances possibles entre deux graphes. Chaque clique dans le GC correspond à une sous-structure commune dont la taille est égale au nombre de nœuds de la clique. Une k-clique est une clique de taille k, c'est-à-dire une clique contenant k nœuds. En conséquence, rechercher une k-clique dans le GC revient à rechercher une sous-structure commune de taille k entre les deux graphes. Dans la figure 3.9 la clique de taille 3 formée des nœuds a, b et c correspond à l'occurrence du motif complet P dans le peptide G. Le problème général de la détection de clique, c'est-à-dire savoir si une k-clique est présente dans un graphe est un problème NP-complet [Garey and Johnson, 1979].

En nous inspirant de ces définitions et en les adaptant aux spécificités de notre problème, nous avons développé une nouvelle méthode efficace pour la recherche d'un motif au sein des peptides non-ribosomiaux.

3.3.3 Notre méthode

Notre but est de détecter efficacement et de façon exacte si un sous-graphe connexe de taille au minimum k d'un motif P est une sous-structure du graphe G modélisant un peptide. k est une variable de l'algorithme fixée par l'utilisateur. Si k est égal à la taille du motif, le problème revient à vérifier si le motif P est entièrement présent dans le graphe G. Dans la définition classique des GC, une clique correspond à un sous-graphe induit commun aux deux graphes d'entrée. Un sous-graphe G' d'un graphe G est un sous-graphe induit si et seulement si toutes les arêtes de G ayant leurs extrémités parmi les nœuds communs sont toutes présentes dans G'. Dans notre cas, nous voulons autoriser l'association entre un nœud de G présentant plus d'arêtes que le nœud associé dans le motif P. Par exemple nous voulons que le motif P de la figure 3.10 soit identifié dans le graphe G de la figure 3.10, bien qu'il n'existe pas d'arête entre les nœuds 0 et 4 de P alors qu'une arête est présente entre les nœuds correspondants, 3 et 4, dans G.

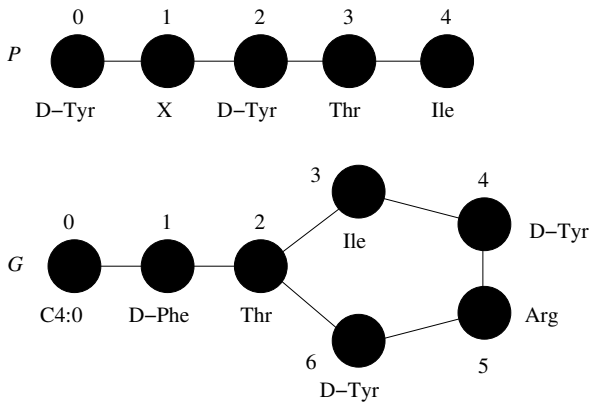


FIG. 3.10 – Exemple d'un motif et d'un peptide

En termes mathématiques, nous recherchons un ensemble de k nœuds dans P pour lesquels le sous-graphe induit dans P est connexe et isomorphe à un sous-graphe, pas forcément induit, de G. Cette asymétrie entre P et G empêche l'utilisation des méthodes classiques pour l'identification d'une sous-structure commune entre deux graphes. Nous avons donc développé une méthode adaptée à ce problème. Notre méthode est basée sur la redéfinition des règles de construction d'un GC et la recherche d'une k-clique au sein du GC.

Nouvelles règles de construction

Avant la construction d'un GC nous vérifions deux conditions assurant qu'une sous-structure de taille k du motif P puisse être présente au sein du graphe G. Premièrement, la taille de G doit être supérieure ou égale à k. Deuxièmement, au moins k nœuds du motif doivent être associés à des nœuds de G. Le GC est construit si ces deux conditions sont respectées.

Lorsque le motif entier est recherché, et dans le but de diminuer le nombre de nœuds dans le GC, nous associons un nœud u de P avec un nœud u' de G si et seulement si, en plus de présenter des étiquettes compatibles, le degré du nœud u' est supérieur ou égal au degré du nœud u. Le

degré d'un nœud est le nombre d'arêtes qui contiennent ce nœud. Par exemple, si un nœud de P a un degré de 3, c'est-à-dire que ce nœud compte 3 arêtes, alors il est impossible de l'associer à un nœud de G ayant seulement 2 arêtes puisqu'un nœud adjacent à un autre nœud de P doit aussi être adjacent dans G.

Afin d'adapter la méthode à notre problème nous devons considérer le cas suivant :

– u est adjacent à v dans P et u' est adjacent à v' dans G (1)

– u n'est pas adjacent à v dans P et u' est ou n'est pas adjacent à v' dans G (2')

En d'autres termes, si les nœuds sont adjacents dans P, les nœuds associés dans G doivent également être adjacents pour avoir une arête dans le GC (condition 1), par contre la réciproque n'est plus nécessairement vraie (condition 2). Ainsi, la condition (2) de la méthode classique est étendue ce qui augmente le nombre d'arêtes dans le GC. Or la taille du GC, en nombre de nœuds et d'arêtes, est le facteur déterminant pour l'efficacité de la méthode. En effet, plus le GC est grand, plus il faut de temps pour rechercher une clique dans ce dernier. Nous devons donc réduire le nombre de nœuds et d'arêtes au sein du GC afin d'obtenir une méthode efficace.

L'extension de la condition (2) dans la méthode classique nous permet d'identifier des sous-graphes communs non obligatoirement induits. Cependant, le nombre d'arêtes dans le GC est augmenté et, par conséquent, le temps de recherche de k-cliques au sein de ce dernier. La figure 3.11 montre l'évolution du nombre d'arêtes en fonction du nombre de nœuds au sein des GC construits soit en utilisant les conditions (1) et (2) de la méthode classique soit en utilisant les conditions (1) et (2').

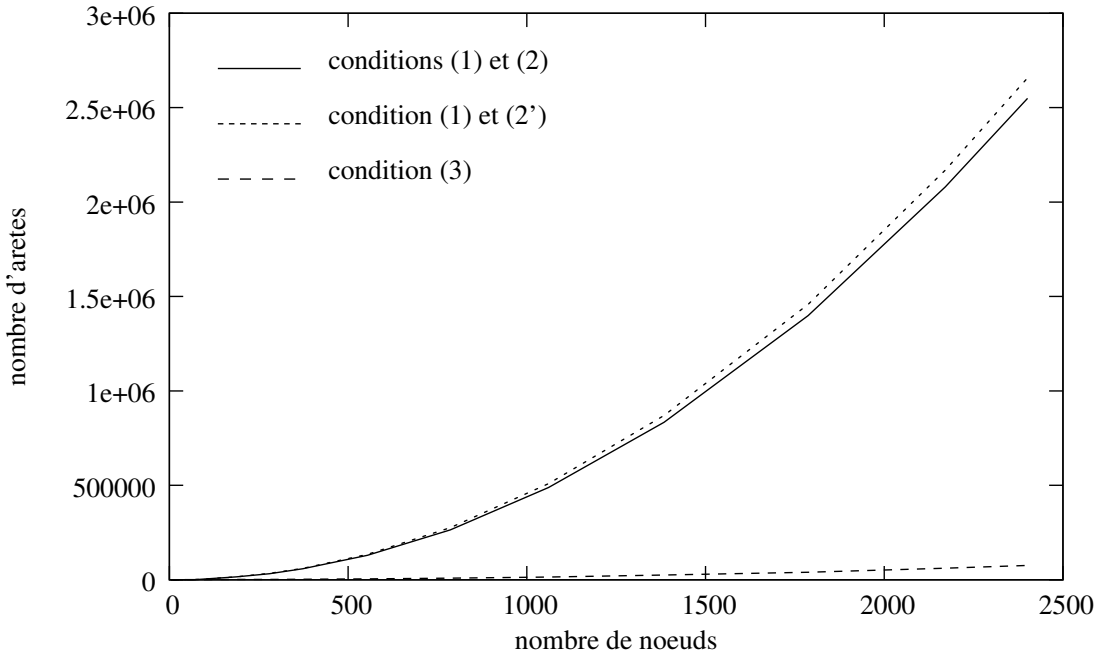


FIG. 3.11 – Nombre d'arêtes en fonction du nombre de nœuds dans les GC en utilisant soit les conditions (1) et (2), soit les conditions (1) et (2'), soit la condition (3).

Cette figure a été obtenue en considérant des motifs correspondants à des peptides réels et en faisant varier le nombre de jokers (« X ») dans ces motifs. Nous constatons que le nombre d'arêtes augmente lorsque la condition (1) est utilisée avec la condition (2'). Cette figure montre également que le nombre d'arêtes dans le GC croît rapidement. En effet, un GC contenant 1 061

nœuds possède 488 724 arêtes avec les conditions (1) et (2) et 509 595 arêtes avec les conditions (1) et (2'). Lorsque le nombre de nœuds du GC est de 2 401, le nombre d'arêtes dépasse les 2 millions que ce soit avec les conditions (1) et (2) ou les conditions (1) et (2'). Rechercher une clique dans un GC aussi dense devient très long. Nous avons donc cherché à réduire le nombre d'arêtes dans le GC.

Le nombre important d'arêtes dans le GC est dû au cas où deux nœuds ne sont pas adjacents dans P. Ce cas amène beaucoup d'associations non significatives car de nombreux nœuds ne sont pas adjacents les uns aux autres dans P. L'idée est donc de filtrer les associations de nœuds ne pouvant pas participer à la sous-structure commune durant la construction du GC. Pour ce faire, nous avons utilisé la taille des chemins élémentaires au sein des deux graphes d'entrée. Un chemin élémentaire (CE) dans un graphe est un chemin sans boucle. Dans chaque graphe P et G, nous calculons la taille de tous les CE entre un nœud donné et tous les autres et ceci pour chaque nœud du graphe. Dans notre cas, nous recherchons des sous-graphes de taille k ce qui nous permet de borner la taille des CE à $k-1$, qui est le nombre maximum de nœuds qui peuvent être visités lors d'un parcours dans un graphe de taille k. Pour un graphe G, nous stockons la taille des CE dans une matrice EPS_G, où $EPS_G[i,j]$ contient le multi-ensemble des tailles de tous les CE entre les nœuds i et j (cf. figure 3.12). Pour calculer la taille des CE entre les nœuds d'un graphe, nous utilisons un algorithme naïf basé sur le parcours en profondeur du graphe. Cet algorithme est quadratique mais est très rapide en pratique du fait de la petite taille des graphes traités dans notre cas.

EPS_P

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

EPS_G

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

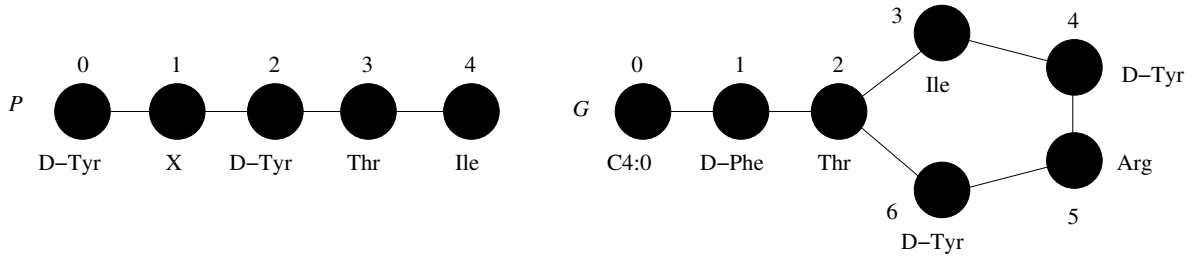


FIG. 3.12 – Exemples des matrices des tailles des CE pour un motif P et un peptide G avec k égal à la taille de P

Par exemple, il existe deux CE entre les nœuds 1 et 4 de G. Le premier est de taille 3 (chemin passant par les nœuds 1, 2, 3 et 4) et le second est de taille 4 (chemin 1-2-6-5-4). Les nœuds 0 et 4 dans G sont connectés par deux CE, l'un de taille 4 et l'autre de taille 5. Le CE de taille 5

n'apparait pas car il est supérieur à $k-1$ qui est ici égal à 4 (dans l'exemple P est de taille 5). A l'aide de ces matrices, nous avons défini de nouvelles règles de construction pour le GC.

Les nœuds $U(u,u')$ et $V(v,v')$ dans le GC sont reliés par une arête si et seulement si le multi-ensemble des tailles des CE entre u et v dans P est inclus dans le multi-ensemble des tailles des CE entre u' et v' dans G. Cela signifie que les distances entre deux nœuds dans P doivent être présentes dans les distances respectives de G pour avoir une arête dans le GC. Dans la figure 3.12, le motif P est linéaire ce qui implique qu'il ne peut y avoir qu'un CE entre deux nœuds ; alors qu'il peut y en avoir deux entre deux nœuds de G qui contient un cycle. Or, nous voulons trouver l'occurrence de P dans G. En résumé, un GC entre un graphe P et un graphe G est défini comme suit :

- chaque nœud $U(u,u')$ du GC correspond à l'association d'un nœud u de P et d'un nœud u' de G avec $f(u)$ compatible avec $f(u')$.
- deux nœuds $U(u,u')$ et $V(v,v')$ sont adjacents dans le GC si et seulement si :
  - $u \neq v$ et $u' \neq v'$ et $EPS_P[u,v] \subseteq EPS_G[u',v']$. (3)

Nous avons remplacé les conditions (1) et (2) de la méthode classique par la condition (3). L'introduction de cette nouvelle condition réduit le nombre d'arêtes dans le GC sans perte d'information sur l'occurrence possible du motif. La figure 3.11 montre l'évolution du nombre d'arêtes en fonction du nombre de nœuds au sein des GC construits soit en utilisant les conditions (1) et (2) de la méthode classique soit en utilisant les conditions (1) et (2'), soit en utilisant la condition (3) de la nouvelle méthode. Nous constatons que le nombre d'arêtes dans le GC diminue fortement lorsque la condition (3) est utilisée. En effet, un GC contenant 1 061 nœuds possède 488 724 arêtes avec les conditions (1) et (2), 509 595 arêtes avec les conditions (1) et (2') et 14 564 arêtes avec la condition (3). Le nombre d'arêtes est divisé par plus de 33 grâce à la condition (3).

En plus de réduire le nombre de nœuds et le nombre d'arêtes au sein du GC, cette nouvelle méthode détecte des sous-graphes communs induits dans P et pas forcément dans G. La figure 3.13 montre un exemple de construction d'un GC avec la méthode classique et avec notre méthode.

Dans un GC, un nœud est identifié par une lettre et représente une association entre un nœud de P et un nœud de G ayant des étiquettes compatibles. Par exemple, le nœud « a » représente l'association entre le nœud 0 de P et le nœud 4 de G. Dans la figure 3.13, les arêtes en pointillés correspondent aux arêtes qui diffèrent entre les deux méthodes de construction. Les arêtes en gras forment une clique de taille 5 (la taille de P). Nous constatons qu'il n'existe pas d'arête entre les nœuds a et l dans le GC construit avec la méthode classique. En effet, les nœuds 0 et 4 ne sont pas adjacents dans P alors que les nœuds 4 et 3 sont adjacents dans G. Par contre, une arête existe entre ces nœuds dans le GC construit avec la nouvelle méthode. Cette arête forme une clique de taille 5 ce qui signifie que le motif P est présent dans le graphe G. Ainsi, avec la nouvelle méthode, il est possible d'identifier le motif P dans le graphe G, ce qui n'est pas possible avec la méthode classique. Par ailleurs, cet exemple montre également une réduction du nombre de nœuds. En effet, le nœud z du GC construit avec la méthode classique disparaît dans le GC construit avec la nouvelle méthode. Ce nœud associe le nœud 1 de P qui est le caractère joker et le nœud 0 de G. Or, le degré du nœud 1 de P (2) est strictement supérieur au degré du nœud 0 de G (1). Ainsi, le nœud z disparait du GC construit avec la nouvelle méthode. Enfin, cet exemple montre également une réduction du nombre d'arêtes dans le GC construit avec la nouvelle méthode. En effet, le GC construit avec la méthode classique contient 22 arêtes alors que celui construit avec la nouvelle méthode n'en contient que 19. Par exemple, dans le GC construit avec la méthode classique, il existe une arête entre les nœuds $b(0,6)$ et $l(4,3)$. Cette arête disparaît lorsqu'on utilise la nouvelle méthode car les tailles des CE entre les nœuds 0 et

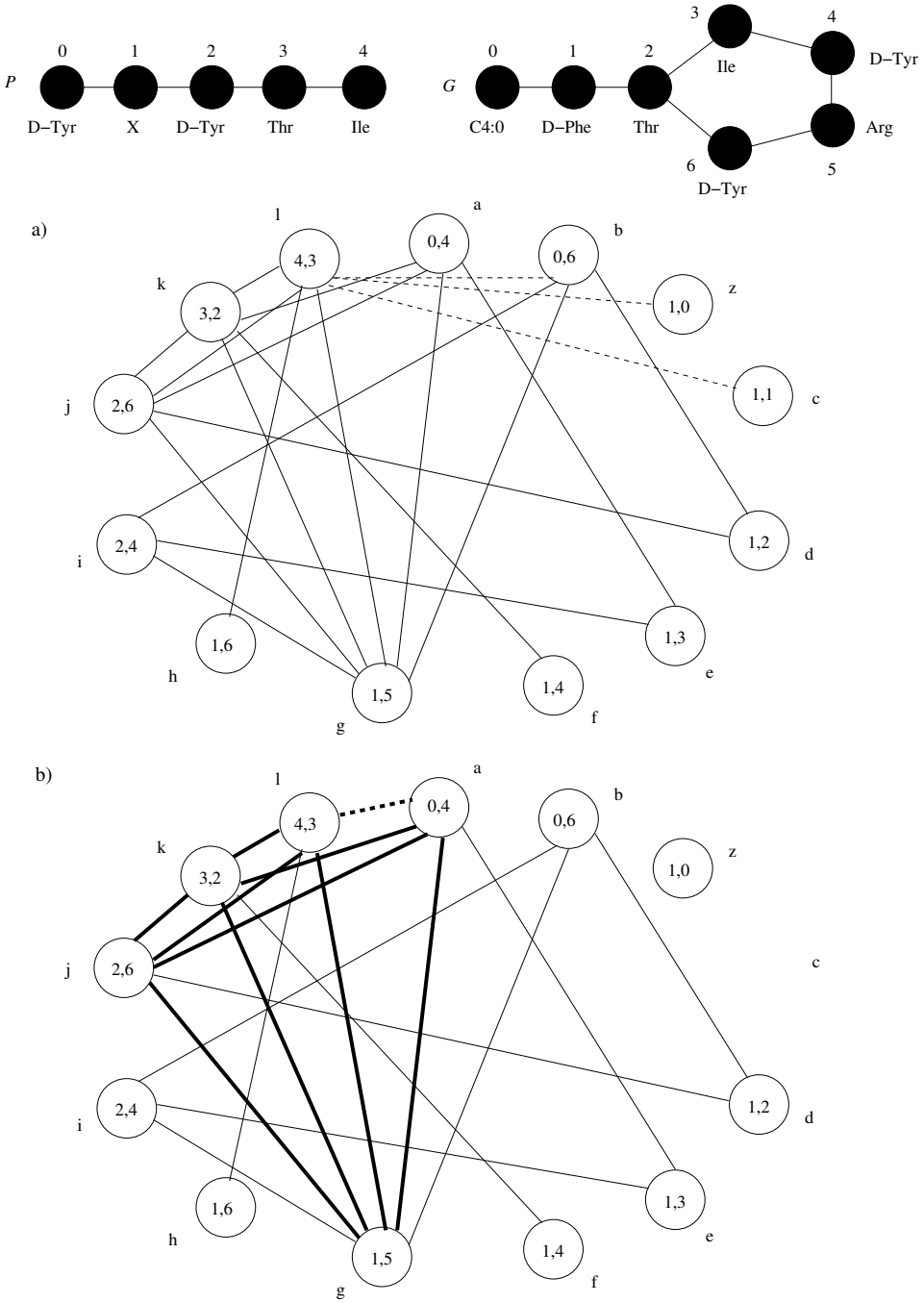


FIG. 3.13 – Exemple de construction du GC des graphes P et G avec a) la méthode classique et b) notre méthode. Les arêtes en pointillés sont celles qui diffèrent entre les GC construits avec les deux méthodes. Les arêtes en gras forment une clique de taille 5.

4 dans P (ici $\{4\}$) ne sont pas incluses dans les tailles des CE entre les nœuds 6 et 3 (ici $\{2,3\}$) dans G. La nouvelle méthode exclut ce type d'arêtes et réduit ainsi le nombre d'arêtes dans le GC correspondant.

Une fois le GC construit, l'étape suivante est la recherche d'une k-clique au sein de ce GC.

Recherche d'une k-clique

La présence d'une k-clique au sein d'un GC signifie qu'il existe un sous-graphe induit de P qui est un sous-graphe de G. Dans le cas où k est inférieur à la taille de P, c'est-à-dire quand une sous-structure de taille k commune entre P et G est recherchée, nous devons vérifier en plus, que le sous-graphe correspondant est connexe dans P, et par conséquent dans G.

Pour rechercher une k-clique, nous utilisons un algorithme de séparation et évaluation (*branch and bound*) inspiré de [Mehlhorn, 1984]. C'est un algorithme exhaustif qui explore l'arbre de recherche du graphe en profondeur. Le pseudo-code est donné dans la procédure 1.

Procédure 1 k-clique (Liste solution,Liste compatible)

Entrées: Liste solution, Liste compatible

 # solution contient la liste des nœuds de la clique courante
 # compatible contient la liste des nœuds compatibles, c'est-à-dire les nœuds qui sont adjacents à tous les nœuds de solution
 # k est la taille de la clique à rechercher
 # clique est un booléen égal à vrai quand une k-clique est trouvée
 # continu est un booléen égal à vrai tant que $|V_P| - k$ nœuds du motif n'ont pas été éliminés de la solution

1: **si** *solution.taille* $\geq k$ **alors**
2: **si** $k = P.taille$ **alors**
3: *clique* $= vrai$
4: **else**
5: **si** solution est connexe **alors**
6: *clique* $= vrai$
7: **finsi**
8: **finsi**
9: **finsi**
10: **si** (*compa.taille* $\neq 0$) ET (*clique* $\neq vrai$) ET (*continu* $\neq faux$) **alors**
11: int element=premier élément de compatible
12: enleve element de compatible
13: Liste nouveau= intersection de compatible et de la liste d'adjacence de element
14: ajout de element dans solution
15: **si** *solution.taille* + *nouveau.taille* $\geq k$ **alors**
16: k-clique(solution, nouveau)
17: le dernier élément de solution est retiré
18: **finsi**
19: **finsi**

A chaque nœud de profondeur h, nous essayons d'étendre la clique courante de taille h avec un nouveau nœud dans le but d'obtenir une clique de taille $h+1$. Chacun de ces nœuds ajoutés doit être relié à tous les nœuds de la solution courante. Le processus est réitéré récursivement jusqu'à ce qu'aucun nœud ne puisse plus être ajouté ou qu'une k-clique soit trouvée. L'arbre de recherche est élagué en n'explorant pas les branches dont la longueur est inférieure à k (ligne 15 de la procédure 1). Une fois qu'une clique de taille k est trouvée, la recherche est arrêtée et le peptide possédant le motif est retourné. Lorsque plus de $(|V_P| - k)$ nœuds du motif ne participent pas à la k-clique, avec V_P le nombre de nœuds du motif, la recherche peut être stoppée. Dans le cas où $V_P = k$, c'est-à-dire que le motif complet est recherché au sein de G, tous les nœuds du

motif doivent contribuer à la clique. Dans l'exemple de la figure 3.13, le nœud 0 de P est impliqué dans les nœuds a et b du GC. Si aucun de ces deux nœuds n'est impliqué dans la clique alors la recherche d'une clique de taille 5 peut s'arrêter. En effet, si ni le nœud a ni le b du GC n'est impliqué dans la clique, il est impossible de trouver une clique de taille 5 dans ce GC, c'est-à-dire une occurrence complète de P dans G. Enfin, une k-clique contenant le nœud de degré maximal du motif P est recherchée dans le GC afin de détecter rapidement une non-occurrence du motif. Initialement, la procédure est lancée avec la liste *solution* vide et la liste *compatible* contenant tous les nœuds du GC.

3.3.4 Tests d'efficacité

Tous les tests de cette section ont été réalisés sur un PC avec un processeur de 1,73 GHz et 512 MB de RAM. L'algorithme de recherche de k-clique utilisé est le même dans les deux cas, seule la méthode de construction du GC varie. Les tests ont été réalisés sur un ensemble contenant 711 structures de peptides non-ribosomiaux modélisées par des graphes étiquetés non-orientés (peptides NOR00001 à NOR00711 de Norine). La figure 3.14 a) montre que le motif le plus fréquent est le motif linéaire. La figure 3.14 b) montre que plus de 70% des 711 peptides étudiés ont au moins 7 monomères. Ainsi, en recherchant un motif composé de 7 caractères « joker », plus de 70% des 711 peptides sont testés. Les autres sont rejetés dès le test portant sur la comparaison de la taille du peptide par rapport celle du motif.

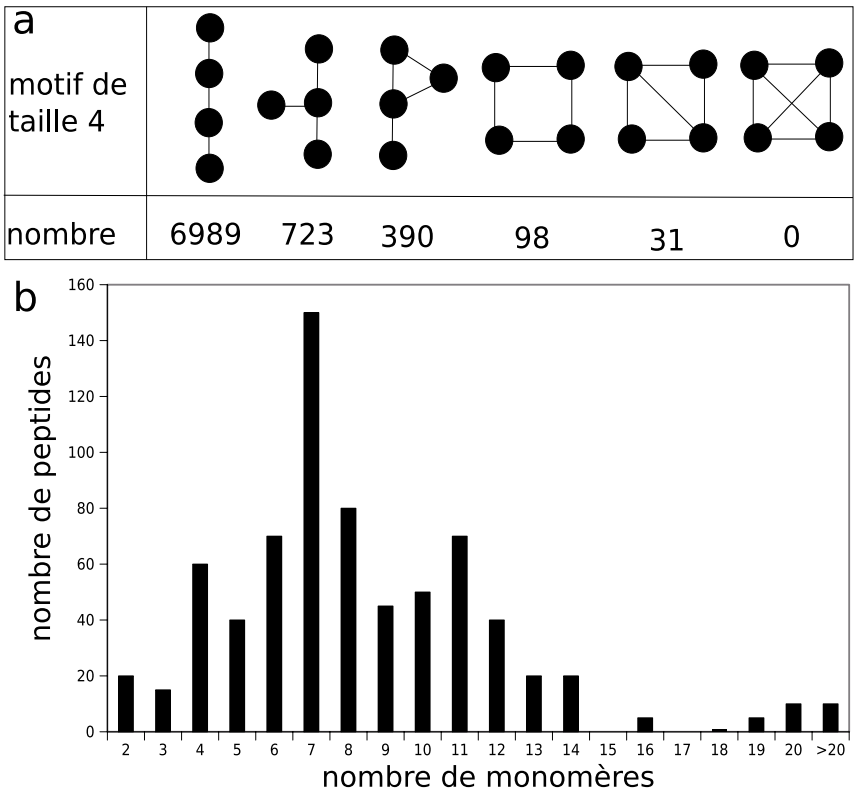


Fig. 3.14 – Etude des caractéristiques structurales des 711 peptides NOR00001 à NOR00711 de Norine : a) distribution des motifs de taille 4 et b) distribution des tailles.

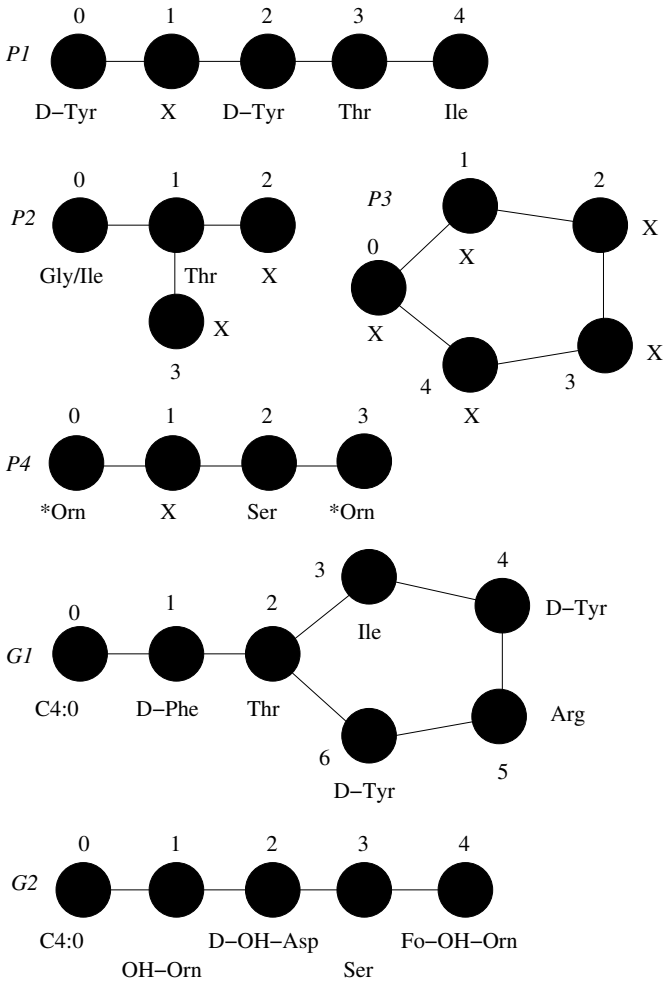


FIG. 3.15 – Quelques exemples de motifs structuraux et de graphes modélisant des peptides

Dans le but de tester l'efficacité de notre méthode, nous avons dans un premier temps, comparé le nombre de nœuds et d'arêtes dans des GC construits avec la méthode classique et la nouvelle méthode, lors de recherche de motifs entiers au sein des différents peptides de la figure 3.15.

motif	peptide	nombre de nœuds	nombre d'arêtes
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**

TAB. 3.1 – Nombre de nœuds et d'arêtes au sein des GC construits avec la méthode classique et la nouvelle. Le premier nombre correspond à la méthode classique et le second, en gras, correspond à la nouvelle méthode.
$^{(a)}$ motif correspondant à l'alamethicine moins le dernier monomère
$^{(b)}$ motif correspondant à l'alamethicine contenant les 20 monomères
$^{(c)}$ motif linéaire composé de 19 « X »

Les résultats de la table 3.1 montrent que le nombre de nœuds et d'arêtes de la nouvelle méthode est toujours inférieur ou égal à ceux de la méthode classique. Dans le cas de la recherche du motif $P3$ dans le graphe $G2$, il n'y a pas d'arête dans le GC construit avec la nouvelle méthode car la liste des CE de $P3$ n'est pas incluse dans celle des CE de G. En effet, $P3$ est cyclique et chaque paire de nœuds est connectée par deux CE alors que $G2$ est linéaire, il n'existe donc qu'un seul CE entre chaque paire de nœuds. Dans ce cas, notre méthode trouve directement le résultat, sans rechercher de k-clique. Lorsque le motif composé de 19 « X » est recherché dans le graphe correspondant à l'alamethicine, le nombre de nœuds est de 380 avec la méthode classique et de 346 avec la nouvelle méthode. Cette diminution du nombre de nœuds est due à l'ajout de la condition sur le degré des nœuds. De plus, dans cet exemple, le nombre d'arêtes passe de 53 010 à 3 948 avec la nouvelle méthode ce qui représente plus de 13 fois moins d'arêtes qu'avec la méthode classique.

Ces exemples montrent que la nouvelle méthode mise en place réduit le nombre de nœuds et d'arêtes au sein des GC, offrant ainsi une recherche efficace d'une k-clique dans ce GC compact.

Dans le but de vérifier la diminution du temps d'exécution de recherche d'un motif au sein d'un ensemble de graphes, nous avons mesuré le temps de recherche de différents motifs entiers au sein des 711 peptides NOR00001 à NOR00711 de NORINE.

La table 3.2 montre les résultats obtenus avec la méthode classique comparés à ceux obtenus avec la nouvelle méthode. La première constatation est que le nombre de résultats obtenus, c'est-à-dire le nombre de peptides contenant le motif recherché, est souvent plus élevé avec la nouvelle méthode. Les peptides trouvés en plus contiennent soit un cycle qui n'est pas présent dans le motif, soit une double liaison entre deux monomères. Dans l'exemple 6, un motif composé de deux monomères quelconques est recherché. Les 711 graphes testés contiennent tous au minimum deux monomères et par conséquent les deux méthodes devraient retourner les 711 graphes. Cependant seuls 698 graphes sont retournés par la méthode classique à cause de la présence de 13 dipeptides cycliques. En effet, certains peptides peuvent contenir un hétérocycle formé entre deux monomères adjacents, ce qui se traduit par deux nœuds reliés par deux arêtes. La méthode classique ne trouve pas ces cas particuliers.

Ensuite, nous pouvons constater que la nouvelle méthode est beaucoup plus rapide que la

	motif	nombre de résultats	temps
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	ND / **0**	ND / **280 ms**

TAB. 3.2 – Temps de recherche de différents motifs complets au sein d'un ensemble de 711 graphes. Le premier nombre est celui concernant la méthode classique et le second, en gras, est celui de la nouvelle méthode. ND signifie que le temps est supérieur à 8 heures.

méthode classique. Dans l'exemple 13, la méthode classique ne produit pas de résultat après 8 heures de calculs alors que seulement 394 ms sont nécessaires à la nouvelle méthode pour retourner la liste des graphes contenant le motif recherché. Pour le motif linéaire composé de 7 « X » (exemple 7), présent dans plus de 70% des 711 peptides, le temps de recherche est 8 fois plus long lorsque la méthode classique est utilisée. L'exemple 14 représente un test négatif car il n'est pas présent dans les 711 peptides. Une fois encore la méthode classique ne produit pas de résultat après 8 heures de calculs alors que la nouvelle méthode ne prend que 280 ms.

Cette expérience, ainsi que les précédentes prouvent que la nouvelle méthode est particulièrement efficace et adaptée à notre problème. En effet, la méthode développée recherche une sous-structure commune de taille k entre un motif P et un graphe G. Il faut moins d'une seconde pour rechercher un motif dans un ensemble de 711 graphes. Cette méthode peut ensuite être facilement modifiée pour comparer strictement deux peptides de façon efficace.

3.3.5 Comparaison stricte

Rechercher si un peptide est présent dans un ensemble de peptides connus est souvent nécessaire lors de l'identification de nouvelles molécules. En effet, après l'identification de la structure d'un peptide isolé à partir d'un organisme étudié, la première étape est de savoir si ce peptide a déjà été identifié. Pour ce faire nous avons besoin de comparer strictement ce peptide à l'ensemble des peptides connus. Dans cette section, le but est de savoir si deux peptides modélisés par des graphes sont strictement identiques. Nous avons donc modifié la recherche de

motifs dans le but de comparer strictement deux graphes.

Un GC entre un graphe $G1$ et un graphe $G2$ est défini comme suit :

- chaque nœud du GC $U(u, u')$ correspond à l'association d'un nœud u de $G1$ et d'un nœud u' de $G2$ avec $deg(u) = deg(u')$ et $f(u) = f(u')$.
- deux nœuds $U(u, u')$ et $V(v, v')$ sont adjacents dans le GC si et seulement si $u \neq v$ et $u' \neq v'$ et $EPS_{G1}[u, v] = EPS_{G2}[u', v']$.

Cette fois, nous associons un nœud de $G1$ avec un nœud de $G2$ si et seulement si les degrés de ces nœuds sont égaux et s'ils présentent exactement les mêmes étiquettes. Deux nœuds $U(u, u')$ et $V(v, v')$ sont adjacents dans le GC, si la liste des tailles des CE entre les nœuds u et v de $G1$ est identique à celle de la taille des CE entre les nœuds u' et v' de $G2$. Toutes ces conditions d'égalité aboutissent à la détection de graphes identiques en un temps très court.

3.4 Extension de la méthode à la recherche de similarités

Il peut être intéressant de rechercher des peptides similaires à un peptide donné. En effet, des peptides possédant des structures similaires présentent souvent des activités biologiques similaires. Une comparaison par recherche de similarités aide à mettre en évidence des caractéristiques structurelles communes à deux peptides. Dans un premier temps, nous avons adapté la méthode de recherche de motifs pour mettre en évidence la plus grande sous-structure commune entre les deux peptides. Pour estimer la similarité entre deux peptides nous avons besoin d'une mesure chiffrée permettant de comparer les similarités trouvées entre différentes paires de peptides. Nous avons mis en place une distance permettant d'estimer la similarité entre deux peptides. Enfin, au vu du grand nombre de monomères pouvant être incorporés au sein des peptides non-ribosomiaux, nous avons mis en place plusieurs niveaux de regroupement des monomères.

3.4.1 Recherche de similarités

Dans le but de trouver une similarité entre deux peptides, nous avons modifié l'algorithme conçu pour la recherche de motifs. Pour ce faire, nous devons identifier la sous-structure commune maximale entre deux peptides, c'est-à-dire nous ne recherchons plus une k-clique au niveau du GC, mais la clique maximale. Le GC est toujours construit sur le même principe que précédemment, cependant, quelques modifications ont été réalisées. Tout d'abord, il faut avoir une comparaison commutative. En effet, la similarité calculée entre le graphe $G1$ et le graphe $G2$ doit être la même que celle calculée entre les graphes $G2$ et $G1$. Le GC entre les graphes $G1$ et $G2$ est défini comme suit :

- chaque nœud $U(u, u')$ du GC correspond à l'association d'un nœud u de $G1$ et d'un nœud u' de $G2$ avec $f(u) = f(u')$
- deux nœuds $U(u, u')$ et $V(v, v')$ du GC sont adjacents si et seulement si $u \neq v$ et $u' \neq v'$ et $EPS_{G1}[u, v] \cap EPS_{G2}[u', v'] \neq \emptyset$

La première différence entre la version adaptée à la recherche de similarités et la version de recherche d'un motif est qu'il n'y a plus de contrainte sur le degré d'un nœud. En effet, comme la comparaison doit être commutative, il est impossible de mettre une condition sur le degré du nœud. Pour que deux nœuds soient associés, il suffit qu'ils portent la même étiquette. La seconde modification est que l'intersection entre la liste des tailles des CE entre deux nœuds d'un graphe et la liste des tailles des CE entre les deux nœuds de l'autre graphe doit être non-nulle pour obtenir une arête dans le GC. Dans ce cas, la condition d'intersection est plus permissive que la

condition d'inclusion précédente car nous recherchons ici une sous-structure commune maximale et non plus un motif dans un peptide.

La clique maximale représente alors la sous-structure commune maximale. Dans ce cas, la sous-structure maximale n'est pas forcément connexe. En effet, un nœud de $G1$ est mis en correspondance avec un nœud de $G2$ si ces deux nœuds se trouvent dans des environnements comparables, même si ces derniers ne sont pas connectés aux autres nœuds de la sous-structure commune identifiée.

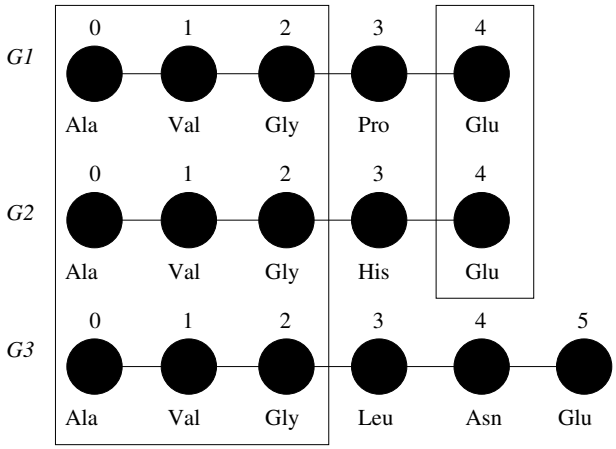


FIG. 3.16 – Exemple de la similarité entre différents graphes

La figure 3.16 montre la sous-structure commune maximale non-connexe entre plusieurs graphes. Nous pouvons constater que les nœuds $0, 1, 2, 4$ de $G1$ sont associés aux nœuds $0, 1, 2, 4$ de $G2$ car ils présentent les mêmes étiquettes. De plus, il existe un CE de taille 2 entre les nœuds 2 et 4 de $G1$ et entre les nœuds 2 et 4 de $G2$. Donc, l'intersection entre la liste des tailles des chemins de ces nœuds n'est pas vide et une arête est tracée dans le GC. Au contraire, les nœuds 4 de $G1$ et 5 de $G3$ ne font pas partie de la sous-structure commune maximale car la taille des CE entre le nœud 5 et le nœud 2 de $G3$ est plus grande que celle des chemins entre les nœuds 4 et 2 de $G1$. En résumé, deux nœuds sont mis en correspondance s'ils se trouvent dans des environnements similaires. Cette méthode autorise donc des substitutions non connues *a priori*, mais pas des insertions ou délétions de nœuds.

Une fois le GC construit, il faut rechercher la clique maximale, c'est à dire la clique ayant le nombre de nœuds maximum au sein de ce GC. Lorsque deux cliques présentent la même taille, c'est-à-dire le même nombre de nœuds, celle qui implique le plus grand nombre d'arêtes au sein des graphes à comparer est conservée. La figure 3.17 montre un cas particulier où deux cliques de même taille existent dans le GC mais n'impliquent pas le même nombre d'arêtes dans les graphes d'entrée.

En effet, nous constatons que deux cliques de taille 3 existent dans le GC. La première comprend les nœuds $0, 1, 3$ de $G1$ et les nœuds $0, 1, 3$ de $G2$. Cette clique de taille 3 correspond à une sous-structure commune non connexe composée de 3 nœuds et d'une seule arête (l'arête entre les nœuds 0 et 1) entre les graphes $G1$ et $G2$. Une seconde clique de taille 3 contient les nœuds $5, 6, 7$ de $G1$ et les nœuds $4, 5, 6$ de $G2$. Cette seconde clique correspond à une sous-structure connexe commune composée de 3 nœuds et de 2 arêtes. Dans ce cas, si la taille de la clique est mesurée uniquement en comptant le nombre de nœuds, l'une ou l'autre des deux cliques peuvent être retournées. Cependant, la seconde implique un nombre d'arêtes supérieur au sein

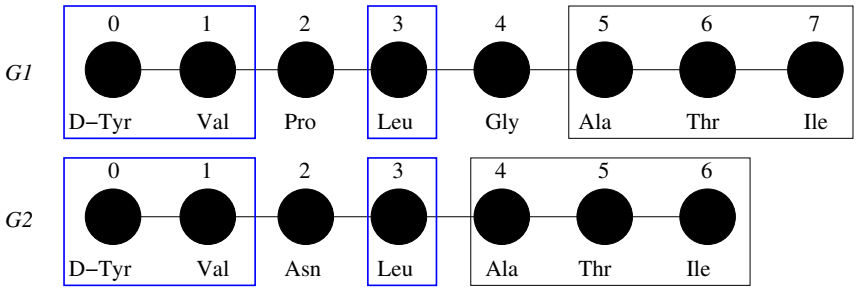


FIG. 3.17 – Exemple de comparaison de deux graphes

des graphes d'entrée et est donc meilleure. Nous considérons donc une clique maximale comme une clique contenant un nombre maximal de nœuds, mais également impliquant un nombre maximal d'arêtes dans les graphes d'entrée.

Ce modèle prend en compte des substitutions de nœuds non connues *a priori*, mais pas des insertions-délétions de nœuds.

3.4.2 Distance entre deux peptides

Une distance est une mesure numérique reflétant les différences entre deux objets. Une distance entre deux graphes est souvent normalisée pour être comprise entre 0 et 1. Plus la distance est proche de 0 et plus les graphes se ressemblent et inversement, plus la distance est proche de 1 et plus les deux graphes sont différents. Il existe différentes façons de calculer une distance entre deux graphes. L'utilisation de ces différentes mesures dépend des graphes sur lesquels elles s'appliquent et de ce que les auteurs cherchent à mettre en évidence [Schenker et al., 2003].

Une distance très utilisée lors de la comparaison de deux graphes $G1$ et $G2$ est celle basée sur le sous-graphe commun maximal ou *Maximal Common Subgraph* (MCS), introduite par Bunke et Shearer [Bunke and Shearer, 1998] :

$$d(G1, G2) = 1 - \frac{|mcs(G1, G2)|}{Max(|G1|, |G2|)}$$

où $|mcs(G1, G2)|$ est la taille (nombre de nœuds) du sous-graphe commun induit maximal de $G1$ et $G2$ et $Max(|G1|, |G2|)$ est la taille (nombre de nœuds) du plus grand graphe entre $G1$ et $G2$.

Soient deux graphes $G1 = (V1, E1)$ et $G2 = (V2, E2)$. D'après notre méthode, nous calculons un sous-graphe commun maximal $mcs(G1, G2) = (V_{mcs}, E_{mcs})$. Cependant, ici le sous-graphe n'est pas forcement un sous-graphe induit, comme dans la plupart des cas dans la littérature. Ceci est dû aux règles de définition du GC. Le mot « maximal » est également à prendre avec précaution, car dans le cas de sous-graphes non-connexes, la maximalité est conditionnée par le fait que les composantes connexes soient reliées par des chemins de même longueur.

Autrement dit, le sous-graphe $mcs(G1, G2)$ apparaît en tant que sous-graphe (pas forcement induit) de G1 et de G2 , c'est-à-dire :

$$V_{mcs} \subseteq V1, E_{mcs} \subseteq E1$$

$$V_{mcs} \subseteq V2, E_{mcs} \subseteq E2$$

Cependant, si nous considérons le sous-graphe induit par V_{mcs} du graphe $G1$, c'est-à-dire le sous-graphe de $G1$ contenant les nœuds V_{mcs}, et le sous-graphe induit par V_{mcs} du graphe

$G2$, chacun d'eux contient les arêtes E_{mcs} mais éventuellement des arêtes supplémentaires. Par exemple, si $G1$ est un cycle de n nœuds et $G2$ est un graphe linéaire de n nœuds, $mcs(G1, G2)$ est $G2$ lui-même et le sous-graphe induit de $G1$ contient une arête supplémentaire. Nous proposons de « corriger » cette mesure pour notre cas comme suit :

$$d(G1, G2) = 1 - \frac{|mcs(G1, G2)|}{Max(|G1|, |G2|)} \cdot \delta(G1[V_{mcs}], G2[V_{mcs}])$$

Ici, $Gi[V_{mcs}]$ est le sous-graphe de Gi induit par V_{mcs} . $G1[V_{mcs}]$ et $G2[V_{mcs}]$ ont le même ensemble de nœuds (V_{mcs}), ils contiennent les arêtes E_{mcs} mais chacun peut contenir des arêtes supplémentaires. δ peut donc être défini en terme d'ensemble d'arêtes :

$$\delta(G1[V_{mcs}], G2[V_{mcs}]) = 1 - \frac{|E(G1[V_{mcs}])| + |E(G2[V_{mcs}])| - 2|E_{mcs}|}{\frac{|V_{mcs}|(|V_{mcs}|-1)}{2}}$$

L'intuition ici est de compter le nombre d'arêtes par lesquels $G1[V_{mcs}]$ et $G2[V_{mcs}]$ diffèrent, rapporté au nombre d'arêtes possibles sur V_{mcs}. Si $\delta = 1$, c'est-à-dire que $G1[V_{mcs}]$ et $G2[V_{mcs}]$ ont le même nombre d'arêtes, alors la distance calculée est celle de Bunke et Shearer.

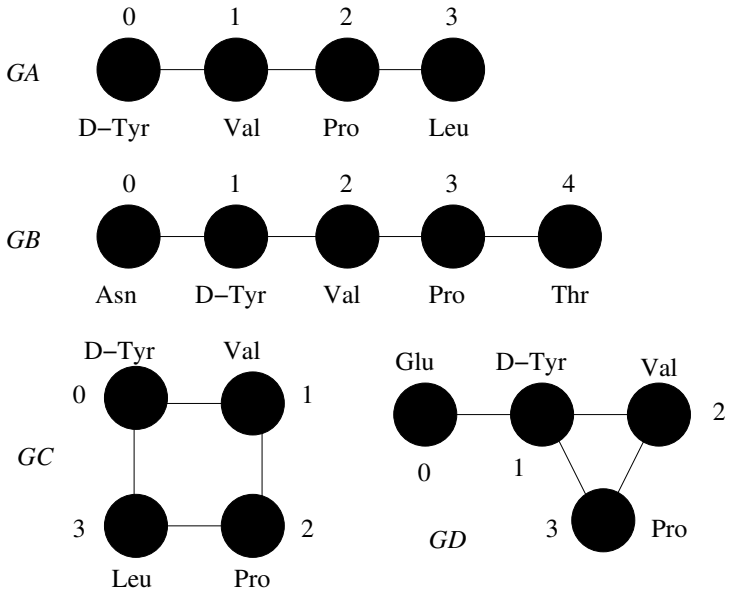


FIG. 3.18 – Quelques exemples de graphes

Le tableau 3.3 montre les différentes valeurs intervenant dans le calcul des distances pour les graphes de la figure 3.18. Les distances obtenues reflètent bien les ressemblances entre peptides. En effet, les graphes les plus proches sont GA et GC qui ne diffèrent que par une arête. En observant les distances, le graphe GA est plus proche du graphe GB que du graphe GD. En effet, les structures, GA et GB sont toutes les deux linéaires alors que GD contient un cycle. La distance la plus grande est observée entre les graphes GB et GD ce qui est cohérent puisque la graphe GB contient un nœud de plus que les autres graphes et que le graphe GD contient une arête en plus entre les nœuds communs. La distance mise en place semble donc bien adaptée aux cas des peptides non-ribosomiaux.

Cette distance est utilisée pour évaluer la similarité entre deux peptides modélisés par des graphes étiquetés non-orientés.

$G1$	$G2$	V_{mcs}	$E(G1[V_{mcs}])$	$E(G2[V_{mcs}])$	E_{mcs}	δ	$\frac{\lvert mcs(G1,G2)\rvert}{Max(\lvert G1\rvert,\lvert G2\rvert)}$	$d(G1,G2)$
GA	GB	3	2	2	2	1	$3/5$	$0,4$
GA	GC	4	3	4	3	$5/6$	1	$0,16$
GA	GD	3	2	3	2	$2/3$	$3/4$	$0,5$
GB	GC	3	2	2	2	1	$3/5$	$0,4$
GB	GD	3	2	3	2	$2/3$	$3/5$	$0,6$
GC	GD	3	2	3	2	$2/3$	$3/4$	$0,5$

TAB. 3.3 – Calcul des distances pour les graphes de la figure 3.18

3.4.3 Clustering des monomères

Comme nous l'avons vu précédemment, il existe un très grand nombre de monomères qui peuvent être incorporés au sein des peptides non-ribosomiaux. Ce très grand nombre de monomères possibles rend difficile la recherche de similarités au sein des structures. En effet, si nous ne considérons que les monomères exacts, il y a peu de chances de mettre en évidence des similarités entre structures. Nous avons donc mis en place un regroupement des monomères, dans le but de pouvoir mettre en évidence des similarités de structures entre des peptides assez éloignés. Deux niveaux de clustering ont été développés. Ces deux niveaux sont utilisés pour rechercher des peptides similaires au sein de la base de donnée, en remplaçant un monomère spécifique par un autre monomère moins spécifique.

Clustering de niveau 1

Le clustering de niveau 1 consiste à regrouper dans un premier temps tous les acides gras. En effet, différents acides gras peuvent être associés à une même chaîne peptidique. Parfois, la nature exacte de l'acide gras n'est pas déterminante pour la recherche de similarités. Le fait de savoir qu'il y a un acide gras à une position donnée peut suffire dans certaines études. De la même manière, les sucres sont regroupés en une seule catégorie. Les différents chromophores rencontrés au sein des peptides forment un seul groupe. Enfin, tous les dérivés d'un même monomère sont regroupés. Par exemple, le groupe valine contiendra la valine et ses dérivés comme la N-méthylvaline ou encore la D-valine. Ce type de clustering est basé sur l'incorporation des acides aminés par les synthétases. En effet, dans le produit final, la valine ou un de ses dérivés peut être observé, mais c'est généralement la valine qui est sélectionnée et activée par la synthétase. Ce clustering nous permet de passer de 506 monomères à 149 groupes de monomères.

Clustering de niveau 2

Le clustering de niveau 2 consiste à former des groupes plus grands en utilisant les groupes issus du clustering de niveau 1. Dans ce niveau de clustering, les groupes de monomères et leurs dérivés sont regroupés en fonction des propriétés physico-chimiques des monomères. Nous nous sommes inspirés des « small clusters » de [Rausch et al., 2005] utilisés pour la prédiction de la spécificité des domaines d'adénylation (section 2.2). Par exemple, la glycine et l'alanine, ainsi que leurs dérivés, sont regroupées en une même catégorie car elles sont toutes les deux de petite taille. Les acides aminés aliphatiques présentant une chaîne carbonée hydrophobe (Val, Leu, Ile, Abu, Iva), ainsi que leurs dérivés, sont regroupés dans une même classe. Les monomères très rares sont regroupés afin de minimiser l'impact de leur présence au sein d'un peptide. Ce niveau de

clustering mène à de grands groupes de monomères partageant des propriétés physico-chimiques similaires. Avec le clustering de niveau 2, les 506 monomères sont répartis dans 31 groupes.

Dans cette section, nous avons présenté une nouvelle méthode pour la recherche de motifs structuraux au sein des peptides non-ribosomiaux. Cette nouvelle méthode est basée sur la redéfinition des règles de construction des graphes de compatibilité, en utilisant la taille des chemins élémentaires, et sur la recherche classique d'une k-clique au sein du graphe de compatibilité construit. Cette méthode est adaptée au cas des peptides non-ribosomiaux et est très efficace. En effet, la recherche d'un motif donné au sein d'un ensemble contenant plus de 700 graphes nécessite moins d'une seconde. Cette méthode a été étendue à la comparaison de structures. Il est possible de réaliser une recherche exacte de la structure d'un peptide donné au sein d'un ensemble de structures. Il est également possible de rechercher des peptides similaires grâce au calcul d'une distance basée sur la sous-structure commune maximale. Nous avons donc mis en place des méthodes informatiques efficaces permettant l'analyse des structures des peptides non-ribosomiaux. Ces méthodes ont été implémentées dans Norine*, une ressource publique pour les peptides non-ribosomiaux.*

Chapitre 4

Norine

Au commencement de ce travail, il n'existait aucune ressource spécifique aux produits issus de la voie non-ribosomiale, bien que le besoin d'une ressource centralisant les informations sur ces peptides et permettant leur analyse était de plus en plus important. Pour pallier ce manque, nous avons développé Norine, la première ressource publique dédiée aux peptides non-ribosomiaux. Le nom de Norine provient de l'association du préfixe « nor- », pour *NOnRibosomal* et du suffixe « -ine », suffixe fréquemment utilisé pour le nom des peptides non-ribosomiaux. Norine est accessible librement via le web (`http://bioinfo.lifl.fr/norine`) et son contenu est en anglais afin de permettre une audience internationale. Une version de Norine a été publiée dans le numéro spécial dédié aux bases de données de *Nucleic Acid Research* en 2008 [Caboche et al., 2008]. Norine contient une base de données regroupant diverses informations sur un grand nombre de peptides non-ribosomiaux. Par exemple, les organismes producteurs, les activités biologiques ou encore la structure d'un peptide sont stockés dans Norine. Les informations sont collectées et vérifiées manuellement. Norine offre également une interface web riche, permettant aux utilisateurs une interrogation facile de la base de données.

Dans ce chapitre, nous présentons dans un premier temps la base de données, c'est-à-dire les différentes informations qu'elle contient ainsi que son alimentation. Dans un second temps, nous décrivons l'interface web. Enfin, nous donnons quelques statistiques d'utilisation de Norine par la communauté scientifique.

4.1 Base de données

Norine est la première ressource entièrement dédiée aux peptides non-ribosomiaux. Elle contient actuellement plus de 1 000 peptides. Pour chacun de ces peptides, différentes informations sont disponibles. Les données contenues dans Norine sont extraites manuellement de la littérature scientifique.

4.1.1 Contenu

Les informations contenues dans Norine sont stockées dans une base de données relationnelle gérée par le SGBD PostgreSQL. Les données peuvent être divisées en deux catégories. La première partie des données concerne les peptides non-ribosomiaux, l'autre les monomères incorporés au sein de ces peptides. La figure 4.1 montre le schéma conceptuel simplifié de la base de données. La table centrale est la table contenant les informations sur les peptides. Elle est reliée aux tables

organisme, *référence*, *lien externe* et *activité*. La table *monomères* contient les données sur les monomères et n'est reliée à aucune autre table de la base.

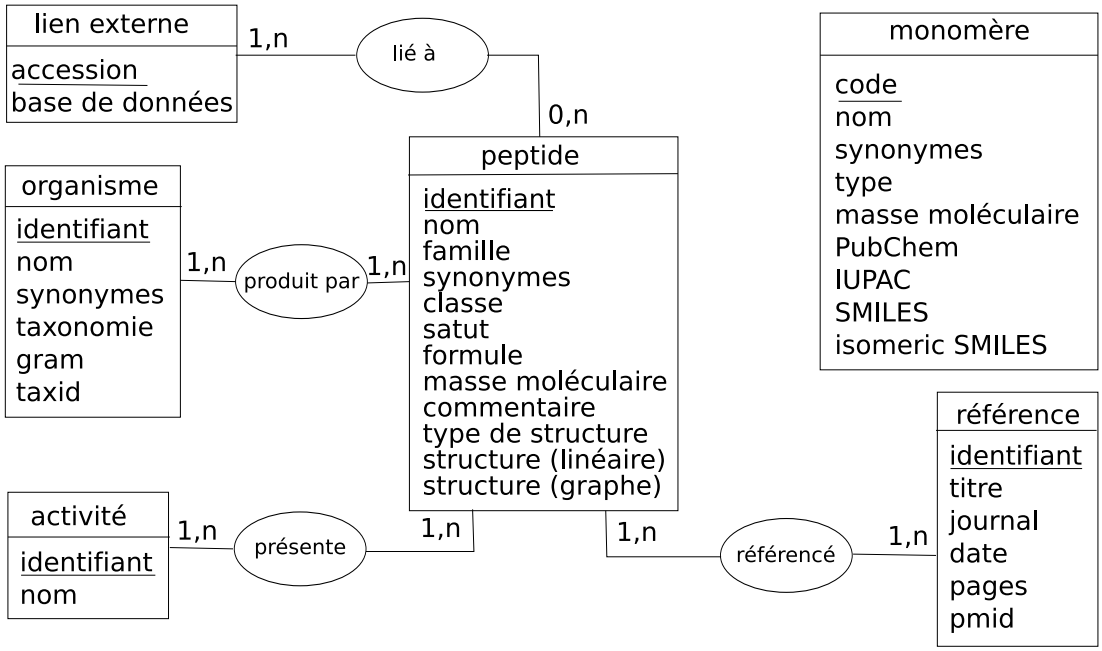


FIG. 4.1 – Schéma conceptuel simplifié de la base de données

Les peptides

NORINE contient deux catégories de peptides. La première regroupe les peptides dits « curated », c'est-à-dire ceux pour lesquels la synthèse non-ribosomiale est admise, soit parce qu'une synthétase a été identifiée soit parce que la synthèse NRPS est admise par la communauté scientifique. La seconde catégorie regroupe les peptides dits « putative ». Pour cette catégorie de peptides, aucune information sur la voie de synthèse n'est connue à ce jour. Cependant, la structure non-linéaire et la présence d'acides aminés non-protéogéniques ou modifiés, laissent à penser que ces peptides sont synthétisés par la voie non-ribosomiale. NORINE contient également tous les variants connus d'un peptide (voir section 1.2.4.0). Par exemple, 57 variants de pyoverdine sont répertoriés.

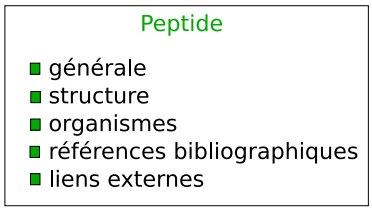


FIG. 4.2 – Schéma général de l'organisation des informations pour un peptide

Pour un peptide, différentes informations sont disponibles. Elles sont consultables via l'interface web et sont divisées en plusieurs catégories (figure 4.2).

Partie « générale »

La figure 4.3 montre une capture d'écran partielle centrée sur la partie « générale » de la fiche de l'actinomycine D.

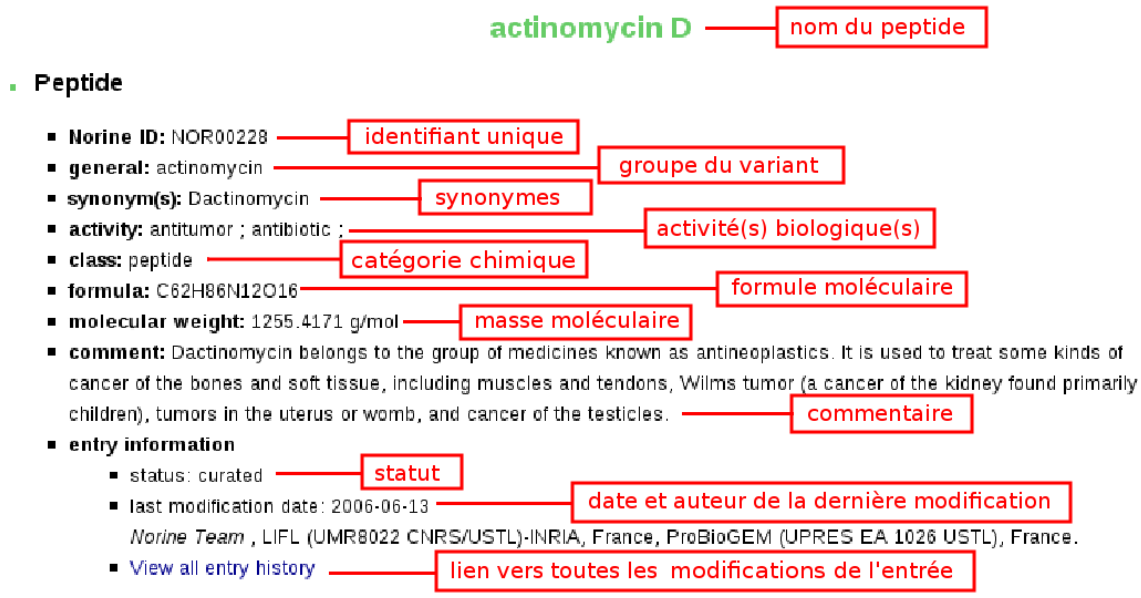


FIG. 4.3 – Fiche de l'actinomycine D : partie « générale »

Dans la partie « générale », l'identifiant (NOR00228) distingue de façon unique tous les peptides contenus dans NORINE. Ensuite, le nom général, ici « actinomycin », donne le groupe auquel appartient le variant. Les synonymes du nom pour un peptide sont également présents. Les activités biologiques connues sont également répertoriées dans cette partie. Par exemple, l'actinomycine D présente à la fois une activité antitumorale et une activité antibiotique. Les peptides contenus dans NORINE sont regroupés en quatre catégories chimiques : les peptides, les lipopeptides, les glycopeptides et les peptaibols. La catégorie à laquelle appartient le peptide est mentionnée dans cette partie, qui donne également la formule et la masse moléculaire du peptide. Des commentaires sont parfois ajoutés à la fiche pour donner des informations supplémentaires sur le peptide. Enfin, des informations concernant l'entrée sont données. Le statut du peptide, « curated » ou « putative », est donné ici. Des informations sur la date de création, de modification, ainsi que sur l'auteur de l'entrée sont également disponibles. En effet, NORINE étant une ressource publique, la plupart des entrées sont créées et mises à jour par notre équipe, mais nous permettons également aux utilisateurs de soumettre des peptides que nous intégrons nous-même dans NORINE. La soumission de peptides par les utilisateurs sera présentée ultérieurement.

Partie « structure »

La seconde partie d'une fiche regroupe les informations concernant la structure du peptide. La figure 4.4 montre une capture d'écran partielle centrée sur la partie « structure » de la fiche de l'actinomycine D.

Dans cette partie, les informations sont basées sur la structure primaire du peptide. Le type de structure, ici bi-cyclique, est donné. Six types de structures sont répertoriées : linéaire, branchée, totalement cyclique, partiellement cyclique (un cycle avec des branchements), bi-cyclique (con-

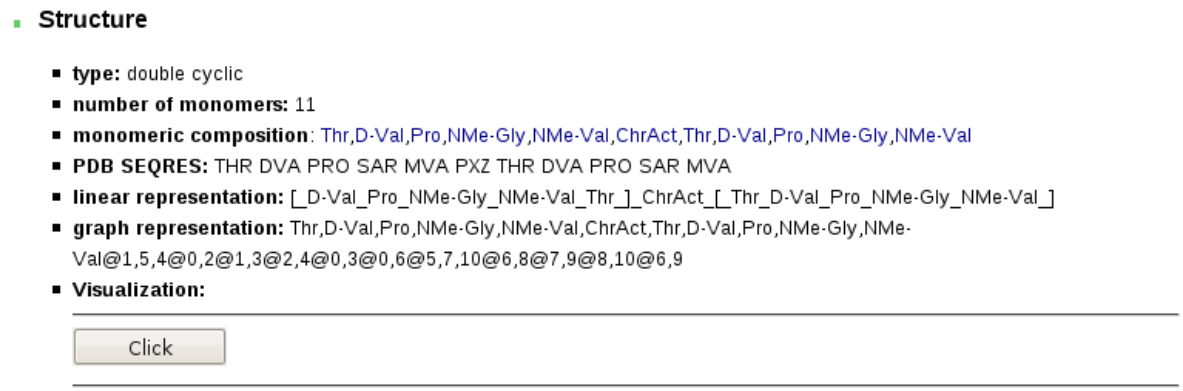


FIG. 4.4 – Fiche de l'actinomycine D : partie « structure »

tenant deux cycles) et autre. Le type « autre » regroupe des structures primaires complexes contenant des cycles chevauchants et des branchements. Le nombre de monomères, ainsi que la composition monomérique sont disponibles. Il est possible de cliquer sur chaque monomère composant le peptide afin d'obtenir des informations sur les monomères.

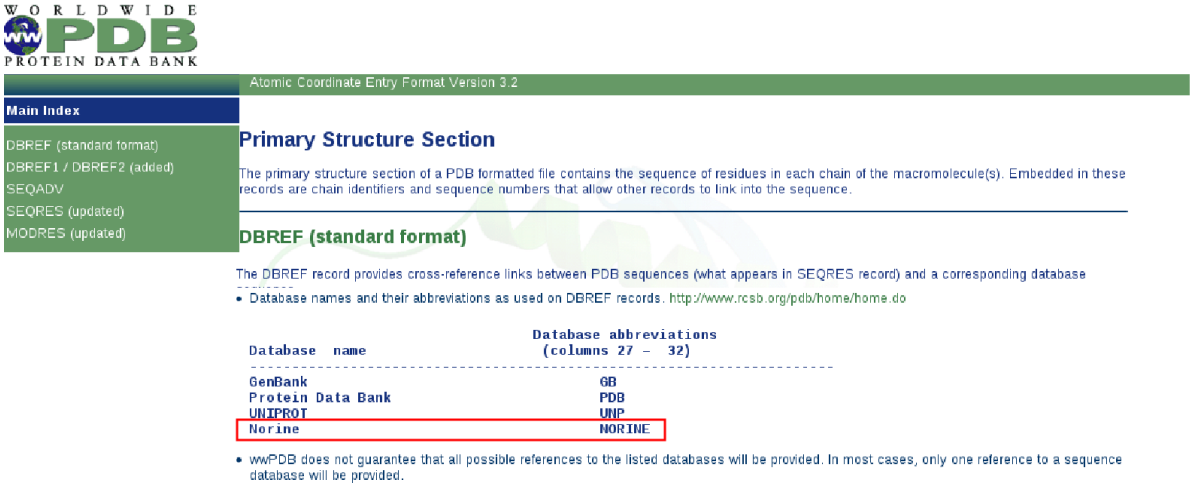


FIG. 4.5 – NORINE : base référence de wwPDB

Le champ « PDB SEQRES » a été ajouté en avril 2009 et donne la séquence monomérique d'un peptide telle qu'elle est représentée dans wwPDB. wwPDB est la ressource mondiale pour la structure 3D de macromolécules (voir section 2.1.1). Les entrées contenues dans wwPDB sont corrélées à d'autres bases de données : GenBank pour les séquences nucléiques, UniProt pour les séquences protéiques et maintenant NORINE pour les peptides non-ribosomiaux. L'équipe de wwPDB nous a contacté début 2009 pour que NORINE devienne la base de référence de wwPDB pour les peptides non-ribosomiaux, ce qui est maintenant réalisé comme le montre la figure 4.5 issue du site web de wwPDB (`http://www.wwpdb.org/documentation/format32/sect3.html`). La ligne « PDB SEQRES » a été ajoutée aux entrées ayant une structure 3D connue afin de mettre en place des liens croisés entre les deux bases de données.

La représentation linéaire (voir section 3.1.2) est donnée lorsque celle-ci est calculable. De

même, la représentation par les graphes (voir section 3.1.3), développée au cours de ce travail, est également donnée dans cette partie. Enfin, nous avons développé une applette Java permettant la visualisation de la structure monomérique en 2 dimensions. Lorsque l'utilisateur clique sur le bouton, la fenêtre de visualisation apparaît. La figure 4.6 montre la visualisation de l'actinomycine D. Cette visualisation est particulièrement utile pour les biologistes qui sont habitués à se référer à une représentation graphique.

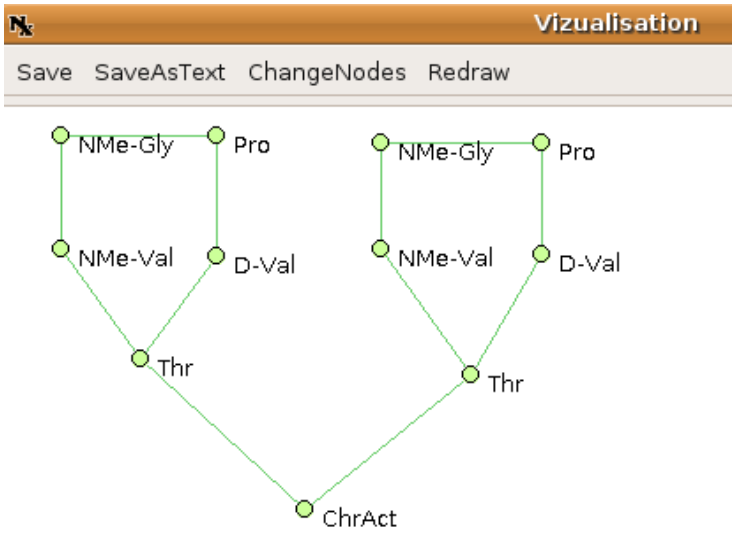


FIG. 4.6 – Visualisation de la structure de l'acinomycine D telle qu'elle est représentée dans NORINE

La visualisation de graphes est un problème complexe. Il existe pour chaque graphe une infinité de représentations possibles car chaque nœud peut être placé en n'importe quel point d'un plan. Il existe des paramètres et des contraintes permettant de réduire le nombre de possibilités, en fonction de ce qui veut être mis en évidence. Par exemple, le but peut être de minimiser le nombre de croisements entre les arêtes ou encore de mettre en évidence la symétrie au sein d'un graphe. Différentes classes d'algorithmes de visualisation de graphes ont été développées. Chacune met en évidence certaines propriétés du graphe et, par conséquent, il faut choisir une classe d'algorithmes en fonction de ce que la visualisation doit mettre en évidence. Une des classes regroupe des algorithmes dits « *Force-Directed* ». Ces algorithmes sont particulièrement bien adaptés au cas des graphes non-orientés et ont pour but de minimiser le nombre de croisements entre les arêtes. Nous avons décidé d'opter pour cette classe d'algorithmes. Les algorithmes de placement « *Force-Directed* » partagent une idée centrale qui est d'introduire un modèle physique pour le graphe. Ces algorithmes se déroulent en deux parties. La première consiste à établir un modèle de force. Le graphe est traduit en un modèle physique comprenant des ressorts, des masses et divers champs et forces appliqués au système. Dans la seconde partie, le but est de déterminer un état stable du système parmi une infinité de compressions et d'extensions possibles. Pour ce faire, le système subit un processus de simulation dans lequel le but est de trouver son état stable minimisant l'énergie. Beaucoup de variations de cet algorithme ont été publiées. Une des plus anciennes est l'algorithme de Eades publié en 1984 [Eades, 1984]. Dans ce dernier, les nœuds sont remplacés par des anneaux et les arêtes par des ressorts pour former un système mécanique. A partir d'un état initial, les forces exercées par les ressorts sur les anneaux

amènent le système vers un état stable. Dans ce modèle, les forces répulsives sont calculées entre tous les nœuds du système, alors que les forces attractives le sont uniquement entre les nœuds voisins, ce qui signifie que, selon Eades, l'important est qu'un nœud soit proche de son voisin immédiat. Cette contrainte sur les forces attractives réduit également la complexité en temps de l'algorithme. Plusieurs autres variations de l'algorithme de Eades ont ensuite été développées. En 1991, Fruchterman et Reingold [Fruchterman and Reingold, 1991] ont proposé une variante de l'algorithme de Eades. Elle permet de dessiner des graphes non-orientés en respectant certains critères tels que la minimisation du nombre de croisements entre les arêtes, la répartition uniforme des nœuds au sein de la fenêtre et également la longueur uniforme des arêtes. Ces différents critères correspondent à ceux recherchés pour la visualisation de peptides non-ribosomiaux. En effet, nous voulons obtenir des graphes proches de formes géométriques avec un minimum de croisements entre les liaisons. L'algorithme de Fruchterman-Reingold semble être bien adapté à notre cas et c'est ce dernier que nous avons choisi d'implémenter pour la visualisation de peptides NRPS. Cet algorithme a été implémenté dans une applette Java permettant la visualisation des peptides non-ribosomiaux. A chaque ouverture de l'applette, les positions des différents nœuds du graphe sont calculées grâce à l'algorithme de Fruchterman-Reingold ce qui implique que la visualisation du graphe est différente. En effet, l'algorithme est non-déterministe à cause du placement aléatoire des nœuds au début de l'algorithme. L'utilisateur peut relancer l'algorithme et obtenir une nouvelle visualisation en cliquant sur le bouton « redraw » si la visualisation actuelle ne lui convient pas. Il peut également déplacer les différents nœuds pour obtenir sa propre visualisation. Enfin, l'applette de visualisation permet d'enregistrer un peptide sous forme d'une image (jpg) ou en format texte afin de le charger dans l'éditeur présenté ultérieurement.

Partie « organismes producteurs »

Une partie de la fiche rassemble ensuite des données sur les organismes à partir desquels le peptide a été isolé. La figure 4.7 montre une capture d'écran partielle de la fiche de l'actinomycine D, centrée sur la partie « organismes producteurs ».

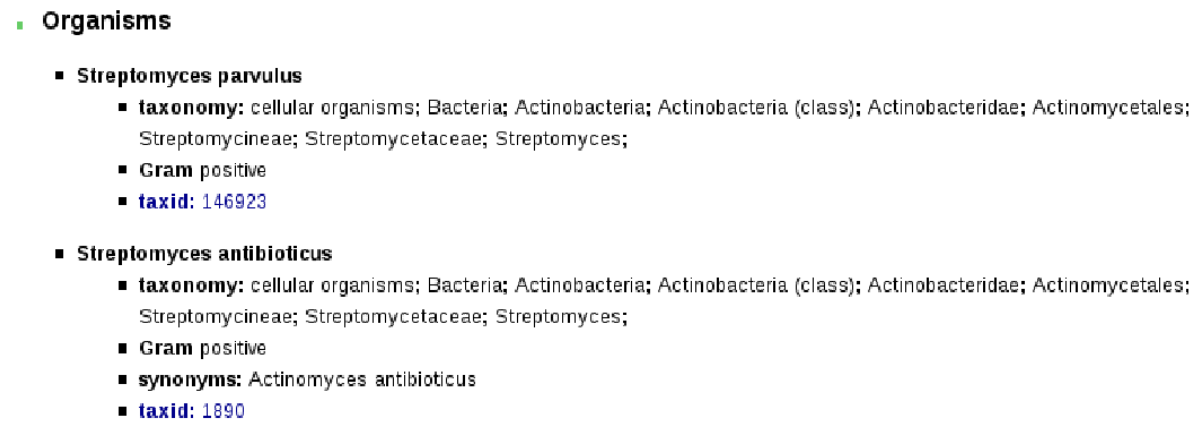


FIG. 4.7 – Fiche de l'actinomycine D : partie « organismes producteurs »

Pour chaque organisme, différentes informations sont disponibles : sa taxonomie complète, le Gram dans le cas des bactéries, les synonymes connus et un lien vers la banque de données « Taxonomy » du NCBI ([Sayers et al., 2009], `http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root`) s'il existe.

Partie « références bibliographiques »

La partie suivante d'une fiche est dédiée aux références bibliographiques. La figure 4.8 montre une capture d'écran partielle de la fiche de l'actinomycine D, centrée sur la partie « références ».

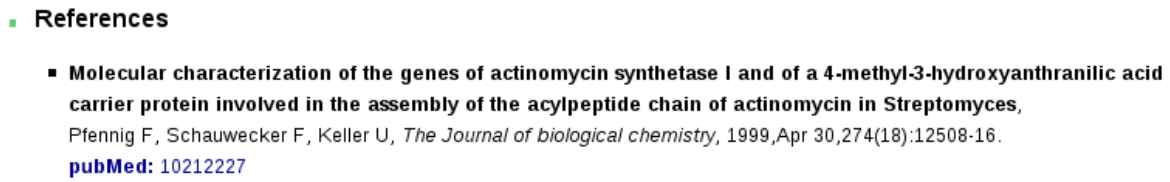


FIG. 4.8 – Fiche de l'actinomycine D : partie « références »

Cette partie contient les références bibliographiques des articles majeurs utilisés pour obtenir les informations sur le peptide. Un lien vers la banque de données bibliographique PubMed du NCBI ([Sayers et al., 2009], `http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed`), est disponible.

Partie « liens »

Enfin, la dernière partie d'une fiche regroupe des liens vers d'autres banques de données. La figure 4.9 montre une capture d'écran partielle de la fiche de l'actinomycine D, centrée sur la partie « liens ».

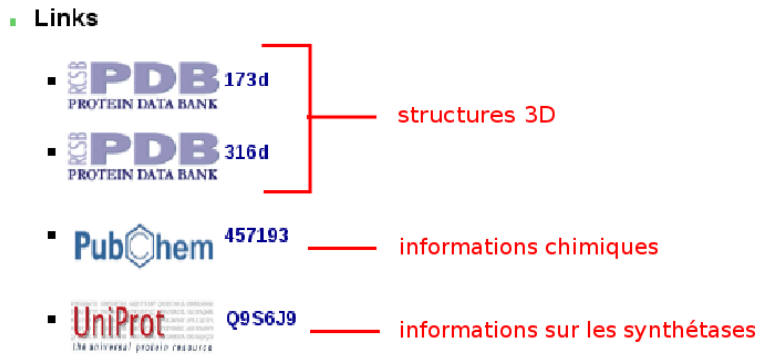


FIG. 4.9 – Fiche de l'actinomycine D : partie « liens »

Cette partie peut contenir, pour l'instant, trois types de liens en fonction de la présence du peptide dans les différentes banques de données concernées. Tout d'abord, un lien vers wwPDB permet d'obtenir des informations sur la structure 3D du peptide. Ensuite, un lien vers « PubChem Compound » du NCBI permet l'obtention d'informations chimiques sur le peptide. Enfin, un lien vers UniProt permet d'obtenir des informations sur les synthétases intervenant dans la synthèse du peptide, notamment la séquence protéique de ces dernières. Ces différentes bases de données ont été présentées dans la section 2.1.

Les monomères

NORINE contient également des informations sur les différents monomères rencontrés au sein des peptides non-ribosomiaux. Actuellement, plus de 500 monomères différents sont répertoriés

parmi lesquels des acides aminés non-protéogéniques, des acides aminés protéogéniques modifiés, des acides gras, des polykétides ou encore des sucres. Pour chacun des monomères, diverses informations sont disponibles. La figure 4.10 montre la capture d'écran de la fiche de l'ornithine.

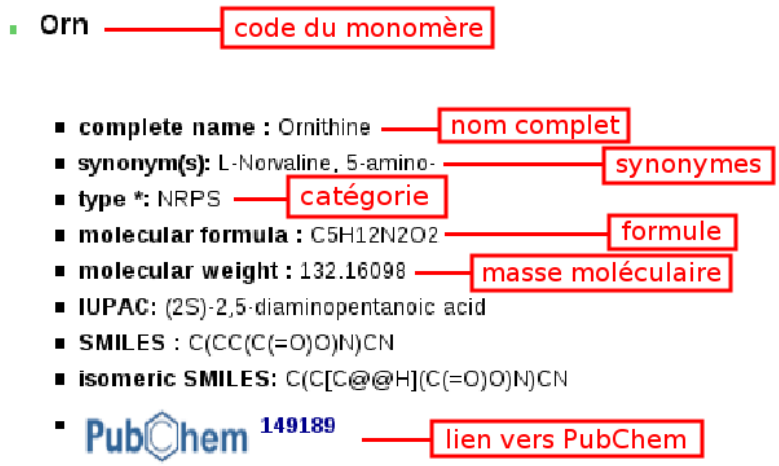


FIG. 4.10 – Fiche d'un monomère : exemple de l'ornithine

Chaque fiche contient le code du monomère utilisé dans NORINE (voir section 3.1.1) ainsi que son nom complet et la nomenclature IUPAC correspondante. Une fiche contient également les SMILES, représentation linéaire de la formule chimique couramment utilisée en chimie, ainsi que la formule et la masse moléculaire. Un lien vers l'entrée de « PubChem Compound » est également donné, si cette dernière existe. Enfin, la catégorie du monomère est donnée, ici NRPS, c'est-à-dire la voie de synthèse par laquelle le monomère est incorporé ou synthétisé. Dans NORINE, cinq voies de synthèse sont répertoriées :

- **FAS** (*fatty acid Synthesis*) qui regroupe les monomères issus de la synthèse des acides gras
- **CS** (*Carbohydrate Synthesis*) qui regroupe les monomères issus de la synthèse des sucres
- **NRPS** (*NonRibosomal Peptide Synthesis*) qui regroupe les monomères issus de la synthèse NRPS
- **PKS** (*PolyKetide Synthesis*) qui regroupe les monomères issus de la synthèse PKS
- **unknown** qui regroupe les monomères ne faisant pas partie des catégories précédentes

Chaque monomère contenu dans NORINE est unique. Etant donné le nombre important de noms et de nomenclatures utilisés pour un même monomère, nous avons vérifié l'unicité de chaque monomère en comparant les différentes structures entre elles afin de ne pas obtenir de redondances, c'est-à-dire une même structure présentant deux noms différents.

4.1.2 Alimentation de la base de données

Toutes les informations contenues dans NORINE sont collectées manuellement au sein de la littérature scientifique, de livres ou encore de comptes-rendus de congrès. 469 articles sont actuellement répertoriés dans NORINE. Les différentes informations sur les peptides sont obtenues à partir de plusieurs articles ce qui rend impossible l'automatisation de l'extraction des données. De plus, dans certains cas, les articles sont contradictoires sur une information donnée, par exemple des structures différentes pour une même molécule, et seule l'expérience humaine permet de trancher.

Les noms des peptides non-ribosomiaux sont obtenus dans des revues dédiées à la synthèse non-ribosomiale ou à partir de recherche dans les bases de données bibliographiques à l'aide de mots-clés (par exemple « nonribosomal peptide » ou NRPS). D'autres articles, sur les peptides bioactifs, sont également étudiés et permettent l'identification de peptides potentiellement synthétisés par la voie non-ribosomiale (peptides « putative » dans NORINE).

Pour chaque peptide, la première étape est la recherche de données au sein de la littérature scientifique. Pour ce faire, le nom du peptide est recherché dans un premier temps au sein de PubMed au NCBI dans le but d'obtenir tous les articles concernant le peptide d'intérêt. Ensuite, ces articles sont étudiés dans le but de rechercher les différentes informations dont nous avons besoin. Dans la plupart des cas, un grand nombre d'articles doivent être traités pour obtenir toutes les informations nécessaires. En effet, généralement un article est axé sur un point précis, comme par exemple les activités biologiques ou l'identification de la structure. Les organismes producteurs sont recherchés dans la banque « Taxonomy » de NCBI afin de récupérer la taxonomie pour chacun d'entre eux.

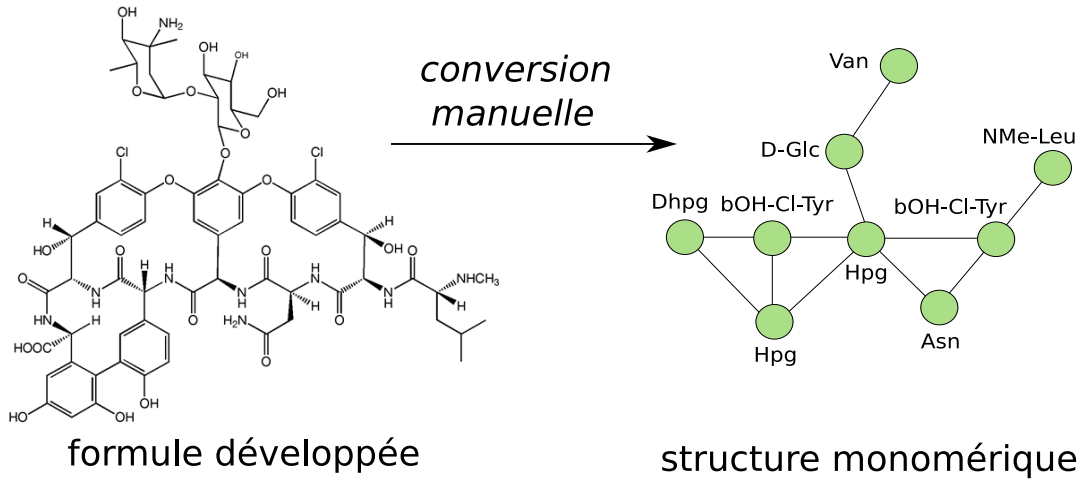


FIG. 4.11 – Obtention de la structure monomérique à partir de la formule chimique développée : exemple de la vancomycine

L'étape la plus fastidieuse est l'obtention de la structure monomérique du peptide. En effet, dans la littérature scientifique, la structure d'un peptide peut être donnée sous divers formats, et même quand celle-ci est donnée sous forme de structure monomérique, le nom des différents monomères varie d'un article à l'autre. Le format le plus fréquent est la formule chimique développée. Dans ce cas, nous devons convertir manuellement cette dernière en structure monomérique, en identifiant les monomères au sein de la formule développée. La figure 4.11 montre l'exemple de conversion de la formule chimique de la vancomycine en structure monomérique.

Ensuite, les différentes bases de données (UniProt, wwPDB et PubChem Compound) sont interrogées afin de pouvoir ajouter les liens vers ces bases de données.

L'alimentation de la base de données est une étape longue et souvent difficile à cause de certaines incohérences rencontrées au sein de la littérature. Par exemple, les articles concernant la lichenysine synthétisée par *Bacillus licheniformis* présentent des incohérences. La lichenysine A a été identifiée en 1995 [Yakimov et al., 1995]. Dans cet article, la structure contient l'acide aminé Glu en position 1 et Asn en position 5. Dans un article de 1999 sur l'étude de l'opéron responsable de la synthèse de la lichenysine [Konz et al., 1999], les auteurs mettent en évidence un

nouveau variant, la lichenysine D qui contient Gln en position 1 et Asp en position 5. Les gènes identifiés dans cet article sont nommés *lic*. Un autre article, publié à la même période, étudie également les synthétases intervenant dans la synthèse de la lichenysine A [Yakimov et al., 1998]. Dans cet article, les gènes sont nommés *lchA* et la structure de la lichenysine A est rectifiée avec Gln en position 1 et Asp en position 5. Cette structure est confirmée dans un autre article [Yakimov et al., 1999]. Finalement, les lichenysines A et D sont identiques et deux noms sont utilisés pour nommer une même molécule. De la même manière, les opérons *lic* et *lchA* contiennent des gènes homologues codant des synthétases produisant le même peptide. L'existence de plusieurs noms pour une même molécule entraîne ensuite une confusion dans certains articles et les erreurs sont répercutées d'article en article. Notre outil permet d'éviter ces erreurs, ce qui est presque impossible tant que toutes les informations ne sont pas disponibles dans un même endroit.

L'entrée de peptides dans la base de données est une étape longue et fastidieuse, mais aboutit à des informations validées de grande qualité.

4.2 Interface web

Une interface Web a été développée, afin de permettre aux l'utilisateurs de rechercher rapidement et facilement des informations au sein de la base de données, puis de visualiser les résultats (figure 4.12).

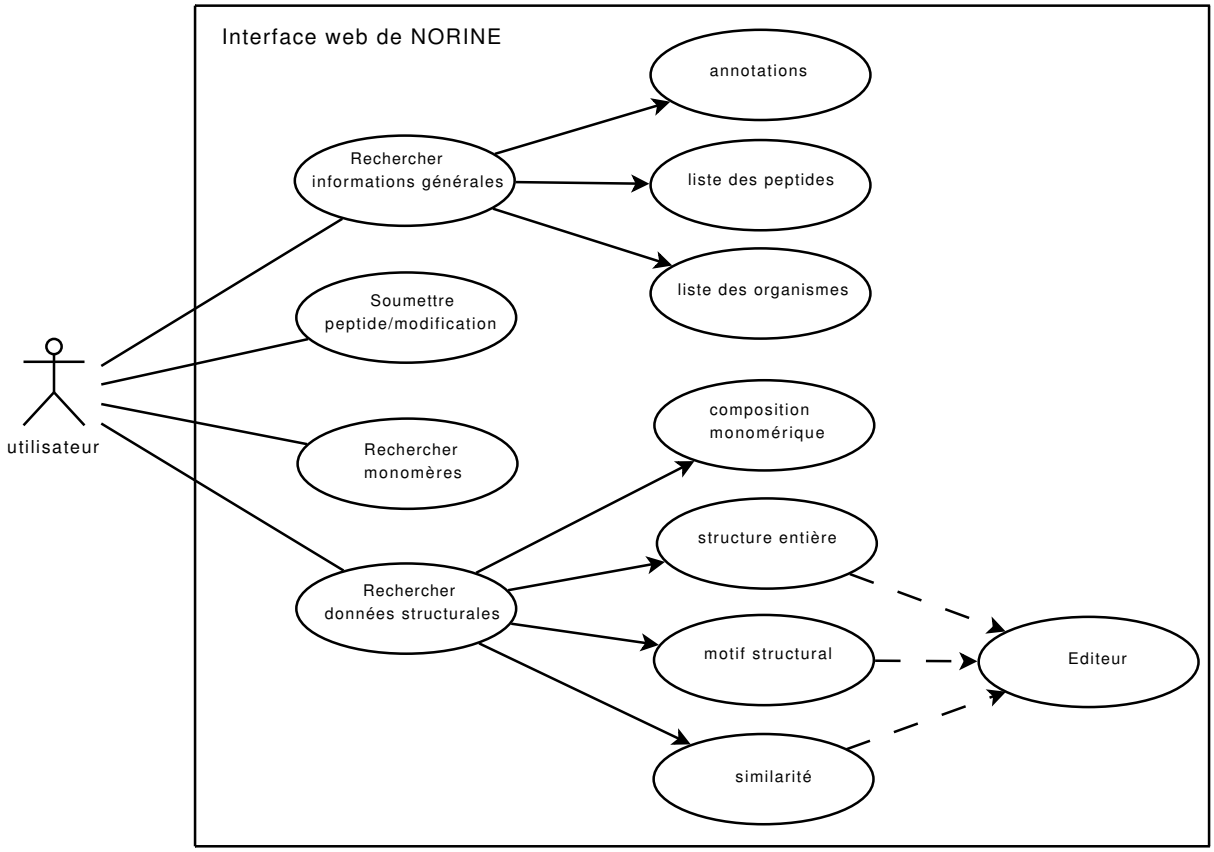


FIG. 4.12 – Diagramme UML de cas d'utilisation de l'interface Web de NORINE

L'interrogation de la base de données peut se faire via les annotations, c'est-à-dire selon les différents champs contenus dans la base, mais aussi, via les informations sur la structure des peptides. Cette interface est disponible à l'adresse suivante : `http://bioinfo.lifl.fr/norine` sur le serveur de l'équipe SEQUOIA, hébergé au LIFL.

4.2.1 Technologies utilisées

Le protocole utilisé pour les communications sur le web est le protocole HTTP (*HyperText Transfer Protocol*) qui permet un transfert de fichiers entre le client, c'est-à-dire l'utilisateur, et le serveur, qui héberge la base de données et les programmes associés. Le serveur web est un logiciel permettant à des clients d'accéder à des pages Web. Il interprète les requêtes HTTP et fournit une réponse. Apache est le serveur web le plus répandu sur Internet et utilisé par l'équipe SEQUOIA. Un serveur d'application est un serveur qui assure la médiation entre un logiciel client et un serveur de base de données. Pour certaines applications complexes, l'architecture client-serveur à deux niveaux (client et serveur) gagne à être subdivisée en trois niveaux, elle est alors appelée architecture trois tiers (figure 4.13).

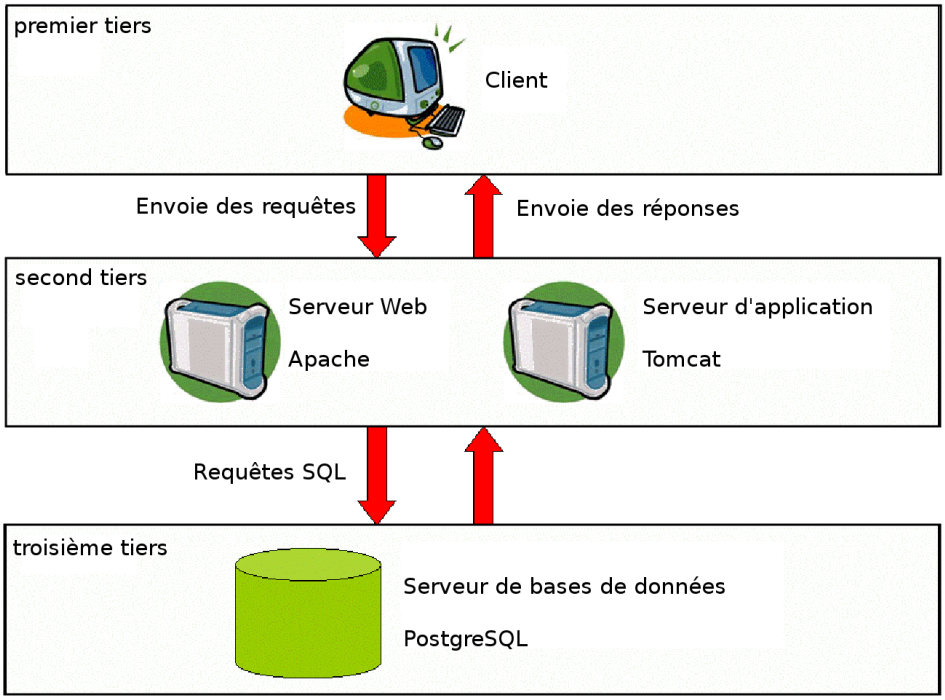


FIG. 4.13 – Architecture trois-tiers

Le serveur d'application est l'intermédiaire entre le logiciel client qui gère uniquement l'interface utilisateur et le logiciel serveur de base de données. Tomcat est un serveur d'application interrogeable via HTTP. Il est combiné au serveur web Apache pour former la couche centrale de l'architecture trois tiers. Enfin, la dernière couche est le serveur de base de données. C'est un logiciel permettant de gérer plusieurs bases de données réparties sur une ou plusieurs machines. Dans notre cas, c'est PostgreSQL qui est utilisé. PostgreSQL est un système de gestion de bases de données relationnelles (SGBDR) qui exécute rapidement des requêtes imbriquées et complexes. Pour agir sur les données stockées, un langage d'interaction, SQL (Structured query

language), est utilisé. C'est un langage de requêtes standard et normalisé, destiné à interroger ou manipuler une base de données relationnelles.

HTML (*Hypertext Markup Language*) est le langage universel utilisé pour communiquer sur le Web, il est donc par conséquent, incontournable lors de la mise en place d'une interface. HTML est un langage de balisage hypertexte. Il permet de décrire un document en utilisant des balises qui sont interprétées par le logiciel de navigation, afin d'assurer une bonne visualisation du document. Cependant, HTML permet uniquement la création de pages statiques. Or, nous avons besoin de générer des pages différentes selon la requête de l'utilisateur. De plus, nous avons également besoin de formulaires d'interrogation de la base et de pouvoir récupérer les données du formulaire entrées par l'utilisateur. C'est pourquoi nous avons décidé d'utiliser la technologie JSP (*Java Server Pages*)/servlets qui est basée sur Java et permet aux développeurs de générer dynamiquement du code HTML.

La technologie JSP/servlets facilite la réalisation de sites web combinant contenu statique et contenu dynamique. Elle désigne des pages comprenant des programmes Java qui génèrent du code HTML du côté du serveur. Le résultat de leur compilation réside en mémoire du serveur d'application. Les JSP/servlets se comparent aux scripts CGI, avec la distinction fondamentale suivante : les JSP/servlets sont des programmes compilés et chargés au préalable (alors que les scripts CGI sont rechargés systématiquement), ce qui réduit le temps entre une requête et la réponse du serveur HTTP. Les JSP/servlets produisent des pages web qui sont transmises au navigateur d'où provient la requête. Une librairie, JSTL (*Java server page Standard Tag Library*), est associée à la technologie JSP/servlets et permet l'utilisation de tags qui proposent des fonctionnalités souvent rencontrées dans les JSP/servlets. L'ensemble des tags de cette librairie permet d'insérer facilement des résultats de requêtes SQL au sein des pages.

4.2.2 Recherche basique

Dans cette partie, l'utilisateur peut interroger la base de données selon les différents champs disponibles. La figure 4.14 montre une capture d'écran du formulaire correspondant. Chaque recherche à partir de ce formulaire génère des requêtes SQL imbriquées et parfois complexes.

La base peut être interrogée en fonction de l'identifiant, du nom ou encore du statut (« curated » ou « putative ») du peptide recherché. L'utilisateur peut également rechercher les peptides présentant une activité donnée, appartenant à une catégorie précise de peptides ou présentant un type de structure spécifique. Par exemple, l'utilisateur peut obtenir facilement la liste des peptides cycliques. Il est possible de rechercher des peptides dont la masse moléculaire est comprise dans un intervalle donné par l'utilisateur. Il peut également rechercher des peptides contenant un monomère donné ou les peptides contenant un monomère ou ses dérivés. Par exemple, en recherchant tous les peptides contenant le monomère « Ala », 272 peptides sont renvoyés, ceux qui contiennent exactement le monomère « Ala ». En revanche, en recherchant les peptides contenant « Ala » ou un de ses dérivés, 461 peptides sont obtenus (dont les 272 de la requête précédente) contenant « Ala » ou un de ses dérivés, comme par exemple « D-Ala » ou « NMe-Ala ».

L'utilisateur a également la possibilité de rechercher des peptides en fonction de références bibliographiques. Les différents champs d'une référence bibliographique, c'est-à-dire les auteurs, le journal, l'année de publication, le titre de l'article ou l'identifiant PubMed, peuvent être utilisés pour la recherche.

Enfin, l'utilisateur peut rechercher des peptides en fonction de l'organisme producteur. Cette recherche se fait sur le nom de l'organisme ou sur sa taxonomie. Le Gram peut être précisé pour la recherche de peptides produits par des bactéries. Par exemple, il est possible de rechercher tous

FIG. 4.14 – Formulaire de recherche basique

les peptides produits par le genre *Bacillus*. Cette requête donne 127 résultats. Il est également possible de descendre au niveau de l'espèce. En tapant *Bacillus subtilis* les 76 peptides produits par cette espèce sont obtenus.

Lorsque plusieurs champs sont complétés, les peptides retournés sont ceux qui correspondent à l'ensemble des champs complétés par l'utilisateur. L'interrogation de plusieurs champs permet

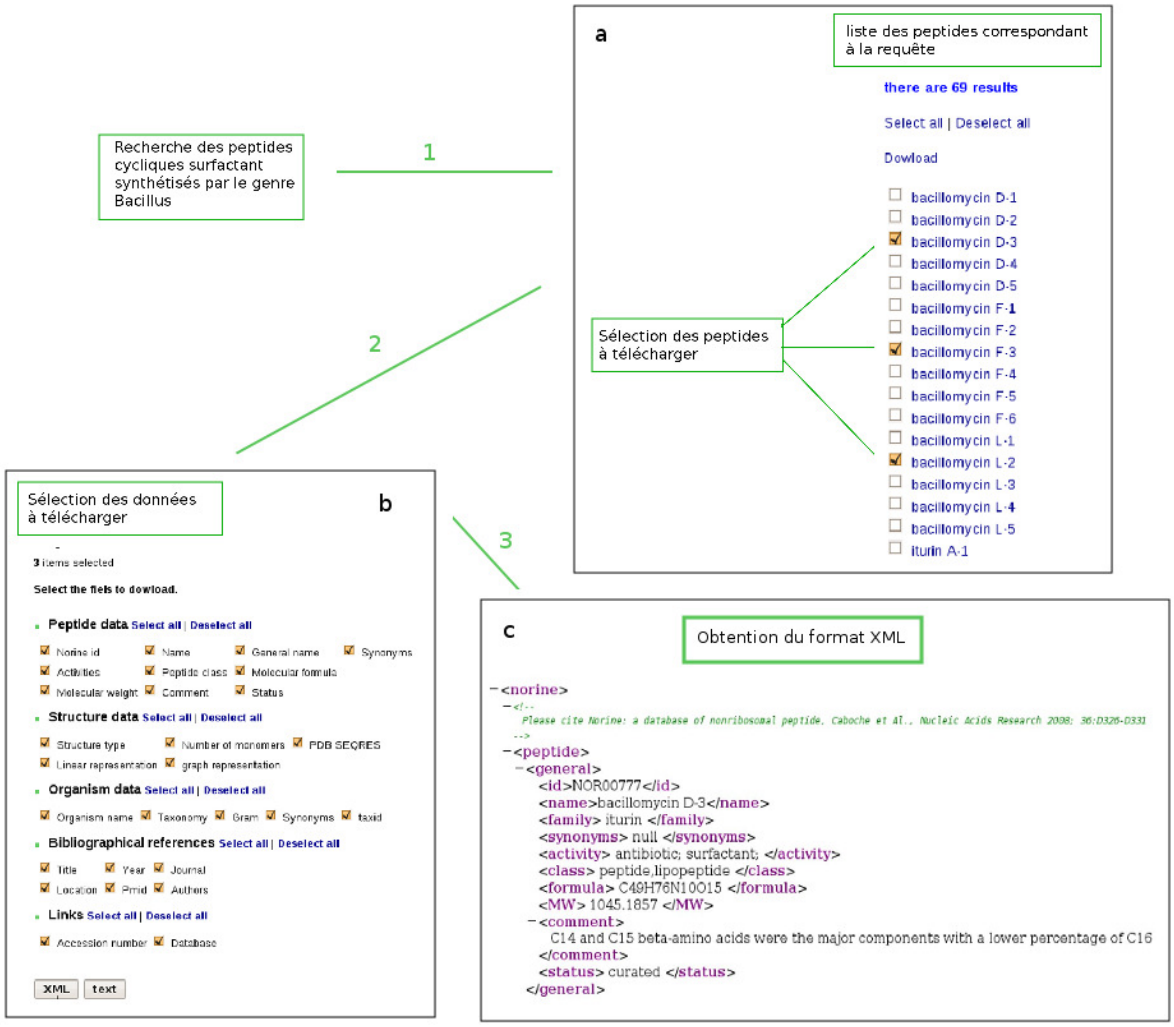


FIG. 4.15 – Téléchargement de données sur les peptides. a) liste des peptides correspondant aux critères de recherche b) formulaire de téléchargement c) format XML des données demandées par l'utilisateur

de réaliser des requêtes multi-critères complexes. Par exemple, l'utilisateur peut obtenir la liste des peptides cycliques étant des surfactants et produits par le genre *Bacillus*. Cette requête donne la liste des 69 peptides de NORINE correspondants à ces critères. La liste de résultats montre les peptides « curated » en bleu et les peptides « putative » en vert. L'utilisateur peut ensuite cliquer sur les noms des différents peptides de la liste pour obtenir la fiche descriptive correspondante. L'utilisateur peut télécharger les informations qui l'intéressent sur les peptides qu'il sélectionne dans la liste des résultats. La figure 4.15 montre les captures d'écran des différentes étapes pour le téléchargement de données. Les données peuvent être téléchargées au format XML (*Extensible Markup Language*) ou au format texte. Le format XML est un format de stockage des données de type texte structuré en champs arborescents. C'est un format très utilisé par les banques de données.

Le formulaire de recherche basique permet également l'affichage de la liste de tous les peptides contenus dans NORINE en cliquant simplement sur le bouton correspondant qui se situe en bas

de la page. Le bouton « organisms » affiche la liste de tous les organismes présents dans la base de données.

4.2.3 Recherche en fonction des données structurales

Dans cette section, la recherche est basée sur les caractéristiques structurales des peptides. Quatre types de recherches peuvent être effectués et sont présentés dans cette section.

Recherche basée sur la composition monomérique

La méthode présentée dans la section 3.2 est utilisée pour la recherche basée sur la composition en monomères. Dans ce type de recherche, l'utilisateur entre une liste de monomères, séparés par des virgules, ainsi que le nombre d'erreurs autorisées. La figure 4.16 montre une capture d'écran de l'interface correspondante.

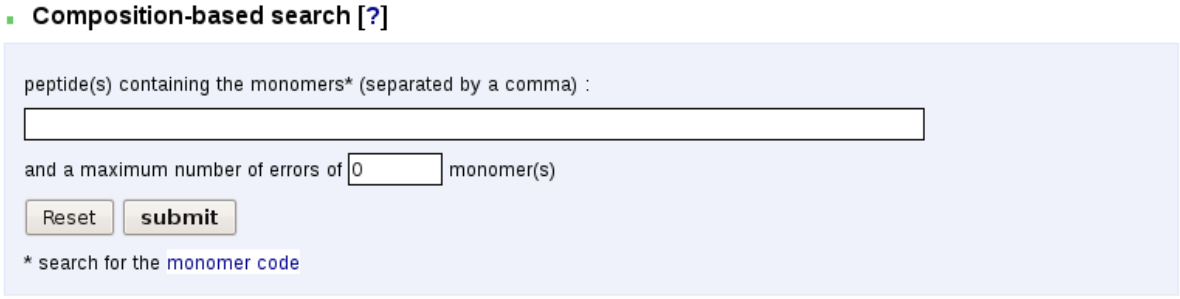


FIG. 4.16 – Formulaire de recherche en fonction d'une composition en monomères

Si le peptide recherché doit contenir plusieurs monomères identiques, il faut répéter le monomère dans la liste le nombre de fois voulu. Par exemple, pour rechercher les peptides contenant au moins deux prolines il faut entrer « Pro,Pro » dans le champ adéquat. Le nombre d'erreurs autorisées correspond au nombre maximum de monomères de la liste non présents dans les peptides retournés. En d'autres termes, une erreur est l'absence d'un monomère de la liste recherchée au sein du peptide testé. Par exemple, en recherchant les peptides contenant « Pro,Pro,Leu,Ala » sans aucune erreur autorisée, nous obtenons 33 peptides contenant ces quatre monomères. En recherchant cette même liste de monomères mais en autorisant une erreur, nous obtenons 138 résultats. Parmi ces 138 résultats, les 33 résultats précédents contenant quatre monomères (0 erreur) sont présents et 105 peptides contenant au moins trois des quatre monomères donnés en entrée, ce qui correspond à exactement une erreur. Parmi les 105 peptides, les trois monomères présents ne sont pas forcément les mêmes. Par exemple, parmi les résultats, la paracelsine D contient Ala, Pro et Leu alors que l'harzianine HC III contient Pro, Pro et Leu.

La recherche par composition permet d'identifier les peptides contenant une certaine liste de monomères, mais dans ce cas, aucune information sur la structure, c'est-à-dire la position des différents monomères les uns par rapport aux autres, n'est prise en compte. Cette fonction s'avère très utile dans le cas où la prédiction de la spécificité des domaines d'adénylation fournit une liste de monomères, sans information sur la structure du peptide produit, comme c'est le cas lors d'une biosynthèse non-linéaire.

Recherche en fonction de la structure

La méthode présentée dans la section 3.3.5 permet de rechercher de façon exacte une structure dans Norine. La figure 4.17 montre une capture d'écran de l'interface correspondante.

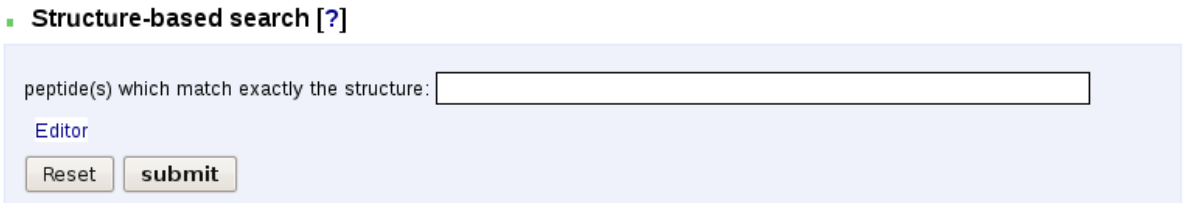


Fig. 4.17 – Formulaire de recherche d'une structure exacte

Lorsque des scientifiques mettent en évidence un nouveau peptide, ils ont besoin de savoir si ce composé n'a pas déjà été décrit. Pour ce faire, il doit être possible de rechercher un peptide à partir de sa structure. L'utilisateur peut rechercher une structure en l'encodant grâce à la représentation linéaire ou à la représentation par les graphes. Cependant, dans certains cas, il peut être difficile de convertir la structure d'un peptide dans l'une ou l'autre des représentations. Nous avons développé un éditeur qui calcule la représentation par un graphe d'un peptide que l'utilisateur dessine. Une capture d'écran de l'éditeur est montrée dans la figure 4.18.

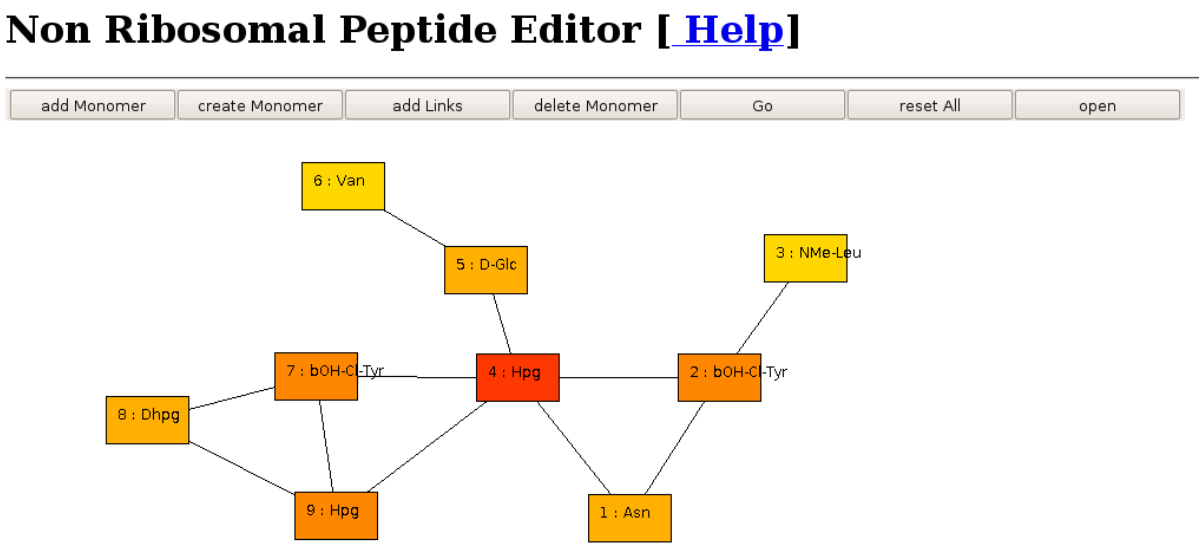


Fig. 4.18 – Capture d'écran de l'éditeur

L'éditeur est une applette Java. Il permet à l'utilisateur de dessiner, grâce à une interface conviviale, la structure du peptide qu'il souhaite rechercher dans Norine. Pour ce faire, l'utilisateur peut créer deux types d'objets à l'aide de l'interface graphique : des monomères, c'est-à-dire des nœuds, et des laisons entre les différents monomères, c'est-à-dire des arêtes. Chaque nœud ajouté par l'utilisateur est associé à un indendifiant unique et stocké dans un tableau avec sa position dans le panneau (coordonnées x et y). Chaque arête dessinée permet de mettre à jour la liste d'adjacence des nœuds impliqués par cette dernière. Pour un nœud, nous stockons donc un identifiant unique, le nom du monomère, sa position dans le panneau et la liste d'adjacence

qui permet de connaître les nœuds auxquels il est lié. La couleur du nœud change en fonction du dégré de celui-ci (obtenu par le nombre d'éléments dans la liste d'adjacence) : la couleur évolue du jaune, pour un nœud sans arête, vers le rouge, pour les nœuds avec un degré important. Lorsqu'un nœud est supprimé par l'utilisateur, les arêtes impliquant ce nœud sont automatiquement supprimées. Lorsque l'utilisateur a terminé son peptide, la liste des nœuds et les listes d'adjacence correspondantes sont utilisées pour convertir le graphe dessiné selon la représentation introduite précédemment. Cette applette a été développée avec l'interface graphique Java qui permet à l'utilisateur d'intéragir simplement avec le programme grâce à des boutons.

L'utilisateur peut rechercher un monomère donné en cliquant sur le bouton « add monomer » ou créer lui-même un monomère en cliquant sur le bouton « create monomer ». Une fois les monomères placés, l'utilisateur sélectionne le bouton « add links » et peut ainsi créer des liaisons entre les différents monomères. Un monomère peut être supprimé à l'aide du bouton « delete monomer » ou l'ensemble du peptide créé peut être effacé avec le bouton « reset all ». Une fois que l'utilisateur a terminé la structure du peptide, il clique sur « Go ». La fenêtre de recherche en fonction des données structurales s'ouvre et le champ correspondant (ici celui de recherche d'une structure exacte) est complété avec la représentation sous forme de graphe de la structure du peptide dessiné dans l'éditeur.

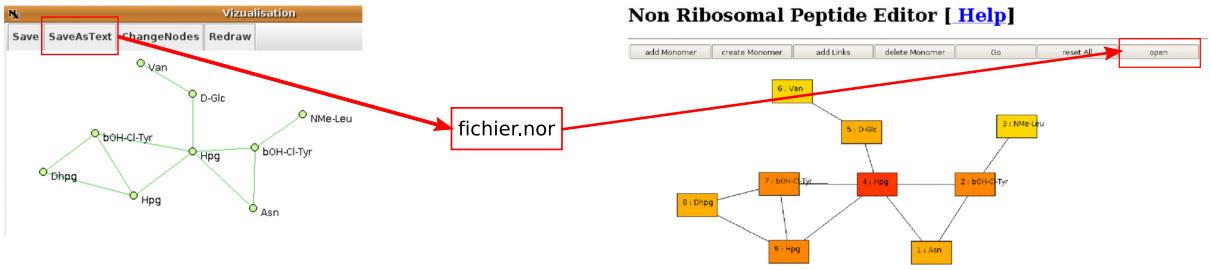


FIG. 4.19 – Interaction entre l'éditeur et le visualisateur : l'éditeur permet d'enregistrer un peptide sous la forme d'un fichier avec l'extension .nor qui peut être chargé dans l'éditeur.

L'éditeur offre également la possibilité de charger un fichier contenant la structure d'un peptide avec l'extension « .nor ». Ce fichier peut être obtenu à partir de l'applette de visualisation en cliquant sur le bouton « save as text » (figure 4.19). Cette fonctionnalité permet d'obtenir rapidement le dessin d'un peptide qui peut être modifié ensuite par l'utilisateur. Par exemple, si l'utilisateur recherche un variant d'un peptide présent dans NORINE, il charge le fichier correspondant obtenu avec l'éditeur et peut apporter les modifications souhaitées sans avoir à redissiner entièrement le peptide.

Recherche de motifs structuraux

La méthode présentée dans la section 3.3 est utilisée pour rechercher des motifs structuraux au sein des peptides contenus dans NORINE. Contrairement aux structures des peptides, les motifs peuvent contenir des caractères spéciaux. Le caractère joker « X » correspond à n'importe quel monomère à une position fixée. Le caractère « * » permet de définir un monomère et ses dérivés. Enfin, le caractère « / » permet de donner une liste de monomères possibles à une position fixée. Pour rechercher un motif, l'utilisateur peut donner la structure sous forme de graphe ou de représentation linéaire, mais peut aussi utiliser l'éditeur vu précédemment. Ensuite, l'utilisateur

a le choix entre rechercher parmi les peptides de la base le motif entier ou une sous-structure de taille donnée du motif. En effet, il peut parfois être intéressant de ne pas rechercher le motif entier mais un motif plus petit, comme dans l'exemple de la syringomycine vu dans la section 3.3. De plus, cette méthode ne nécessite aucun *a priori* sur les monomères du motif qui ne seront pas considérés. La figure 4.20 montre une capture d'écran de l'interface correspondante.

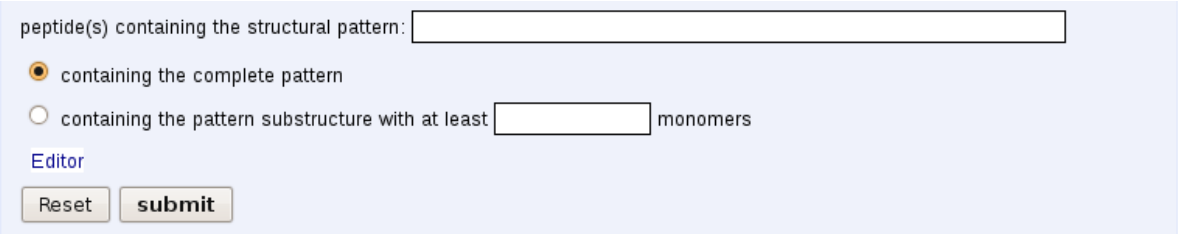


FIG. 4.20 – Formulaire de recherche d'un motif

La recherche d'un motif donné peut permettre la mise en exergue de certaines caractéristiques. Par exemple, la recherche du motif complet cyclique composé de 7 « X » et d'un acide gras (symbolisé par « * : »), c'est-à-dire en représentation linéaire [_* :_X_X_X_X_X_X_X_], retourne 71 peptides de la famille des iturines (iturines, bacillomycines et mycosubtilines), surfactines et lichenysines. La recherche de ce motif met en évidence des peptides qui partagent des caractéristiques structurales.

Comme nous l'avons vu dans la section 3.3, il existe une relation forte entre la structure et la fonction d'un peptide. Par exemple, rechercher le motif *Orn_X_Ser_*Orn dérivé de l'ornibactine peut être intéressant. L'ornibactine est un sidérophore, c'est-à-dire une molécule capable de chélater le fer. Les sidérophores contiennent des fonctions bidentates capables de lier le fer. L'ornithine et ses dérivés sont capables d'assurer cette fonction. La recherche du motif complet dérivé de l'ornibactine retourne 14 peptides. Ces peptides sont des amphibactines, des pyoverdines, des ornibactines et la foroxymithine qui sont tous des sidérophores. La recherche du motif * :_*OH-Orn_*Asp_*Ser_*Orn, dérivé également de l'ornibactine, avec au moins deux monomères communs entre le motif et un peptide, retourne une liste de 90 peptides. Parmi ces 90 peptides, 74 sont des sidérophores. La recherche de motifs structuraux permet donc de mettre en évidence des relations entre la structure et la fonction des peptides.

Recherche par similarité

La méthode présentée dans la section 3.4 est utilisée pour la recherche de structures similaires au sein de NORINE. L'utilisateur peut entrer sa structure avec la représentation sous forme de graphes, la représentation linéaire ou il peut utiliser l'éditeur. Il a également le choix du niveau de groupement (clustering) des monomères qu'il souhaite utiliser (voir section 3.4.3). La figure 4.21 montre une capture d'écran de l'interface correspondante.

Les résultats sont présentés dans un tableau et sont triés par distance croissante, c'est-à-dire du peptide le plus similaire vers le moins similaire. La figure 4.22 montre une capture d'écran partielle des résultats de la recherche de structures similaires à l'ornibactine C4 (NOR00403), avec les clustering de niveau 1.

Pour chaque comparaison, la distance obtenue est donnée, ainsi que le nombre de monomères communs et le nom du peptide avec lequel la comparaison est effectuée. L'utilisateur peut obtenir la fiche descriptive de chaque peptide en cliquant simplement sur le nom de celui-ci. En cliquant

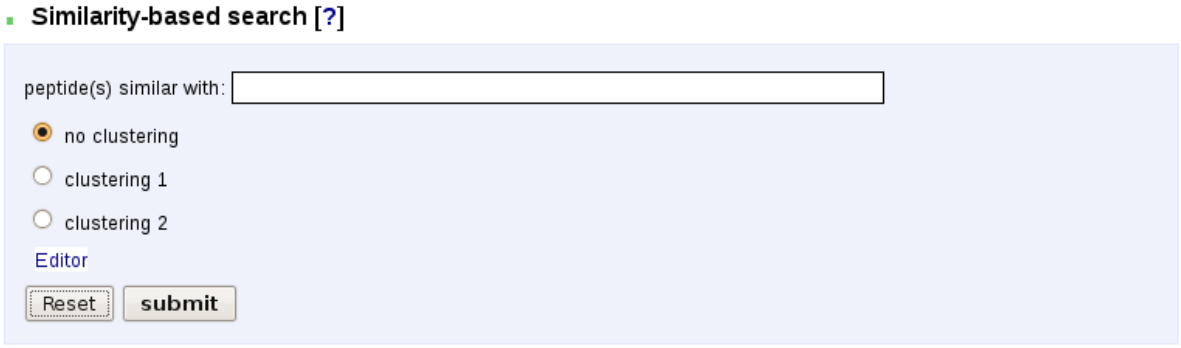


FIG. 4.21 – Formulaire de recherche de structures similaires

distance	common monomers	peptide	download
0.0	6	ornibactin C6	☐
0.0	6	ornibactin C4	☐
0.0	6	ornibactin C8	☐
0.33333333333333337	4	amphibactin H	☐
0.33333333333333337	4	amphibactin E	☐
0.33333333333333337	4	amphibactin I	☐
0.33333333333333337	4	amphibactin B	☐
0.33333333333333337	4	amphibactin G	☐
0.33333333333333337	4	amphibactin D	☐
0.33333333333333337	4	amphibactin F	☐
0.33333333333333337	4	amphibactin C	☐
0.5	3	pyoverdin 6-10	☐
0.5	3	foroxymithine	☐
0.5714285714285714	3	marinobactin B	☐
0.5714285714285714	3	marinobactin E	☐

FIG. 4.22 – Tableau des résultats de la recherche par similarité

sur une distance, une page détaillée de la comparaison est obtenue. La figure 4.23 montre une capture d'écran d'une comparaison détaillée.

Dans la fiche descriptive d'une comparaison, les monomères mis en relation apparaissent en rouge. Il est possible de visualiser les deux structures comparées, ce qui permet d'identifier facilement la partie commune entre les peptides.

Cette recherche permet à l'utilisateur de trouver des peptides présentant des structures similaires et donc, certainement, des activités biologiques similaires. Reprenons l'exemple de l'ornibactine. En recherchant des peptides similaires à la structure de l'ornibactine en utilisant le clustering de niveau 1, les peptides les plus proches sont également des sidérophores. En effet, dans le tableau de résultats de la recherche correspondante (figure 4.22), les peptides les plus similaires sont les amphibactines, les pyoverdines, les marinobactines et la foroxymithine qui sont tous des sidérophores.

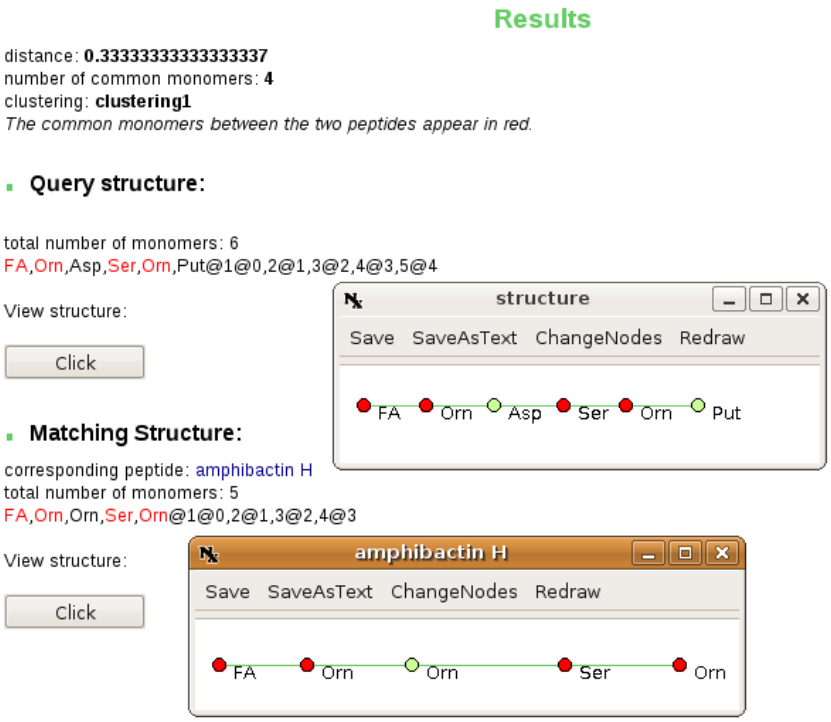


FIG. 4.23 – Fiche détaillée d'une comparaison entre deux structures

4.2.4 Recherche de monomères

L'information sur les monomères intégrés dans les peptides non-ribosomiaux sont des données importantes contenues dans NORINE.

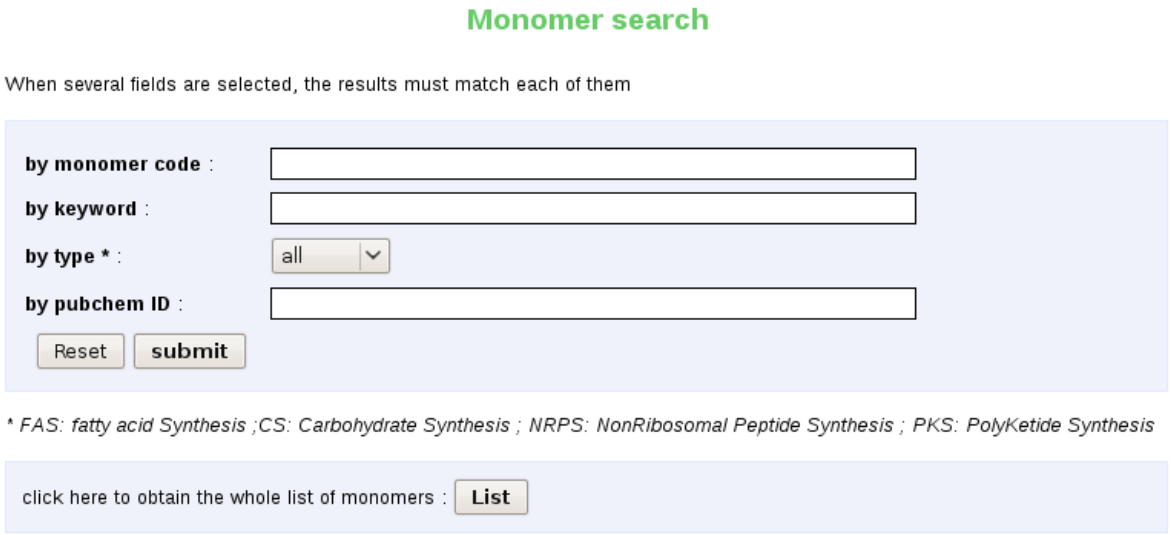


FIG. 4.24 – Capture d'écran de la recherche de monomères

Il est possible de rechercher des informations sur les monomères à l'aide d'un formulaire

dédié. La figure 4.24 montre une capture d'écran de la recherche de monomères. L'utilisateur peut rechercher des monomères en fonction de leur code, par mots-clés (c'est-à-dire le nom, la nomenclature IUPAC ou les synonymes), ou encore par l'identifiant PubChem. Il peut également rechercher les monomères appartenant à un type donné. Par exemple, la liste des monomères incorporés par les NRPS peut être demandée. Enfin, la liste de tous les monomères présents dans NORINE peut être obtenue en cliquant sur le bouton correspondant. Une fois la requête lancée, la liste des monomères correspondant à la requête est affichée et la fiche descriptive de chaque monomère peut être consultée (figure 4.10).

4.2.5 Autres fonctionnalités

NORINE contient tout d'abord une page d'accueil présentant la base de données ainsi que les informations sur les diverses modifications apportées. Elle contient également des pages d'aide très détaillées pour guider l'utilisateur lors des premières consultations. Les pages d'aide concernent les différentes recherches, mais aussi l'éditeur et la visualisation.

NORINE permet également aux utilisateurs de soumettre des peptides. En effet, une page de soumission de nouveaux peptides est proposée. Après vérification des différentes informations par notre équipe, les peptides soumis sont intégrés dans la base. Actuellement, six peptides de NORINE proviennent de soumissions directes. L'utilisateur peut également apporter des modifications sur les informations concernant un peptide donné grâce à un formulaire de modification. Les modifications apportées par un utilisateur sont vérifiées et intégrées dans NORINE. Pour chaque peptide, les utilisateurs peuvent également laisser des commentaires.

4.3 Statistiques d'utilisation de Norine

A partir de juin 2008, nous avons utilisé *Google Analytics* pour obtenir des statistiques d'utilisation de NORINE. *Google Analytics* est une solution d'analyse d'audience Internet gratuite qui indique comment les visiteurs sont arrivés sur le site et la façon dont ils l'ont exploré. Dans *Google Analytics*, l'unité de mesure est la visite ce qui correspond à une session. Une session est une période d'interaction entre le navigateur du visiteur et un site Web particulier, qui se termine lorsque le visiteur ferme la fenêtre ou le programme de navigation, ou lorsqu'il n'effectue aucune action sur le site pendant 30 minutes. Une session correspond donc à un utilisateur qui effectue plusieurs requêtes sur le site de NORINE.

La région Nord-Pas-de-Calais (où l'équipe qui travaille sur NORINE est située) a été retirée des statistiques afin de ne pas biaiser les résultats. Entre juin 2008 et mai 2009, NORINE a reçu 2 281 visites. Le nombre de visites par mois est donnée dans la figure 4.25.

Sur cette période, NORINE reçoit, en moyenne, 200 visites par mois. Une nette augmentation du nombre de visites est observée en mars. Cette augmentation peut s'expliquer par la publication de l'article dans *BMC Structural Biology*, le 18 mars 2009.

20 885 pages ont été consultées, ce qui correspond à une moyenne de 9 pages par visite. Le temps moyen passé sur le site est de 8 minutes. Les visiteurs proviennent du monde entier, comme le montre la figure 4.26 donnant la répartition géographique des visites sur NORINE.

Le plus grand nombre de visites provient des Etats-Unis, avec 718 visites, puis de la Chine (203 visites) et de l'Allemagne (183 visites). La répartition géographique des visites correspond à la répartition des équipes travaillant sur la synthèse non-ribosomiale dans le monde. Par exemple, 272 visites proviennent du Michigan où une équipe de l'Université du Michigan (*College of pharmacy*) travaille sur les NRPS. Les pages les plus consultées sont les formulaires de recherche

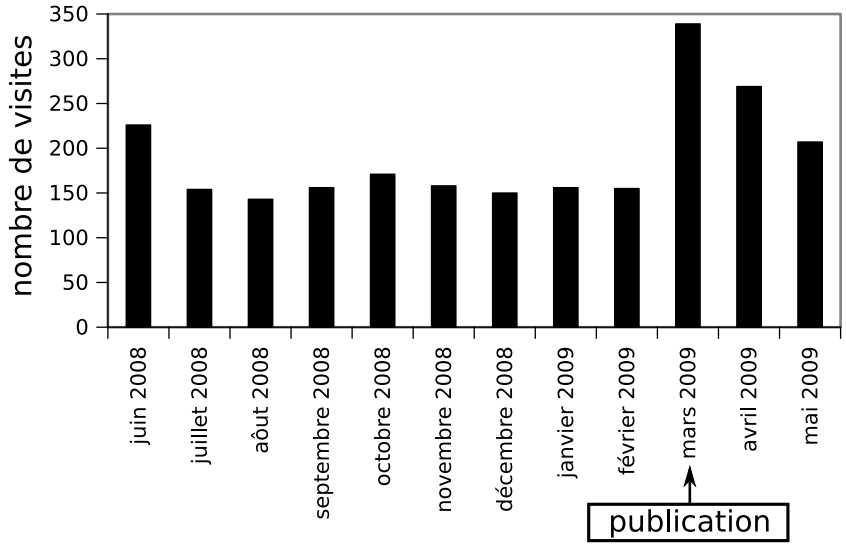


FIG. 4.25 – Nombre de visites par mois entre juin 2008 et mai 2009

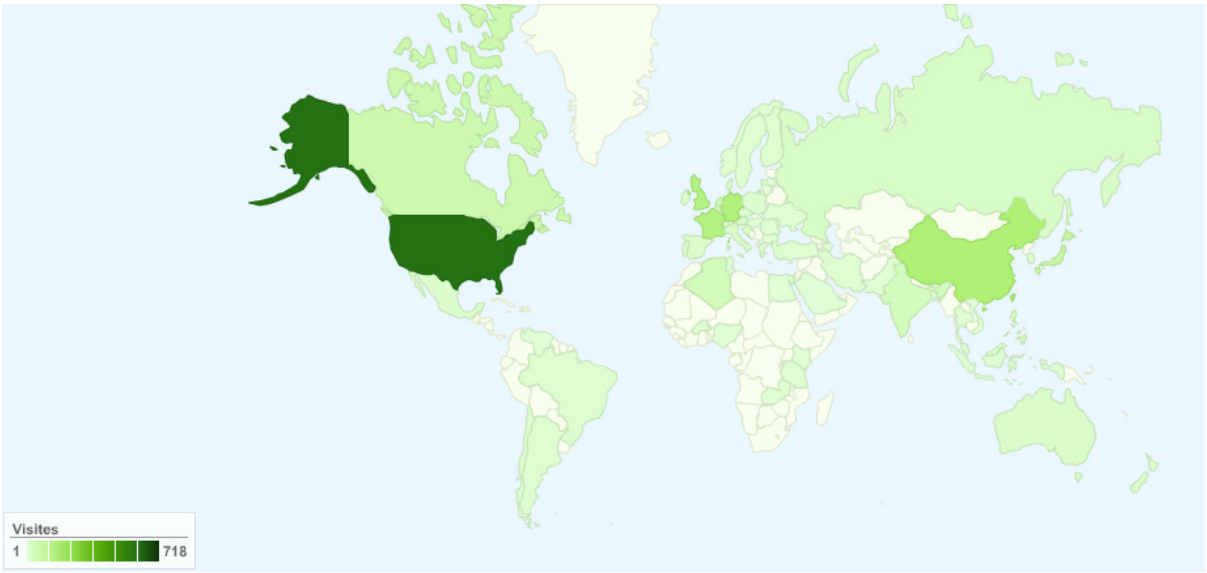


FIG. 4.26 – Répartition géographique des visites sur NORINE entre juin 2008 et mai 2009

(basique, structure et monomères) ainsi que la page d'accueil et la liste des peptides contenus dans NORINE. 48% des visites proviennent d'un accès direct, c'est-à-dire que l'utilisateur a enregistré NORINE dans ses favoris. 14% des visites proviennent de sites référents, c'est-à-dire de sites qui ont un lien vers NORINE tels que les sites sur lesquels nos articles sont disponibles ou d'autres sites institutionnels qui référencent NORINE. Enfin, 38% des visites proviennent d'une recherche par mots-clés dans Google.

NORINE *est la seule ressource publique entièrement dédiée aux peptides non-ribosomiaux. Elle contient actuellement plus de 1 000 peptides et est disponible à l'adresse suivante* `http://bioinfo.lifl.fr/norine/`. NORINE *propose à l'utilisateur une interface web qui permet de rechercher facilement une liste de peptides répondant à certains critères.* NORINE *est la ressource de référence mondiale pour les peptides non-ribosomiaux, notamment par le lien direct avec wwPDB. Elle a également permis d'obtenir des connaissances biologiques sur les peptides non-ribosomiaux.*

Chapitre 5

Statistiques sur la base de données

Dans ce chapitre, nous présentons diverses analyses statistiques réalisées sur les peptides et les monomères contenus dans Norine. Ces expériences constituent la première étude d'aussi grande ampleur réalisée depuis la découverte du mécanisme non-ribosomial. En effet, même si Norine ne contient pas encore l'ensemble des peptides non-ribosomiaux identifiés à ce jour, elle reste l'unique ressource contenant un nombre aussi important de peptides et variants, représentatifs de la diversité de ces molécules. Les résultats de ces analyses sont regroupés dans un article en cours de rédaction. Nous avons réalisé diverses études nous permettant, dans un premier temps, d'extraire des informations d'ordre général. Dans un second temps, nous avons comparé les protéines classiques et les peptides non-ribosomiaux. Ensuite, nous avons étudié les caractéristiques en fonction des organismes producteurs. Enfin, nous nous sommes intéressés aux caractéristiques des peptides en fonction des activités biologiques, ce qui nous a amené au développement d'un outil de prédiction de l'activité biologique d'un peptide à partir de sa composition.

5.1 Méthodes et définitions

Dans cette partie, nous introduisons les diverses méthodes et définitions utilisées lors des analyses statistiques.

Distribution des monomères

Nous avons réalisé un script qui prend en entrée une liste de peptides et une liste de monomères. Le nombre d'occurrences de chaque monomère dans l'ensemble des peptides de la liste d'entrée est calculé.

Cohortes de peptides utilisées et notion de variants

Comme nous l'avons vu précédemment, Norine contient tous les variants identifiés d'une famille de peptides (section 1.2.4.0). En tout, 184 familles de peptides sont répertoriées dans Norine, dont 62 constituées d'un seul peptide. Au sein d'une famille, les variations peuvent toucher uniquement l'acide gras pour les lipopeptides ou des monomères pour d'autres familles ou encore se situer au niveau de la structure. Au sein d'une famille, les variants peuvent être plus ou moins proches les uns des autres. Lors d'une étude statistique, considérer tous les variants d'une même famille peut biaiser les résultats. En effet, le nombre d'occurrences des monomères ubiquitaires dans une famille de peptides sera surestimé. Nous avons donc étudié les 122 familles

de NORINE contenant au moins deux variants peptidiques afin de mettre en évidence les familles présentant des variants peu similaires. Nous avons calculé la distance moyenne entre les peptides d'une famille donnée. Pour ce faire, nous avons utilisé la méthode de calcul de similarité présentée dans la section 3.4. Nous avons réalisé ces tests sans clustering sur les monomères, mais aussi avec le clustering de niveau 1 et celui de niveau 2 (section 3.4.3). Nous avons considéré comme non-conservées, les familles présentant une distance moyenne supérieure à $0,4$ avec le clustering de niveau 2. Nous avons choisi un seuil égal à $0,4$ car, après une analyse des distances moyennes au sein des différentes familles, il semble le plus approprié. Nous avons tracé l'histogramme des distances moyennes avec le clustering de niveau 2, ordonnées par ordre croissant, au sein des 80 familles présentant une distance moyenne non nulle (voir figure 5.1).

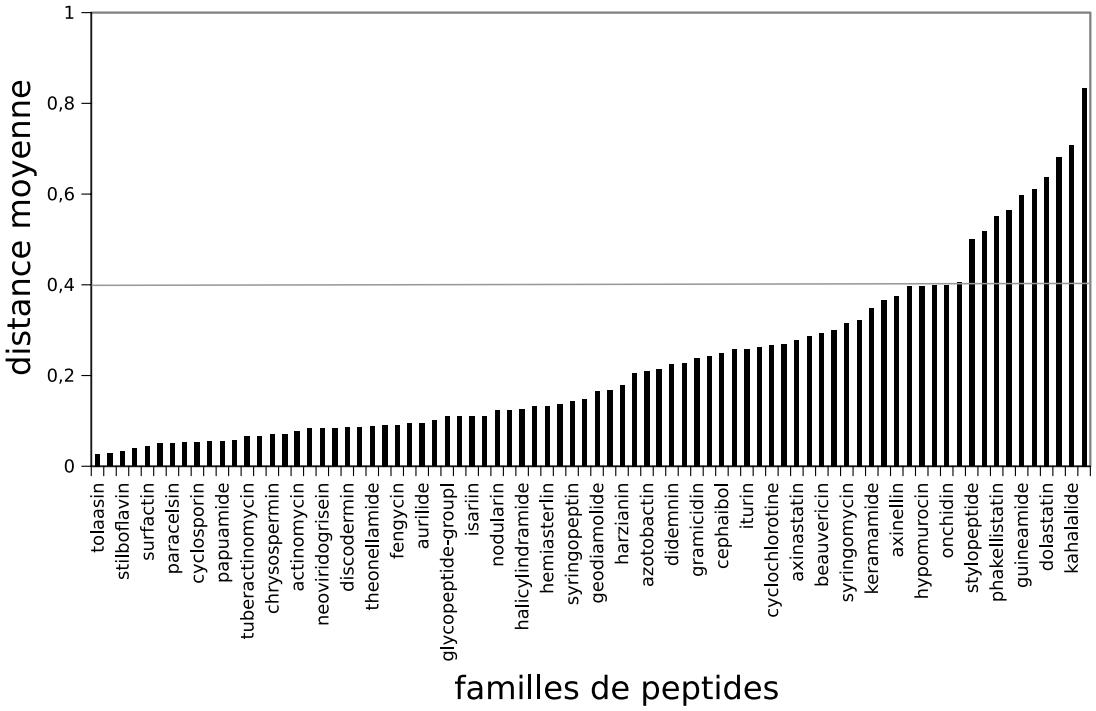


FIG. 5.1 – Distances moyennes calculées au sein des 80 familles contenant au moins deux peptides et présentant une distance non nulle, avec un clustering sur les monomères de niveau 2. Les distances moyennes sont ordonnées par ordre croissant. Pour une raison de lisibilité, le nom des 80 familles n'apparaît pas sur la figure.

Nous pouvons remarquer qu'il existe un plateau au niveau de la distance moyenne égale à $0,4$ puis à partir de ce seuil, la distance moyenne augmente fortement, reflétant des familles de plus en plus hétérogènes. Cette observation justifie le choix d'un seuil égal à $0,4$. Nous avons ainsi obtenu une liste de familles pour lesquelles les variants sont très différents les uns des autres et doivent donc être considérés comme indépendants les uns des autres. La table 5.1 montre les familles pour lesquelles les variants sont différents.

Les peptaibols ne forment pas une famille mais un ensemble de familles présentant des caractéristiques communes. Cette catégorie de peptides présente une importante dissimilarité. Par contre, les familles composant cette catégorie contiennent des peptides similaires (données non présentées ici). Il faut donc considérer les peptaibols comme plusieurs familles et non pas comme une unique famille de peptides.

TAB. 5.1 – Familles de peptides présentant des variants distincts

famille	statut	distance moyenne clustering 2	nombre de variants
dolastatines	curated	0,64	4
guineamides	putative	0,60	6
hymenamides	putative	0,61	10
kahalalides	curated	0,71	16
kapakahines	putative	0,52	5
peptaibols	curated	0,56	130
phakellistatines	putative	0,55	14
pyoverdines	curated	0,68	57
serrawettines	curated	0,83	2
stylopeptides	putative	0,5	2

Pour une famille dont les variants sont similaires, nous prenons un seul variant au hasard pour représenter cette famille. Au contraire, pour les familles dont les variants sont distincts (citées précédemment) nous allons prendre tous les variants composant cette famille. Dans la suite, c'est cette notion qui est utilisée lorsque le terme **« sans variants »** est employé.

NORINE contient deux types de peptides : « curated » et « putative ». Les peptides « curated » sont ceux pour lesquels la synthèse non-ribosomiale est prouvée ou admise. Les peptides « putative » sont ceux pour lesquels la voie de biosynthèse n'est pas encore confirmée. Il peut être intéressant de distinguer ces deux cohortes lors d'études statistiques dans le but de ne pas biaiser les résultats obtenus avec des peptides ne provenant pas nécessairement de la voie non-ribosomiale. Nous réalisons donc les analyses statistiques sur quatre cohortes principales de peptides :

- *total*, contenant l'ensemble des peptides (1071 peptides)
- *curated*, contenant uniquement les peptides « curated » (790 peptides)
- *sans variants*, contenant un seul variant par famille (sauf pour les familles vues précédemment) (290 peptides)
- *sans variants curated*, contenant les peptides « curated » et en considérant un seul variant par famille (175 peptides)

Coefficient de corrélation

Etudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. Le coefficient de corrélation (CC) r permet de chiffrer la liaison entre deux distributions. Il est compris entre -1 et 1. Plus le coefficient est proche des valeurs extrêmes (-1 et 1), plus la corrélation entre les variables est forte, les variables sont dites « fortement corrélées ». Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes. Le CC r se calcule comme suit :

Soit deux séries $X(x_1, ..., x_n)$ et $Y(y_1, ..., y_n)$, on a :

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

Dans notre cas le CC est calculé entre deux distributions monomériques de deux ensembles de peptides. Une distribution monomérique est le nombre d'occurrences de chaque monomère au sein d'un ensemble de peptides appartenant au groupe étudié.

Apprentissage automatique

L'apprentissage automatique (ou *machine learning* en anglais) est un des champs d'étude de l'intelligence artificielle. Le but des méthodes d'apprentissage automatique supervisé est de déterminer le meilleur classifieur permettant de classer correctement un maximum de données étiquetées (c'est-à-dire dont la classe est connue) d'un ensemble d'apprentissage (*training set*) dans le but de pouvoir classer des données non-étiquetées (c'est-à-dire dont la classe est inconnue).

WEKA est un logiciel open-source permettant de réaliser de l'apprentissage automatique, très utilisé en bioinformatique [Frank et al., 2004]. Il met à disposition différents algorithmes utilisés en apprentissage automatique. Ceux que nous avons utilisés sont les suivants :

- *SMO*, algorithme basé sur les machines à vecteur de support (SVM) multi-classe
- *BayNet*, basé sur les réseaux Bayésiens
- *zero*, basé sur des règles de décision
- *J48*, basé sur la création d'un arbre de décision
- *IBK*, basé sur l'algorithme des k plus proches voisins

Pour chacune des méthodes, WEKA propose des paramètres par défaut. Ce sont ces paramètres que nous avons utilisés.

WEKA prend en entrée un fichier « .arff » contenant les données d'apprentissage. Dans notre cas, le même fichier est utilisé pour les différentes méthodes testées. La structure du fichier est donnée dans la figure 5.2.

```
@RELATION monomers
@ATTRIBUTE Asn real
@ATTRIBUTE Val real
.
.
.
@ATTRIBUTE Gly real
@ATTRIBUTE Pro real
@ATTRIBUTE class {antibiotic,antitumor,immuno,sidero,surfactant,toxin}
0,2,...,0,1,antibiotic
1,0,...,0,0,antitumor
0,1,...,0,3,toxin
0,0,...,1,0,antibiotic
1,0,...,0,0,sidero
```

FIG. 5.2 – Structure d'un fichier WEKA.

Les premières lignes, commançant par le symbole « @ », définissent les différents attributs, ainsi que leur type, qui permettent de décrire les données, qui sont ici des peptides. Dans notre cas, nous avons deux types d'attributs : les monomères et la classe d'activité biologique. La totalité des monomères de l'étude sont considérés comme des attributs et sont représentés par un nombre réel correspondant au nombre d'occurences de ce monomère dans le peptide considéré. L'attribut « class » représente la classe d'activité biologique à laquelle appartient le peptide. Un peptide est donc représenté par un vecteur constitué de $n + 1$ valeurs, avec n le nombre de monomères étudiés et $+1$ pour la classe. Dans le fichier d'entrée, après la déclaration des

attributs, une ligne représente un peptide. Dans l'exemple de la figure 5.2, le fichier d'entrée contient cinq peptides. Le premier peptide du fichier contient 2 valines et 1 proline et présente une activité antibiotique. Lorsqu'un peptide présente plusieurs activités biologiques, par exemple antibiotique et anti-tumorale, ce peptide est répété deux fois dans le fichier, une fois avec la valeur « antibiotic » pour l'attribut « class », et une fois avec l'attribut « class » égal à « antitumoral ».

Vrais positifs, faux positifs et taux de prédiction

Lors de tests pour mesurer la validité de la méthode de prédiction d'activité biologique, nous avons besoin de mesures chiffrées. Dans notre cas, nous cherchons à classer des peptides dans une classe d'activité. Un vrai positif (VP) est un peptide dont l'activité prédite correspond à celle connue. Le taux de vrais positifs dans une classe X est le rapport entre le nombre d'éléments classés dans X qui sont réellement X et le nombre total d'éléments classés dans X.

Un faux positif (FP) est un peptide dont l'activité prédite est différente de l'activité réelle de ce dernier. Le taux de faux positifs dans une classe X est le rapport entre le nombre d'éléments classés X alors qu'ils n'appartiennent pas à X et le nombre total d'éléments classés X.

Nous ne pouvons pas calculer les taux de vrais négatifs (peptides non classés dans une classe et n'appartenant réellement pas à cette classe) et de faux négatifs (peptides non classés dans une classe et appartenant à cette classe) car certains peptides présentent plusieurs activités biologiques et par conséquent appartiennent à plusieurs classes et dans ce cas, nous ne savons pas quelle classe considérer.

Le taux de prédictions correctes est simplement le rapport entre le nombre d'éléments classés correctement, c'est-à-dire le nombre de vrais positifs, et le nombre total d'éléments dans le test.

5.2 Statistiques générales

Dans cette première partie, nous donnons des statistiques d'ordre général sur les peptides et monomères contenus dans Norine. Norine contient actuellement 1071 peptides répartis en 184 familles. 790 peptides ont le statut « curated » et 281 le statut « putative ». Tout d'abord, nous nous sommes intéressés à la distribution de la taille (nombre de monomères) des peptides non-ribosomiaux. La figure 5.3 montre la distribution de la taille des peptides « curated » sans variants (175 peptides).

Nous pouvons voir que tous les peptides non-ribosomiaux, sauf un, ont une taille comprise entre 2 et 23 monomères, avec une grande proportion de peptides, environ 30%, ayant une taille de 7 ou 8 monomères. Seul un peptide, le polythéonamide B, présente une taille de 49 monomères [Hamada et al., 2005]. Son origine non-ribosomiale n'a pas été validée expérimentalement, mais est largement admise par les spécialistes du domaine, et il représente donc, à ce jour, le plus grand peptide non-ribosomial identifié.

Dans un second temps, nous avons étudié la répartition des différents organismes producteurs. Les organismes producteurs sont divisés en 3 grandes catégories : les bactéries, les champignons et les autres. La catégorie « autres » regroupe divers organismes eucaryotes tels que les éponges ou les nématodes. La répartition des organismes est similaire que tous les variants soient considérés ou non. Nous présentons ici la répartition des peptides en considérant tous les variants. La figure 5.4 montre la répartition des peptides en fonction des organismes producteurs.

Lorsque l'ensemble des peptides est utilisé (figure 5.4a), les bactéries représentent le règne le plus fréquent d'organismes producteurs. En effet, plus de la moitié des peptides contenus dans Norine sont identifiés chez des bactéries. La proportion de peptides isolés au sein des

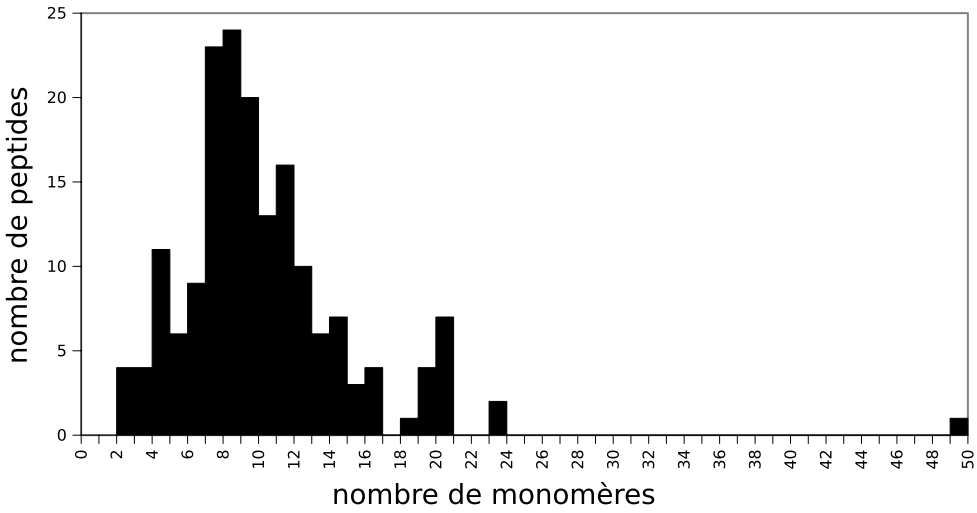


FIG. 5.3 – Distribution de la taille des peptides « curated » sans variants (175 peptides)

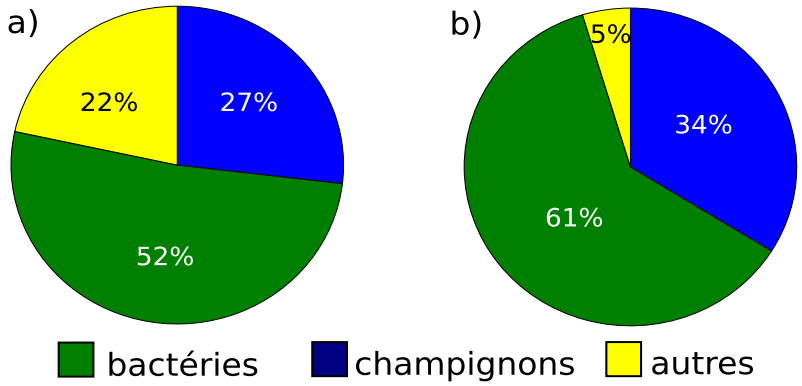


FIG. 5.4 – Répartition des peptides en fonction des groupes d'organismes producteurs en considérant a) la totalité des peptides (1071 peptides) b) les peptides « curated » (790 peptides)

champignons est de 27% et celle des peptides identifiés au sein d'autres espèces eucaryotes est de 22%. Lorsque seuls les peptides « curated » sont utilisés (figure 5.4b), les proportions varient. Le règne des bactéries reste majoritaire (plus de 60%) et la proportion des organismes eucaryotes « autres » diminue (moins de 5%). Les 5% de peptides produits par des eucaryotes « autres » n'ont pas de gènes codant pour des synthétases identifiés expérimentalement mais sont reconnus comme des peptides non-ribosomiaux par la communauté scientifique. Le règne des champignons représente 34% des organismes producteurs. Ces résultats s'expliquent par le fait que les bactéries sont généralement bien étudiées, par conséquent les gènes des synthétases sont étudiés expérimentalement et le peptide produit a le statut « curated ». A l'inverse, la majeure partie des peptides identifiés dans la catégorie « autres » est potentiellement produite par la voie NRPS mais peu d'études concernant la biosynthèse de ces peptides sont réalisées, plaçant ces peptides dans la catégorie « putative ».

Nous avons ensuite étudié la répartition des types de structures rencontrés au sein des peptides présents dans la base de données. Les résultats sont similaires que tous les variants soient considérés ou non. Nous présentons ici les résultats obtenus en considérant tous les variants d'une

même famille. La figure 5.5 montre la répartition des types de structures rencontrées au sein des peptides non-ribosomiaux.

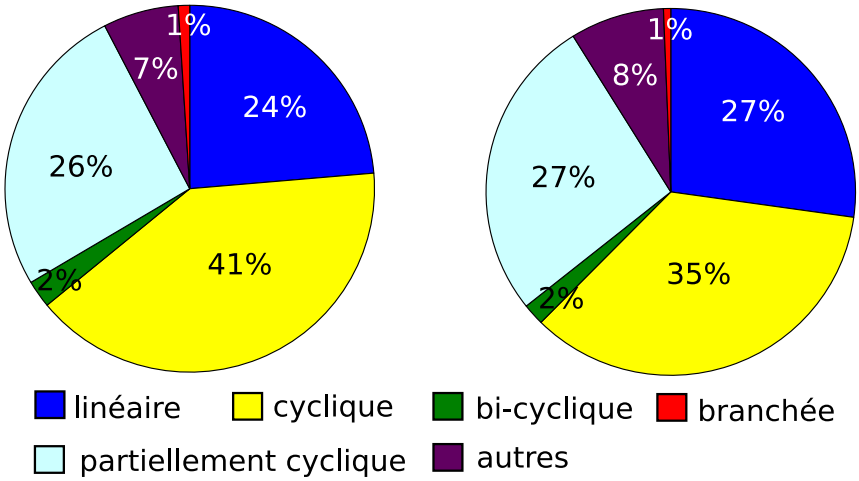


FIG. 5.5 – Répartition des types de structures des peptides en considérant a) la totalité des peptides (1071 peptides) b) les peptides « curated » (790 peptides)

Nous pouvons voir que les peptides ayant une structure primaire linéaire comparable à celle des peptides classiques représentent seulement un quart des structures des peptides de NORINE. Les structures cycliques (totalement, partiellement ou bi-cycliques) représentent 65% des structures primaires. Les structures branchées sont rares au sein des peptides non-ribosomiaux. Environ 10% des peptides ont une structure primaire complexe contenant des cycles chevauchants (catégorie « autres »). La figure 5.5 montre également une répartition similaire des types structuraux entre les peptides totaux et les peptides « curated », ce qui signifie que les peptides « putative » présentent une répartition des types structuraux similaire à celles des peptides « curated ».

5.3 Protéines ribosomiales *versus* peptides non-ribosomiaux

Dans cette section, nous comparons la distribution des acides aminés protéogéniques au sein des protéines classiques et au sein des peptides non-ribosomiaux. La répartition des vingt acides aminés protéogéniques pour les protéines et peptides ribosomiaux est obtenue à partir des statistiques de la banque de données UniProtKB/TrEMBL qui sont disponibles à l'adresse suivante : `http://www.ebi.ac.uk/uniprot/TrEMBLstats/`. La figure 5.6 montre la répartition des acides aminés protéogéniques au sein des protéines et peptides ribosomiaux.

Nous avons calculé la distribution des acides aminés protéogéniques au sein des peptides non-ribosomiaux « curated » et sans variants (175 peptides). Les pourcentages sont calculés par rapport au nombre d'acides aminés protéogéniques, et non par rapport au nombre total de monomères. La figure 5.7 montre les résultats obtenus. Les vingt acides aminés protéogéniques représentent 40% des monomères incorporés au sein des peptides non-ribosomiaux.

En comparant le pourcentage des acides aminés protéogéniques au sein des protéines classiques et des peptides non-ribosomiaux, nous pouvons remarquer que dans les deux cas, les acides aminés aliphatiques et hydroxylés sont les plus fréquents. De même, les acides aminés aromatiques et soufrés sont les moins fréquents dans les deux cas. En revanche, la sérine (Ser) et la

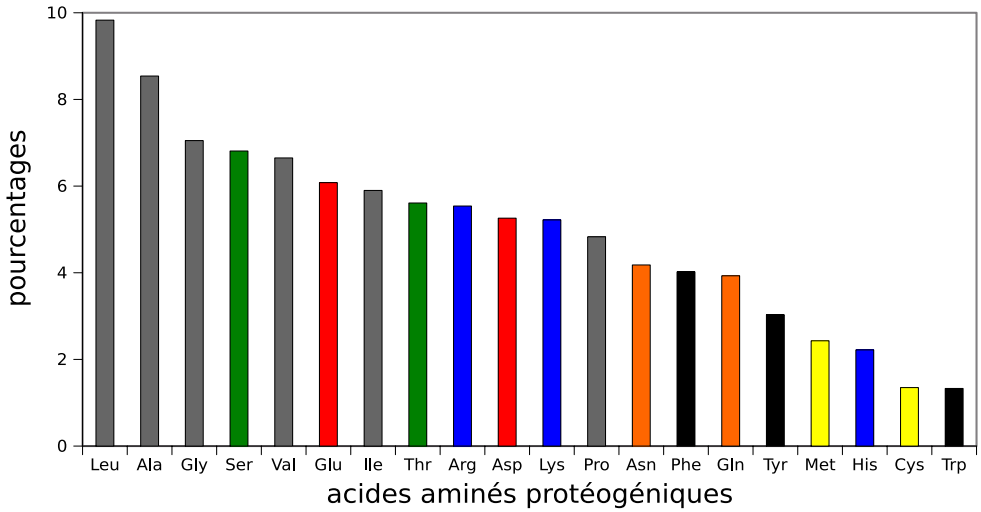


FIG. 5.6 – Répartition des acides aminés protéogéniques au sein des protéines et peptides ribosomiaux contenus dans UniProtKB/TrEMBL. (gris = aliphatique, rouge = acide, vert = petit hydroxy, bleu = basique, noir = aromatique, orange = amidé, jaune = soufré)

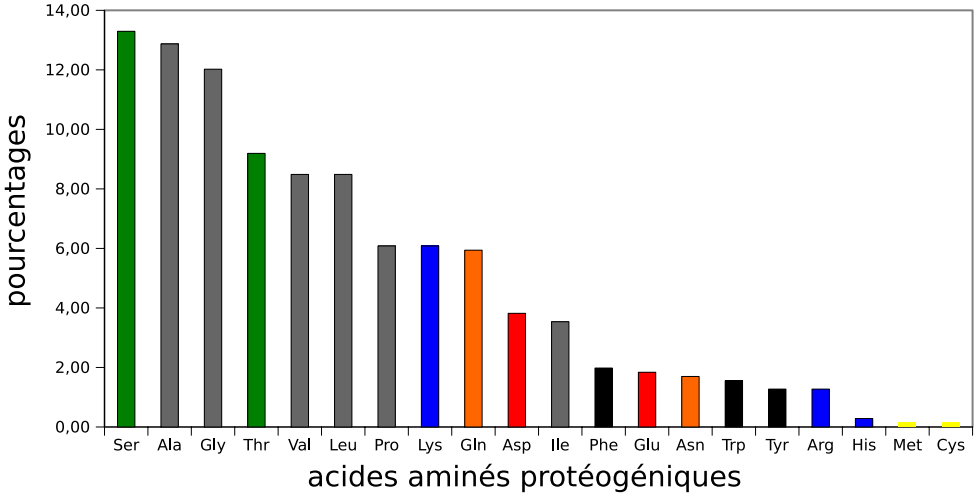


FIG. 5.7 – Répartition des acides aminés protéogéniques au sein des peptides « curated » sans variants de NORINE. (gris = aliphatique, rouge = acide, vert = petit hydroxy, bleu = basique, noir = aromatique, orange = amidé, jaune = soufré)

thréonine (Thr) sont plus fréquentes pour les peptides non-ribosomiaux. Cela peut s'expliquer par le fait que ces acides aminés possèdent un groupement hydroxyl permettant la formation d'une liaison covalente supplémentaire sur la chaîne latérale et ainsi l'obtention de structures primaires non-linéaires. Une autre caractéristique intéressante est que pour les protéines classiques, les acides aminés acides (Glu et Asp) sont plus fréquents que les acides aminés portant une fonction amine sur leur chaîne latérale (Gln et Asn). Au contraire, dans le cas des peptides non-ribosomiaux, la glutamine (Gln) semble être préférée à l'asparagine (Asn) et l'acide aspartique (Asp) semble être préféré à l'acide glutamique (Glu). Enfin, ces expériences montrent également que l'acide glutamique (Glu) et l'arginine (Arg) sont moins fréquents au sein des pep-

tides non-ribosomiaux qu'au sein des protéines synthétisées par la voie ribosomiale. Egalement, l'arginine est en neuvième position pour les protéines classiques alors qu'elle se place en seizième position pour les peptides non-ribosomiaux. Enfin nous pouvons constater que l'histidine (His), la méthionine (Met) et la cystéine (Cys) sont des acides aminés peu présents chez les peptides non-ribosomiaux.

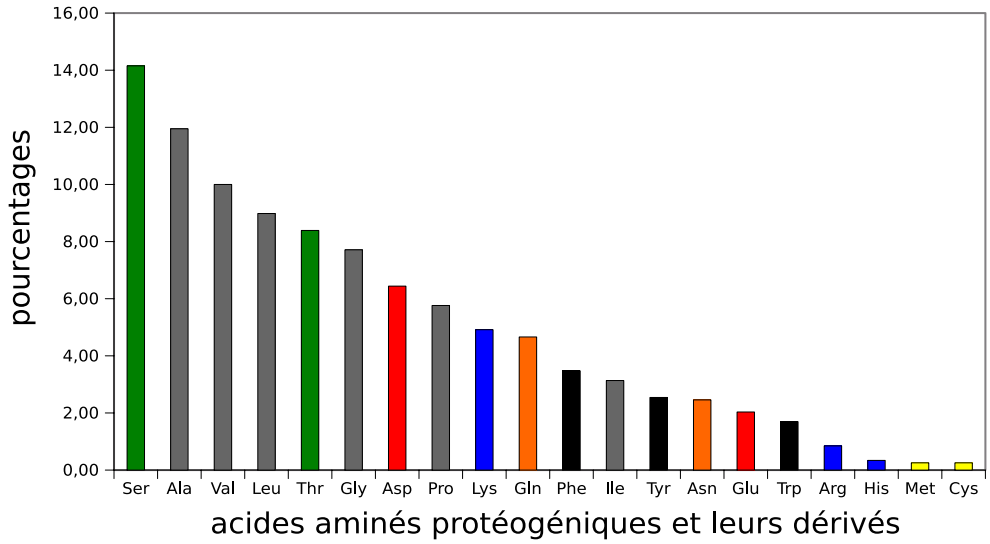


FIG. 5.8 – Répartition des acides aminés protéogéniques et leurs dérivés au sein des peptides « curated » sans variants de NORINE. (gris = aliphatique, rouge = acide, vert = petit hydroxy, bleu = basique, noir = aromatique, orange = amidé, jaune = soufré)

Dans un dernier temps, nous avons regroupé les acides aminés avec leurs dérivés (par exemple Ala, D-Ala, OH-Ala formeront le groupe Ala), ce qui correspond au clustering de niveau 1, et avons calculé la distribution de ces groupes dans le cas des peptides non-ribosomiaux. Les acides aminés protéogéniques et leurs dérivés représentent 70% des monomères rencontrés au sein des peptides non-ribosomiaux. Les 30% restants sont principalement d'autres acides aminés non-protéogéniques, mais aussi d'autres composés chimiques tels que des acides gras, des sucres ou des polykétides. La figure 5.8 montre la répartition des acides aminés protéogéniques et leurs dérivés au sein des peptides non-ribosomiaux « curated » sans variants (175 peptides). Les pourcentages sont calculés par rapport au nombre d'acides aminés et de leurs dérivés, et non par rapport au nombre total de monomères.

Les remarques effectuées précédemment sur les acides aminés protéogéniques sont toujours vraies lorsque les dérivés sont ajoutés. Cette expérience confirme la différence d'utilisation des acides aminés protéogéniques au sein des peptides non-ribosomiaux par rapport aux protéines ribosomiales.

5.4 En fonction des organismes producteurs

238 organismes producteurs produisant des peptides non-ribosomiaux sont répertoriés dans NORINE. Ces organismes sont soit des organismes pour lesquels les gènes codant les synthétases impliquées dans la synthèse du peptide ont été identifiés soit des organismes à partir desquels un peptide a été isolé, sans identification des gènes ou de la voie de biosynthèse. En effet, beaucoup

de peptides sont étudiés et recherchés pour leurs propriétés biologiques, ce qui implique que les génomes et les gènes intervenant dans la synthèse de ces molécules ne sont toujours pas connus. Dans un premier temps, nous avons étudié les caractéristiques des peptides synthétisés par les différents phyla présents dans NORINE. Nous avons mis en évidence des caractéristiques différentes entre les peptides synthétisés par les bactéries et ceux synthétisés par les champignons. Nous avons confirmé ces différences en étudiant la répartition des acides aminés protéogéniques et celle de la taille des peptides synthétisés par ces deux groupes d'organismes.

Distribution des monomères en fonction des phyla

Les organismes producteurs sont divisés en différents domaines, comme par exemple les bactéries et les eucaryotes. Cependant, les eucaryotes sont divisés en règnes dont les champignons (fungi) et les métazoaires (metazoa) qui comprennent, dans notre cas, principalement des éponges. Il est possible d'affiner les catégories d'organismes en descendant plus bas dans la classification du vivant. Nous avons étudié les principaux phyla présents dans Norine. Il sont au nombre de cinq :

- actinobacteria (131 peptides)
- cyanobacteria (106 peptides)
- firmicutes (127 peptides)
- proteobacteria (175 peptides)
- ascomycota (256 peptides)

La figure 5.9 montre l'arbre phylogénique des phyla utilisés dans l'étude.

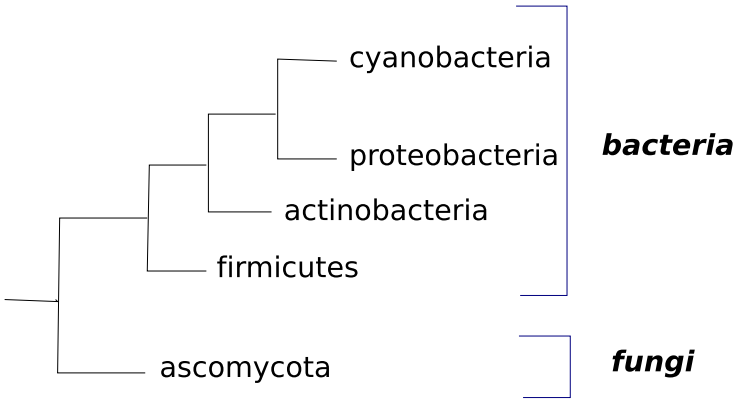


FIG. 5.9 – Arbre phylogénique des différents phyla étudiés

Pour les études statistiques, nous utilisons, en plus des 5 catégories correspondant aux phyla cités précédemment, les catégories suivantes qui rassemblent plusieurs phyla :

- bacteria
- eukaryota (fungi+plantes+metazoa)
- fungi
- metazoa

Nous avons calculé le coefficient de corrélation (CC) entre les distributions monomériques des différentes catégories. Nous avons réalisé ces tests sur les quatre cohortes de peptides (total, « curated », sans variants et sans variants « curated »). Les résultats obtenus sont semblables quelle que soit la cohorte observée. Nous présentons donc ici les résultats obtenus avec la totalité des peptides. Nous avons calculé la distribution de l'ensemble des monomères (clustering 0), mais

aussi celle des groupes de monomères issues du clustering de niveau 1 et celle des groupes de monomères issues du clustering de niveau 2. La table 5.2 montre les résultats obtenus.

TAB. 5.2 – Coefficients de corrélation entre les différentes distributions de monomères en fonction des organismes producteurs. Les chiffres entre parenthèses donnent le nombre de peptides dans chaque groupe.

organisme 1	organisme 2	clustering 0	clustering 1	clustering 2
bacteria (539)	eukaryota (507)	0,386	0,533	0,684
bacteria (539)	fungi (281)	**0,273**	0,388	0,535
bacteria (539)	metazoa (218)	**0,566**	0,725	0,838
fungi (281)	metazoa (218)	0,369	0,510	0,654
actinobacteria (131)	cyanobacteria (106)	0,085	0,335	0,475
actinobacteria (131)	firmicutes (127)	0,302	0,515	0,506
actinobacteria (131)	proteobacteria (175)	0,274	0,449	0,490
actinobacteria (131)	ascomycota (256)	0,138	0,244	0,312
cyanobacteria (106)	firmicutes (127)	0,167	0,385	0,605
cyanobacteria (106)	proteobacteria (175)	0,242	0,491	0,532
cyanobacteria (106)	ascomycota (216)	0,135	0,291	0,453
firmicutes (127)	proteobacteria (175)	**0,480**	0,605	0,723
firmicutes (127)	ascomycota (256)	0,260	0,424	0,521
proteobacteria (175)	ascomycota (256)	0,237	0,351	0,518

La première remarque est que plus le clustering est important plus le CC augmente. Cela s'explique par le fait que de plus en plus de monomères sont regroupés et par conséquent, la probabilité d'avoir l'utilisation des mêmes classes de monomères augmente également. Ensuite, il existe une grande différence entre les monomères observés chez les champignons et ceux isolés chez les bactéries. En effet, le CC faible reflète une mauvaise corrélation entre les deux distributions et cela quel que soit le niveau de clustering utilisé. Cette expérience montre donc que les monomères utilisés par les bactéries sont différents de ceux utilisés par les champignons.

Une autre remarque très importante est que le CC est plus élevé entre les distributions monomériques des bactéries et des métazoaires $(0,566)$ qu'entre celles des bactéries et des champignons $(0,273)$. Ce CC est élevé quel que soit le niveau de clustering considéré. Cela montre qu'il existe une corrélation entre les deux distributions monomériques correspondantes. Cette expérience semble donc confirmer l'hypothèse selon laquelle les peptides non-ribosomiaux isolés à partir des éponges et autres eucaryotes supérieurs seraient en fait synthétisés par des bactéries symbiotiques et non par les eucaryotes eux-mêmes. La symbiose est une interaction spécifique et à long terme entre deux organismes ou plus. Ce résultat est donc très intéressant, puisqu'il vient appuyer une hypothèse largement acceptée par la communauté scientifique. En effet, plusieurs articles montrent qu'il existe de nombreuses symbioses entre des bactéries et des eucaryotes supérieurs. Par exemple, une revue récente recense les métabolites, dont certains peptides non-ribosomiaux, synthétisés par des bactéries symbiotiques [Piel, 2009]. Dans une autre étude, les auteurs ont identifié des gènes NRPS au sein de génomes de bactéries isolées dans quatre espèces d'éponges [Zhang et al., 2009]. Ces études tendent à prouver que les peptides non-ribosomiaux sont synthétisés par des bactéries symbiotiques plutôt que par les éponges elles-mêmes. De nombreux micro-organismes ont été identifiés au sein des éponges [Lee et al., 2001]. Une éponge peut contenir plusieurs micro-organismes symbiotiques très différents. Ainsi, les différents phyla bactériens étudiés ici sont identifiés au sein des éponges et aucun phylum n'est spécifique à la

symbiose. Par exemple, des protéobactéries symbiotiques ont été idenbtifiées chez *Theonella swinhoei*, des cyanobactéries chez *Dysidea herbacea*, des actinobactéries chez *Rhopaloeides odorabile* ou encore des firmicutes chez *Aplysina sp.*.

Chaque phylum semble avoir un ensemble spécifique de monomères. En effet, les différents CC observés sont faibles, ce qui montre une distribution monomérique spécifique. Le phylum le plus éloigné des autres est celui des cyanobactéries. En effet, les CC observés entre ce phylum et les autres phyla sont très proches de 0. Les cyanobactéries sont des bactéries vivant dans divers milieux aquatiques. Leur environnement est différent de celui des autres organismes et par conséquent, les ressources présentes dans le milieu peuvent être spécifiques. La spécificité du milieu peut expliquer la spécificité des monomères incorporés.

De la même manière, le phylum des ascomycota montre des CC très faibles. Ce phylum est le seul représentant des champignons (les 4 autres étant des bactéries). Il semble donc logique que ce phylum montre une distribution monomérique propre, ce qui est cohérent avec la première remarque effectuée soulignant une différence d'utilisation des monomères entre bactéries et champignons.

Enfin, les firmicutes et les protéobactéries semblent utiliser un ensemble de monomères proches (CC=$0,480$).

Distribution des acides aminés protéogéniques

Nous avons étudié la distribution des acides aminés protéogéniques dans les peptides non-ribosomiaux synthétisés par les bactéries et ceux synthétisés par les champignons. Dans ces expériences, un seul variant par famille est considéré. Que la totalité des peptides ou seulement ceux dont le statut est « curated » soient pris en compte, les résultats sont similaires. La figure 5.10 montre la distribution des acides aminés protéogéniques obtenue sur l'ensemble des peptides « curated » sans variants produits par les bactéries (121 peptides). Les acides aminés protéogéniques représentent environ 40% des monomères incorporés dans ces peptides.

Les acides aminés protéogéniques les plus fréquents dans les peptides synthétisés par les bactéries sont la sérine, la glycine, la thréonine, l'alanine et la lysine. La méthionine n'a pas été identifiée au sein des peptides bactériens présents dans Norine. Nous avons réalisé la même expérience en regroupant les acides aminés protéogéniques et leurs dérivés. Ils représentent environ 70% des monomères constituant les peptides bactériens. Les groupes d'acides aminés les plus fréquents restent sensiblement les mêmes que dans l'expérience précédente et sont, dans l'ordre décroissant des fréquences : sérine, thréonine, alanine, glycine et valine.

Au sein des peptides non-ribosomiaux synthétisés par les champignons, les acides aminés protéogéniques représentent plus de 40% des monomères incorporés. La figure 5.11 montre la distribution des acides aminés protéogéniques au sein des peptides « curated » et sans variants synthétisés par les champignons (34 peptides).

Pour les peptides synthétisés par les champignons, les acides aminés les plus fréquents sont la glutamine, la proline, la leucine, l'alanine et la valine. L'histine, la lysine, la méthionine et l'arginine ne sont pas identifiés au sein des peptides synthétisés par les champignons présents dans Norine. En regroupant les acides aminés protéogéniques et leurs dérivés, les groupes les plus fréquents, par ordre décroissant, sont : valine, proline, leucine, alanine et glutamine. Cette expérience montre que les acides aminés protéogéniques incorporés au sein des peptides sont très différents selon l'organisme producteur. En effet, les cinq acides aminés les plus fréquents chez les bactéries sont complètement différents de ceux identifiés pour les champignons, sauf l'alanine, qui est fréquente chez les bactéries et les champignons. Par exemple, la glutamine, acide aminé le

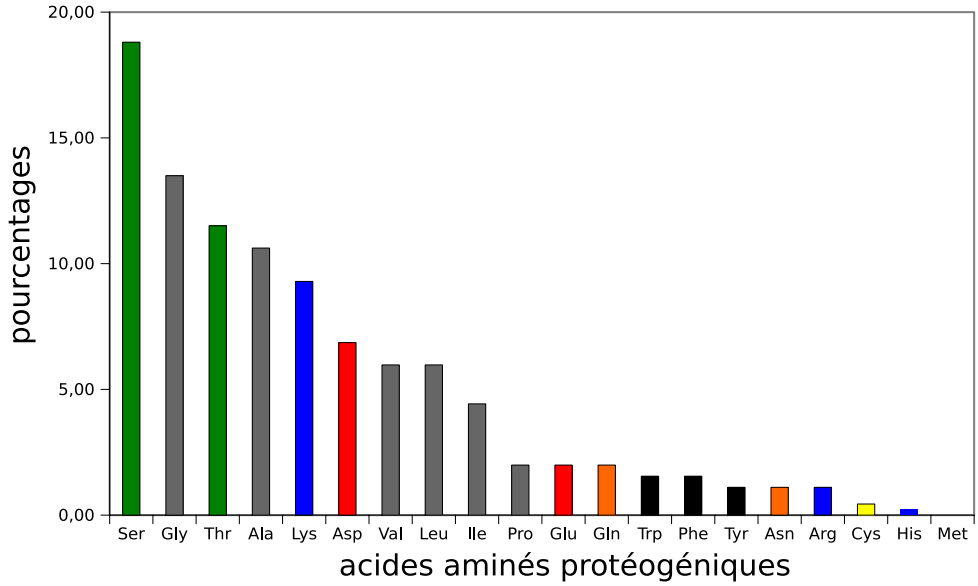


FIG. 5.10 – Répartition des acides aminés protéogéniques au sein des peptides non-ribosomiaux (« curated » et sans variants) produits par les bactéries.(gris = aliphatique, rouge = acide, vert = petit hydroxy, bleu = basique, noir = aromatique, blanc = amidé, jaune = soufré)

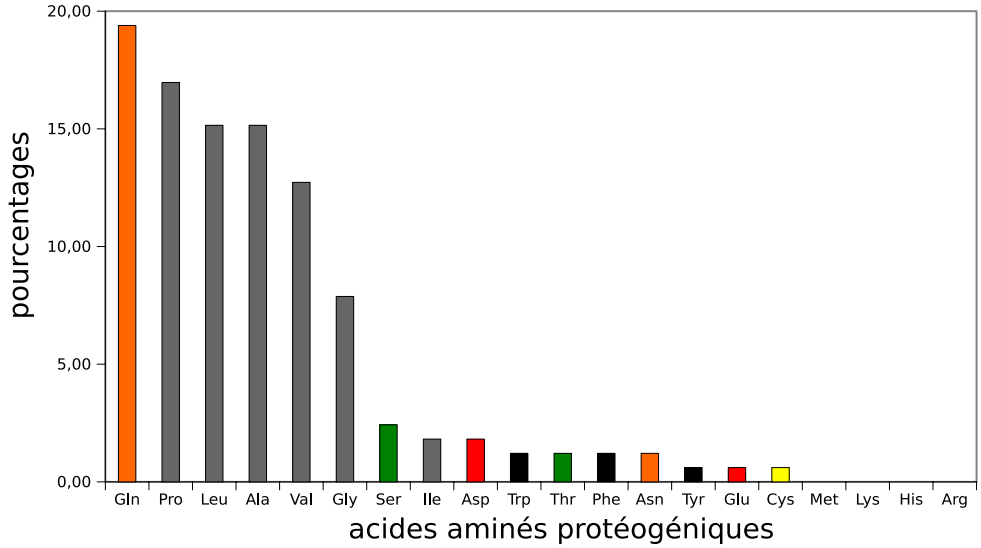


FIG. 5.11 – Répartition des acides aminés protéogéniques au sein des peptides non-ribosomiaux (« curated » et sans variants) produits par les champignons. gris = aliphatique, rouge = acide, vert = petit hydroxy, bleue = basique, noir = aromatique, blanc = amide, jaune = soufré

plus fréquent chez les champignons est beaucoup moins fréquent chez les bactéries. De même, la lysine fait partie des cinq acides aminés les plus fréquents dans le cas des bactéries alors qu'elle n'est pas identifiée au sein des peptides synthétisés par les champignons présents dans NORINE. Une fois de plus, nous avons mis en évidence une différence significative des caractéristiques des peptides non-ribosomiaux produits par les bactéries et ceux produits par les champignons.

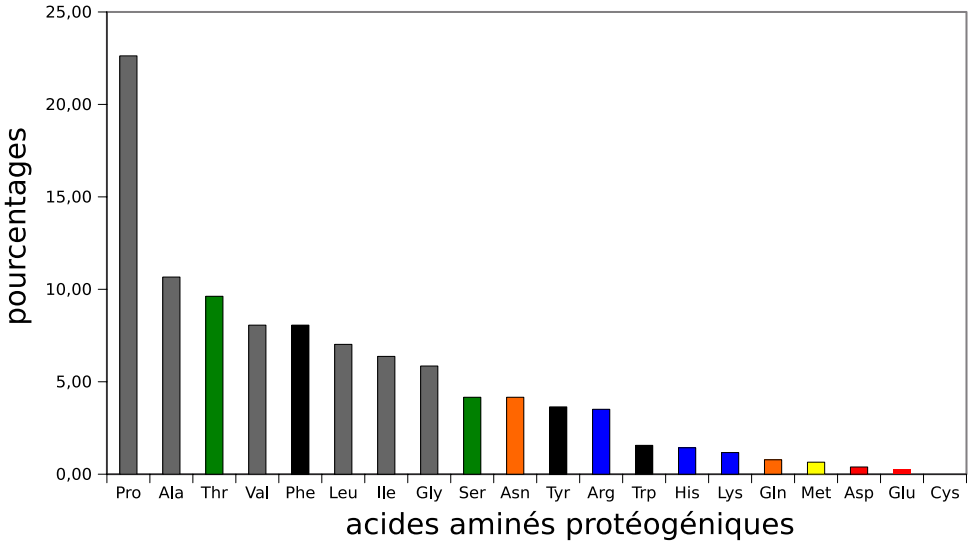


FIG. 5.12 – Répartition des acides aminés protéogéniques au sein des peptides non-ribosomiaux produits par les métazoaires. gris = aliphatique, rouge = acide, vert = petit hydroxy, bleue = basique, noir = aromatique, blanc = amide, jaune = soufré

Nous avons calculé la distribution des acides aminés protéogéniques au sein des peptides synthétisés par les métazoaires (figure 5.12). Nous présentons les résultats obtenus avec l'ensemble des peptides isolés chez les métazoaires, mais des résultats similaires sont obtenus lorsque seuls les peptides « curated » sont considérés ou uniquement un variant par famille. Nous pouvons noter que les peptides synthétisés par les métazoaires présentent des monomères qui leurs sont spécifiques. Par exemple, la phénylalanine (Phe) est beaucoup plus fréquente dans les peptides synthétisés par les métazoaires par rapport à ceux synthétisés par les bactéries ou les champignons. Cette spécificité de certains monomères explique un coefficient de corrélation différent de 1 entre les différents groupes d'organismes. Cependant, la distribution des acides aminés protéogéniques au sein des peptides isolés chez les métazoaires présente des similitudes avec celle observée chez les bactéries. Par exemple, la thréonine (Thr), l'alanine (Ala), la valine (Val), la leucine (Leu) et l'isoleucine (Ile) sont aussi fréquentes et la lysine (Lys), absente pour les champignons, est observée au sein des peptides isolés chez les métazoaires. Cependant, la proline très fréquente chez les champignons, l'est également au sein des peptides isolés chez les métazoaires. Le coefficient de corrélation plus élevé entre les bactéries et les métazoaires observé précédemment est certainement du aux autres monomères, plutôt qu'aux acides aminés protéogéniques.

Distribution de la taille des peptides

Nous avons calculé la distribution des tailles des peptides pour les peptides synthétisés par les bactéries et ceux synthétisés par les champignons. Les profils de distributions sont les mêmes quelle que soit la cohorte de peptides utilisée (total, « curated », sans variants et sans variants « curated »). Nous donnons les résultats avec l'ensemble des peptides « curated » synthétisés par les bactéries (figure 5.13) et par les champignons (figure 5.14).

Nous pouvons voir d'après les deux figures que les profils de distribution des tailles des peptides synthétisés par les bactéries et les champignons sont différents. En effet, pour les bactéries, la

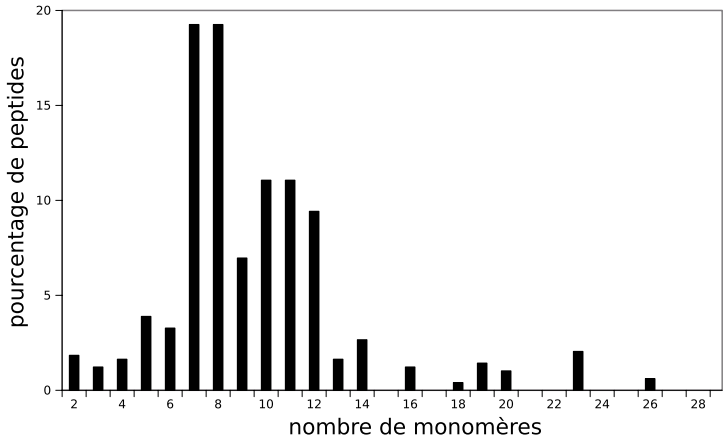


FIG. 5.13 – Distribution des tailles des peptides « curated » synthétisés par les bactéries (488 peptides).

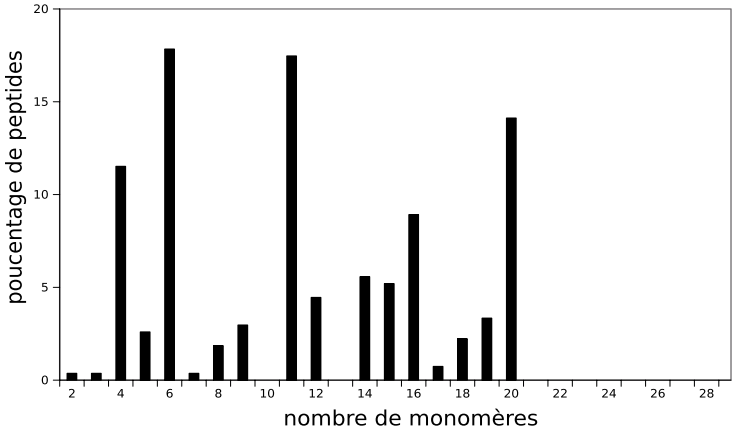


FIG. 5.14 – Distribution des tailles des peptides « curated » synthétisés par les champignons (269 peptides).

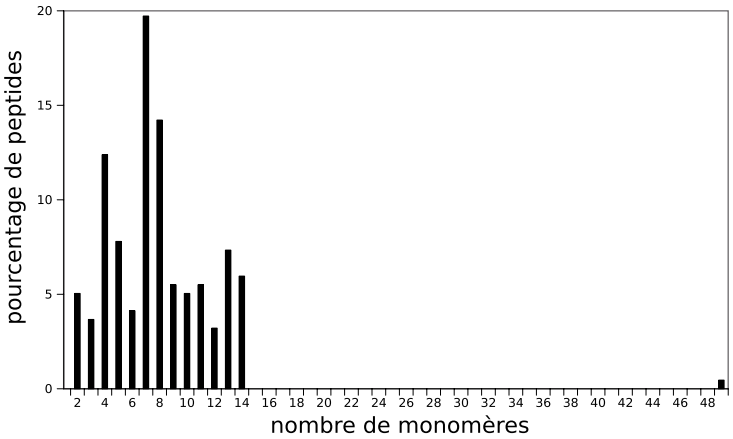


FIG. 5.15 – Distribution des tailles des peptides synthétisés par les métazoaires (218 peptides).

distribution des tailles des peptides semble être relativement homogène, avec une grande proportion des peptides présentant une taille de 7 ou 8 monomères. A l'inverse, les peptides synthétisés par les champignons semblent avoir des tailles sous-représentées et d'autres sur-représentées. En moyenne, les peptides synthétisés par les champignons sont plus grands que ceux synthétisés par les bactéries. Cette expérience montre donc qu'il existe une différence entre la répartition de la taille des peptides synthétisés par les bactéries et ceux synthétisés par les champignons.

Nous avons ensuite calculé la distribution des tailles des peptides isolés chez les métazoaires (figure 5.15). Ils ont une taille souvent située entre 2 et 14 monomères, ce qui est comparable aux tailles des peptides synthétisés par les bactéries. Les tailles les plus fréquentes observées chez les métazoaires sont 7 et 8 monomères, comme dans le cas des bactéries. La distribution des tailles des peptides synthétisés par les métazoaires est similaire à celle des tailles des peptides synthétisés par les bactéries, ce qui renforce une fois de plus, l'hypothèse selon laquelle les peptides isolés chez les métazoaires sont en fait synthétisés par des bactéries symbiotiques.

5.5 En fonction des catégories chimiques

Dans cette section, nous avons étudié la distribution des monomères au sein des différentes catégories chimiques de peptides présents dans Norine : les glycopeptides, les lipopeptides, les peptides et les peptaibols. Nous avons calculé la répartition des peptides au sein de ces différentes catégories (figure 5.16). Environ la moitié des peptides non-ribosomiaux sont des peptides et un quart sont des lipopeptides.

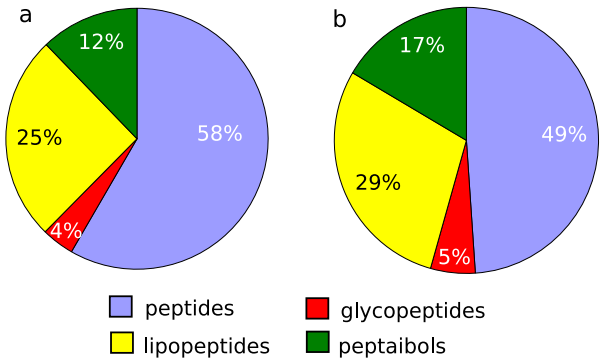


Fig. 5.16 – Répartition des catégories chimiques a) sur l'ensemble des peptides et b) sur les peptides « curated ».

Peptides

La figure 5.17 montre la distribution des trente monomères les plus fréquents au sein des peptides, c'est-à-dire sans les lipopeptides, glycopeptides et peptaibols. Seuls sont considérés les peptides « curated » et sans variants (127 peptides).

Une grande proportion de monomères apparaît sous la forme D. La sérine et la thréonine sont très fréquentes. Ces deux acides aminés permettent l'obtention des structures primaires non-linéaires. Le chromophore identifié dans les pyoverdines (ChrP) apparaît comme fréquent. En effet, dans notre étude, les pyoverdines présentant des structures non-similaires, ne sont pas considérées comme des variants et sont donc toutes prises en compte et chacune d'elle contient un chromophore. Elles représentent 45% de la catégorie chimique étudiée.

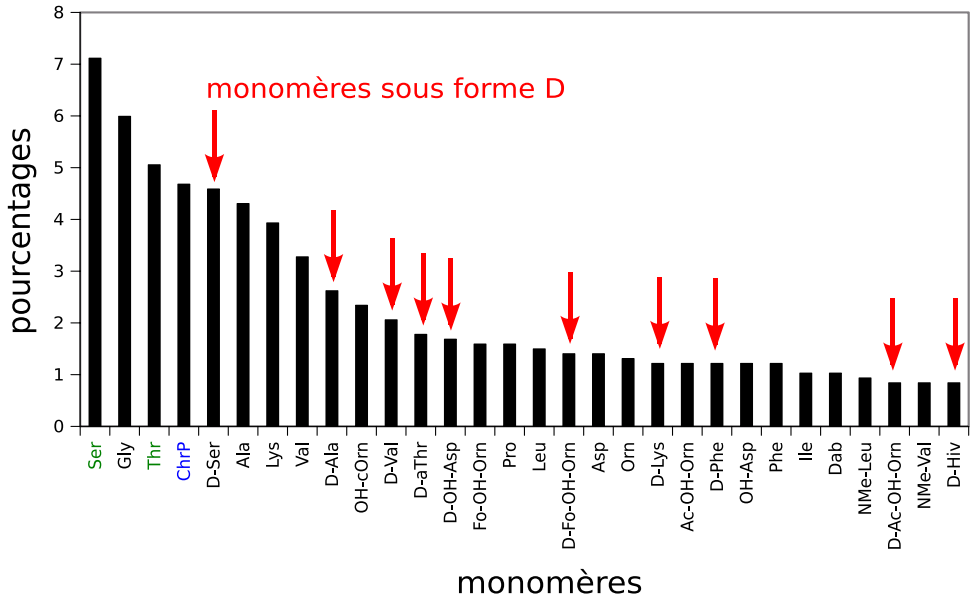


FIG. 5.17 – Distribution des trente monomères les plus fréquents au sein des peptides (non lipopeptides, glycopeptides et peptaibols) « curated » et sans variants (127 peptides)

Glycopeptides

La figure 5.18 montre la répartition des 31 monomères incorporés au sein des glycopeptides. En effet, seuls 31 monomères sont identifiés au sein des glycopeptides contenus dans NORINE.

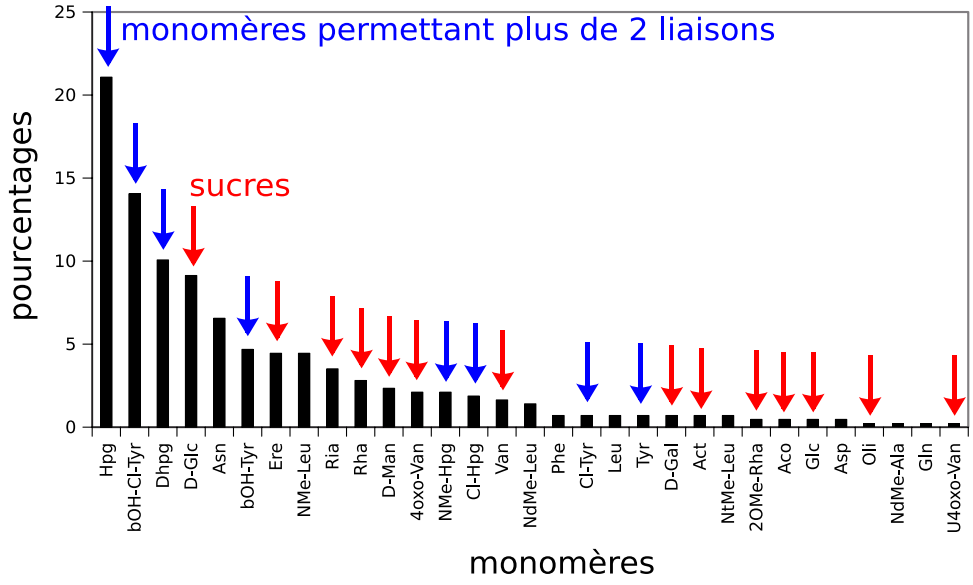


FIG. 5.18 – Distribution des 31 monomères incorporés au sein des glycopeptides « curated » (43 peptides)

Les glycopeptides forment une grande catégorie de peptides présentant des structures simi-

laires. Si nous ne considérons qu'un seul variant par famille, seuls deux peptides seront considérés. Tous les variants « curated » sont donc considérés dans ce cas (43 peptides).

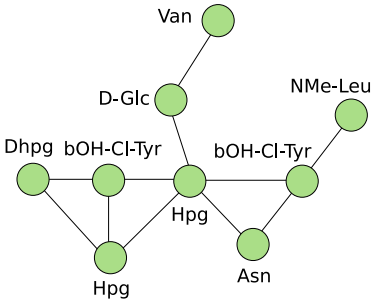


FIG. 5.19 – Structure monomérique de la vancomycine

Les glycopeptides sont des antibiotiques. L'exemple le plus connu est celui de la vancomycine (figure 5.19), un antibiotique utilisé lors d'infections causées par des bactéries à Gram positif. Les glycopeptides présentent une structure primaire complexe composée de cycles chevauchants et de branchements. Le monomère le plus fréquent est le Hpg (HydroxyPhenylGlycine). Ce monomère permet la formation de 5 liaisons covalentes, nécessaires pour la structure primaire des glycopeptides. D'autres monomères fréquemment identifiés au sein des glycopeptides sont des monomères permettant la formation de plus de deux liaisons covalentes (Tyr, Hpg ...). Bien évidemment, les sucres (D-Glc, Van, Ere, ...) sont également très fréquents au sein des glycopeptides.

Lipopeptides

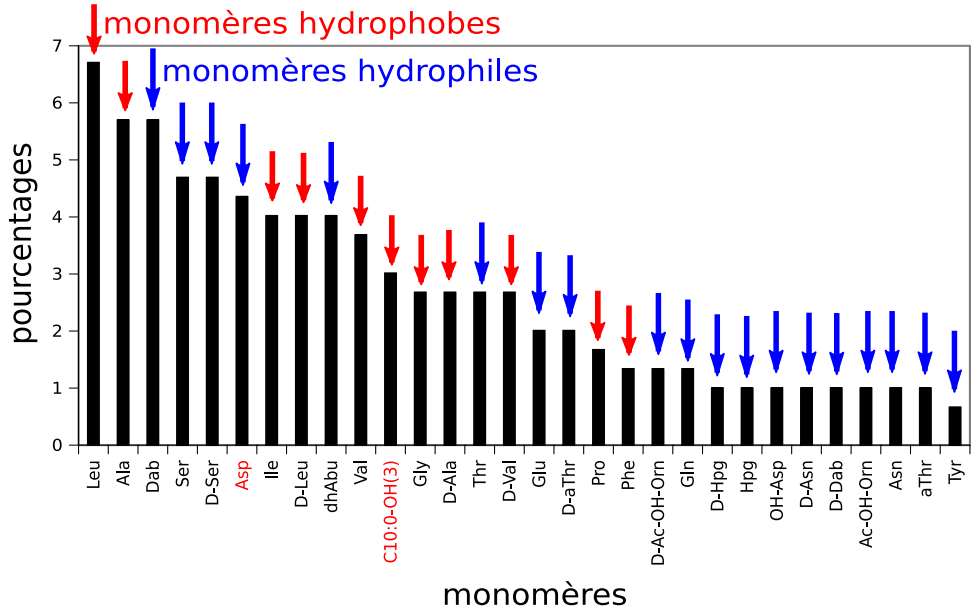


FIG. 5.20 – Distribution des trente monomères les plus fréquents au sein des lipopeptides « curated » et sans variants (26 peptides)

La figure 5.20 montre la distribution des trente monomères les plus fréquents au sein des lipopeptides. Seuls les lipopeptides « curated » et sans variants (26 peptides) sont considérés. Les lipopeptides sont constitués d'un acide gras et d'une partie peptidique. L'acide gras le plus fréquemment rencontré est l'acide 3-hydroxy-décanoïque (C10 :0-OH(3)). Les monomères les plus fréquents sont la leucine (Leu), l'Alanine (Ala) et l'acide 2,4-diaminobutyrique (Dab). L'acide aminé acide aspartique (Asp) est plus fréquent dans les lipopeptides que dans les autres peptides non-ribosomiaux. Les lipopeptides ont souvent une partie hydrophobe (dont l'acide gras) et une partie hydrophile, cette amphiphilie étant nécessaire à leur fonction. La distribution des monomères montre également une répartition selon ces critères d'hydrophobicité.

Peptaibols

La figure 5.21 montre la distribution des 27 monomères identifiés au sein des peptaibols. En effet, seuls 27 monomères sont présents au sein des peptaibols contenus dans NORINE. Seuls les peptaibols « curated » et sans variants sont considérés (20 peptides) dans cette étude.

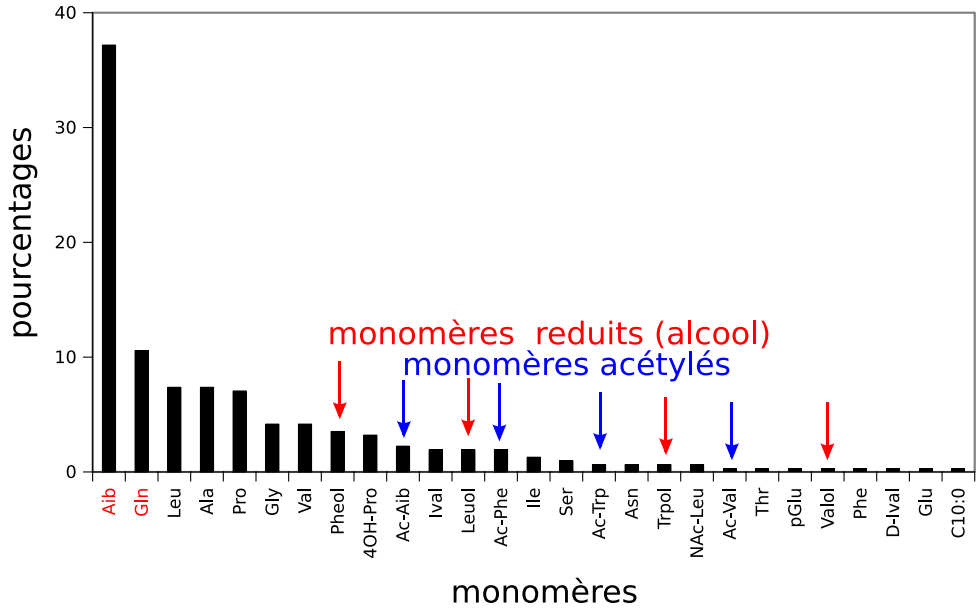


FIG. 5.21 – Distribution des 27 monomères identifiés au sein des peptaibols « curated » et sans variants (20 peptides)

Les peptaibols sont des antibiotiques linéaires synthétisés par voie non-ribosomiale par des champignons. Il sont riches en Aib (acide 2-Aminoisobutyrique) et présentent un résidu alcool à l'extrémité C-terminale. Le nom « peptAIBol » provient de la richesse en Aib et le suffixe -ol provient de la fonction alcool portée de l'extrémité C-terminale. L'Aib est, de loin, le monomère le plus fréquent au sein des peptaibols. Il est présents dans tous les peptaibols de NORINE et, en moyenne, il est présent plus de six fois dans un même peptide. L'Aib est présent dans 132 peptides de NORINE dont 130 sont des peptaibols. Les deux peptides, non peptaibols, contenant l'Aib sont la chlamydocine (NOR00623), un peptide « putative » synthétisé par le champignon *Pochonia chlamydosporia* et la microcystine LAib (NOR00816), une toxine synthétisée par les cyanobactéries *Microcystis aeruginosa*. La glutamine (Gln) est plus fréquente dans les peptaibols que dans les autres peptides non-ribosomiaux. Les acides aminés dont la fonction car-

boxyle est réduite en alcool (Leuol, Pheol, Valol, Trpol) ne sont identifiés que dans le cas des peptaibols. Nous pouvons également noter que les monomères acétylés (Ac-Aib, Ac-Phe, Ac-Val, Ac-Trp) sont fréquents dans cette catégorie de peptides. En effet, au sein des peptaibols, le premier monomère est acétylé. Très peu de données sont disponibles sur les synthétases impliquées dans la biosynthèse des peptaibols. Une étude a montré que lorsque de l'Aib est ajouté dans le milieu de culture, la production des peptaibols augmente [Chutrakul et al., 2008]. Cette observation appuie donc l'hypothèse selon laquelle l'Aib est directement incorporé par la synthétase et ne résulte pas d'une modification d'un acide aminé préalablement incorporé. Une autre étude a permis l'identification du cluster de gènes responsable de la production de peptaibols chez *Trichoderma harzianum* [Vizcaíno et al., 2006]. Cette étude a mis en évidence un domaine déhydrogénase/réductase qui remplace le domaine terminal de la thioestérase. Ce domaine est responsable de la réduction de la fonction carboxyle en fonction alcool du monomère terminal qui accompagne la libération du peptide. Dans un autre article [Wiest et al., 2002], les auteurs ont mis en évidence des domaines kétoacyl synthase et acétyltransférase qui seraient responsables de l'acétylation du premier monomère des peptaibols. Une étude sur la base de données des peptaibols présentée dans la section 2.1.2 a été menée [Whitmore and Wallace, 2004a]. Le rôle important de l'Aib dans la formation de la structure en hélice des longs peptaibols (dont le nombre de monomères et supérieur ou égal à 16) explique la fréquence élevée de ce monomère au sein des peptaibols. Cette étude a également mis en évidence que la glutamine est le seul monomère polaire identifié dans les peptaibols et que sa position au sein de la structure est cruciale pour l'activité antibiotique.

5.6 En fonction des activités biologiques

Les peptides non-ribosomiaux présentent des activités biologiques importantes et variées. Dans cette section, nous étudions les caractéristiques des peptides contenus dans Norine en fonction des différentes activités biologiques qu'ils présentent. Dans cette section, tous les variants des familles de peptides sont pris en compte car tous les variants d'une même famille ne présentent pas les mêmes activités. Par exemple, l'actinomycine D présente des activités antibiotiques et anti-tumorales alors que les autres actinomycines n'ont pas été identifiées comme des anti-tumoraux. Onze activités sont répertoriées dans Norine, dont six principales présentés par les peptides « curated » (antibiotique, immuno-modulatrice, anti-tumorale, toxine, surfactant, sidérophore). Le terme « antibiotique » regroupe les anti-bactériens, les anti-fongiques et les anti-viraux. La figure 5.22 montre la répartition des activités biologiques des peptides « curated » de Norine.

Un grand nombre de peptides sont des antibiotiques. La plupart des peptides présente plusieurs activités biologiques. Les antibiotiques peuvent également présenter une activité immuno-modulatrice ou anti-tumorale et sont parfois également des toxines ou des surfactants. L'association antibiotique/immuno-modulateur/anti-tumoral est fréquente. Par contre, les sidérophores semblent ne posséder qu'une seule activité. En effet, les sidérophores sont souvent uniquement des sidérophores et peuvent éventuellement être des surfactants, mais ils ne montrent aucune intersection avec les autres activités biologiques.

Nous avons étudié la distribution des monomères au sein des peptides en fonction des activités biologiques qu'ils présentent.

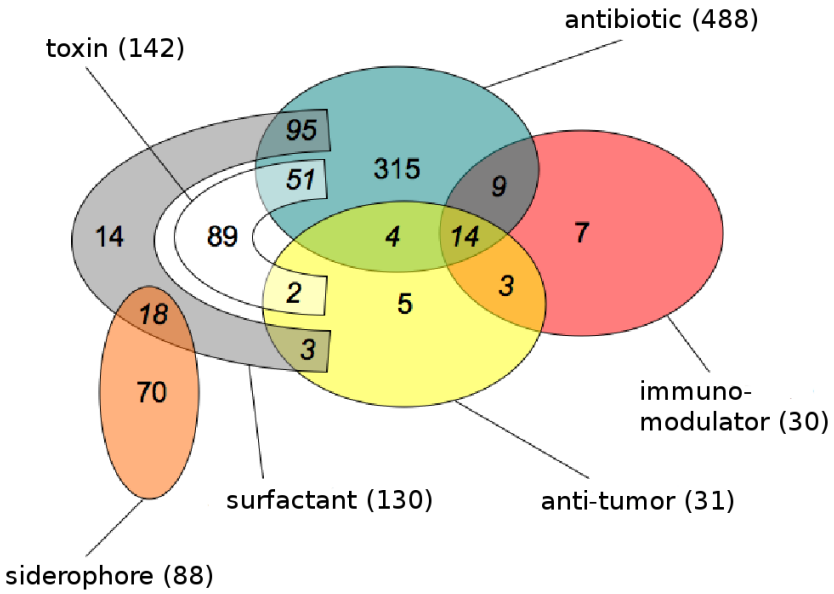


FIG. 5.22 – Répartition des activités biologiques des peptides non-ribosomiaux « curated » de NORINE

Antibiotiques

La figure 5.23 montre la distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité antibiotique (488 peptides).

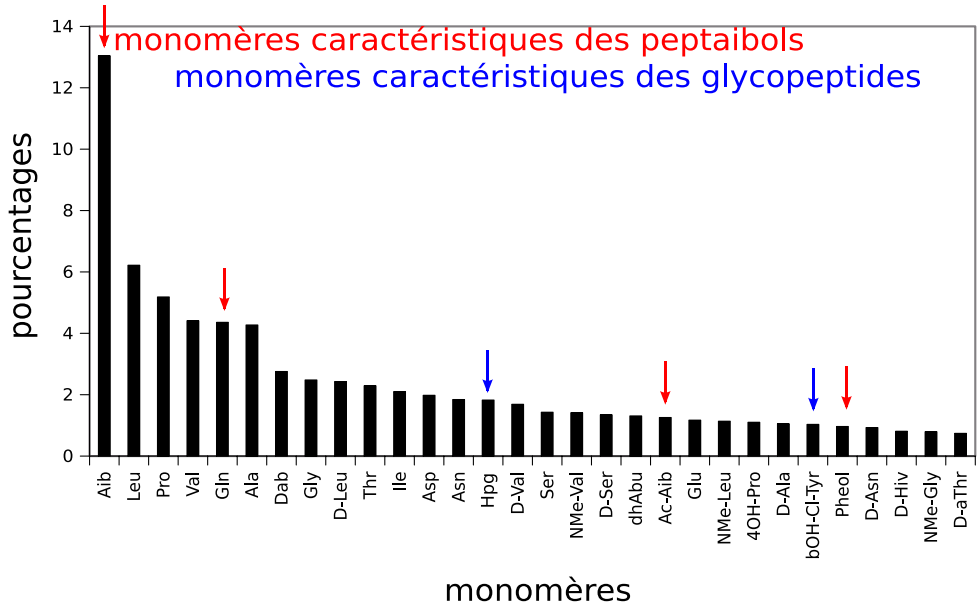


FIG. 5.23 – Distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité antibiotique (488 peptides)

L'Aib est le monomère le plus fréquent. Comme nous l'avons vu précédemment, il est caractéristique des peptaibols, une grande catégorie de peptides antibiotiques. Il n'est donc pas étonnant d'identifier ce monomère comme le plus fréquent parmi les peptides présentant une activité antibiotique. D'autres monomères caractéristiques des peptaibols sont également identifiés parmi les monomères les plus fréquents. De la même manière, certains monomères caractéristiques des glycopeptides, comme par exemple Hpg ou bOH-Cl-Tyr, apparaissent parmi les monomères les plus fréquents observés pour les peptides antibiotiques.

Nous avons recommencé l'expérience en enlevant les glycopeptides et les peptaibols de l'étude (figure 5.24).

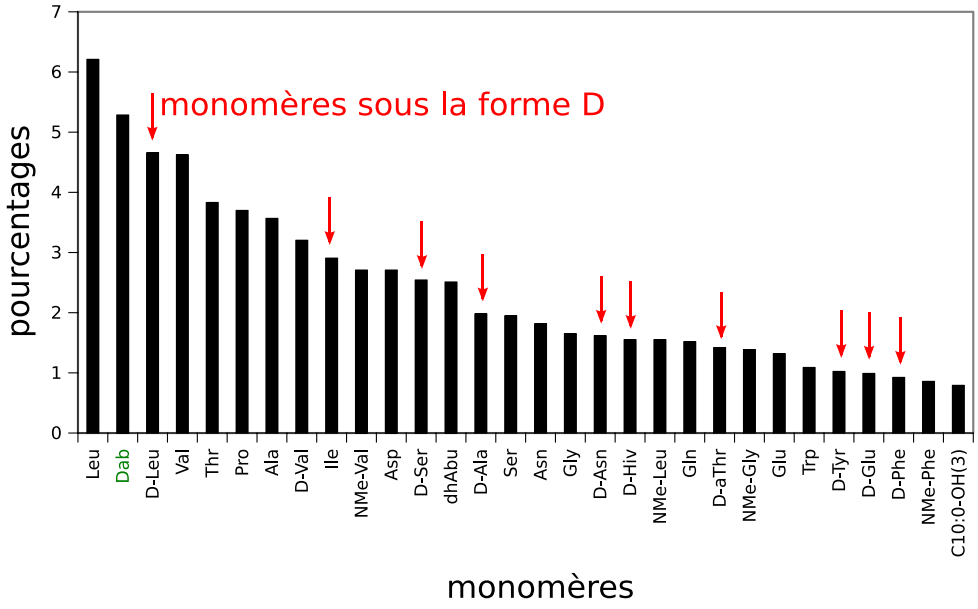


FIG. 5.24 – Distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité antibiotique en enlevant les peptaibols et les glycopeptides (295 peptides)

Le monomère Dab (acide 2,4-diaminobutyrique) est très fréquent au sein des peptides antibiotiques. En effet, les trois-quarts des peptides contenant ce monomère présentent une activité antibiotique. Nous pouvons également noter que de nombreux monomères apparaissent sous la forme D. Il a été montré que les isomères D empêchent la dégradation enzymatique de la molécule, fait essentiel pour un antibiotique, sans altérer la structure et la fonction de la molécule [Wade et al., 1990].

Anti-tumoraux

La figure 5.25 montre la distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité anti-tumorale (31 peptides).

Certains monomères apparaissent comme fréquents au sein des peptides anti-tumoraux alors qu'ils ne le sont pas dans les autres peptides non-ribosomiaux. Par exemple, l'acide lactique (Lac) ou encore la N,O-diméthyl-tyrosine (NMe-OMe-Tyr) semblent caractéristiques de cette classe de peptides. De manière générale, la distribution monomérique pour les peptides anti-tumoraux montre que beaucoup de monomères incorporés dans ces peptides subissent de nombreuses modifications. Le nombre de peptides anti-tumoraux considérés étant petit (31 peptides), ces résultats

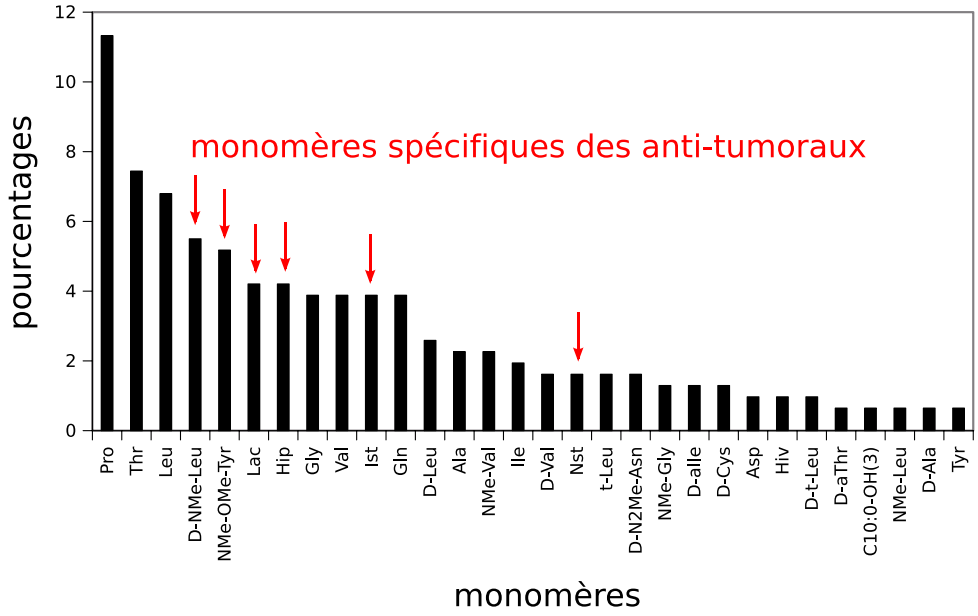


FIG. 5.25 – Distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité anti-tumorale (31 peptides)

sont à prendre avec précaution.

Immuno-modulateurs

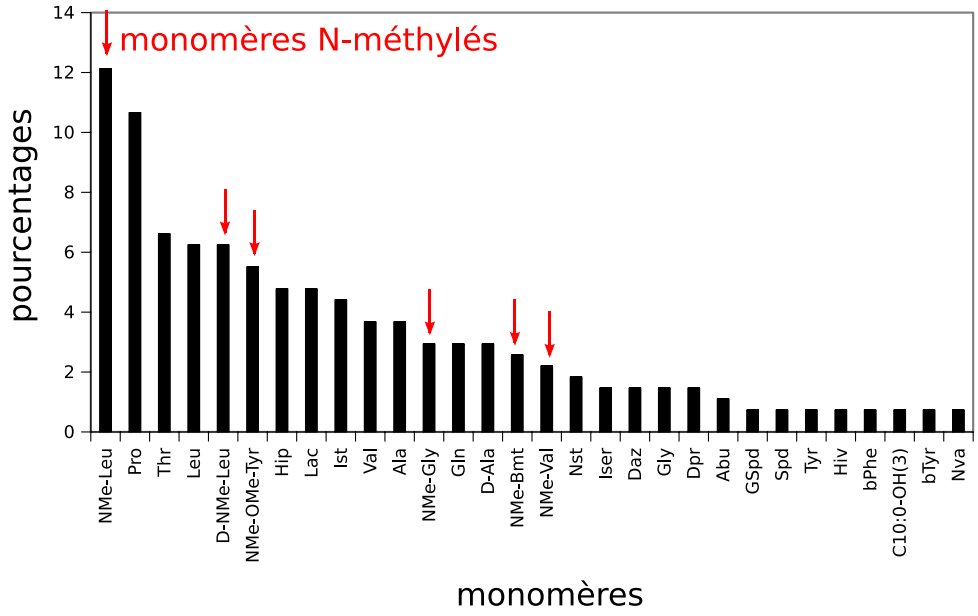


FIG. 5.26 – Distribution des trente monomères les plus fréquents au sein des peptides « curated » présentant une activité immuno-modulatrice (30 peptides)

La figure 5.26 montre la distribution des trente monomères les plus fréquents au sein des

peptides « curated » présentant une activité immuno-modulatrice (30 peptides). La distribution monomérique des immuno-modulateurs montre des monomères communs avec les anti-tumoraux (Lac par exemple). Cela s'explique par le fait qu'il existe 17 peptides communs entre ces deux classes d'activité. De nombreux monomères sont N-méthylés au sein des immuno-modulateurs. Le nombre de peptides immuno-modulateurs « curated » identifiés étant petit (30 peptides), ces résultats sont à prendre avec précaution.

Sidérophores

La figure 5.27 montre la distribution des trente monomères les plus fréquents au sein des sidérophores « curated » (88 peptides).

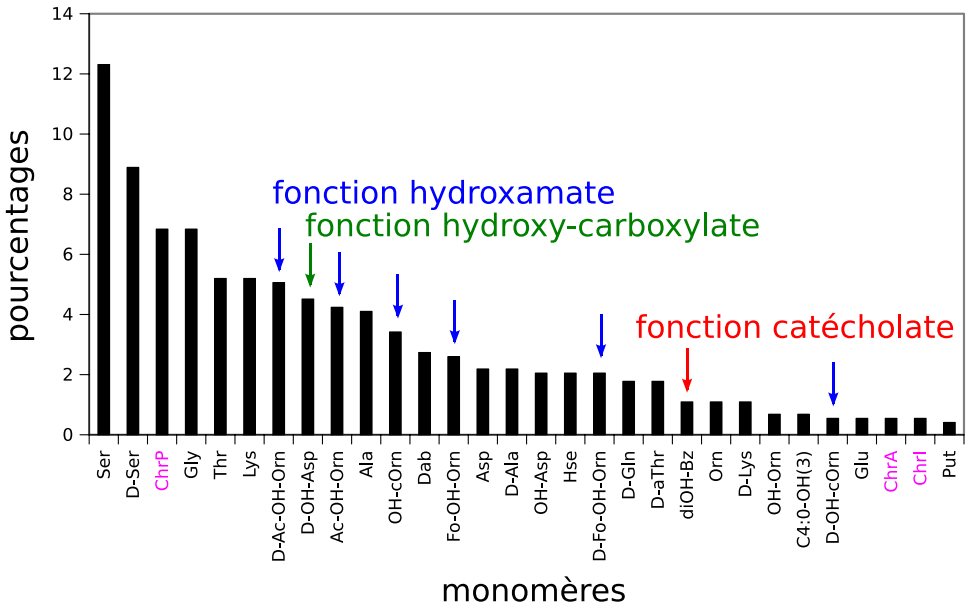


FIG. 5.27 – Distribution des trente monomères les plus fréquents au sein des sidérophores « curated » (88 peptides)

Les sidérophores sont des molécules capables de fixer le fer. Pour ce faire, elles ont besoin de six fonctions bidentates. Les catécholates, les hydroxamates et les hydroxy-carboxylates sont les fonctions les plus fréquemment rencontrées pour jouer ce rôle dans la coordination du fer. Différents monomères dans les peptides non-ribosomiaux contiennent ces fonctions :

- catécholate : 2,3-dihydroxybenzoate (Dhb ou diOH-Bz)
- hydroxamate : N-formyl-N-hydroxyornithine (Fo-OH-Orn), N-acétyl-N-hydroxyornithine (Ac-OH-Orn), N-hydroxy-cycloornithine (OH-cOrn), hydroxylysine (OH-Lys)
- hydroxy-carboxylate : acide hydroxyaspartique (OH-Asp)

Nous pouvons remarquer que tous ces monomères apparaissent parmi les monomères les plus fréquents au sein des sidérophores de NORINE. Le chromophore des pyoverdines (ChrP) est également très fréquent car les pyoverdines représentent un ensemble hétérogène de sidérophores (57 peptides) dont chaque membre contient un chromophore qui rend la molécule fluorescente.

Surfactants

La figure 5.28 montre la distribution des trente monomères les plus fréquents au sein des surfactants « curated » (130 peptides).

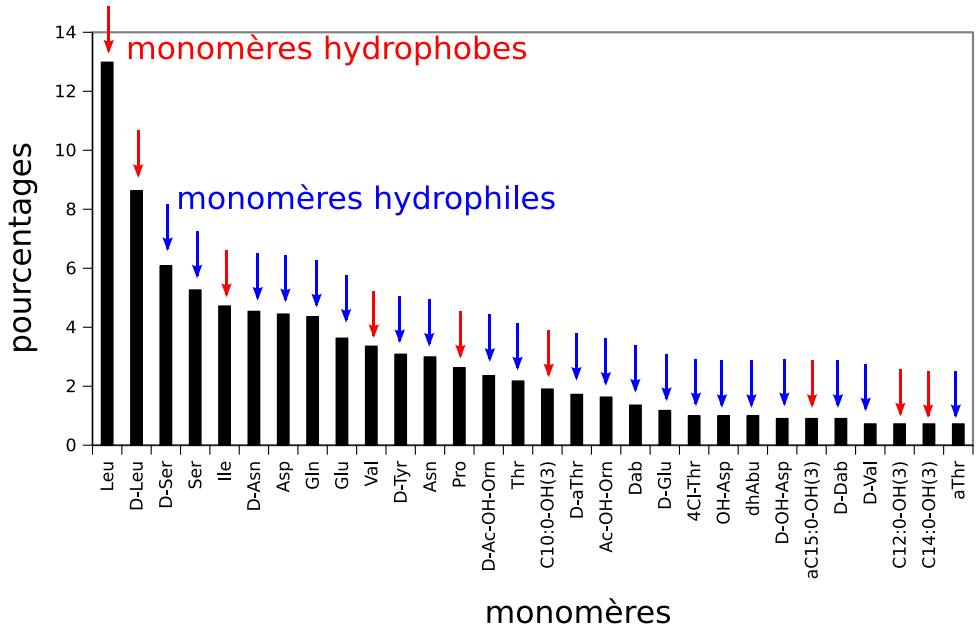


FIG. 5.28 – Distribution des trente monomères les plus fréquents au sein des surfactants « curated » (130 peptides)

Au sein des surfactants, des monomères hydrophobes (Leu, Val) et des monomères hydrophiles (Ser, Glu, Asp) figurent parmi les monomères les plus fréquents. Les surfactants ont besoin d'une partie hydrophile et d'une partie hydrophobe pour assurer leur fonction.

Toxines

La figure 5.29 montre la distribution des trente monomères les plus fréquents au sein des toxines « curated » (142 peptides).

Les monomères les plus fréquents dans les toxines sont souvent sous leur forme D. L'Adda (l'acide 3-amino-9-methoxy-2,6,8-trimethyl-10-phenyldeca-4,6-dienoïque), apparaissant comme fréquent, est le monomère reconnu comme caractéristique des microcystines [Harada et al., 2004]. Ce monomère est également présent dans les nodularines. Les microcystines et les nodularines sont des toxines produites par les cyanobactéries et présentent des structures similaires. 56 microcystines et 8 nodularines sont répertoriées dans NORINE. Certaines références citent plus de 70 microcystines différentes, mais il est difficile de trouver ces molécules. Cette étude a mis en évidence le monomère D-bMe-Asp (acide β-méthyl-D-aspartique) qui est présent dans 46 peptides de la famille des microcystines et des nodularines.

En analysant les monomères les plus fréquents au sein des différentes classes d'activités biologiques, il semble que chaque classe présente un certain nombre de monomères spécifiques. Pour

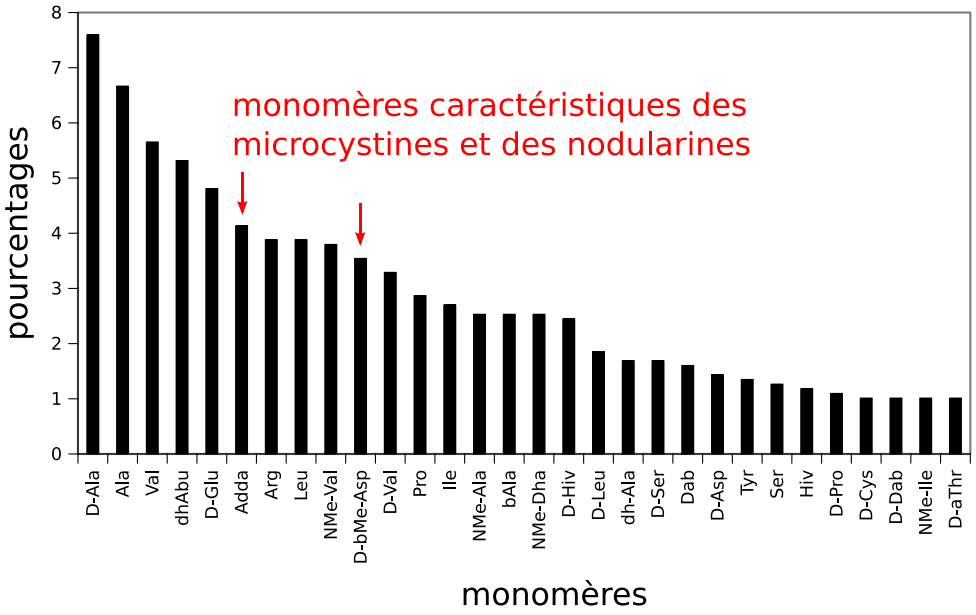


FIG. 5.29 – Distribution des trente monomères les plus fréquents au sein des toxines « curated » (142 peptides)

vérifier ces observations, nous avons calculé les coefficients de corrélation entre les distributions monomériques de chaque classe d'activité.

Corrélation entre les distributions monomériques

Nous avons calculé le CC entre les distributions des monomères en fonction des différentes activités.

Dans la table 5.3 nous pouvons remarquer que les CC sont plutôt bas, ce qui signifie que l'on a une mauvaise corrélation entre les distributions des monomères en fonction des différentes activités. En d'autres termes, les distributions monomériques sont différentes dans les groupes d'activités comparés. La colonne et la ligne « complément » correspondent à l'ensemble des peptides ne présentant pas l'activité étudiée. Nous notons aussi que les sidérophores semblent avoir une composition en monomères qui leur est propre (CC proche de 0). Nous pouvons également remarquer que les activités qui présentent des intersections avec d'autres activités ont des CC qui varient dans le même sens. Ce qui est cohérent avec l'observation des intersections puisque certains peptides sont dans plusieurs classes d'activités.

Voyons maintenant les résultats des mêmes expériences lorsque seuls les peptides présentant une activité unique sont considérés. Les résultats sont présentés dans la table 5.4.

La table 5.4 montre des résultats similaires à ceux obtenus précédemment mais avec des CC souvent plus faibles (plus proches de 0). Ces expériences semblent confirmer le fait que les sidérophores présentent une distribution monomérique qui leur est propre. Par exemple, un CC de $0,075$, qui révèle une très mauvaise corrélation, est observé entre la distribution monomérique des sidérophores et celle des toxines. Les sidérophores doivent fixer le fer et pour cela, ils ont besoin de fonctions spécifiques, et par conséquent, de monomères spécifiques.

	complément	antibio (488)	anti-tumor (31)	immuno (30)	sidero (88)	surfactant (130)	toxine (142)
complément	X	0,412	0,478	0,414	**0,203**	0,411	0,378
antibio		X	0,484	0,389	0,191	0,504	0,398
anti-tumor			X	0,799	0,143	0,442	0,317
immuno				X	0,092	0,268	0,291
sidero					X	0,344	0,177
surfactant						X	0,342

TAB. 5.3 – Coefficients de corrélation entre les différentes distributions des monomères en fonction des activités présentées par les peptide « curated ». Les chiffres entre parenthèses donnent le nombre de peptides de chaque classe.

	complément	antibio (315)	anti-tumor (5)	immuno (4)	sidero (70)	surfactant (14)	toxine (89)
complément	X	0,377	0,350	0,259	0,238	0,474	0,206
antibio		X	0,247	0,118	0,148	0,330	0,113
anti-tumor			X	0,264	0,329	0,047	0,110
immuno				X	0,060	-0,004	0,122
sidero					X	0,140	**0,075**
surfactant						X	0,194

TAB. 5.4 – Coefficients de corrélation entre les différentes distributions des monomères en fonction des activités exclusives présentées par les peptides « curated ». Le nombre de peptides de chaque classe est donné entre parenthèses.

Cette section nous a permis de mettre en évidence une distribution des monomères différente en fonction de l'activité biologique des peptides. En effet, chaque classe d'activité semble montrer une distribution spécifique des monomères. L'exemple le plus flagrant est celui des sidérophores. Nous avons également calculé la similarité des structures, c'est-à-dire les distances deux à deux, au sein des différentes classes d'activités mais aucune similarité structurale n'a pu être mise en évidence (données non présentées ici). Ces observations nous ont conduit au développement d'un outil aidant à la prédiction de l'activité biologique des peptides à partir de leur composition.

5.7 Aide à la prédiction de l'activité biologique

Dans cette section, nous présentons un outil aidant à la prédiction de l'activité biologique d'un peptide à partir de sa composition en monomères. Cet outil s'appuie sur l'hypothèse selon laquelle chaque classe d'activité présente des monomères spécifiques. Nous nous basons sur la fréquence des monomères au sein de chaque classe d'activité pour évaluer la composition des peptides présentant une activité avec lesquels la composition monomérique est la plus proche. Après avoir présenté notre méthode, nous la comparons aux prédictions obtenues grâce à diverses méthodes d'apprentissage automatique. Enfin, nous donnons quelques exemples de prédiction d'activités sur un ensemble de peptides non-ribosomiaux, qui ne sont pas encore intégrés à NORINE.

Cet outil de prédiction d'activité permet à l'utilisateur d'identifier l'activité potentielle d'un peptide prédit à partir de la synthétase ou isolé à partir d'un organisme, et ainsi orienter les validations expérimentales à entreprendre.

5.7.1 Méthode

Tout d'abord, nous avons regroupé les acides gras en trois groupes distincts :

- le groupe « OH-FA » contient tous les acides gras hydroxylés dont le nombre de carbones est supérieur à 9. Ils correspondent aux acides gras souvent présents dans les lipopeptides.
- le groupe « FA » contient tous les acides gras, ainsi que les chaînes contenant un groupement amide et/ou un groupement hydroxyl (dont le nombre de carbones est inférieur ou égal à 9)
- le groupe « FAD » contient tous les autres acides gras

Nous avons également regroupé les sucres dans un seul groupe, ainsi que les chromophores.

Pour chaque monomère, nous avons calculé le nombre d'occurrences de ce dernier au sein des six classes principales d'activité biologique : antibiotique, immuno-modulatrice, anti-tumorale, surfactant, toxine et sidérophore. Nous avons normalisé le nombre d'occurrences de chaque groupe de monomères dans le but d'obtenir un nombre total de monomères équivalent au sein de chacune des activités. En effet, comme nous l'avons vu dans la section précédente, la classe des antibiotiques est sur-représentée par rapport aux autres classes. De même, les peptides immuno-modulateurs et anti-tumoraux sont beaucoup moins représentés. Nous devons donc normaliser chaque classe d'activité dans le but d'obtenir des chiffres comparables. Après avoir normalisé le nombre d'occurrences de chaque monomère, nous calculons sa fréquence au sein des différentes classes d'activité. Cela signifie que pour un monomère donné, la somme des fréquences de ce monomère au sein des six classes est égale à 1. Nous obtenons donc un fichier contenant, pour chaque monomère, la fréquence au sein des six classes principales d'activité.

Pour prédire l'activité biologique d'un peptide, nous calculons la moyenne des fréquences de chaque monomère qui le compose, au sein des six classes d'activités. En d'autres termes, la fréquence moyenne FM dans la classe C pour un peptide composé de n monomères est de :

$$FM^C = \frac{\sum_{i=1}^{n} f_i^C}{n}$$

Dans le but de rester cohérent avec les données de départ, nous appliquons ensuite un coefficient correcteur à chacune des fréquences moyennes. Il est de $1,4$ pour les antibiotiques, ce qui revient à augmenter la fréquence moyenne obtenue pour cette classe. En effet, les antibiotiques étant sur-représentés, nous considérons que les données sont plus fiables dans cette classe et la privilégions. A l'inverse, nous appliquons un coefficient correcteur de $0,5$ aux fréquences moyennes obtenues pour les anti-tumoraux et les immuno-modulateurs car ces deux classes contenaient peu de données au départ et la normalisation peut biaiser les résultats obtenus. Le coefficient correcteur des surfactants et des toxines est de 1, ce qui signifie que l'on conserve la valeur obtenue. Enfin, nous appliquons un coefficient correcteur de $0,75$ pour les sidérophores car les effectifs de départ étaient légèrement inférieurs à ceux des surfactants et des toxines. La classe d'activité pour laquelle la fréquence moyenne est la plus élevée est celle dont le peptide étudié a la composition monomérique la plus proche et, par conséquent, la classe la plus probable à laquelle le peptide étudié appartient. Dans la suite, lorsque la fréquence moyenne est inférieure ou égale à $0,16$, nous considérons que la prédiction n'est pas fiable et ne retournons pas de résultats (« no prediction »). Nous avons fixé le seuil à $0,16$ car il représente la fréquence moyenne pour laquelle un monomère est réparti de façon égale dans les six classes d'activité (1/6).

5.7.2 Tests

Nous avons réalisé un ensemble de tests en comparant les prédictions obtenues avec notre méthode et celles obtenues avec les diverses méthodes d'apprentissage automatiques de Weka introduite dans la section 5.1. Nous avons réalisé ces expériences sur trois jeux de données différents.

Premier jeu de données

Le premier jeu de données contient des peptides « curated » de NORINE ne présentant qu'une seule activité. Bien entendu, ces peptides sont retirés des données utilisées pour calculer les fréquences ou construire le modèle d'apprentissage. Ce jeu de données contient 32 peptides répartis dans les six classes d'activité.

TAB. 5.5 – Résultats obtenus pour la prédiction de 32 peptides « curated » présentant une seule activité

	antibiotique 18	anti-tumor 1	immuno 1	sidero 4	surfactant 2	toxin 6	total 32
taux de vrais positifs							
notre méthode	1	1	1	1	1	1	1
zero	0,56	0	0	0	0	0	0,56
BayNet	0,84	0	1	0,75	1	1	0,84
J48	0,84	0	0	1	0,5	1	0,84
IBK	1	1	1	1	1	1	1
SMO	1	1	1	1	1	1	1
taux de faux positifs							
notre méthode	0	0	0	0	0	0	0
zero	0,43	0	0	0	0	0	0,43
BayNet	0,16	1	0	0,25	0	0	0,16
J48	0,16	0	0	0	0,5	0	0,16
IBK	0	0	0	0	0	0	0
SMO	0	0	0	0	0	0	0
taux de prédictions correctes							
notre méthode	1	1	1	1	1	1	1
zero	1	0	0	0	0	0	0,56
BayNet	0,89	0	1	0,75	1	0,83	0,84
J48	0,89	0	0	0,75	1	1	0,84
IBK	1	1	1	1	1	1	1
SMO	1	1	1	1	1	1	1

La table 5.5 montre les résultats obtenus pour ce jeu de données. Nous pouvons remarquer que notre méthode ainsi que la méthode basée sur les k-plus proches voisins (IBK) et celle basée sur les SVM (SMO) permettent de prédire correctement l'activité biologique des 32 peptides du test. Au contraire, la méthode basée sur la construction de règles de décision (zero) classe tous les peptides dans la classe des antibiotiques (taux de vrais positifs égal à 0 pour les autres classes) ce qui donne 56% des peptides du test classés correctement, correspondant au nombre d'antibiotiques présents dans le test. La méthode basée sur les réseaux bayésiens (BayNet) ne parvient pas à

classer correctement le peptide anti-tumoral et classe correctement 84% des données. La méthode basée sur la construction d'un arbre de décision (J48) ne permet pas de classer correctement le peptide anti-tumoral ni le peptide immuno-modulateur et classe correctement 84% des peptides du test.

Ce premier jeu de données est issu des données validées de NORINE. Nous voulons donc que la méthode de prédiction soit capable de classer correctement l'ensemble de ces données. Seules trois méthodes, dont la notre, parviennent à classer correctement l'ensemble des peptides.

Second jeu de données

Dans un second temps, nous avons réalisé les mêmes expériences avec un jeu de données composé de 69 peptides pouvant présenter plusieurs activités. Cela signifie que pour un peptide du test, différentes activités biologiques sont connues. Par exemple, un peptide peut présenter en même temps une activité antibiotique et une activité immuno-modulatrice. Lorsque l'activité prédite correspond à l'une des activités identifiées pour le peptide, la prédiction est évaluée comme correcte. Les résultats obtenus avec le jeu de données contenant 69 peptides « curated » pouvant présenter plusieurs activités sont données dans la table 5.6.

TAB. 5.6 – Résultats obtenus pour la prédiction de 69 peptides « curated »

	antibiotique	anti-tumor	immuno	sidero	surfactant	toxin	total
	54	8	6	8	21	10	69
taux de vrais positifs							
notre méthode	1	1	1	1	1	0,92	0,98
zero	0,78	0	0	0	0	0	0,78
BayNet	0,96	1	0	0,875	1	1	0,96
J48	0,97	0	1	1	0,90	1	0,96
IBK	1	1	1	1	1	0,875	0,98
SMO	0,97	1	1	1	1	1	0,98
taux de faux positifs							
notre méthode	0	0	0	0	0	0,08	0,01
zero	0,22	0	0	0	0	0	0,28
BayNet	0,04	0	0	0,125	0	0	0,03
J48	0,03	0	0	0	0,10	0	0,04
IBK	0	0	0	0	0	0,125	0,01
SMO	0,03	0	0	0	0	0	0,01
taux de prédictions correctes							
notre méthode	0,98						
zero	0,78						
BayNet	0,96						
J48	0,96						
IBK	0,98						
SMO	0,98						

Dans cette expérience, il nous est impossible de calculer le taux de prédiction correct pour chacune des classes car un peptide peut être dans plusieurs classes et, par conséquent, nous ne pouvons pas calculer le nombre de peptides présentant une activité donnée. Nous donnons donc le taux global de prédictions correctes. Nous pouvons remarquer que notre méthode permet

de classer correctement 98% des peptides du jeu de données. En d'autres termes, une seule prédiction est fausse, avec un peptide classé comme toxine alors qu'il n'est pas annoté comme tel. Les mêmes résultats sont obtenus avec la méthode IBK. La méthode SMO permet également la classification de 98% des peptides, avec un peptide classé à tort dans les antibiotiques. Nous pouvons également remarquer que ces trois méthodes classent des peptides dans chacune des six classes (taux de vrais positifs différent de 0). La méthode zero classe tous les peptides dans la classe des antibiotiques, ce qui représente 78% des peptides présentant une activité antibiotique. La méthode BayNet ne classe aucun peptide comme immuno-modulateur et classe correctement 96% des peptides (2 peptides mal classés). Enfin, la méthode J48 ne parvient pas à prédire correctement deux peptides et ne classe aucun peptide dans la classe des anti-tumoraux.

Dans ce jeu de données, nous voulons que la méthode de prédiction soit capable de prédire un maximum de peptides, mais aussi un maximum de classes. Les trois méthodes (notre méthode, IBK et SMO), déjà identifiées comme meilleures avec le jeu de données précédent, sont aussi les meilleures pour ce jeu de données. Notre méthode présente un avantage supplémentaire par rapport aux deux autres méthodes. En effet, en observant toutes les fréquences moyennes obtenues pour un peptide, et non plus la plus grande, il est souvent possible d'identifier les différentes activités présentées par un même peptide. Par exemple, la surfactine iC15 (NOR00856) est un surfactant présentant également une activité antibiotique. Les fréquences moyennes de chaque classe sont : antibiotique=0,23, anti-tumoral=0,07, immuno-modulateur=0,03, sidérophore=0,03, surfactant=0,46 et toxine=0,10. Ce peptide est donc prédit surfactant, mais la fréquence moyenne obtenue dans la classe antibiotique est également élevée ce qui nous permet de prédire ce peptide comme un surfactant présentant également une activité antibiotique. Un autre exemple est celui de [Hap2]didemnin B (NOR00387) annoté comme présentant des activités antibiotiques, immuno-modulatrices et anti-tumorales. Les fréquences moyennes obtenues sont les suivantes : antibiotique=0,12, anti-tumoral=0,22, immuno-modulateur=0,17, siderophore=0,02, surfactant=0,07,toxine=0,03. L'activité prédite sera l'activité anti-tumorale, mais les activités antibiotiques et immuno-modulatrices peuvent également être retenues car elles présentent également des fréquences moyennes élevées. Un dernier exemple est celui l'amphibactin I (NOR00726) qui est un sidérophore et un surfactant. Les fréquences moyennes obtenues sont : antibiotique=0,06, anti-tumoral=0,003, immuno-modulateur=0,0, siderophore=0,44, surfactant=0,32, toxine=0,05. Le peptide sera donc prédit comme sidérophore mais en observant les fréquences moyennes, nous pouvons penser qu'il est aussi un surfactant.

Les tests sur ce jeu de données ont permis de montrer la validité de notre méthode ainsi que son avantage par rapport aux autres méthodes.

Troisième jeu de données

Un dernier test a été réalisé sur l'ensemble des peptides de NORINE dont le statut est « putative » et dont les activités sont connues et figurent parmi les six catégories principales d'activité, c'est-à-dire 164 peptides. La figure 5.7 montre les résultats obtenus avec ce jeu de données.

Nous pouvons noter que dans cette expérience, le taux de vrais positifs n'est pas égal au taux de prédictions correctes dans le cas de notre méthode. En effet, avec ce jeu de données, notre méthode ne retourne pas de prédiction pour 17 peptides car la fréquence moyenne la plus élevée est inférieure à $0,16$, seuil que nous avons fixé au départ (section méthode). Ce jeu de données contient des peptides non-ribosomiaux putatifs et, par conséquent, il est possible que certains d'entre eux ne soient pas synthétisés par la voie non-ribosomiale, et peuvent donc fausser les résultats de prédiction. De plus, cet ensemble de peptides ne présente aucune corrélation avec les

TAB. 5.7 – Résultats obtenus pour la prédiction de 164 peptides « putative »

	antibiotique 87	anti-tumor 29	immuno 3	sidero 1	surfactant 1	toxin 89	total 164
taux de vrais positifs							
notre méthode	0,73	1	0	0,11	0	0,93	0,73
zero	0,53	0	0	0	0	0	0,53
BayNet	0,5	0	0	0,04	0	0,71	0,44
J48	0,74	0	0	1	0	0,57	0,35
IBK	0,53	1	0	0	0	0,46	0,46
SMO	0,50	1	0	0,07	0	0,9	0,52
taux de faux positifs							
notre méthode	0,27	0	1	0,89	1	0,07	0,27
zero	0,47	0	0	0	0	0	0,47
BayNet	0,5	1	0	0,96	1	0,29	0,56
J48	0,26	0	1	0	1	0,43	0,65
IBK	0,47	0	1	0	1	0,54	0,54
SMO	0,50	0	1	0,93	0	0,1	0,48
taux de prédictions correctes							
notre méthode	0,65						
zero	0,53						
BayNet	0,44						
J48	0,35						
IBK	0,46						
SMO	0,52						

données utilisées pour l'apprentissage ou le calcul des fréquences, ce qui fait de ce jeu de données un test robuste. Nous pouvons remarquer que le taux de vrais positifs est élevé avec notre méthode $(0,73)$ et est très supérieur à ceux obtenus avec les différentes méthodes d'apprentissage. En effet, cela signifie que dans 73% des cas où notre méthode prédit une activité, celle-ci est bien l'une identifiée pour le peptide. Pour ce jeu de données, les méthodes d'apprentissage prédisent de 35 à 53 % des activités correctement. De plus, ce jeu de données contient un sidérophore, or, nous avons vu dans la section précédente que les sidérophores présentent des monomères très spécifiques. Nous voulons donc que la prédiction du sidérophore soit correcte, ce qui est le cas avec notre méthode, SMO, BayNet et J48.

Ces tests permettent de valider notre méthode de prédiction basée sur la fréquence des monomères au sein des différentes classes d'activités. Nous avons également procédé à d'autres validations en utilisant notre méthode de prédiction sur des peptides non-ribosomiaux qui ne sont pas encore intégrés dans NORINE.

5.7.3 Exemples de prédictions

Dans cette section, nous avons recherché des peptides non-ribosomiaux qui ne sont pas encore intégrés à NORINE, puis nous avons soumis ces peptides à notre outil de prédiction d'activité dans le but de valider notre méthode.

Les **trichotoxines** sont des peptaibols synthétisés par le champignon *Trichoderma asperellum* et présentant des activités antibiotiques [Chutrakul et al., 2008]. Nous avons soumis la tri-

chotoxine 1704E à notre outil de prédiction. Les résultats sont les suivants : antibiotique=0,89, anti-tumoral=0,04, immuno-suppresseur=0,04, siderophore=0,04, surfactant=0,09, toxine=0,06. D'après la fréquence moyenne très élevé obtenue pour la classe antibiotique, nous pouvons conclure que ce peptide est un antibiotique, ce qui correspond à l'activité mise en évidence de manière expérimentale.

Les **fuscachelines** sont des sidérophores synthétisés par les bactéries *Thermobifida fusca* [Dimise et al., 2008]. Nous avons soumis la fuscacheline A à notre outil de prédiction et avons obtenu les résultats suivants : antibiotique=0,37, anti-tumoral=0,06, immuno-modulateur=0,02, siderophore=0,40, surfactant=0,03, toxine=0,01. D'après ces résultats, ce peptide est correctement prédit comme un sidérophore. Nous pouvons également remarquer que ce peptide contient des monomères identifiés au sein des antibiotiques, cependant, nous n'avons trouvé aucune information concernant cette activité au sein de la littérature, bien que certains sidérophores, comme par exemple l'asterobactine [Nemoto et al., 2002], présentent une activité antibiotique.

La **thiocoraline** est un peptide synthétisé par les bactéries du genre *Micromonospora* et connue pour ses propriétés anti-tumorales et antibiotiques [Romero et al., 1997]. Les résultats obtenus sont les suivants : antibiotique=0,55, anti-tumoral=0,06, immuno-modulateur=0,01, siderophore=0,33, surfactant=0,0, toxine=0,0 Ces résultats sont cohérents avec l'activité antibiotique identifiée pour ce peptide, cependant, l'activité anti-tumorale n'est pas prédite.

Les **mannopeptimycines** sont des glycopeptides antibiotiques synthétisés par la bactérie *Streptomyces hygroscopicus* [Magarvey et al., 2006]. Les résultats obtenus pour l'un des variants sont : antibiotique=0,83, anti-tumoral=0,03, immuno-modulateur=0,01, siderophore=0,08, surfactant=0,03, toxine=0,06. Ces résultats sont cohérents avec l'activité biologique des mannopeptimycines.

La **glidobactine** est un peptide présentant des activités antiobiotiques et anti-tumorales synthétisé par les bactéries *Polyangium brachysporum* [Schellenberg et al., 2007]. Ce peptide est composé d'un acide gras, une thréonine, une hydroxylysine et l'acide 2-amino pentenoïque. Les résultats obtenus sont les suivants : antibiotique=0,17, anti-tumoral=0,06, immuno-modulateur=0,04, siderophore=0,19, surfactant=0,29, toxine=0,11. La prédiction effectuée pour ce peptide est fausse puisqu'elle le prédit surfactant. Cela s'explique par le fait que le peptide est petit (seulement quatre monomères) et que l'un des monomère n'est pas inclus dans NORINE (hydroxylysine) et n'est donc pas prise en compte. La prédiction se fait donc sur seulement trois monomères, ce qui rend la tâche difficile.

Dans cette section, nous avons recherché des peptides non-ribosomiaux « curated » qui ne sont pas encore intégrés dans NORINE afin de valider notre méthode de prédiction d'activité biologique. Comme le montrent ces résultats, notre méthode permet de prédire correctement, dans la plupart des cas, l'activité biologique d'un peptide non-ribosomial. Cependant, pour obtenir une prédiction correcte, un nombre suffisant de monomères inclus dans NORINE sont nécessaires. Les peptides traités ici seront prochainement ajoutés à NORINE.

Dans ce chapitre, nous avons réalisé un ensemble d'analyses statistiques sur les données contenues dans NORINE. *Ces analyses sont les premières à traiter un nombre aussi important de peptides non-ribosomiaux et de monomères. Elles nous ont permis de mettre en évidence des différences, de composition monomérique et de taille, entre les peptides synthétisés par les*

bactéries et ceux synthétisés par les champignons. Nous avons également montré que la composition en monomères des peptides isolés chez les métazoaires, notamment des éponges, est similaire à celle des peptides synthétisés par les bactéries, ce qui est cohérent avec l'hypothèse selon laquelle les peptides isolés chez les métazoaires sont en fait synthétisés par des bactéries symbiotiques. Enfin, les expériences ont montré que la composition monomérique peut être spécifique d'une activité biologique. Cela nous a conduit à l'élaboration d'un outil aidant à la prédiction de l'activité biologique à partir de la composition en monomères. Cet outil permet d'identifier l'activité potentielle d'un peptide étudié et ainsi orienter les validations expérimentales à entreprendre.

Chapitre 6

Analyse de peptides non-ribosomiaux putatifs dans les génomes

De nombreuses études portent sur la découverte de molécules bioactives synthétisées par la voie non-ribosomiale. Dans ces études, le protocole consiste tout d'abord à identifier un cluster de gènes codant potentiellement des synthétases NRPS, puis prédire le peptide produit. Ensuite, les expériences biologiques confirment ou non la production de ce peptide, sa structure et sa fonction biologique (pour des exemples récents, voir [Challis, 2008, Tobiasen et al., 2007, de Bruijn et al., 2007]).

Dans cette section, nous avons suivi une démarche comparable, allant de la recherche de gènes NRPS putatifs jusqu'à la prédiction du produit final, puis nous avons utilisé les divers outils de Norine pour mettre en évidence des caractéristiques biologiques des peptides prédits. Norine nous a permis d'identifier très rapidement certaines caractéristiques très intéressantes qui peuvent orienter les expériences biologiques de l'étape suivante. Après avoir introduit le protocole général, nous donnerons quelques exemples pour lesquels Norine nous a aidé à déterminer des caractéristiques biologiques intéressantes pour des peptides putatifs chez *Lactococcus lactis* et *Pseudomonas entomophila*.

6.1 Protocole général

La première étape est l'identification de gènes NRPS putatifs. Pour ce faire, nous recherchons au sein de génomes dont la séquence complète est disponible dans les banques EMBL/Genbank/DDBJ, des gènes codant potentiellement des synthétases NRPS soit par différents mots-clés (par exemple « non-ribosomal », « NRPS » ou encore « adenylation »), soit en recherchant grâce à BLAST, les régions des génomes présentant des similarités avec une synthétase connue. En étudiant l'environnement d'un gène, nous identifions les autres gènes pouvant appartenir au cluster. Une fois le cluster de gènes identifié, nous stockons les séquences protéiques correspondantes pour chacun des gènes du cluster.

La seconde étape est la prédiction du peptide synthétisé à partir de la séquence protéique de la synthétase. Nous utilisons les différents logiciels de prédiction présentés dans la section 2.2 : NRPSpredictor, PKS/NRPS analysis Web site (AWS) et NRPS/PKS. A partir des résultats obtenus avec ces trois logiciels, nous déduisons la structure des synthétases (c'est-à-dire les modules et les domaines qui la composent) ainsi que le peptide putatif le plus probable, en intégrant les connaissances sur les domaines présents. En effet, les logiciels de prédiction donnent les monomères

sélectionnés par les domaines d'adénylation, mais il faut ajouter les modifications apportées par les domaines secondaires, comme par exemple le domaine d'épimérisation qui donnera la forme D du monomère dans le produit final.

Une fois le peptide putatif obtenu, nous utilisons NORINE pour mettre en évidence des caractéristiques biologiques pour ce peptide. Tout d'abord, nous recherchons la structure au sein de NORINE afin de savoir si le peptide a déjà été identifié. Ensuite, nous recherchons les peptides présentant une composition en monomères proche de celle du peptide putatif, en augmentant progressivement le nombre d'erreurs possibles. Nous recherchons ensuite la structure ou une sous-structure commune entre le peptide étudié et les peptides de la base de données, à l'aide de la recherche de motifs. La recherche de peptides similaires nous permet ensuite d'identifier des peptides présentant des caractéristiques (composition monomérique et structure) proches du peptide étudié. Enfin, nous utilisons notre outil d'aide à la prédiction de l'activité biologique pour identifier l'activité potentielle du peptide d'intérêt. Nous pouvons ainsi comparer les informations obtenues à l'aide des différents logiciels afin d'améliorer la qualité de nos prédictions.

6.2 Analyse du génome de *Lactococcus lactis*

Chez *Lactococcus lactis*, aucun cluster de gènes codant des synthétases dont le produit est connu n'a été identifié. Dans l'article [Siezen et al., 2008], les auteurs comparent les génomes de différentes souches de *L. lactis* disponibles dans les banques de données, ainsi que des séquences génomiques de deux autres souches, *L. lactis KF147 et L. lactis KF282*, qu'ils ont séquencés. Ils ont mis en évidence des gènes codant potentiellement des synthétases NRPS au sein des deux souches KF147 et KF282. Ils ont réalisé une étude préliminaire sur ces séquences. Elles présentent une forte homologie avec des synthétases de *Bacillus* (jusqu'à 40% d'homologie) qui produisent de la bacitracine (un antibiotique), la bacillibactine (un sidérophore) et la bacillomycine (un lipopeptide antibiotique et surfactant). Ils ont également identifié un cluster de 5 gènes qui code pour un système de transport de type ABC, présentant une forte homologie avec le système de *B. subtilis* impliqué dans le transport de l'entérobactine et de la bacillibactine. L'homologie de séquence avec la synthétase d'un sidérophore ainsi que la présence d'un cluster pour le système de transport ABC amènent les auteurs à la conclusion que le peptide synthétisé par le cluster NRPS est certainement un sidérophore. Cependant, il est couramment admis que chez les *Lactococcus*, il n'y a pas de sidérophores [Pandey et al., 1994], par conséquent, l'hypothèse d'un sidérophore est peu probable. Ils précisent également que le peptide produit peut également être un lipopeptide surfactant, sans avancer aucun argument supplémentaire.

Les séquences du cluster NRPS étudié dans cet article ne sont malheureusement pas disponibles. Cependant, nous avons obtenu par les auteurs de l'article la prédiction partielle du peptide produit qui est la suivante : Leu_D-Leu_Asp_D-Asn_D-Asp. Grâce à des recherches dans NORINE, nous avons pu étudier ce peptide partiel. Les surfactines présentent une sous-structure commune de 3 monomères avec la structure de départ. Les surfactines sont des lipopeptides antibiotiques et des surfactants. Trois peptides, l'amphisine, l'arthrofactine et la lokisine, ont 4 monomères en commun avec les 5 monomères de la structure étudiée. Ces trois peptides sont des lipopeptides antibiotiques et des surfactants. Notre outil de prédiction d'activité biologique fournit les résultats suivants : antibiotique=0,24, anti-tumoral=0,05, immuno-modulateur=0,02, siderophore=0,04, surfactant=0,48, toxine=0,16. D'après ces résultats, le peptide serait un surfactant et un antibiotique. Tous les résultats obtenus à l'aide de NORINE aboutissent à la conclusion que le peptide produit par le cluster n'est probablement pas un sidérophore, mais plutôt

un surfactant présentant des activités antibiotiques.

Dans cette section, nous avons analysé un peptide putatif partiel. Dans l'article correspondant, les auteurs concluent que ce peptide est certainement un sidérophore, sur base de l'homologie de séquences, bien que les *Lactococcus* ne soient pas connus pour synthétiser des sidérophores. D'après les résultats obtenus avec les divers outils de NORINE, nous sommes parvenus à la conclusion que ce peptide est probablement un antibiotique et un surfactant, plutôt qu'un sidérophore, ce qui semble plus cohérent avec les caractéristiques des *Lactococcus*. Cet exemple montre donc qu'il est difficile de prédire la fonction d'un peptide, uniquement à partir de l'analyse de la séquence des synthétases. En effet, des séquences éloignées peuvent conduire à la synthèse de peptides similaires et inversement, des séquences similaires peuvent conduire à la synthèse de peptides différents. En effet, les auteurs ont identifié trois synthétases de *B. subtilis* présentant une forte homologie avec leur séquence. Ces trois synthétases produisent des peptides présentant des caractéristiques très différentes (un sidérophore et deux antibiotiques dont un surfactant), ce qui prouve que des séquences de synthétases similaires ne produisent pas forcément des peptides similaires. De plus, dans l'article de Stachelhaus *et al.* [Stachelhaus et al., 1999] décrit dans la section 2.2.2, les auteurs ont constaté que les séquences des domaines d'adénylation se regroupent en fonction de l'espèce et non de la spécificité des domaines. Nous avons donc démontré à travers cet exemple, l'utilité de NORINE pour l'analyse d'un peptide putatif.

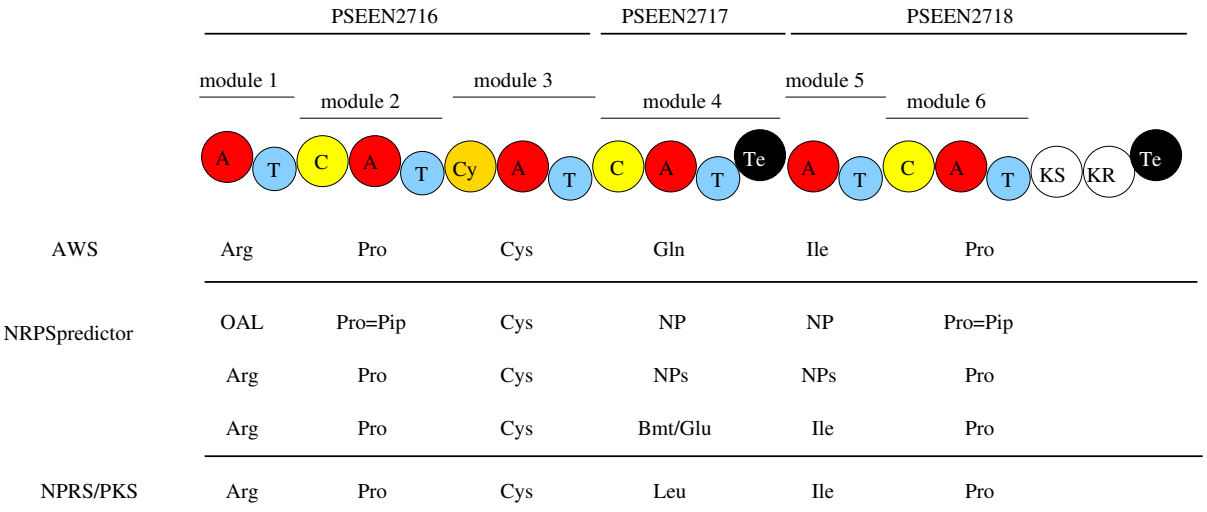
6.3 Analyse du génome de *Pseudomonas entomophila*

Nous avons recherché au sein de UniprotKB des synthétases putatives et en avons identifié deux (Q1I964 et Q1I963) chez *Pseudomonas entomophila*. Nous avons donc décidé d'étudier ce génome en détail. Le génome de *Pseudomonas entomophila* a été séquencé [Vodovar et al., 2006] (GenBank CT573326) et des gènes putatifs codant des NRPS ont été prédits. Nous avons étudié ce génome et avons identifié quatre clusters de gènes codant potentiellement des synthétases NRPS, que nous avons analysés.

6.3.1 Cluster 1

Le premier cluster identifié est composé de trois gènes situés sur le brin sens aux loci PSEEN2716, PSEEN2717 et PSEEN2718. Les gènes situés aux loci PSEEN2716 et PSEEN 2717 sont annotés comme *non-ribosomal peptide synthetase* et le gène situé au locus PSEEN2718 est annoté comme *polyketide synthase (terminal component)*. Nous avons obtenu les séquences protéiques correspondant à ces trois gènes et avons utilisé les logiciels pour prédire l'organisation des synthétases et le peptide produit (figure 6.1).

D'après les résultats obtenus, le produit final est un hybride NRPS/PKS et contient au moins 6 monomères. Nous pouvons noter la présence d'un domaine de cyclisation (Cy) qui forme un groupement thiazole (Thz) par hétérocyclisation d'une cystéine. Des domaines Cy sont identifiés dans les synthétases de sidérophores (pyoverdines, pyocheline), d'anti-tumoraux (épothilone, bléomycines) ou encore de toxines (curacine A). La structure prédite à l'aide de l'analyse des synthétases est la suivante : Arg_Pro-Thz_X_Ile_Pro. Aucun peptide de NORINE ne présente la même structure. Toutefois, un peptide cyclique possédant une structure très proche ([Phe_Pro-Thz_Phe_MeOx-Ile_Pro]), la ceratospongamide (NOR00578), a été identifié. La ceratospongamide est un peptide « putative » anti-inflammatoire isolé à partir d'une algue vivant en symbiose avec une éponge [Tan et al., 2000] et est hypothétiquement synthétisé par la voie non-ribosomiale [Deng and Taunton, 2002]. Les résultats obtenus pour la prédiction de l'activité



AWS	Arg	Pro	Cys	Gln	Ile	Pro
NRPSpredictor	OAL	Pro=Pip	Cys	NP	NP	Pro=Pip
	Arg	Pro	Cys	NPs	NPs	Pro
	Arg	Pro	Cys	Bmt/Glu	Ile	Pro
NPRS/PKS	Arg	Pro	Cys	Leu	Ile	Pro

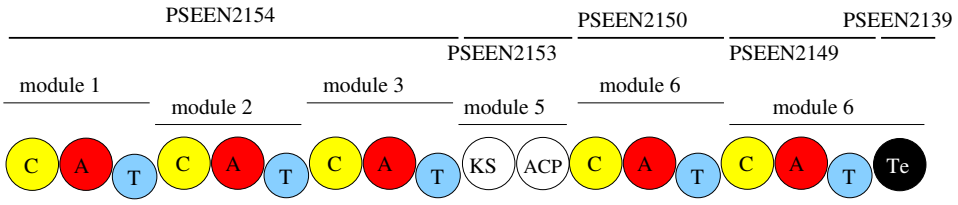
FIG. 6.1 – Prédiction du peptide synthétisé par le cluster 1. NP=Gly,Ala,Val,Leu,Ile,Abu,Iva ; NPs=Val,Leu,Ile,Abu,Iva ; AOL=Arg,Orn,Lys

biologique sont les suivants : antibiotique=0,12, anti-tumoral=0,06, immuno-modulateur=0,03, siderophore=0,0, surfactant=0,14, toxine : 0,30. La composition du peptide prédit est proche de celle des toxines. Ce cluster de trois gènes a également été identifié chez *Pseudomonas syringae pv. phaseolicola 1448A* (PSPPH1749-1751) [Joardar et al., 2005], mais aussi chez *Pseudomonas fluorescens Pfo-1* et chez *Pseudomonas syringae pv. syringae B728a* (Psyr1792–1794) [Feil et al., 2005]. Cependant, le produit synthétisé par ce cluster n'est toujours pas identifié. Dans l'article concernant *P. entomophila*, ce cluster est annoté comme produisant un lipopeptide (lipopeptide III), ce qui n'est pas cohérent avec l'analyse que nous avons effectuée, ni avec l'organisation des synthétases (absence de domaine de condensation au sein du premier module, souvent caractéristique de la synthèse d'un lipopeptide et présence d'un domaine Cy).

6.3.2 Cluster 2

Le second cluster contient trois gènes annotés *non-ribosomal peptide synthetase* et situés aux loci PSEEN2149, PSEEN2150 et PSEEN2154, ainsi qu'un gène annoté *polyketide synthase* situé au locus PSEEN2153 et un gène annoté *thioesterase* au locus PSEEN2139. Tous les gènes sont situés sur le brin anti-sens. Grâce aux séquences protéiques correspondantes, nous avons prédit l'organisation des synthétases et le peptide produit (figure 6.2).

D'après les résultats obtenus, le peptide synthétisé par le cluster 2 aurait la structure suivante : Gly_Thr_Ile_X_Glu. Cette structure n'est pas présente dans NORINE mais 26 peptides présentent une structure commune de 3 monomères avec la structure étudiée. La plupart de ces peptides sont des lipopeptides antibiotiques (surfactines, syringafactines, A54145, putisolvines). Trois peptides (A54145 A, A54145 A1 et A54145 D) contiennent les 4 monomères prédits (Gly, Thr, Ile et Glu). Ces trois peptides sont des lipopeptides antibiotiques. La prédiction de l'activité biologique donne les résultats suivants : antibiotique=0,22, anti-tumoral=0,09, immuno-modulateur=0,04, siderophore=0,14, surfactant=0,29, toxine=0,07. D'après les résultats, le peptide produit serait un surfactant et un antibiotique, ce qui est cohérent avec les résultats précédents. Dans l'article concernant *P. entomophila*, ce cluster est présenté comme permettant la synthèse d'un lipopeptide (lipopeptide II), pas encore caractérisé. Tous les résultats obtenus



AWS	Gly	Thr	Ile	?	?
NRPSpredictor	Cys	?	NP	?	?
	Cys ou Gly/Ala	Thr/dht	NPs	?	Glu=Gln
	Gly	Thr/dht	Ile	Arg	Glu/Leu
NPRS/PKS	Gly	Thr	Ile	Ser	D–Glu

FIG. 6.2 – Prédiction du peptide synthétisé par le cluster 2. NP=Gly,Ala,Val,Leu,Ile,Abu,Iva ; NPs=Val,Leu,Ile,Abu,Iva

semblent confirmer la synthèse d'un lipopeptide, certainement antibiotique, par le cluster 2.

6.3.3 Cluster 3

Le troisième cluster identifié contient deux gènes situés aux loci PSEEN3044 et PSEEN3045. Ces gènes sont situés sur le brin anti-sens. La protéine codée par le gène du locus PSEEN3044 est annotée comme *non-ribosomal peptide synthetase, terminal component* et celle codée par le locus PSEEN3045 est annotée comme *non ribosomal peptide synthetase*. Nous avons récupéré les séquences protéiques correspondantes et les avons utilisées pour prédire le peptide synthétisé par ce cluster (figure 6.3).

D'après les résultats de prédiction obtenus, la première synthétase contient 8 modules et la seconde en contient 4, conduisant à la synthèse d'un peptide composé de 12 monomères. Nous pouvons noter l'existence de deux domaines de thioestérase (Te). Deux domaines Te en tandem sont observés au sein des synthétases codant des peptides cycliques comme par exemple l'arthrofactine [Roongsawang et al., 2003]. Nous pouvons donc penser que notre peptide contient un cycle. La séquence consensus du peptide produit est : X_Val_Leu_X_Val_Leu_X_Ser_Val_Leu_Ser_Leu/Ile. Le peptide produit par le cluster 3 n'est pas dans NORINE. En recherchant le motif correspondant à la partie produite par la seconde protéine, c'est-à-dire Val_Leu_Ser_Ile/Leu, nous obtenons un peptide contenant ce motif, la putisolvine I. La putisolvine I (NOR00361) est un lipopeptide produit par *Pseudomonas putida* et est un surfactant antibiotique. En utilisant le clustering de niveau 1, les peptides les plus similaires obtenus sont donnés dans la table 6.1.

Nous pouvons noter que tous les peptides similaires au peptide étudié sont des lipopeptides antibiotiques et des surfactants. De plus, tous ces peptides sont produits par des bactéries du genre *Pseudomonas*, genre auquel l'organisme étudié appartient également. Nous avons ensuite recherché dans NORINE, les peptides présentant une composition monomérique proche de celle de notre peptide. Les putisolvines I et II contiennent 7 monomères (sur les 9 prédits) en commun avec ceux de notre peptide. Les putisolvines contiennent 12 monomères et un acide

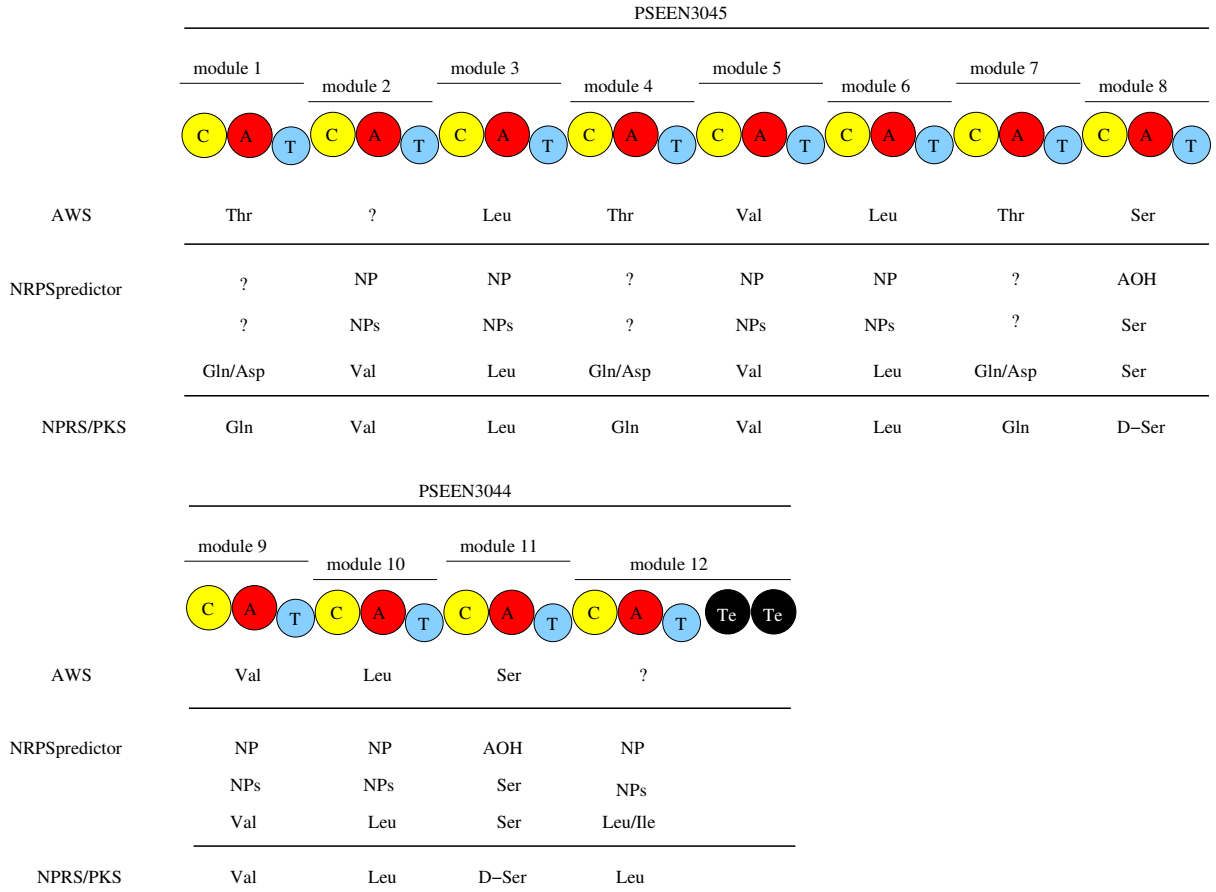


FIG. 6.3 – Prédiction du peptide synthétisé par le cluster 3. NP=Gly,Ala,Val,Leu,Ile,Abu,Iva ; NPs=Val,Leu,Ile,Abu,Iva ; AOH=Ser,Thr,Dht,Dhpg,Hpg,Dpg

TAB. 6.1 – Recherche des peptides similaires avec le clustering de niveau 1

nom	type	activité(s)	organisme producteur
WLIP	lipopeptide	antibiotique, surfactant	*Pseudomonas reactans*
massetolide F	lipopeptide	antibiotique, surfactant	*Pseudomonas sp.*
massetolide E	lipopeptide	antibiotique, surfactant	*Pseudomonas sp.*
viscosin	lipopeptide	antibiotique, surfactant	*Pseudomonas fluorescens*
massetolide G	lipopeptide	antibiotique, surfactant	*Pseudomonas Sp.*
massetolide H	lipopeptide	antibiotique, surfactant	*Pseudomonas Sp.*
viscosinamide	lipopeptide	antibiotique, surfactant	*Pseudomonas fluorescens*
putisolvin III	lipopeptide	antibiotique, surfactant	*Pseudomonas putida*
putisolvin II	lipopeptide	antibiotique, surfactant	*Pseudomonas putida*
putisolvin I	lipopeptide	antibiotique, surfactant	*Pseudomonas putida*

gras et présentent un cycle formé par le groupement hydroxyl de l'une des sérines. Le peptide étudié contient également une sérine qui pourrait permettre l'obtention d'un cycle. Enfin, l'outil de prédiction d'activité biologique fournit les résultats suivants : antibiotique=0,22, anti-tumoral=0,06, immuno-modulateur=0,06, siderophore=0,11, surfactant=0,28, toxine=0,15.

D'après les résultats obtenus, le peptide serait surfactant et très certainement antibiotique. Toutes les expériences menées grâce à NORINE convergent et nous conduisent à conclure que le peptide putatif synthétisé par le cluster 3 est très certainement un lipopeptide antibiotique et un surfactant. L'hypothèse du lipopeptide est renforcée par la présence d'un domaine de condensation sur le premier module. En effet, ce domaine permet de lier l'acide gras à la partie peptidique. De plus, les bactéries du genre *Pseudomonas* sont connues pour synthétiser des lipopeptides cycliques [Raaijmakers et al., 2006].

Dans l'article concernant *P. entomophila*, ce cluster est annoté comme produisant un lipopeptide (lipopeptide I). Cependant, les auteurs ajoutent à ce cluster deux gènes putatifs codant des synthétases NRPS (PSEEN0132 et PSEEN 3332). Selon eux, le gène PSEEN0132 coderait la synthétase du module d'initiation (domaines A et domaine T), qui manque à ce cluster. Cependant, comme nous l'avons vu précédemment, le module d'initiation des synthétases des lipopeptides, contient généralement un domaine de condensation (C) en première position pour lier l'acide gras à la chaîne peptidique. Nous pensons donc que ce gène ne fait pas partie du cluster. Ils ajoutent également le gène PSEEN3332 au cluster car son homologue a été identifié au sein d'un cluster de *P. fluorescens Pf-5* contenant également les homologues de PSEEN3044 et PSEEN3045 (qui serait dupliqué). Cependant, ce gène est très éloigné des deux autres gènes chez *P. entomophila* et nous pensons que les deux gènes identifiés (PSEEN3044 et PSEEN3045) suffisent à la synthèse du lipopeptide.

6.3.4 Cluster 4

Nous avons identifié un autre cluster de gènes putatifs codant des NRPS, situé sur le brin anti-sens. Le cluster 4 contient 6 gènes situés entre le locus PSEEN3229 et le locus PSEEN3234. Parmi ces gènes, le premier est annoté *« peptide synthase »* (PSEEN3229), 4 sont annotés comme *« pyoverdine sidechain peptide synthetase »* et le dernier est annoté *« SyrP protein »* . D'après les annotations, le peptide produit par ce cluster serait donc une pyoverdine. Les pyoverdines sont des sidérophores produits par le genre *Pseudomonas*. Les sidérophores sont des molécules synthétisées lors d'une carence en fer et qui permettent la chélation du fer. Cependant, ces annotations sont obtenues à partir d'homologies de séquences avec le génome de *Pseudomonas fluorescens* et, par conséquent, nous considérons dans un premier temps que le peptide produit est inconnu. Nous avons donc récupéré les séquences protéiques correspondant aux gènes NRPS putatifs et effectué la prédiction du peptide produit à l'aide des trois logiciels (figure 6.4).

D'après les résultats obtenus, le peptide produit contiendrait 10 monomères, dont au moins 3 sous la forme D car 3 domaines d'épimérisation sont prédits au sein des différentes synthétases. D'après les résultats obtenus, la structure prédite est la suivante : D-Ala_X_X_D-Asp_Gly_Gly_Ser_Thr_D-Ser_X. Ce peptide n'est pas présent dans NORINE. La pyoverdine 1.2 (NOR00180) montre une sous-structure commune de 4 monomères avec notre peptide. Lorsque la structure est étendue aux monomères et leurs dérivés, c'est-à-dire que chaque monomère remplace M par le monomère $*M$, deux pyoverdines (GM et 51W) présentent une sous-structure commune de 6 monomères avec le peptide étudié. Dans le peptide prédit, un motif particulier composé de deux Glycines apparaît. Ce motif est présent uniquement dans 5 pyoverdines et aucun autre peptide de NORINE. Les quatre peptides les plus similaires (sans clustering) sont également 4 pyoverdines. La prédiction de l'activité biologique fournit les résultats suivants : antibiotique=0,15, anti-tumoral=0,06, immuno-modulateur=0,04, siderophore=0,26, surfactant=0,12, toxin=0,18. Le peptide putatif est donc prédit sidérophore, ce qui est cohérent avec la fonction des pyoverdines. Les recherches effectuées dans NORINE confirment le fait que le cluster 4 permettrait

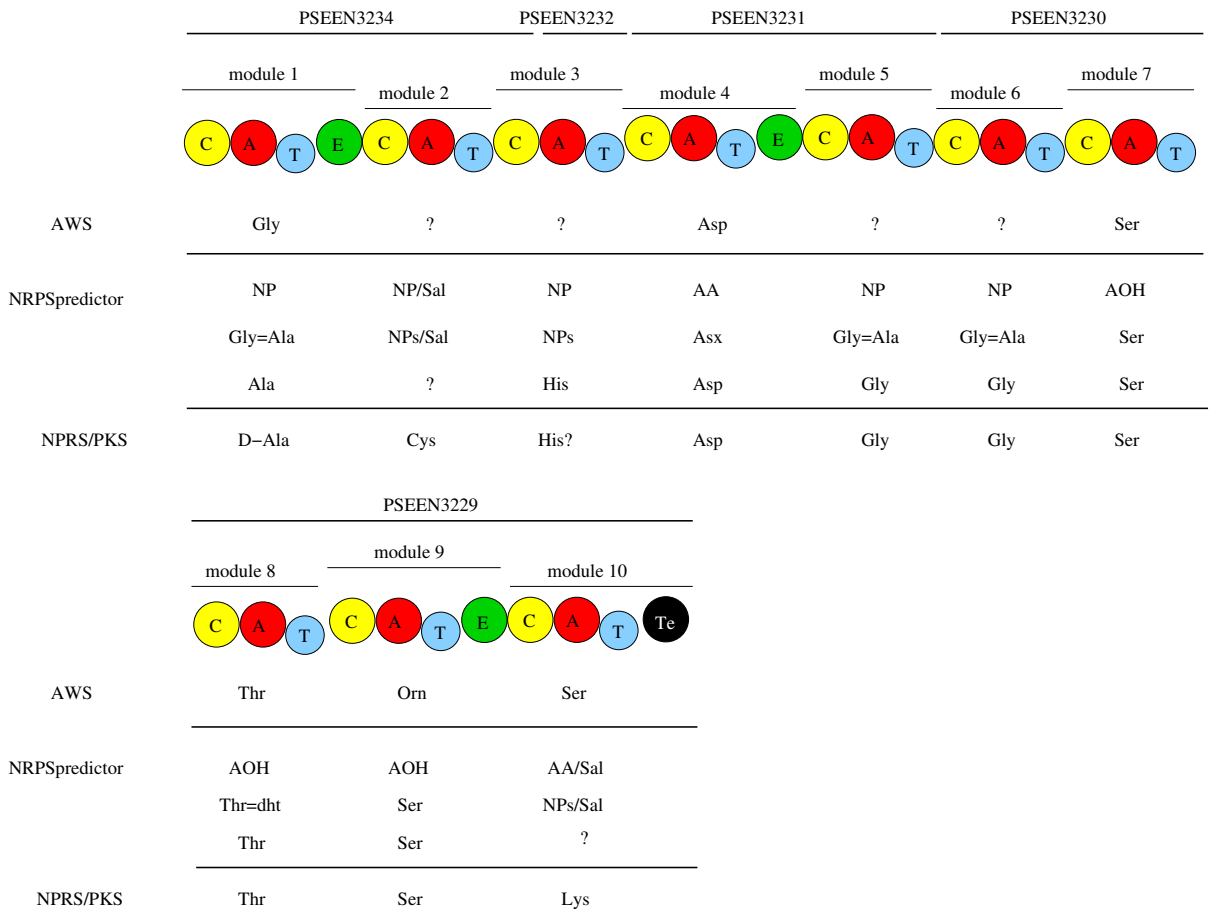


FIG. 6.4 – Prédiction du peptide synthétisé par le cluster 4. NP=Gly,Ala,Val,Leu,Ile,Abu,Iva ; NPs=Val,Leu,Ile,Abu,Iva ; AOH=Ser,Thr,Dht,Dhpg,Hpg,Dpg ; AA=Asp,Asn,Glu,Gln,Aad ; Asn=Asp,Asn

la production d'une pyoverdine. Cependant, la structure de cette pyoverdine ne correspond à aucune des structures des pyoverdines déjà identifiées. Nous avons donc décidé d'approfondir notre étude, en recherchant au sein du génome tous les gènes impliqués dans la synthèse des pyoverdines. Tout d'abord, nous avons recherché les gènes codant les synthétases impliquées dans la synthèse du chromophore. En effet, il a été montré qu'une synthétase, très conservée au sein des différentes espèces, est impliquée dans la synthèse du chromophore [Mossialos et al., 2002]. Cette synthétase contient trois modules qui sélectionnent l'acide glutamique (Glu), la tyrosine (Tyr) et l'acide 2,4-diaminobutyrique. Nous avons recherché les séquences nucléiques similaires à *pvdL*, le gène codant la synthétase responsable de la synthèse du chromophore chez *Pseudomonas aeruginosa*, au sein du génome de *P. entomophila* avec BLAST. Le gène situé au locus PSEEN1815 présente une grande similarité avec *pvdL*. Nous avons donc soumis la séquence protéique de PSEEN1815 aux logiciels de prédiction et avons obtenu 3 modules prédits incorporant les monomères attendus. Nous pouvons donc conclure que le gène du locus PSEEN1815 code la synthétase responsable de la synthèse du chromophore de la pyoverdine.

A part les gènes codant les synthétases, de nombreux autres gènes, répartis sur la totalité du génome, interviennent dans la synthèse, la régulation et le transport des pyoverdines.

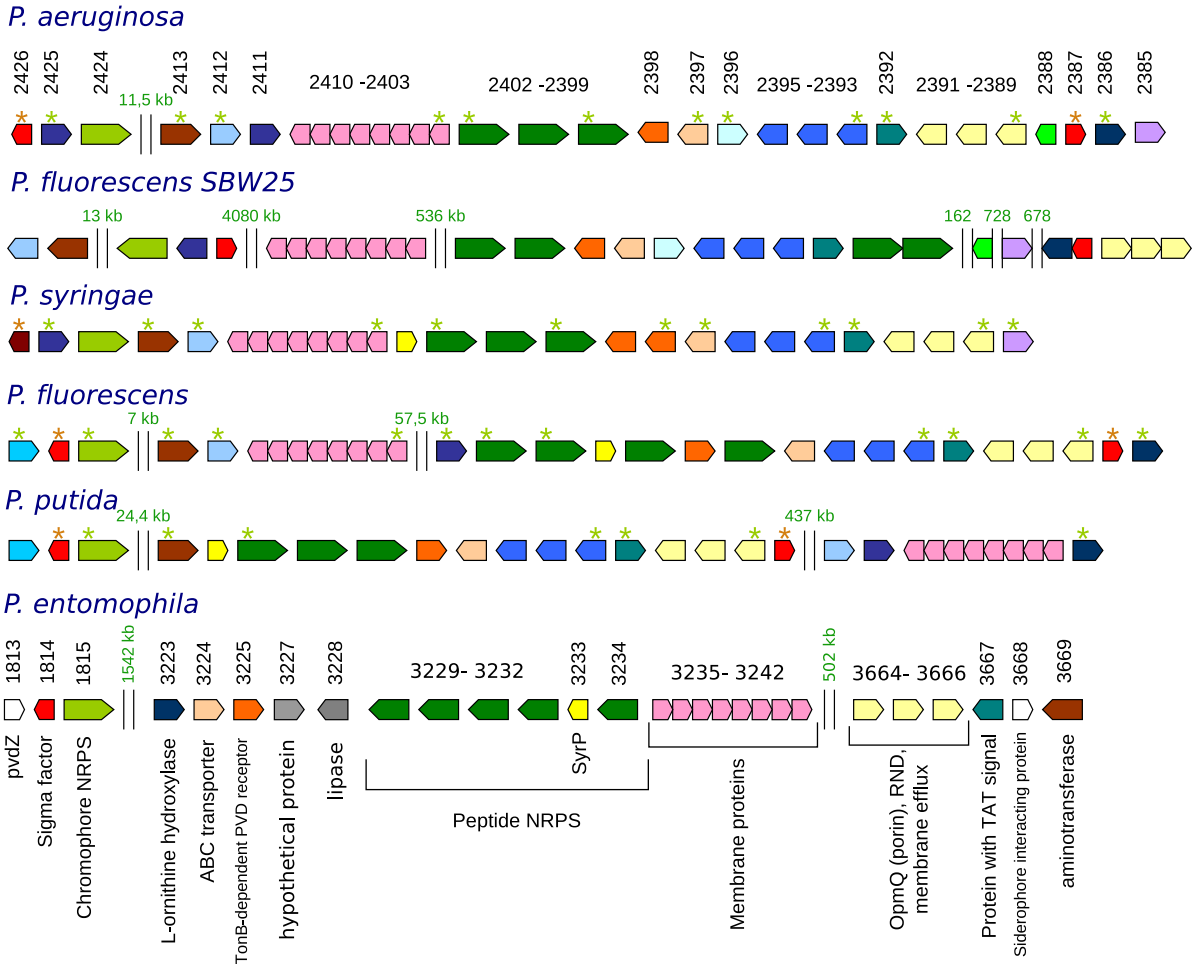


FIG. 6.5 – Organisation des clusters des gènes de la pyoverdine chez différentes espèces de *Pseudomonas*. Les gènes homologues apparaissent dans la même couleur. Ce schéma est basé sur celui de J. Ravel et P. Cornelis [Ravel and Cornelis, 2003].

L'organisation de ces gènes est très différente d'une espèce à l'autre. J. Ravel et P. Cornelis [Ravel and Cornelis, 2003] ont étudié l'organisation de ces gènes au sein de différentes espèces de *Pseudomonas* synthétisant des pyoverdines. A partir des résultats de cet article, nous avons recherché les différents gènes impliqués dans la synthèse de la pyoverdine potentielle dans le génome de *P. entomophila*. L'organisation obtenue pour les gènes, ainsi que celles des clusters identifiés chez les autres espèces, est donnée dans la figure 6.5 et est basée sur le schéma de J. Ravel et P. Cornelis [Ravel and Cornelis, 2003].

Cette figure montre que l'organisation des clusters de la pyoverdine est très différente d'une espèce à l'autre. Le cluster de *P. entomophila* possède une organisation qui lui est propre et qui confirme l'hypothèse de la production d'une nouvelle pyoverdine par *P. entomophila*. En effet, seul le cluster de *P. entomophila* contient 5 gènes NRPS potentiellement impliqués dans la synthèse de la partie peptidique de la pyoverdine, ce qui suggère la synthèse d'une nouvelle pyoverdine. Cependant, le cluster identifié chez *P. entomophila* contient un grand nombre de gènes homologues avec les autres clusters identifiés au sein des autres espèces de *Pseudomonas*.

Dans un article très récent [Matthijs et al., 2009], les auteurs ont étudié les sidérophores de *P. entomophila L48*, dont la pyoverdine produite par le cluster 4. Ils ont montré que le cluster impliqué dans la synthèse de la pyoverdine contient 28 gènes distribués sur trois loci différents du génome (cf. figure 6.5). Ils ont ensuite déterminé la structure de la pyoverdine par une analyse de masse. La partie peptidique de la pyoverdine produite par *P. entomophila* contient les 10 résidus suivants : Ala_Asn_Dab_OH-His_Gly_Gly_Ala_Thr_Ser_cOH-Orn, avec cOH-Orn pour la cyclo-hydroxy-ornithine. La partie peptidique que nous avons prédite, c'est-à-dire D-Ala_X_X_D-Asp_Gly_Gly_Ser_Thr_D-Ser_X, est très similaire à celle identifiée expérimentalement dans cet article. En effet, cinq monomères ont été prédits correctement (Ala, Gly, Gly, Thr et Ser) et seulement deux monomères sont incorrects, l'acide aspartique qui est en fait une hydroxy-histidine et la première sérine qui est une alanine. Ces expériences confirment donc la synthèse d'une nouvelle pyoverdine par *P. entomophila*.

Lorsque nous avons étudié ce cluster, l'article [Matthijs et al., 2009] n'était pas encore paru, aucune donnée expérimentale sur la production de pyoverdine par *P. entomophila* n'était alors disponible. Nous avons donc validé expérimentalement la synthèse d'une nouvelle pyoverdine par cette souche.

6.4 Validations expérimentales

Les bactéries du genre *Pseudomonas* sont des bacilles aérobies à Gram négatif. Elles sont généralement mobiles grâce à des flagelles polaires. Ce genre contient plus d'une centaine d'espèces ubiquitaires, largement répandues et vivant dans le sol et l'eau. Certaines espèces de *Pseudomonas* sont des pathogènes des plantes, principalement *P. syringae* [Kim et al., 2008], ou de l'homme et l'animal, principalement *P. aeruginosa* [Morrison and Wenzel, 1984]. La figure 6.6 montre des bactéries *P. aeruginosa* visualisées au microscope électronique à balayage.

L'espèce *P. entomophila* a été isolée pour la première fois à partir d'espèces de *Drosophila melanogaster*, insecte également appelé mouche du vinaigre [Vodovar et al., 2005]. Une fois ingérée, elle cause la mort des adultes et des larves. Elle est hautement pathogène et cause la mort de 70% des larves en moins de 24 heures. Le génome de la souche de *P. entomophila* L48 a été séquencé en 2006 [Vodovar et al., 2006].

Le fer est indispensable dans de nombreux processus tels que la synthèse de l'ADN ou la respiration. Malgré le fait qu'il soit l'élément le plus abondant sur la planète, la disponibilité du fer dans certains environnements comme le sol ou la mer est limitée par la faible solubilité des ions Fe^{3+}. Les micro-organismes synthétisent des sidérophores pour récupérer le fer par la formation de complexes d'ions Fe^{3+} solubles qui sont ensuite captés par des mécanismes de transport actif. L'étude de la synthèse des sidérophores présente un intérêt médical puisqu'en inhibant leur synthèse, la croissance des micro-organismes est également inhibée. Les micro-organismes produisent des sidérophores spécifiques ce qui permettrait le développement d'antibiotiques ciblés.

Les pyoverdines sont des sidérophores qui possèdent un chromophore, qui rend ces molécules fluorescentes, une chaîne latérale, généralement un acide dicarboxylique ou une amide et une chaîne peptidique variable, spécifique à une souche productrice [Ravel and Cornelis, 2003]. Le chromophore et la chaîne peptidique sont synthétisés par la voie non-ribosomiale. Ce sont des pigments de couleur vert-jaune. La présence de pyoverdine est mise en évidence par culture sur milieu King B, qui est limité en fer et permet donc la production de sidérophores. Plus de 60

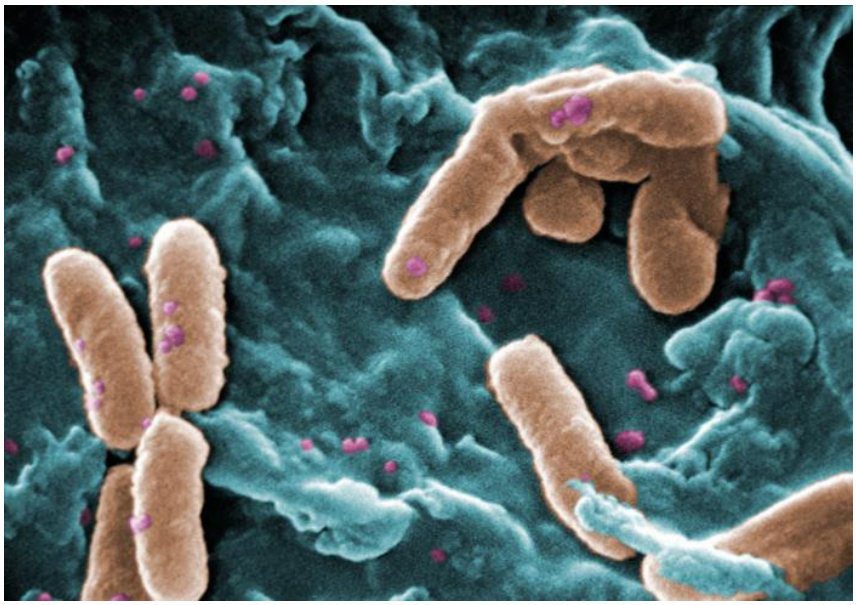


FIG. 6.6 – Bactéries *Pseudomonas aeruginosa* visualisées au microscope électronique à balayage. Image provenant du site *Public Health Image Library*.

pyoverdines, présentant des structures très diverses, ont été identifiées au sein de différentes espèces de *Pseudomonas* [Budzikiewicz, 2004]. NORINE répertorie 57 pyoverdines, la structure de certaines pyoverdines n'étant encore que partielle. Les pyoverdines montrent un spectre d'absorption caractéristique, avec un pic à 400 nm (pH 7) [MEYER and ABDALLAH, 1978].

6.4.1 Matériels et méthodes

Souches et milieux de culture

La Souche *P. entomophila* L48 nous a gentiment été envoyée par Frédéric Boccard (Gif-sur-Yvette, France) et la souche *P. aeuginosa* 7NSK2 nous a gracieusement été envoyée par Monica Höfte (Université de Gand, Belgique). Les souches isolées sur milieu King B (Biokar) sont incubées à 30 ° C pendant 24 heures. Le milieu King B permet la production de sidérophores, dont les pyoverdines. Pour la production de pyoverdine, les souches sont cultivées dans du milieu CAA (casaminoacides 5g/L ; K_2HPO_4 0,9g/L ; $MgSO_4$ 0,25g/L) stérilisé par autoclavage à 120 ° C pendant 20 minutes. Toute la vaisselle est débarrassée des traces de fer par traitement à l'acide nitrique 10% pendant au moins 2 heures et abondamment rincée par de l'eau ultra pure.

Détection des pyoverdines

Les échantillons (boîtes King B ou surnageant de culture en milieu CAA) peuvent être placés sous UV pour être photographiés. La production des pyoverdines est suivie par mesure de l'absorbance à 400 nm (Spectrophotomètre Uvikon-Kontron instrument) après dilution dans du tampon phosphate de sodium 50 mM pH7 (1/1, V/V).

Purification des pyoverdines

Quelques colonies isolées sur King B sont mises en suspension dans 1 mL de CAA pour ensemencer 200 mL du même milieu dans une fiole d'un litre dépourvu de toute trace de fer. La culture est conduite pendant 20 heures à 30 ° C sous agitation orbitale (130 rpm). La culture est alors centrifugée à 10000 g pendant 30 minutes. Le surnageant est alors filtré sur membrane de nitro-cellulose 0,22 μm. Les pyoverdines sont alors purifiées sur colonne C18 maxi-clean (Alltech). La colonne est conditionnée par 200 mL de méthanol pur, rincée par 90 mL d'eau, l'échantillon (100 mL) est alors déposé. La colonne est ensuite lavée par 100 mL d'eau et séchée à l'air. Les pyoverdines sont éluées par 100 mL de méthanol à 80%. L'échantillon peut ensuite être concentré sous vide.

6.4.2 Résultats - Discussion

Production de pyoverdine par *Pseudomonas entomophila*

La souche a été ensemencée sur boîtes King B. Deux témoins sont également ensemencés sur la même boîte. Le premier témoin, positif, est la souche de *P. aeruginosa* 7NSK2 qui produit de la pyoverdine [Höfte et al., 1991] et le second témoin, négatif, est *Bacillus subtilis* 168 qui, en condition de carence en fer, produit la bacillibactine (NOR00330), un sidérophore non fluorescent.

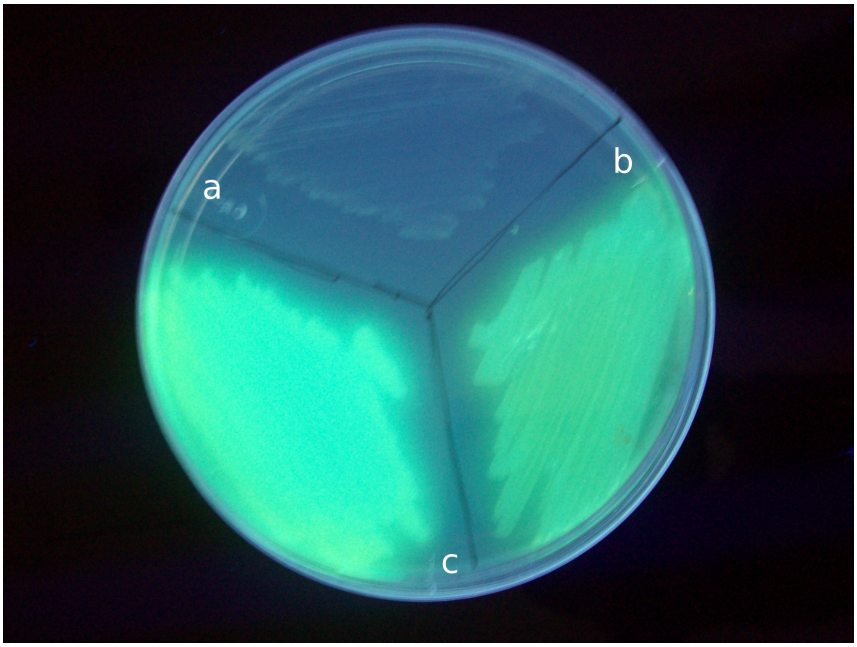


FIG. 6.7 – Boîte King B photographiée sous ultra-violets ensemencée avec a) *B. subtilis* 168 b) *P. entomophila* L48 c) *P. aeruginosa* 7NSK2

La boîte contenant les trois espèces a été photographiée sous rayonnement ultra-violet (figure 6.7). Les résultats montrent bien que les souches *P. entomophila* L48 et *P. aeruginosa* produisent bien une molécule qui fluoresce en réponse à une carence en fer, contrairement à la souche *B. subtilis* 168. Ces résultats confirment l'hypothèse de la synthèse d'une pyoverdine par *P. entomophila* L48.

Cinétique de production

P. entomophila L48 a été ensemencée en milieu CAA liquide. Les échantillons ont été prélevés à différents temps. Un spectre d'absorption est réalisé entre 350 et 450 nm. La mesure de densité optique à 600 nm nous renseigne sur la biomasse et la mesure d'absorbance à 400 nm nous renseigne sur la quantité de pyoverdine présente dans le milieu de culture.

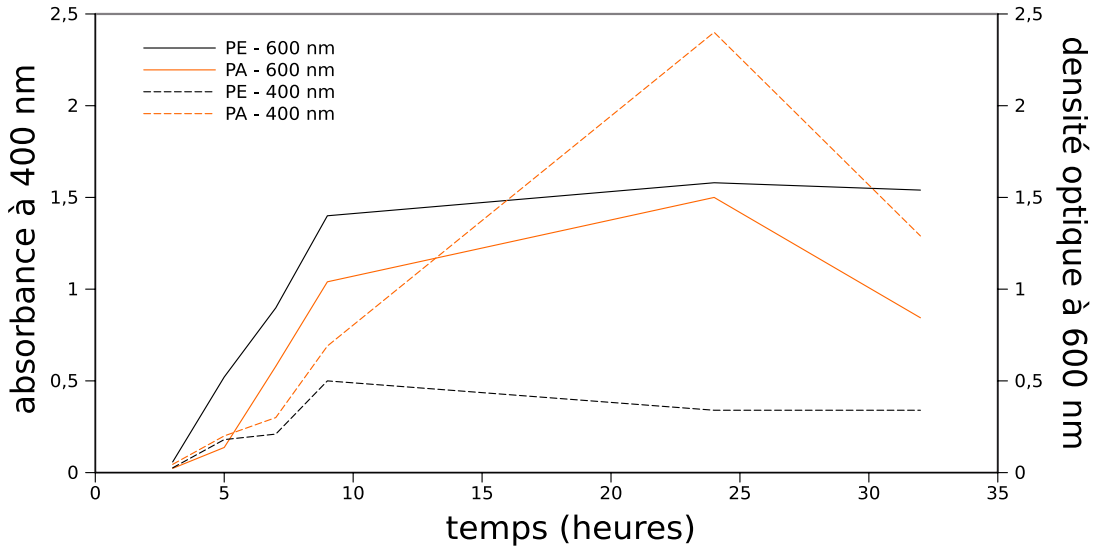


FIG. 6.8 – Cinétique de production de pyoverdines par *P. entomophila* L48 (PE) et *P. Aeruginosa* 7NSK2 (PA)

La figure 6.8 montre la cinétique de production de pyoverdine. Nous constatons qu'à partir de 24 heures la croissance, ainsi que la production de pyoverdine, stagnent chez *P. entomophila* L48. Il n'est donc pas nécessaire de laisser pousser les souches plus longtemps. Nous pouvons également remarquer que *P. entomophila* L48 semble produire une quantité plus faible de pyoverdine que la souche témoin *P. aeruginosa* 7NSK2. Il est très difficile d'interpréter des résultats de courbe de croissance et de production de pyoverdine pour différentes raisons. Tout d'abord, la sensibilité de la régulation par la concentration en fer de la synthèse de pyoverdine varie d'une souche à l'autre. Dans un même milieu, les productivités peuvent s'avérer très différentes en fonction des souches. Dans l'article [Leclère et al., 2009], il a été montré que dans des conditions de culture identiques, la souche *P. aeruginosa 7NSK2* produit des quantités beaucoup plus importantes de pyoverdine que les deux autres souches considérées dans l'article (*P. fluorescens* et *P. putida*). Par ailleurs la quantité de pyoverdine produite étant souvent très largement excessive par rapport aux besoins, la corrélation avec la cinétique de croissance est difficile. La troisième hypothèse est que dans des conditions importantes de stress en fer, une dihydropyoverdine peut être produite, qui n'absorbe pas à 400 nm ([Jacques et al., 2003]). La production inférieure de pyoverdine dans le cas de *P. entomphila* pourrait donc être causée par la formation de dihydropyoverdine, qui n'est pas détectée par spectrométrie. Une autre hypothèse est que la souche peut aussi produire d'autres sidérophores que la pyoverdine. Dans l'article très récent [Matthijs et al., 2009], les auteurs ont mis en évidence la synthèse de deux sidérophores par *P. entomophila* : une pyoverdine et la pseudomonine, ce qui pourrait expliquer la production moindre de pyoverdine chez *P. entomophila* puisque cette souche synthétise un second sidérophore. Enfin, la dernière hypothèse

est qu'il exite un besoin particulier en monomère qui ne serait pas assuré par le milieu, cependant, vérifier cette hypothèse est particulièrement difficile car il faut contrôler les traces de fer ajoutées qui pourraient contaminer les solutions d'acides aminés ajoutés.

Purification

La souche a été cultivée pendant 24 heures dans le milieu CAA sous agitation et la pyoverdine purifiée sur colonne C18. Nous avons réalisé le spectre d'absorption entre 350 et 450 nm du produit purifié (figure 6.9).

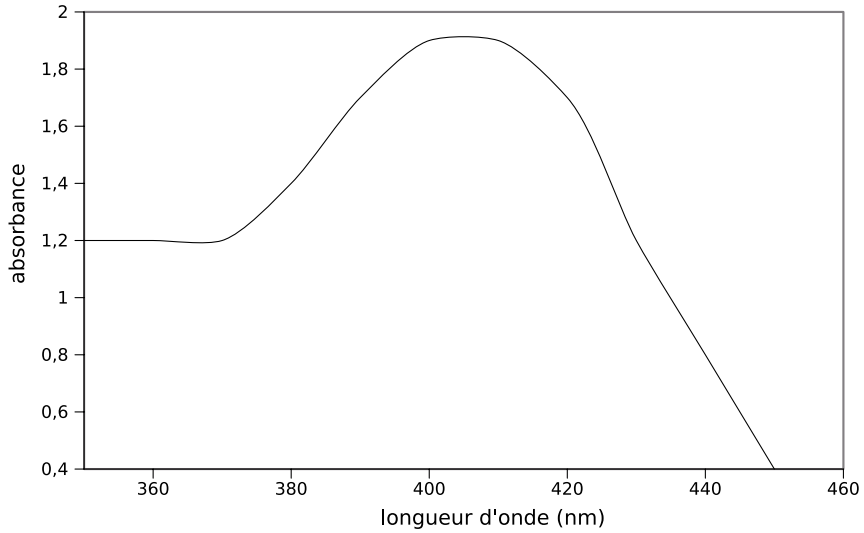


FIG. 6.9 – Spectre d'absorption entre 350 et 450 nm du produit purifié

Le spectre d'absorption montre un spectre caractéristique des pyoverdines, avec un pic à 400 nm, ce qui prouve que la pyoverdine est présente après la purification. Un échantillon sera envoyé à l'équipe de Mathias Schäfer (Institut für Organische Chemie der Universität zu Köln, Allemagne) pour déterminer la structure de la pyoverdine synthétisée par *P. entomophila* par spectrométrie de masse [Moon et al., 2008].

Dans l'article [Matthijs et al., 2009] paru très récemment, les auteurs ont prouvé expérimentalement la synthèse d'une nouvelle pyoverdine par *P. entomophila*. Après avoir purifié la pyoverdine sur colonne C18, une analyse par spectrométrie de masse a permis l'identification des dix résidus composant la partie peptidique. Ils ont ensuite utilisé la méthode de mutagénèse plasposon pour générer des insertions de transposons dans le chromosome de *P. entomophila*. Cinq mutants contenant des insertions dans différents gènes du cluster de la pyoverdine, dont un dans la synthétase du chromophore et trois dans les synthétases intervenant dans la synthèse de la partie peptidique, ont été isolés et ont montré une perte totale de la flusorescence, prouvant ainsi l'implication de ces différents gènes dans la synthèse de la pyoverdine.

Dans cette section, nous avons dans un premier temps démontré l'utilité de NORINE *pour l'analyse des peptides non-ribosomiaux avec l'étude d'un cluster NRPS chez Lactococcus lactis produisant probablement un lipopeptide antibiotique et surfactant. Ensuite, nous avons étudié les différents clusters NRPS chez Pseudomonas entomophila. Nous avons confirmé la synthèse par voie non-ribosomiale d'un lipopeptide par le cluster 3 et d'une pyoverdine par le cluster 4. Les auteurs nous ont confirmé la synthèse d'un lipopeptide par le cluster 3, qu'ils étudient actuellement. Nous avons donc décidé de nous concentrer sur la pyoverdine et avons validé expérimentalement la synthèse d'une pyoverdine par cette souche.*

Conclusion et perspectives

Les travaux présentés ici ont permis le développement de NORINE, plate-forme logicielle pour l'analyse des peptides non-ribosomiaux dont le but est de centraliser les données sur ces peptides, ainsi que de faciliter leur analyse à l'aide de divers outils informatiques.

Les peptides non-ribosomiaux présentent des activités biologiques importantes et variées (antibiotique, anti-tumorale, immuno-modulatrice, ...). Certains de ces peptides sont utilisés dans le domaine médical comme la cyclosporine, un immuno-suppresseur utilisé après une greffe ou la daptomycine (commercialisée sous le nom de Cubicin), un antibiotique utilisé pour traiter les infections bactériennes à Gram positif. Les activités biologiques de ces peptides en font des molécules de grand intérêt. En effet, l'apparition d'une résistance de certaines bactéries aux antibiotiques mène à la recherche continuelle de nouveaux antibiotiques. De la même manière, un des enjeux de la société actuelle est la lutte contre le cancer par le développement d'agents anti-tumoraux efficaces. Les peptides non-ribosomiaux montrent alors un potentiel énorme pour l'industrie pharmaceutique. La modification génétique des synthétases impliquées dans la synthèse d'un peptide peut mener à la production d'un peptide modifié pouvant présenter une nouvelle ou une meilleure activité biologique. Cependant, le processus d'obtention de nouveaux peptides actifs est difficile et se heurte à de nombreux problèmes. De meilleures connaissance sur la voie de biosynthèse non-ribosomiale et sur les peptides produits, ainsi que des outils informatiques performants permettraient la mise en place de protocoles plus adaptés à l'obtention de nouveaux peptides bioactifs par modifications génétiques des synthétases. En effet, connaître les relations structure/fonction des peptides ou les monomères spécifiques à une activité biologique orienterait les modifications génétiques à entreprendre.

Dans le but de regrouper les connaissances sur les peptides non-ribosomiaux, nous avons mis en place une base de données contenant actuellement plus de 1 000 peptides. Pour chaque peptide, différentes annotations sont disponibles telles que l'activité biologique, la structure ou encore les organismes producteurs. NORINE contient également des informations sur les monomères incorporés au sein des peptides non-ribosomiaux. Nous avons développé une interface Web conviviale, et permettant l'interrogation de la base en fonction des différentes annotations. L'interface offre aux biologistes la possibilité de trouver et consulter les informations concernant des peptides d'intérêt. Par exemple, un biologiste travaillant sur un organisme donné peut facilement identifier les peptides non-ribosomiaux produits par cet organisme grâce à NORINE. De même, lors de l'identification d'un peptide, l'une des premières données obtenues est souvent la masse moléculaire, qui peut être recherchée dans NORINE et ainsi, le biologiste peut obtenir la liste des peptides présentant une masse moléculaire similaire.

NORINE permet également la recherche de peptides en fonction de diverses caractéristiques structurales grâce aux outils informatiques que nous avons développés. Lors de la découverte d'un peptide, la première étape est de savoir si ce peptide a été identifié auparavant. Pour ce faire, l'utilisateur peut rechercher si le peptide d'intérêt est présent dans NORINE. Nous avons également

mis en place une méthode efficace pour la recherche de motifs structuraux au sein des peptides contenus dans NORINE. Cette fonctionnalité s'avère très utile lors de la phase de prédiction d'un peptide à partir de la séquence protéique d'une synthétase identifiée dans un génome d'intérêt. En effet, la prédiction fournit généralement à l'utilisateur une structure partielle, contenant des incertitudes, c'est-à-dire des monomères non prédictibles ou plusieurs monomères possibles à une position donnée. Cette structure partielle peut être recherchée sous la forme d'un motif dans NORINE pour aider l'identification du peptide. La méthode de recherche de motifs structuraux a ensuite été étendue à la recherche de peptides similaires à un peptide d'intérêt. Les peptides possédant des structures similaires présentent souvent des fonctions similaires. Cette fonctionnalité permet donc de mettre en évidence des caractéristiques biologiques pouvant orienter les validations expérimentales.

Les peptides non-ribosomiaux sont modélisés par des graphes étiquetés non-orientés. Les graphes sont utilisés dans de nombreux domaines comme par exemple en chémo-informatique où ils modélisent des molécules, en bio-informatique, où ils peuvent modéliser des réseaux de régulation de gènes ou des réseaux métaboliques, ou encore dans divers domaines de l'informatique tels que l'analyse d'images ou les technologies du Web. La recherche de motifs structuraux et de sous-structures communes maximales sont des opérations utiles et nécessaires dans tous ces domaines. Les méthodes que nous avons développées s'avèrent très efficaces et peuvent être appliquées et étendues pour les autres domaines utilisant les graphes.

Nous avons ensuite exploité les données contenues dans NORINE en réalisant des études statistiques qui nous ont permis de mettre en évidence des propriétés biologiques intéressantes. L'une des plus remarquables est la spécificité des monomères en fonction de l'activité biologique. Cette observation nous a conduit à l'élaboration d'un outil d'aide à la prédiction de l'activité biologique en fonction de la composition monomérique d'un peptide. La mise en évidence d'une spécificité des monomères en fonction de l'activité biologique ouvre de nombreuses perspectives. L'une d'elles est de pouvoir mettre en évidence des motifs structuraux caractéristiques d'une activité. En effet, le fait que certains monomères semblent être caractéristiques d'une activité donnée nous amène à penser que des motifs structuraux caractéristiques d'une activité biologique peuvent être déduits des peptides présents dans NORINE. Une autre perspective ouverte par cette observation est une meilleure compréhension de la relation structure/fonction. En effet, en utilisant une annotation sémantique de l'activité biologique du type *Gene Ontology*, il serait possible de stocker le processus biologique dans lequel le peptide intervient, ainsi que sa cible, ce qui permettrait de mieux comprendre les relations entre la structure d'un groupe de peptides et leurs fonctions biologiques.

Nous avons également montré que les peptides présentent des caractéristiques spécifiques en fonction de l'organisme producteur. En effet, nous avons mis en évidence des différences de composition et de structure entre les peptides synthétisés par les bactéries et ceux synthétisés par les champignons. Ces analyses ont aussi montré une similarité de composition et de structure entre les peptides isolés chez les métazoaires, principalement des éponges, et ceux produits par les bactéries. Cette observation renforce l'hypothèse selon laquelle les peptides isolés chez les métazoaires sont en fait synthétisés par des bactéries symbiotiques. Ces spécificités en fonction de l'organisme producteur pourraient aider à l'identification de l'organisme producteur lors de l'isolement de peptides à partir de surnageants de culture ou de colonies.

Nous avons ensuite prouvé l'utilité et l'efficacité de nos outils en analysant des peptides non-ribosomiaux putatifs. En effet, les différents outils mis en place nous ont permis de mettre en évidence des propriétés biologiques pour ces peptides putatifs. Nous avons validé expérimentalement les propriétés biologiques révélées lors de l'étude d'une nouvelle pyoverdine

synthétisée par *Pseudomonas entomophila*.

En moins de trois ans, NORINE est devenue la ressource de référence internationale pour les peptides non-ribosomiaux. Par exemple, elle est une base de référence pour wwPDB, au même titre que GenBank et UniProt. Ces travaux ont été publiés dans des journaux internationaux, mais aussi présentés dans de nombreuses conférences, nationales et internationales, biologiques et informatiques (voir la liste des publications et communications). Lors de la septième conférence de la Société Française de Microbiologie, nous avons obtenu le prix du meilleur poster sur plus de 300 posters présentés, ce qui montre encore une fois l'intérêt de la communauté scientifique pour notre travail.

NORINE est destinée à croître afin de garder la base de données complète et à jour. De même, certains outils de NORINE peuvent être améliorés, comme par exemple la recherche par similarité qui ne gère pas, pour le moment, les insertions/délétions de monomères. Notre outil de prédiction de l'activité biologique peut être amélioré avec, par exemple, le calcul d'une p-valeur ou la prise en compte d'autres propriétés.

Le but de NORINE est de centraliser toutes les informations et outils facilitant l'analyse des peptides non-ribosomiaux. A terme, NORINE doit proposer des outils pour prédire puis analyser un peptide, à partir de la séquence nucléique ou protéique d'une synthétase. Nous avons commencé une collaboration avec l'équipe de Daslav Hranueli (Université de Zagreb, Croatie). Cette équipe a développé *ClustScan*, un logiciel d'annotation semi-automatique de clusters de gènes de biosynthèse modulaire dans les génomes. Le but de cette collaboration est d'utiliser *ClustScan* pour créer une base de données sur les synthétases, complémentaire et liée à notre base de données sur les peptides.

Notre plate-forme a pour finalité de centraliser un maximum de données sur les peptides non-ribosomiaux. Nous sommes en contact avec l'équipe de Pavel Pevzner (Université de California San Diego, Etats-Unis) qui travaille sur les données de spectrométrie de masse pour les peptides non-ribosomiaux. Cette collaboration devrait aboutir à l'ajout de données de spectrométrie de masse dans NORINE.

Ces travaux ont permis de centraliser les données sur les peptides non-ribosomiaux et de développer des nouveaux outils informatiques pour leur analyse. Malgré les nombreuses perspectives et le travail restant encore à effectuer, NORINE est d'ores et déjà une ressource importante pour le domaine en pleine expansion des peptides non-ribosomiaux.

Liste des publications et communications

Publications internationales

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

- NORINE : a database of nonribosomal peptides,
  Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P and Kucherov G,
  Nucleic Acids Research - Database issue, 2008, 36 :D326-31.

Publications nationales

- La synthèse peptidique non-ribosomiale, source de biodiversité de composés actifs,
  Jacques P, Gancel F, Chollet-Imbert M, Guez JS, Béchet M, Caboche S, Coucheney F, Coutte F, Tapi A and Leclère V,
  SFM letter, 2007.

Séminaires et Communications

- Novel approaches to isolate antimicrobial peptides using bioinformatics and molecular tools
  Jacques P, Froidevaux R, Chollet-Imbert M, Tapi A, Caboche S, Pupin M, Kucherov G, Dhulster P, Le Flem G, Vercaigne-Marko D and Leclère V,
  Second International Symposium on Antimicrobial Peptides, Saint Malo (France) 2009, communication.

- Relationships between producing organisms, activities and monomers incorporated into NRPS peptides,
  Leclère V, Caboche S, Pupin M, Kucherov G and Jacques P,
  FEMS, congress of european microbiologists, Gothenburg (Sweden), 2009, communication et poster.

- Norine : a public resource for nonribosomal peptides,
  Caboche S,
  EuroDocInfo 2009, Mons (Belgium) 2009, communication.

Liste des publications et communications

- Norine : database and efficient algorithms dedicated to nonribosomal peptides,
  Caboche S, Pupin M, Leclère V, Jacques P and Kucherov G,
  ECCB 2008, Cagliari (Italy) 2008, poster.

- Norine : a public resource of nonribosomal peptides,
  Caboche S, Pupin M, Leclère V, Jacques P and Kucherov G,
  EMBnet Conference 2008, Martina Franca (Italy) 2008, poster.

- Norine : une plate-forme dédiée aux peptides non ribosomiaux,
  Caboche S,
  Rencontres PPF, Lille, 2008, communication.

- Norine : a platform dedicated to nonribosomal peptides,
  Caboche S, Pupin M, Leclère V, Jacques P and Kucherov G,
  JOBIM 2008 (Lille, France), poster.

- Norine : une base de données de peptides non-ribosomaux,
  Caboche S,
  rencontres PPF, Lille, 2007, communication

- Norine : une nouvelle base de données qui met en exergue la biodiversité des structures et des activités des peptides synthétisés par la voie non-ribosomale (NRPS),
  Caboche S, Leclère V, Pupin M, Kucherov G and Jacques P,
  Poster 7ième congrés de la Société Francaise de Microbiologie, Nantes, 2007, Prix du meilleur poster.

- NORINE : a recent database highlighting a large biodiversity among NRPS peptide structures and activities,
  Leclère V, Caboche S, Pupin M, Kucherov G and Jacques P,
  RSC conference, Chemical Biology : directing biosynthesis, Cambridge, 2006, poster.

- Database and comparison of non ribosomal peptides,
  Caboche S, Leclère V, Jacques P, Pupin M and Kucherov G,
  JOBIM 2006, (Bordeaux, France), communication et poster.

Bibliographie

[Abdel-Mawgoud et al., 2008] Abdel-Mawgoud, A. M., Aboulwafa, M. M., and Hassouna, N. A. (2008). Characterization of surfactin produced by Bacillus subtilis isolate BS5. *Appl. Biochem. Biotechnol.*, 150 :289–303.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215 :403–410.

[Ansari et al., 2004] 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 Research*, 32(Web-Server-Issue) :405–413.

[authors listed, 1984] authors listed, N. (1984). IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature and symbolism for amino acids and peptides. Recommendations 1983. *Biochem. J.*, 219 :345–373.

[Balibar et al., 2005] Balibar, C. J., Vaillancourt, F. H., and Walsh, C. T. (2005). Generation of D amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. *Chem. Biol.*, 12 :1189–1200.

[Besson and Michel, 1989] Besson, F. and Michel, G. (1989). Action of mycosubtilin, an antifungal antibiotic of Bacillus subtilis, on the cell membrane of Saccharomyces cerevisiae. *Microbios*, 59 :113–121.

[Budzikiewicz, 2004] Budzikiewicz, H. (2004). Siderophores of the Pseudomonadaceae sensu stricto (fluorescent and non-fluorescent Pseudomonas spp.). *Fortschr Chem Org Naturst*, 87 :81–237.

[Bunke and Shearer, 1998] Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4) :255–259.

[Caboche et al., 2009] Caboche, S., Pupin, M., Leclere, V., Jacques, P., and Kucherov, G. (2009). Structural pattern matching of nonribosomal peptides. *BMC Struct. Biol.*, 9 :15.

[Caboche et al., 2008] Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE : a database of nonribosomal peptides. *Nucleic Acids Res.*, 36 :D326–331.

[Challis, 2008] Challis, G. L. (2008). Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiology (Reading, Engl.)*, 154 :1555–1569.

[Challis et al., 2000] Challis, G. L., Ravel, J., and Townsend, C. A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, 7 :211–224.

[Chutrakul et al., 2008] Chutrakul, C., Alcocer, M., Bailey, K., and Peberdy, J. F. (2008). The production and characterisation of trichotoxin peptaibols, by Trichoderma asperellum. *Chem. Biodivers.*, 5 :1694–1706.

[Conti et al., 1997] Conti, E., Stachelhaus, T., Marahiel, M. A., and Brick, P. (1997). Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.*, 16 :4174–4183.

[de Bruijn et al., 2007] de Bruijn, I., de Kock, M. J., Yang, M., de Waard, P., van Beek, T. A., and Raaijmakers, J. M. (2007). Genome-based discovery, structure prediction and functional analysis of cyclic lipopeptide antibiotics in Pseudomonas species. *Mol. Microbiol.*, 63 :417–428.

[Deng and Taunton, 2002] Deng, S. and Taunton, J. (2002). Kinetic control of proline amide rotamers : total synthesis of trans,trans- and cis,cis-ceratospongamide. *J. Am. Chem. Soc.*, 124 :916–917.

[Dimise et al., 2008] Dimise, E. J., Widboom, P. F., and Bruner, S. D. (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.

[Doekel et al., 2008] Doekel, S., Coëffet-Le Gal, M. F., Gu, J. Q., Chu, M., Baltz, R. H., and Brian, P. (2008). Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. *Microbiology (Reading, Engl.)*, 154 :2872–2880.

[Duclohier, 2007] Duclohier, H. (2007). Peptaibiotics and peptaibols : an alternative to classical antibiotics ? *Chem. Biodivers.*, 4 :1023–1026.

[Duerfahrt et al., 2004] Duerfahrt, T., Eppelmann, K., Müller, R., and Marahiel, M. A. (2004). Rational design of a bimodular model system for the investigation of heterocyclization in nonribosomal peptide biosynthesis. *Chem. Biol.*, 11 :261–271.

[Eades, 1984] Eades, P. (1984). A heuristic for graph drawing. *Congressus Numerantium*, 42 :149–160.

[Eddy, 1998] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14 :755–763.

[Eppelmann et al., 2002] Eppelmann, K., Stachelhaus, T., and Marahiel, M. A. (2002). Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry*, 41 :9718–9726.

[Erlanger and Goode, 1967] Erlanger, B. F. and Goode, L. (1967). Mode of action of penicillin. *Nature*, 213 :183–184.

[Feil et al., 2005] Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E., Thiel, J., Malfatti, S., Loper, J. E., Lapidus, A., Detter, J. C., Land, M., Richardson, P. M., Kyrpides, N. C., Ivanova, N., and Lindow, S. E. (2005). Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. *Proc. Natl. Acad. Sci. U.S.A.*, 102 :11064–11069.

[Fellows et al., 2007] Fellows, M., Fertin, G., Hermelin, D., and Vialette, S. (2007). Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. In *Proceedings of the 34th International Colloquium on Automata, Languages and Programming (ICALP), June 9-13, 2007, Wroclaw (Poland)*, volume 4596 of *Lecture Notes in Computer Science*, pages 340–351. Springer Verlag. Available from : `http://www.springerlink.com/content/978-3-540-73419-2`, http ://dx.doi.org/10.1007/978-3-540-73420-8 `doi:10.1007/978-3-540-73420-8`.

[Felnagle et al., 2008] Felnagle, E. A., Jackson, E. E., Chan, Y. A., Podevels, A. M., Berti, A. D., McMahon, M. D., and Thomas, M. G. (2008). Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol. Pharm.*, 5 :191–211.

[Frank et al., 2004] Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20 :2479–2481.

[Fruchterman and Reingold, 1991] Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Softw. Pract. Exper.*, 21(11) :1129–1164. doi:http://dx.doi.org/10.1002/spe.4380211102.

[Garey and Johnson, 1979] Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability : A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

[Geib et al., 2008] Geib, N., Woithe, K., Zerbe, K., Li, D. B., and Robinson, J. A. (2008). New insights into the first oxidative phenol coupling reaction during vancomycin biosynthesis. *Bioorg. Med. Chem. Lett.*, 18 :3081–3084.

[Grangemard et al., 2001] Grangemard, I., Wallach, J., Maget-Dana, R., and Peypoux, F. (2001). Lichenysin : a more efficient cation chelator than surfactin. *Appl. Biochem. Biotechnol.*, 90 :199–210.

[Green, 1981] Green, C. J. (1981). Immunosuppression with cyclosporin A : a review. *Diagn Histopathol*, 4 :157–174.

[Grünewald and Marahiel, 2006] Grünewald, J. and Marahiel, M. A. (2006). Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiol. Mol. Biol. Rev.*, 70 :121–146.

[Guenzi et al., 1998] Guenzi, E., Galli, G., Grgurina, I., Gross, D. C., and Grandi, G. (1998). Characterization of the syringomycin synthetase gene cluster. A link between prokaryotic and eukaryotic peptide synthetases. *J. Biol. Chem.*, 273 :32857–32863.

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

[Hamada et al., 2005] Hamada, T., Matsunaga, S., Yano, G., and Fusetani, N. (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.

[Harada et al., 2004] Harada, K., Imanishi, S., Kato, H., Mizuno, M., Ito, E., and Tsuji, K. (2004). Isolation of Adda from microcystin-LR by microbial degradation. *Toxicon*, 44 :107–109.

[Hofemeister et al., 2004] Hofemeister, J., Conrad, B., Adler, B., Hofemeister, B., Feesche, J., Kucheryava, N., Steinborn, G., Franke, P., Grammel, N., Zwintscher, A., Leenders, F., Hitzeroth, G., and Vater, J. (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.

[Hojati et al., 2002] Hojati, Z., Milne, C., Harvey, B., Gordon, L., Borg, M., Flett, F., Wilkinson, B., Sidebottom, P. J., Rudd, B. A., Hayes, M. A., Smith, C. P., and Micklefield, J. (2002). Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. *Chem. Biol.*, 9 :1175–1187.

[Höfte et al., 1991] Höfte, M., Seong, K. Y., Jurkevitch, E., and Verstraete, W. (1991). Pyoverdin production by the plant growth beneficial pseudomonas strain 7NSK2 : Ecological significance in soil. *Plant and Soil*, 130(1-2) :249–257. Available from : http://www.springerlink.com/content/u3436pu6n2j722u8/, http ://dx.doi.org/10.1007/BF00011880 doi:10.1007/BF00011880.

[Jacques et al., 2003] Jacques, P., Ongena, M., Bernard, F., Fuchs, R., Budzikiewicz, H., and Thonart, P. (2003). Fluorescent Pseudomonas mainly produce the dihydro form of pyoverdine at low specific growth rate. *Lett. Appl. Microbiol.*, 36 :259–262.

[Joardar et al., 2005] Joardar, V., Lindeberg, M., Jackson, R. W., Selengut, J., Dodson, R., Brinkac, L. M., Daugherty, S. C., Deboy, R., Durkin, A. S., Giglio, M. G., Madupu, R., Nelson, W. C., Rosovitz, M. J., Sullivan, S., Crabtree, J., Creasy, T., Davidsen, T., Haft, D. H., Zafar, N., Zhou, L., Halpin, R., Holley, T., Khouri, H., Feldblyum, T., White, O., Fraser, C. M., Chatterjee, A. K., Cartinhour, S., Schneider, D. J., Mansfield, J., Collmer, A., and Buell, C. R. (2005). Whole-genome sequence analysis of Pseudomonas syringae pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition. *J. Bacteriol.*, 187 :6488–6498.

[Kallow et al., 1997] Kallow, W., Neuhof, T., Arezi, B., Jungblut, P., and von Döhren, H. (1997). Penicillin biosynthesis : intermediates of biosynthesis of delta-L-alpha-aminoadipyl-L-cysteinyl-D-valine formed by ACV synthetase from Acremonium chrysogenum. *FEBS Lett.*, 414 :74–78.

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

[Kim et al., 2008] Kim, M. G., Kim, S. Y., Kim, W. Y., Mackey, D., and Lee, S. Y. (2008). Responses of Arabidopsis thaliana to challenge by Pseudomonas syringae. *Mol. Cells*, 25 :323–331.

[Koba and Konopa, 2005] Koba, M. and Konopa, J. (2005). [Actinomycin D and its mechanisms of action]. *Postepy Hig Med Dosw (Online)*, 59 :290–298.

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

[Kohli and Walsh, 2003] Kohli, R. M. and Walsh, C. T. (2003). Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chem. Commun. (Camb.)*, pages 297–307.

[Kohli et al., 2002] Kohli, R. M., Walsh, C. T., and Burkart, M. D. (2002). Biomimetic synthesis and optimization of cyclic peptide antibiotics. *Nature*, 418 :658–661.

[Konz et al., 1999] Konz, D., Doekel, S., and Marahiel, M. A. (1999). Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. *J. Bacteriol.*, 181 :133–140.

[Konz et al., 1997] Konz, D., Klens, A., Schörgendorfer, K., and Marahiel, M. A. (1997). The bacitracin biosynthesis operon of Bacillus licheniformis ATCC 10716 : molecular characterization of three multi-modular peptide synthetases. *Chem. Biol.*, 4 :927–937.

[Kracht et al., 1999] Kracht, M., Rokos, H., Ozel, M., Kowall, M., Pauli, G., and Vater, J. (1999). Antiviral and hemolytic activities of surfactin isoforms and their methyl ester derivatives. *J. Antibiot.*, 52 :613–619.

[Kurmayer et al., 2002] Kurmayer, R., Dittmann, E., Fastner, J., and Chorus, I. (2002). Diversity of microcystin genes within a population of the toxic cyanobacterium Microcystis spp. in Lake Wannsee (Berlin, Germany). *Microb. Ecol.*, 43 :107–118.

[Lam et al., 2000] Lam, P. K., Yang, M., and Lam, M. H. (2000). Toxicology and evaluation of microcystins. *Ther Drug Monit*, 22 :69–72.

[Lawen and Zocher, 1990] Lawen, A. and Zocher, R. (1990). Cyclosporin synthetase. The most complex peptide synthesizing multienzyme polypeptide so far described. *J. Biol. Chem.*, 265 :11355–11360.

[Lazo, 1999] Lazo, J. S. (1999). Bleomycin. *Cancer Chemother Biol Response Modif*, 18 :39–45.

[Leclère et al., 2009] Leclère, V., Beaufort, S., Dessoy, S., Dehottay, P., and Jacques, P. (2009). Development of a biological test to evaluate the bioavailability of iron in culture media. *J. Appl. Microbiol.*

[Leclère et al., 2005] Leclère, V., Béchet, M., Adam, A., Guez, J. S., Wathelet, B., Ongena, M., Thonart, P., Gancel, F., Chollet-Imbert, M., and Jacques, P. (2005). Mycosubtilin overproduction by Bacillus subtilis BBG100 enhances the organism's antagonistic and biocontrol activities. *Appl. Environ. Microbiol.*, 71 :4577–4584.

[Lee et al., 2001] Lee, Y. K., Lee, J.-H., and Lee, H. K. (2001). Microbial symbiosis in marine sponges. *J. Microbiol.*, 39 :254—-264.

[Levine, 2008] Levine, D. P. (2008). Vancomycin : understanding its past and preserving its future. *South. Med. J.*, 101 :284–291.

[Li et al., 2008] Li, Y., Weissman, K. J., and Müller, R. (2008). Myxochelin biosynthesis : direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. *J. Am. Chem. Soc.*, 130 :7554–7555.

[Lin et al., 1999] Lin, T. P., Chen, C. L., Chang, L. K., Tschen, J. S., and Liu, S. T. (1999). Functional and transcriptional analyses of a fengycin synthetase gene, fenC, from Bacillus subtilis. *J. Bacteriol.*, 181 :5060–5067.

[Lipmann et al., 1971] Lipmann, F., Gevers, W., Kleinkauf, H., and Roskoski, R. J. (1971). Polypeptide synthesis on protein templates : the enzymatic synthesis of gramicidin s and tyrocidine. *Adv Enzymol Relat Areas Mol Biol*, 35 :1–34.

[Livingstone and Barton, 1993] Livingstone, C. D. and Barton, G. J. (1993). Protein sequence alignments : a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, 9 :745–756.

[Magarvey et al., 2006] Magarvey, N. A., Haltli, B., He, M., Greenstein, M., and Hucul, J. A. (2006). Biosynthetic pathway for mannopeptimycins, lipoglycopeptide antibiotics active against drug-resistant gram-positive pathogens. *Antimicrob. Agents Chemother.*, 50 :2167–2177.

[Marahiel et al., 1997] Marahiel, M. A., Stachelhaus, T., and Mootz, H. D. (1997). Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem. Rev.*, 97 :2651–2674.

[Matthijs et al., 2009] Matthijs, S., Laus, G., Meyer, J. M., Abbaspour-Tehrani, K., Schäfer, M., Budzikiewicz, H., and Cornelis, P. (2009). Siderophore-mediated iron acquisition in the entomopathogenic bacterium Pseudomonas entomophila L48 and its close relative Pseudomonas putida KT2440. *Biometals*.

[Mehlhorn, 1984] Mehlhorn, K. (1984). *Graph algorithms and NP-completeness*. Springer-Verlag New York, Inc., New York, NY, USA.

[Meyer, 2000] Meyer, J. M. (2000). Pyoverdines : pigments, siderophores and potential taxonomic markers of fluorescent Pseudomonas species. *Arch. Microbiol.*, 174 :135–142.

[MEYER and ABDALLAH, 1978] MEYER, J. M. and ABDALLAH, M. A. (1978). The Fluorescent Pigment of Pseudomonas fluorescens : Biosynthesis, Purifica-

tion and Physicochemical Properties. *J Gen Microbiol*, 107(2) :319–328. Available from : http://mic.sgmjournals.org/cgi/content/abstract/107/2/319, http ://dx.doi.org/10.1099/00221287-107-2-319 doi:10.1099/00221287-107-2-319.

[Miao et al., 2005] Miao, V., Coëffet-Legal, M. F., Brian, P., Brost, R., Penn, J., Whiting, A., Martin, S., Ford, R., Parr, I., Bouchard, M., Silva, C. J., Wrigley, S. K., and Baltz, R. H. (2005). Daptomycin biosynthesis in Streptomyces roseosporus : cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology (Reading, Engl.)*, 151 :1507–1523.

[Miller et al., 2002] Miller, D. A., Luo, L., Hillson, N., Keating, T. A., and Walsh, C. T. (2002). Yersiniabactin synthetase : a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of Yersinia pestis. *Chem. Biol.*, 9 :333–344.

[Minowa et al., 2007] Minowa, Y., Araki, M., and Kanehisa, M. (2007). Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, 368 :1500–1517.

[Molnár et al., 2000] Molnár, I., Schupp, T., Ono, M., Zirkle, R., Milnamow, M., Nowak-Thompson, B., Engel, N., Toupet, C., Stratmann, A., Cyr, D. D., Gorlach, J., Mayo, J. M., Hu, A., Goff, S., Schmid, J., and Ligon, J. M. (2000). The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90. *Chem. Biol.*, 7 :97–109.

[Moon et al., 2008] Moon, C. D., Zhang, X. X., Matthijs, S., Schäfer, M., Budzikiewicz, H., and Rainey, P. B. (2008). Genomic, genetic and structural analysis of pyoverdine-mediated iron acquisition in the plant growth-promoting bacterium Pseudomonas fluorescens SBW25. *BMC Microbiol.*, 8 :7.

[Mootz et al., 2002a] Mootz, H. D., Kessler, N., Linne, U., Eppelmann, K., Schwarzer, D., and Marahiel, M. A. (2002a). Decreasing the ring size of a cyclic nonribosomal peptide antibiotic by in-frame module deletion in the biosynthetic genes. *J. Am. Chem. Soc.*, 124 :10980–10981.

[Mootz et al., 2000] Mootz, H. D., Schwarzer, D., and Marahiel, M. A. (2000). Construction of hybrid peptide synthetases by module and domain fusions. *Proc. Natl. Acad. Sci. U.S.A.*, 97 :5848–5853.

[Mootz et al., 2002b] Mootz, H. D., Schwarzer, D., and Marahiel, M. A. (2002b). Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *Chembiochem*, 3 :490–504.

[Morrison and Wenzel, 1984] Morrison, A. J. and Wenzel, R. P. (1984). Epidemiology of infections due to Pseudomonas aeruginosa. *Rev. Infect. Dis.*, 6 Suppl 3 :S627–642.

[Mossialos et al., 2002] Mossialos, D., Ochsner, U., Baysse, C., Chablain, P., Pirnay, J. P., Koedam, N., Budzikiewicz, H., Fernández, D. U., Schäfer, M., Ravel, J., and Cornelis, P. (2002). Identification of new, conserved, non-ribosomal peptide synthetases from fluorescent pseudomonads involved in the biosynthesis of the siderophore pyoverdine. *Mol. Microbiol.*, 45 :1673–1685.

[Moyne et al., 2004] 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.

[Nemoto et al., 2002] Nemoto, A., Hoshino, Y., Yazawa, K., Ando, A., Mikami, Y., Komaki, H., Tanaka, Y., and Gräfe, U. (2002). Asterobactin, a new siderophore group antibiotic from Nocardia asteroides. *J. Antibiot.*, 55 :593–597.

[Pandey et al., 1994] Pandey, A., , Bringel, F., and Meyer, J.-M. (1994). Iron requirement and search for siderophores in lactic acid bacteria. *Applied Microbiology and Biotechnology*, 40(5) :735–739. Available from : http://www.springerlink.com/content/v52307k2p5845hj8/, http ://dx.doi.org/10.1007/BF00173337 doi:10.1007/BF00173337.

[Pelludat et al., 1998] Pelludat, C., Rakin, A., Jacobi, C. A., Schubert, S., and Heesemann, J. (1998). The yersiniabactin biosynthetic gene cluster of Yersinia enterocolitica : organization and siderophore-dependent regulation. *J. Bacteriol.*, 180 :538–546.

[Peypoux et al., 1999] Peypoux, F., Bonmatin, J. M., and Wallach, J. (1999). Recent trends in the biochemistry of surfactin. *Appl. Microbiol. Biotechnol.*, 51 :553–563.

[Pfennig et al., 1999] Pfennig, F., Schauwecker, F., and Keller, U. (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.

[Piel, 2009] Piel, J. (2009). Metabolites from symbiotic bacteria. *Nat Prod Rep*, 26 :338–362.

[Raaijmakers et al., 2006] Raaijmakers, J. M., de Bruijn, I., and de Kock, M. J. (2006). Cyclic lipopeptide production by plant-associated Pseudomonas spp. : diversity, activity, biosynthesis, and regulation. *Mol. Plant Microbe Interact.*, 19 :699–710.

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

[Ravel and Cornelis, 2003] Ravel, J. and Cornelis, P. (2003). Genomics of pyoverdine-mediated iron uptake in pseudomonads. *Trends Microbiol.*, 11 :195–200.

[Raymond and Willett, 2002] Raymond, J. W. and Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7) :521–533.

[Raymond et al., 2003] Raymond, K. N., Dertz, E. A., and Kim, S. S. (2003). Enterobactin : an archetype for microbial iron transport. *Proc. Natl. Acad. Sci. U.S.A.*, 100 :3584–3588.

[Reverchon et al., 2002] Reverchon, S., Rouanet, C., Expert, D., and Nasser, W. (2002). Characterization of indigoidine biosynthetic genes in Erwinia chrysanthemi and role of this blue pigment in pathogenicity. *J. Bacteriol.*, 184 :654–665.

[Romero et al., 1997] Romero, F., Espliego, F., Pérez Baz, J., García de Quesada, T., Grávalos, D., de la Calle, F., and Fernández-Puentes, J. L. (1997). Thiocoraline, a new depsipeptide with antitumor activity produced by a marine Micromonospora. I. Taxonomy, fermentation, isolation, and biological activities. *J. Antibiot.*, 50 :734–737.

[Roongsawang et al., 2003] Roongsawang, N., Hase, K., Haruki, M., Imanaka, T., Morikawa, M., and Kanaya, S. (2003). Cloning and characterization of the gene cluster encoding arthrofactin synthetase from Pseudomonas sp. MIS38. *Chem. Biol.*, 10 :869–880.

[Rouhiainen et al., 2000] Rouhiainen, L., Paulin, L., Suomalainen, S., Hyytiäinen, H., Buikema, W., Haselkorn, R., and Sivonen, K. (2000). Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in Anabaena strain 90. *Mol. Microbiol.*, 37 :156–167.

[Sayers et al., 2009] Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer,

L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 37 :5–15.

[Schellenberg et al., 2007] Schellenberg, B., Bigler, L., and Dudler, R. (2007). Identification of genes involved in the biosynthesis of the cytotoxic compound glidobactin from a soil bacterium. *Environ. Microbiol.*, 9 :1640–1650.

[Schenker et al., 2003] Schenker, A., Last, M., Bunke, H., and Kandel, A. (2003). Comparison of distance measures for graph-based clustering of documents. In Hancock, E. R. and Vento, M., editors, *Proceedings of the Graph Based Representations in Pattern Recognition, 4th IAPR International Workshop (GbRPR 2003)*, volume 2726 of *Lecture Notes in Computer Science*, pages 202–213. Springer. Available from : `http://springerlink.metapress.com/link.asp?id=bxkjxj1vt9mlxylx`.

[Schwarzer et al., 2003] Schwarzer, D., Finking, R., and Marahiel, M. A. (2003). Nonribosomal peptides : from genes to products. *Nat Prod Rep*, 20 :275–287.

[Schwarzer and Marahiel, 2001] Schwarzer, D. and Marahiel, M. A. (2001). Multimodular biocatalysts for natural product assembly. *Naturwissenschaften*, 88 :93–101.

[Shen et al., 2001] Shen, B., Du, L., Sanchez, C., Edwards, D. J., Chen, M., and Murrell, J. M. (2001). The biosynthetic gene cluster for the anticancer drug bleomycin from Streptomyces verticillus ATCC15003 as a model for hybrid peptide-polyketide natural product biosynthesis. *J. Ind. Microbiol. Biotechnol.*, 27 :378–385.

[Shen et al., 2002] Shen, B., Du, L., Sanchez, C., Edwards, D. J., Chen, M., and Murrell, J. M. (2002). Cloning and characterization of the bleomycin biosynthetic gene cluster from Streptomyces verticillus ATCC15003. *J. Nat. Prod.*, 65 :422–431.

[Sieber and Marahiel, 2003] Sieber, S. A. and Marahiel, M. A. (2003). Learning from nature's drug factories : nonribosomal synthesis of macrocyclic peptides. *J. Bacteriol.*, 185 :7036–7043.

[Sieber and Marahiel, 2005] Sieber, S. A. and Marahiel, M. A. (2005). Molecular mechanisms underlying nonribosomal peptide synthesis : approaches to new antibiotics. *Chem. Rev.*, 105 :715–738.

[Siezen et al., 2008] Siezen, R. J., Starrenburg, M. J., Boekhorst, J., Renckens, B., Molenaar, D., and van Hylckama Vlieg, J. E. (2008). Genome-scale genotype-phenotype matching of two Lactococcus lactis isolates from plants identifies mechanisms of adaptation to the plant niche. *Appl. Environ. Microbiol.*, 74 :424–436.

[Smith et al., 1990] Smith, D. J., Earl, A. J., and Turner, G. (1990). The multifunctional peptide synthetase performing the first step of penicillin biosynthesis in Penicillium chrysogenum is a 421,073 dalton protein similar to Bacillus brevis peptide antibiotic synthetases. *EMBO J.*, 9 :2743–2750.

[Stachelhaus et al., 1999] Stachelhaus, T., Mootz, H. D., and Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, 6 :493–505.

[Stachelhaus et al., 1995] Stachelhaus, T., Schneider, A., and Marahiel, M. A. (1995). Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science*, 269 :69–72.

[Starcevic et al., 2008] Starcevic, A., Zucko, J., Simunkovic, J., Long, P. F., Cullum, J., and Hranueli, D. (2008). 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 Res.*, 36 :6882–6892.

[Strieker and Marahiel, 2009] Strieker, M. and Marahiel, M. A. (2009). The structural diversity of acidic lipopeptide antibiotics. *Chembiochem*, 10 :607–616.

[Tabata et al., 2005] Tabata, K., Ikeda, H., and Hashimoto, S. (2005). ywfE in Bacillus subtilis codes for a novel enzyme, L-amino acid ligase. *J. Bacteriol.*, 187 :5195–5202.

[Tan et al., 2000] Tan, L. T., Williamson, R. T., Gerwick, W. H., Watts, K. S., McGough, K., and Jacobs, R. (2000). cis,cis- and trans,trans-ceratospongamide, new bioactive cyclic heptapeptides from the Indonesian red alga Ceratodictyon spongiosum and symbiotic sponge Sigmadocia symbiotica. *J. Org. Chem.*, 65 :419–425.

[Thompson et al., 2002] Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2 :Unit 2.3.

[Tobiasen et al., 2007] Tobiasen, C., Aahman, J., Ravnholt, K. S., Bjerrum, M. J., Grell, M. N., and Giese, H. (2007). Nonribosomal peptide synthetase (NPS) genes in Fusarium graminearum, F. culmorum and F. pseudograminearium and identification of NPS2 as the producer of ferricrocin. *Curr. Genet.*, 51 :43–58.

[Trauger et al., 2000] Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A., and Walsh, C. T. (2000). Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature*, 407 :215–218.

[Trauger et al., 2001] Trauger, J. W., Kohli, R. M., and Walsh, C. T. (2001). Cyclization of backbone-substituted peptides catalyzed by the thioesterase domain from the tyrocidine nonribosomal peptide synthetase. *Biochemistry*, 40 :7092–7098.

[Tseng et al., 2002] Tseng, C. C., Bruner, S. D., Kohli, R. M., Marahiel, M. A., Walsh, C. T., and Sieber, S. A. (2002). Characterization of the surfactin synthetase C-terminal thioesterase domain as a cyclic depsipeptide synthase. *Biochemistry*, 41 :13350–13359.

[van Wageningen et al., 1998] van Wageningen, A. M., Kirkpatrick, P. N., Williams, D. H., Harris, B. R., Kershaw, J. K., Lennard, N. J., Jones, M., Jones, S. J., and Solenberg, P. J. (1998). Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem. Biol.*, 5 :155–162.

[Vizcaíno et al., 2006] Vizcaíno, J. A., Cardoza, R. E., Dubost, L., Bodo, B., Gutiérrez, S., and Monte, E. (2006). Detection of peptaibols and partial cloning of a putative peptaibol synthetase gene from T. harzianum CECT 2413. *Folia Microbiol. (Praha)*, 51 :114–120.

[Vodovar et al., 2006] Vodovar, N., Vallenet, D., Cruveiller, S., Rouy, Z., Barbe, V., Acosta, C., Cattolico, L., Jubin, C., Lajus, A., Segurens, B., Vacherie, B., Wincker, P., Weissenbach, J., Lemaitre, B., Médigue, C., and Boccard, F. (2006). Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium Pseudomonas entomophila. *Nat. Biotechnol.*, 24 :673–679.

[Vodovar et al., 2005] Vodovar, N., Vinals, M., Liehl, P., Basset, A., Degrouard, J., Spellman, P., Boccard, F., and Lemaitre, B. (2005). Drosophila host defense after oral infection by an entomopathogenic Pseudomonas species. *Proc. Natl. Acad. Sci. U.S.A.*, 102 :11414–11419.

[Voet et al., 2007a] Voet, D., Voet, J. G., and Pratt, C. W. (2007a). *Principles of Biochemistry : Life at the Molecular Level*, chapter 4. John Wiley & Sons Inc.

[Voet et al., 2007b] Voet, D., Voet, J. G., and Pratt, C. W. (2007b). *Principles of Biochemistry : Life at the Molecular Level*, chapter 5. John Wiley & Sons Inc.

[Voet et al., 2007c] Voet, D., Voet, J. G., and Pratt, C. W. (2007c). *Principles of Biochemistry : Life at the Molecular Level*, chapter 6. John Wiley & Sons Inc.

[Voet et al., 2007d] Voet, D., Voet, J. G., and Pratt, C. W. (2007d). *Principles of Biochemistry : Life at the Molecular Level*, chapter 3. John Wiley & Sons Inc.

[Voet et al., 2007e] Voet, D., Voet, J. G., and Pratt, C. W. (2007e). *Principles of Biochemistry : Life at the Molecular Level*, chapter 26. John Wiley & Sons Inc.

[Voet et al., 2007f] Voet, D., Voet, J. G., and Pratt, C. W. (2007f). *Principles of Biochemistry : Life at the Molecular Level*, chapter 27. John Wiley & Sons Inc.

[von Döhren et al., 1997] von Döhren, H., Keller, U., Vater, J., and Zocher, R. (1997). Multifunctional Peptide Synthetases. *Chem. Rev.*, 97 :2675–2706.

[Wade et al., 1990] Wade, D., Boman, A., Wåhlin, B., Drain, C. M., Andreu, D., Boman, H. G., and Merrifield, R. B. (1990). All-D amino acid-containing channel-forming antibiotic peptides. *Proc. Natl. Acad. Sci. U.S.A.*, 87 :4761–4765.

[Walton, 2006] Walton, J. D. (2006). HC-toxin. *Phytochemistry*, 67 :1406–1413.

[Weber et al., 1994] Weber, G., Schörgendorfer, K., Schneider-Scherzer, E., and Leitner, E. (1994). The peptide synthetase catalyzing cyclosporine production in Tolypocladium niveum is encoded by a giant 45.8-kilobase open reading frame. *Curr. Genet.*, 26 :120–125.

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

[Weis et al., 2008] Weis, F., Beiras-Fernandez, A., and Schelling, G. (2008). Daptomycin, a lipopeptide antibiotic in clinical practice. *Curr Opin Investig Drugs*, 9 :879–884.

[Whitmore and Wallace, 2004a] Whitmore, L. and Wallace, B. A. (2004a). Analysis of peptaibol sequence composition : implications for in vivo synthesis and channel formation. *Eur. Biophys. J.*, 33 :233–237.

[Whitmore and Wallace, 2004b] Whitmore, L. and Wallace, B. A. (2004b). The Peptaibol Database : a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Res.*, 32 :D593–594.

[Wiest et al., 2002] Wiest, A., Grzegorski, D., Xu, B. W., Goulard, C., Rebuffat, S., Ebbole, D. J., Bodo, B., and Kenerley, C. (2002). Identification of peptaibols from Trichoderma virens and cloning of a peptaibol synthetase. *J. Biol. Chem.*, 277 :20862–20868.

[Yakimov et al., 1999] Yakimov, M. M., Abraham, W. R., Meyer, H., Giuliano, L., and Golyshin, P. N. (1999). Structural characterization of lichenysin A components by fast atom bombardment tandem mass spectrometry. *Biochim. Biophys. Acta*, 1438 :273–80.

[Yakimov et al., 1998] Yakimov, M. M., Kröger, A., Slepak, T. N., Giuliano, L., Timmis, K. N., and Golyshin, P. N. (1998). A putative lichenysin A synthetase operon in Bacillus licheniformis : initial characterization. *Biochim. Biophys. Acta*, 1399 :141–153.

[Yakimov et al., 1995] Yakimov, M. M., Timmis, K. N., Wray, V., and Fredrickson, H. L. (1995). Characterization of a new lipopeptide surfactant produced by thermotolerant and halotolerant subsurface Bacillus licheniformis BAS50. *Appl. Environ. Microbiol.*, 61 :1706–1713.

[Zhang et al., 2009] Zhang, W., Li, Z., Miao, X., and Zhang, F. (2009). The screening of antimicrobial bacteria with diverse novel nonribosomal peptide synthetase (NRPS) genes from South China sea sponges. *Mar. Biotechnol.*, 11 :346–355.

Résumé

Les peptides non-ribosomiaux sont des molécules produites par les micro-organismes et présentant un large éventail d'activités biologiques et pharmaceutiques. Par exemple, ils peuvent présenter des activités antibiotiques, immuno-modulatrices ou anti-tumorales. Ces peptides sont synthétisés par de grands complexes multi-enzymatiques, appelés synthétases ou NRPS (NonRibosomal Peptide Synthetases). Deux traits caractéristiques distinguent ces peptides des peptides ribosomiaux classiques : le premier est que leur structure primaire n'est pas toujours linéaire mais peut être totalement ou partiellement cyclique, branchée voir même poly-cyclique, et le second est la diversité des monomères incorporés au sein de ces peptides qui dépasse largement les vingt acides aminés protéogéniques. Nous avons développé Norine, la première ressource publique entièrement dédiée aux peptides non-ribosomiaux. Norine contient actuellement plus de 1 000 peptides, modélisés par des graphes étiquetés non-orientés, ainsi que des outils informatiques permettant leur analyse, comme la comparaison de compositions en monomères, la recherche de motifs structuraux ou la recherche par similarité. Des analyses statistiques sur les données contenues dans Norine ont permis de mettre en évidence des caractéristiques biologiques intéressantes comme la spécificité des monomères en fonction de l'activité biologique qui nous a conduit à l'élaboration d'un outil d'aide à la prédiction de la fonction biologique d'un peptide à partir de sa composition monomérique. En trois ans, Norine est devenue la ressource internationale pour les peptides non-ribosomiaux.

Mots clefs : bio-informatique, peptides nonribosomiaux, NRPS, base de données, algorithme, graphe

Abstract

Nonribosomal peptides are molecules produced by microorganisms and displaying a broad spectrum of biological activities and pharmaceutical applications. They can harbor anti-microbial, immunomodulating or anti-tumor activities. These peptides are synthesized by huge multi-enzymatic complexes, called NonRibosomal Peptide Synthetases. Two main structural traits distinguish these peptides from the ribosomally synthesized ones : first, their primary structure is not always linear but is often cyclic (partially or totally), branched or poly-cyclic, and second, the diversity of monomers incorporated into nonribosomal peptides extends far beyond the 20 proteogenic amino acids residues. We have developed Norine, the first public resource entirely dedicated to nonribosomal peptides. Norine currently contains more than 1 000 peptides, modeled by non-oriented labeled graphs, and computational tools allowing their analysis, such as monomer composition comparison, structural pattern matching or similarity search. Statistical analysis of Norine data highlighted interesting biological properties such as a specific monomer composition depending on the biological activity, that led us to develop a tool for helping the prediction of peptide activity from its monomeric composition. In three years, Norine became the international resource for nonribosomal peptides.

Keywords : bioinformatics, nonribosomal peptides, NRPS, database, algorithm, graph