Non Ribosomal Peptides : A monomeric puzzle

Yoann DUFRESNE[1], Valérie LECLÈRE[2], Philippe JACQUES[2], Laurent NOÉ[1] and Maude PUPIN[1]

[1] LIFL, UMR USTL/CNRS 8022, INRIA Lille-Nord Europe, 59655 Villeneuve d'Ascq, France
Yoann Dufresne : yoann.dufresne@etudiant.univ-lille1.fr, fmaude.pupin, laurent.noeg@univ-lille1.fr

[2] ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, Polytech-Lille/IUTA, 59655 Villeneuve d'Ascq, France
valerie.leclere@univ-lille1.fr, philippe.jacques@polytech-lille.fr

Abstract *Nonribosomal peptides (NRPs) are increasingly studied because they harbor activities which can be exploited in various domains. They are often denoted as graphs illustrating their chemical structure, where the atoms are represented by nodes and the chemical bonds by arcs. Another possible representation is the monomeric structure. This structure, inspired by the biosynthetic pathway of these peptides, is effectuated by large enzymatic complexes which assemble together smaller compounds called monomers. Consequently, the nonribosomal peptides are composed of a great variety of monomers (more than 500 are known) including amino acids, lipids and carbohydrates. Likewise, nonpetidic bonds are formed between multiple monomers, producing peptides with cycles and/or branches. Thus, the monomeric structure is a graph formed by the monomers present in the peptide and their interlinking chemical bonds. Until now, there did not exist a tool allowing for the conversion between the atomic and monomeric structures. This article presents a novel algorithm capable of localising the monomers from a reference list in the chemical structures of peptides extracted from the Norine database. The algorithm is based on a heuristic that utilizes chemical information of NRPs. The preliminary results are encouraging, and should lead to further studies.*

Keywords nonribosomal peptides, chemical structures, graphs.

Les peptides non-ribosomiques : un puzzle monomérique

Résumé *Les peptides non-ribosomiques (NRP) sont des molécules de plus en plus étudiées car elles présentent des activités ayant des applications principalement dans le domaine pharmaceutique. Elles sont souvent décrites par leur structure chimique, c'est-à-dire un graphe dont les nœuds sont des atomes et les arêtes les liaisons chimiques. Une autre représentation possible est la structure monomérique. Cette structure, inspirée de la voie de synthèse de ces peptides, est réalisé par de gros complexes enzymatiques qui assemblent les briques de base, appelées monomères. Ainsi, les peptides non-ribosomiques sont composés d'une grande variété de monomères (plus de 500 recensés jusqu'à présent) tels que des acides aminés, mais aussi des lipides ou des sucres. De plus, des liaisons non-peptidiques peuvent être formées entre certains monomères, ce qui produit des peptides contenant des cycles et/ou des branchements. La structure monomérique est donc le graphe formé par les monomères présents dans le peptide et les liaisons qui les relient. A l'heure actuelle, il n'existe pas d'outil permettant de convertir la structure chimique d'un peptide non-ribosomique en sa structure monomérique. Cet article présente un algorithme capable de localiser les monomères d'une liste de référence dans les structures chimiques des peptides de la base de données Norine. Il est basé sur une heuristique gloutonne qui utilise des connaissances sur la chimie des NRP. Les résultats préliminaires sont satisfaisants et devraient conduire à de nouvelles études.*

Mots-clés Peptides non-ribosomiques, structures chimiques, graphes.

1 Introduction

Les peptides non-ribosomiques sont synthétisés par certains micro-organismes (bactéries et fungi) et couvrent un large spectre d'activités biologiques [1]. Leur grande diversité de structures et de propriétés physico-chimiques est due à leur mode de synthèse, alternatif à la voie classique pour les peptides et protéines. Leur

synthèse est mise en œuvre par de grands complexes enzymatiques appelés synthétases peptidiques non-ribosomiques (abrévié NRPS en anglais).

Les peptides non-ribosomiques sont une source, encore sous-exploitée, de principes actifs dans l'industrie pharmaceutique, l'industrie cosmétique, le domaine des pesticides, des détergents et de la dépollution. Plusieurs de ces peptides sont déjà commercialisés tels que la pénicilline et d'autres antibiotiques, la cyclosporine qui réduit les risques de rejet de greffes ou l'actinomycine utilisée dans le traitement de certains cancers. Les étapes préliminaires à l'étude de la mise sur le marché de nouveaux principes actifs est la découverte de nouvelles molécules et l'étude de leur(s) activité(s). La réalisation expérimentale de ces étapes est coûteuse et demande une grande expertise. Il est cependant possible de réaliser une partie du travail via l'analyse bio-informatique des différentes sources de données à disposition afin de réduire le champ de criblage de manière efficace et ainsi diminuer fortement le temps d'investigation et les coûts engendrés.

À la différence des peptides classiques, les peptides non-ribosomiques (NRP) ne sont pas linéaires. Leur synthèse complexe vient modifier la structure linéaire pour ajouter des cycles et des branchements. Les briques de base (monomères) composant les NRP sont une deuxième source de différence avec les peptides classiques. Alors que les peptides classiques se basent sur les 20 acides aminés standards, les NRP s'appuient sur plus de 500 monomères différents dont environ 200 acides aminés, mais aussi des sucres et des lipides [2].

Dans les articles scientifiques et bases de données de petites molécules, les peptides non-ribosomiques sont décrits sous la forme de structures chimiques, c'est-à-dire un graphe dont les sommets sont les atomes et les arêtes sont les liaisons. Cependant, dans la base de données Norine et pour certaines utilisations comme la prédiction d'activité, il est intéressant d'avoir la structure monomérique, c'est-à-dire un graphe dont les sommets sont les monomères, chaque monomère étant composé de plusieurs atomes. La description sous forme de structure monomérique est inspirée du mode de synthèse des NRP. En effet, les synthétases sélectionnent spécifiquement les monomères puis les assemblent pour former un peptide. Ainsi, les monomères sont équivalents aux acides aminés ou aux nucléotides incorporés, respectivement, dans les protéines et les acides nucléiques. Pour l'instant, le passage de la structure chimique à la structure monomérique est réalisée manuellement et nécessite l'expertise de biochimistes qui découpent les monomères grâce à leurs connaissances concernant les NRP.

Dans cet article, nous allons présenter les travaux en cours au sein de l'équipe Bonsai du LIFL, en collaboration avec le laboratoire ProBioGEM, afin de permettre l'automatisation de cette tâche chronophage. Dans une première partie nous expliquerons pourquoi il est nécessaire d'avoir une approche différente de celle appliquée aux peptides classiques. Ensuite nous expliquerons les méthodes mises au point pour créer une première heuristique gloutonne qui a été testée sur la base de données Norine [3]. Enfin, nous parlerons des perspectives envisagées afin d'améliorer les performances de ce premier algorithme.

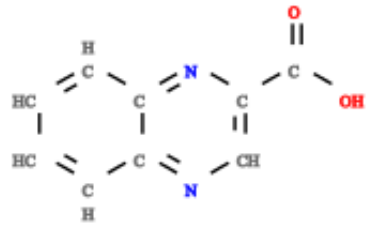
2 Objectifs

La méthode la plus utilisée pour connaître la séquence en acides aminés d'un peptide ribosomique est de traduire, à l'aide d'outils bio-informatiques, la séquence nucléique codant ce peptide. Il est également possible de faire du séquençage de peptides, mais cette technique expérimentale, lourde et coûteuse, est peu employée. Les NRP n'étant pas directement issus de la traduction des ARN, les outils bio-informatiques classiques ne sont pas transposables. Il est donc nécessaire de passer par la détermination expérimentale de la structure qui aboutit à une structure chimique. La décomposition d'un peptide en monomères revient à rechercher les graphes chimiques des monomères dans le graphe chimique du peptide que l'on cherche à décomposer.

La difficulté vient du fait que les monomères ne sont pas intégrés tels quels dans les NRP. En effet, à chaque formation d'une liaison entre deux monomères, certains atomes sont perdus (par exemple, dans le cas d'une liaison peptidique, un H_2O est libéré avec la perte d'un OH^- et d'un H^+). Et surtout, les liaisons effectuées ne sont pas toutes des liaisons peptidiques, ce qu'il faut prendre en compte.

3 Algorithme de conversion d'une structure chimique en structure monomérique

Dans un premier temps, nous avons développé une heuristique gloutonne basée sur une bibliothèque de fonctions pour la chimie appelée Openbabel [4]. Les molécules chimiques peuvent être considérées comme des graphes étiquetés et la bibliothèque intègre un format de représentation de ces graphes appelés SMILES (FIG. 1). Openbabel intègre également un outil de recherche de motifs dans des molécules que nous utilisons pour la recherche des monomères dans les peptides. La syntaxe appelée SMARTS [5] permet d'exprimer les motifs sous la forme d'expressions régulières de type SMILES.



OC(=O)c1cnc2c(n1)cccc2

Figure 1. SMILES et structure chimique de la 2-carboxyquinoxaline

3.1 Radicaux et liaisons

Comme expliqué précédemment, en se liant les uns aux autres, les monomères se transforment. Pour les localiser dans les peptides, il est nécessaire de générer les différents *radicaux* : il s'agit de monomères qui ont perdus certains atomes lors de la formation d'une liaison. Afin d'inférer l'ensemble des radicaux possibles, nous avons étudié les liaisons formées dans les peptides de la base de données Norine.

La liaison peptidique est la plus fréquente. Elle est effectuée entre les groupements NH_2 et COOH (FIG. 2). Cette liaison peut parfois être légèrement modifiée. Par exemple, dans la proline, l'atome d'azote est dans un cycle et forme un groupement NH et non NH_2.

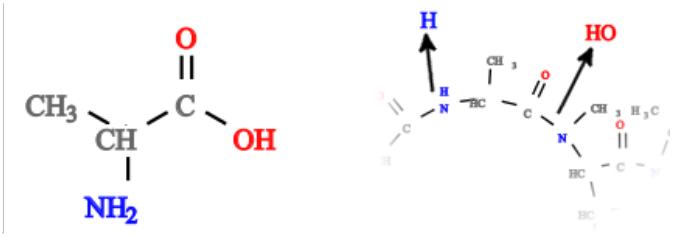


Figure 2. Alanine non liée - Alanine avec deux liaisons peptidiques

Certains acides aminés comme la cystéine possèdent un atome de soufre. Deux monomères portant un soufre peuvent former un pont disulfure (FIG. 3).

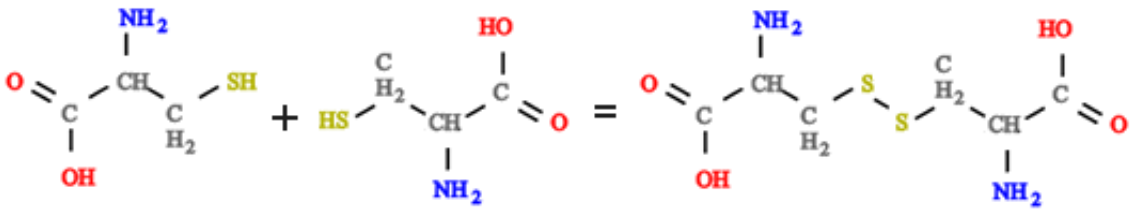


Figure 3. Formation d'un pont disulfure

Enfin, nous avons également observé des liaisons qui s'effectuent sur les cycles aromatiques. L'un des carbones de ces cycles peut perdre un hydrogène afin de former une liaison (FIG. 4).

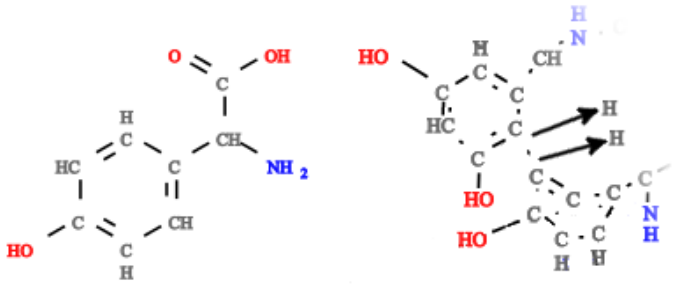


Figure 4. Hpg non lié - Deux Hpg liés dans la vancomicyne

Grâce à l'identification des liaisons présentes dans les NRP, nous sommes en mesure de construire des règles SMARTS pour localiser les atomes susceptibles d'être impliqués dans ces liaisons. Ainsi, nous pouvons générer tous les radicaux possibles pour chaque monomère (au total, 18030 radicaux pour les 531 monomères). Prenons l'exemple de la cystéine (FIG. 5). On peut distinguer sur la molécule trois groupements pouvant former une liaison (COOH, NH_2 et SH). Selon les peptides, ces liaisons ne sont pas toutes utilisées. En énumérant les cas possibles (liaisons uniques, paires et triplet), on peut générer 7 radicaux différents (FIG. 5).

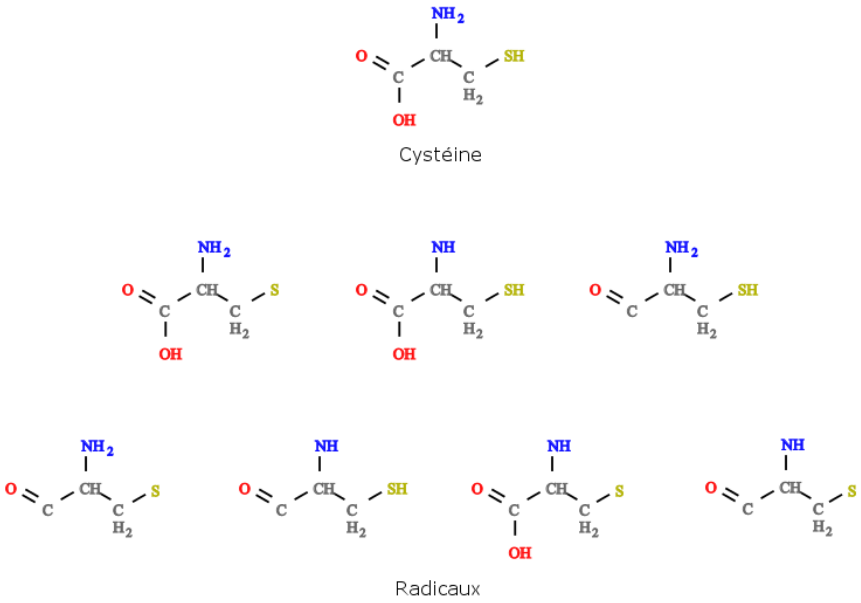


Figure 5. La Cystéine et ses radicaux générés

3.2 Heuristique gloutonne

Grâce aux règles précédemment générées et à l'API d'Openbabel, nous pouvons désormais rechercher les radicaux correspondant à l'ensemble des monomères au sein des peptides. Mais une question subsiste : "Quel monomère choisir lorsque deux radicaux différents peuvent être placés à un même endroit de la molécule ?"

Pour résoudre ce problème, nous avons choisi dans un premier temps d'appliquer une heuristique gloutonne de placement. On recherche séquentiellement les radicaux triés selon les critères définis ci-dessous. Dès que

l'un d'entre eux recouvre une partie du peptide, la partie couverte devient inaccessible pour les suivants (voir Algorithm 1).

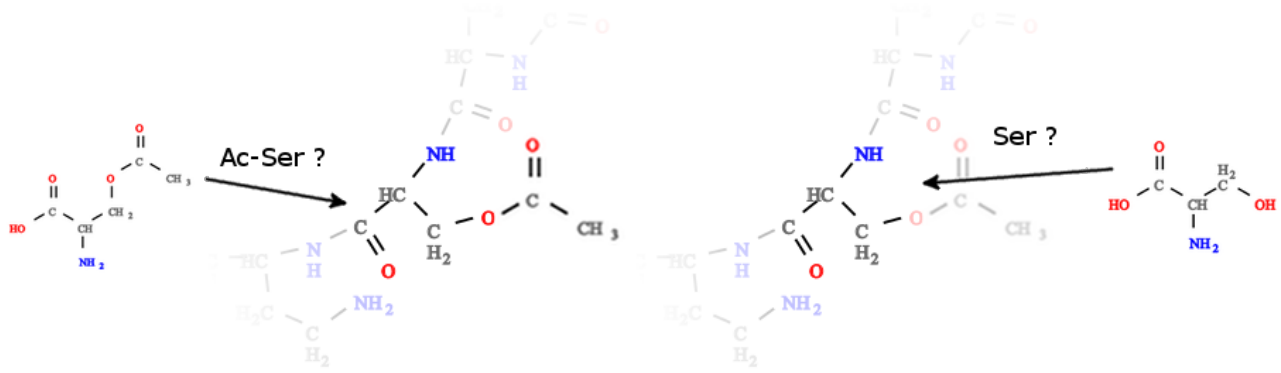


Figure 6. Exemple de peptide dans lequel plusieurs monomères peuvent être placés à un même endroit

Le premier critère discriminant choisi pour le tri est la taille (nombre d'atomes). La principale raison vient du fait que certains monomères et, a fortiori leurs radicaux, sont totalement inclus dans d'autres. Il est donc préférable, en règle générale, de placer les monomères les plus gros en premier pour que les petits ne prennent pas leur place. Par exemple, dans le peptide de la FIG. 6, il est possible de placer deux radicaux de deux monomères différents sur les mêmes atomes. De plus, certains radicaux d'un même monomère peuvent occuper le même emplacement, en couvrant plus ou moins d'atomes. Si le mauvais radical est choisi, il entrave des atomes. Ces atomes entravés peuvent, soit décaler le positionnement d'autres monomères, soit purement et simplement empêcher la pose d'autres monomères.

Un second critère de tri est appliqué pour discriminer les radicaux de même taille. Il s'agit de donner un score à chaque radical en fonction des liaisons qu'il peut former. Le poids de chaque liaison est déterminé en fonction de la fréquence de celle-ci dans les données. Par exemple, la liaison peptidique, étant majoritaire dans les NRP, a le poids le plus élevé. Le score d'un radical est alors défini comme la somme des poids des liaisons qu'il peut former. Ainsi, les radicaux peuvent être triés par taille puis score décroissants.

Algorithm 1: Heuristique gloutonne

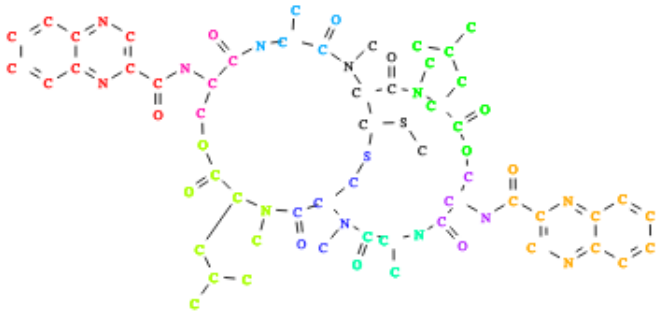
Data: Un peptide P et une liste de monomères M
Result: Couverture de P par des monomères m tels que $m \ 2 \ M$
Soit R une liste initialement vide de radicaux;
forall the $m \ 2 \ M$ **do**
    Ajouter dans R tous les radicaux possibles de m;
end
Trier R selon la taille puis la pondération;
forall the $r \ 2 \ R$ **do**
    if r *match sur des atomes non couverte de* P **then**
        Retenir le monomère m ayant permis de générer r;
        Couvrir les atomes trouvés de P;
    end
end

3.3 Résultats

3.3.1 Description de la sortie du programme. Pour tester cette heuristique, nous avons utilisé les données provenant de la base de donnée Norine qui fait référence pour les NRP. Elle nous a permis d'obtenir les SMILES des 531 monomères qu'elle contient, ainsi que 204 peptides dont à la fois le SMILES et la structure

monomérique sont connus, parmi les 1200 peptides de la base. Dans cette section, nous allons présenter les résultats obtenus en comparant les prédictions du programme avec les annotations manuelles de Norine. La sortie du programme est disponible à l'adresse suivante :
`http://www.lifl.fr/~dufresne/norine/greedysplit/`

Voici comment interpréter ces pages : dans chacun des cadres, nous affichons les informations relatives à un peptide non-ribosomique. La première partie du cadre contient l'image générée par notre programme, c'est-à-dire le peptide coloré en fonction des monomères qui le recouvrent, ainsi que des indicateurs des performances de l'algorithme (FIG. 7). Le taux de couverture atomique (nombre d'atomes couverts / nombre total d'atomes dans le peptide), ainsi que le taux de couverture monomérique (nombre de monomères correctement trouvés / nombre de monomères dans le peptide) sont indiqués.



quinomycin B
Atomic coverage : 0.898
Monomeric coverage : 0.7

Monomers:

found	wrong	not found	total
7	2	3	10

Figure 7. Exemple de résultat pour un peptide

À la suite de cette première partie se trouvent trois listes (FIG. 8) représentant , respectivement :
- les monomères qui n'ont pas été trouvés dans le peptide alors qu'ils en font partie (FN) ;
- les monomères trouvés qui n'en font pas réellement partie (FP) ;
- les monomères trouvés qui sont vraiment présents dans le peptide (VP).

Pour aider le lecteur à localiser les monomères dans le peptide, une légende colorée est donnée.

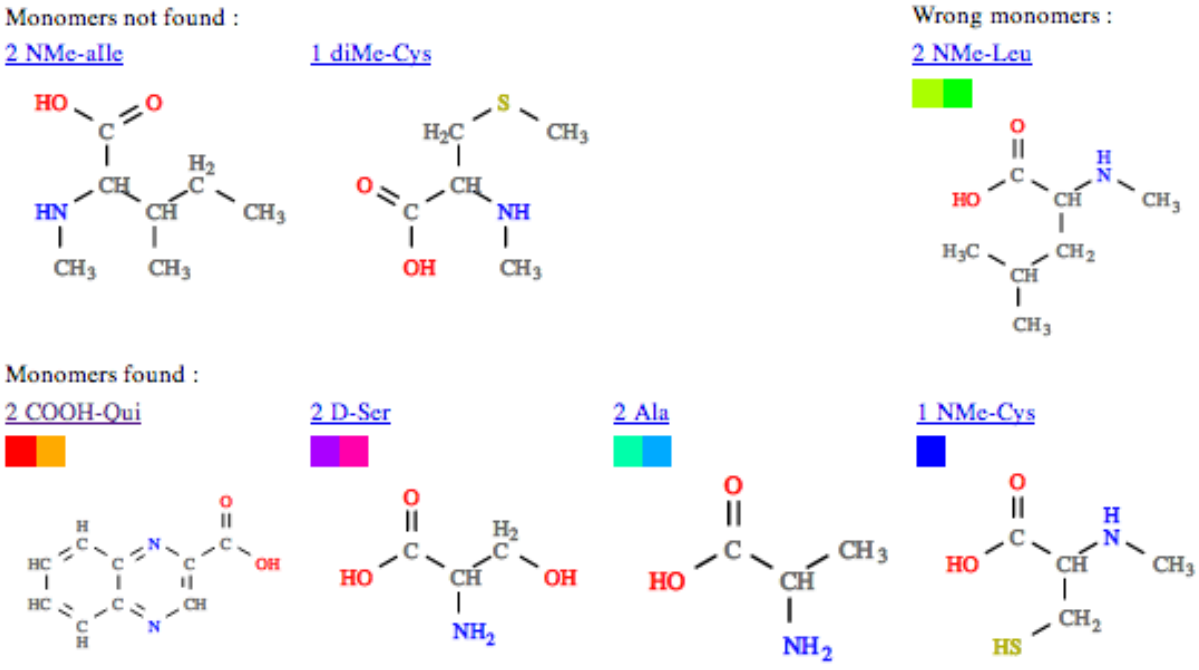


Figure 8. Listes de monomères

3.3.2 Bilan chiffré. Pour évaluer notre heuristique, regardons les couvertures atomiques et monomériques moyennes pour les 204 peptides extraits de la base de données Norine :

- Couverture atomique globale : 91.327%
- Couverture monomérique globale : 1552/1797 = 85.156%

Avec plus de 85% de monomères reconnus, nous pouvons dire que l'heuristique gloutonne donne de très bons résultats préliminaires. En effet, 1552 monomères ont été trouvés sur les 1797 possibles et seuls 188 ont été mal placés (voir TABLE 1). Concevoir un outil à partir de cette heuristique permettra ainsi aux biologistes d'avoir un assistant lors de l'annotation des peptides non-ribosomiques.

VP ratio	FP ratio	FN ratio
1552/1797 = 86,4%	188/1797 = 10,5%	245/1797 = 13,6%

Table 1. Résultats

En examinant les peptides qui ont des monomères mal placés, nous avons observé que beaucoup d'erreurs viennent du fait que les monomères peuvent être observés sous différentes formes chimiques (tautomères, formes ionisées ou non, ...). La variation la plus fréquente est la tautomérie [6]. Il s'agit de doubles liaisons qui changent d'emplacement (FIG. 9). Ainsi, lorsque la double liaison n'est pas à la même position dans le SMILES du monomère et dans celui du peptide, nous ne sommes pas en mesure de localiser le monomère correctement. À cause de cette particularité, la recherche exacte de sous-graphes ne permet pas de trouver toutes les formes d'un monomère. Une équipe a défini 21 règles permettant d'énumérer les tautomères d'une molécule donnée [7]. Nous avons pour objectif d'intégrer ces règles dans notre dispositif, ce qui augmentera encore le nombre de radicaux générés à partir d'un monomère. De plus, il faudra prendre en compte le fait que la tautomérie peut se faire entre deux monomères adjacents dans un peptide.

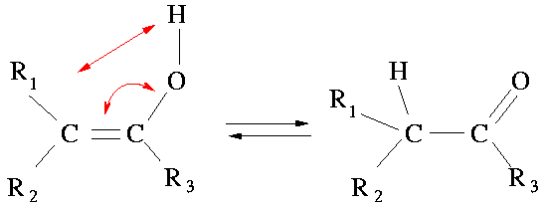


Figure 9. Exemple de tautomérisation

Nous avons également observé que d'autres erreurs proviennent de liaisons qui concernent des carbones à priori quelconques. Par exemple, la diMe-Cys de la quinomycin B (représentée en noir dans la FIG. 7) n'est pas trouvée car elle forme 3 liaisons avec d'autres monomères dont une directement sur un de ses carbones avec perte d'un hydrogène. Nous ne pouvons pas prendre en compte ces liaisons dans notre heuristique actuelle car la combinatoire serait trop élevée.

Enfin le dernier type d'erreurs ne provient pas du programme mais de la base de données Norine. Le test de notre programme nous a permis de mettre en évidence la présence d'erreurs dans certaines structures monomériques contenues dans Norine.

Pour conclure, nous pouvons affirmer que l'heuristique gloutonne donne des résultats de très bonne qualité puisque la plupart des monomères dont nous avons le SMILES sont retrouvés correctement dans les peptides. Il serait cependant intéressant de l'étendre en une heuristique plus souple permettant de réitérer le matching initial dans le but de l'améliorer (*hill-climbing*, algorithmes génétiques, ...).

4 Perspectives

La solution proposée ici repose sur une librairie n'offrant que la recherche exacte de motifs et non une recherche approchée. Le format SMARTS permet d'obtenir plusieurs variantes connues à l'avance d'une même molécule mais pas d'avoir des transformations inattendues. La recherche de SMARTS correspond à de l'isomorphisme de sous-graphe (*Subgraph Isomorphism*) [8]. Or, la catégorie d'algorithmes permettant la résolution

de ce problème n'est pas assez puissante pour résoudre les points bloquants qui sont les variations chimiques des monomères et les monomères inconnus.

Une solution qui permettrait de s'affranchir de la génération des radicaux serait de rechercher la plus grande sous partie commune entre le peptide étudié et le monomère recherché. On se ramène en fait à un autre problème connu dans le domaine de la théorie des graphes qui est le plus grand sous-graphe commun (*Maximum Common Subgraph*) [9]. Cette solution nous permettrait également de prendre en compte les éventuels monomères inconnus car il existe certainement des monomères qui ne sont pas encore dans Norine. Certains sont des dérivés de monomères connus et il serait intéréssant de pouvoir les inférer voir de découvrir dans certains cas par des approches comparatives de nouvelles formes de monomères. La version actuelle du programme ne permet pas de détecter automatiquement de nouveaux monomères. Elle peut cependant constituer une aide visuelle grâce à l'ajout de couleurs sur les images des peptides fournis en sortie du programme.

Enfin, l'ajout d'informations issues des connaissances biologiques concernant les peptides non-ribosomiques et leur mode de synthèse pourrait également permettre d'améliorer la qualité des résultats obtenus. Notamment, nous avons mené une étude statistique sur les données de Norine [2] qui a mis en évidence une différence de distribution entre les monomères présents chez les bactéries et ceux des fungi.

Remerciements

Ce travail a été financé par Inria et le PPF Bioinformatique de l'Université Lille 1.

Références

[1] G. Schoenafinger and M. A Marahiel, Nonribosomal peptides. in *Natural Products in Chemical Biology*, John Wiley & Sons, Inc, 2012.

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

[3] S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques and G. Kucherov. Norine : a database of nonribosomal peptides. *Nucleic Acids Research*, 26(D) :326–331, 2008.

[4] N. M O'Boyle, M. Banck, C. A James, C. Morley, T. Vandermeersch, G. R Hutchison. Open Babel : An open chemical toolbox. *Journal of Cheminformatics*, 3 :33, 2011

[5] Daylight Theory : SMARTS - http ://www.daylight.com/dayhtml/doc/theory/ theory.smarts.html.

[6] M.B. Smith, J. March. *Advanced Organic Chemistry (5th ed. ed.)*, Wiley Interscience, New York, 2001.

[7] M. Sitzmann, W.-D. Ihlenfeldt, M.C. Nicklaus. Tautomerism in large databases. *J Comput Aided Mol Des*, 24, 521–551, 2010.

[8] E. B. Krissinel and K. Henrick. Common subgraph isomorphism detection by backtracking search. *Software - Practice and Experience* 34 :591-607, 2004

[9] J. W. Raymond, P. Willet. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16 : 521–533, 2002