



MASTER 1 <sup>PJI</sup> INFORMATIQUE

UNIVERSITÉ LILLE 1

---

# Prédiction de l'activité des peptides

*CRISTAL - Equipe BONSAI*

---

*Auteur:*  
Emilie ALLART

*Tuteurs:*  
Maude PUPIN  
Laurent NOÉ

*9 mars 2015*

# Contents

Introduction . . . . .	2
<b>1 Contexte et méthodes</b>	<b>3</b>
1.1 Contexte . . . . .	3
1.1.1 NRPs . . . . .	3
1.1.2 Norine . . . . .	3
1.1.3 Problématique . . . . .	3
1.2 Méthodes . . . . .	3
1.2.1 Récupération des données . . . . .	3
1.2.2 Utilisation de méthodes d'apprentissage . . . . .	4
1.2.3 Analyse des mesures de robustesse . . . . .	4
<b>2 Mise en oeuvre</b>	<b>5</b>
2.1 Récupération depuis Norine . . . . .	5
2.2 Traitement des données . . . . .	5
2.3 Choix des critères . . . . .	5
2.4 Lancement des méthodes d'apprentissage . . . . .	6
2.5 mesures de robustesse . . . . .	6
<b>3 Résultats</b>	<b>7</b>
3.1 Résultats obtenus . . . . .	7
3.2 Comparaison avec les résultats précédents . . . . .	7
Conclusion . . . . .	7
Glossaire . . . . .	9
Annexe . . . . .	10

## Introduction

Au cours de la première année de master informatique, il est demandé d'effectuer un projet dans un laboratoire de recherche. J'ai donc effectué le mien au sein de l'équipe Bonsai, une équipe orientée bioinformatique faisant partie de CRISTAL. Encadrée par Maude Pupin et Laurent Noé, j'ai travaillé sur le thème de : ' Rechercher les meilleurs critères pour la prédestination de l'activité d'un peptide'. Il s'agit d'analyser les données d'un ensemble de peptides dont nous connaissons l'activité, les peptides de Norine, et de voir quelles sont les données ou combinaisons de données qui nous permettraient de prédire leur activité.

Le but est de pouvoir par la suite prédire l'activité d'un peptide inconnu, avec pour seule information dessus sa composition en monomères et sa structure. Une étude au sein de l'équipe BONSAI a déjà été faite, il s'agit ici de l'approfondir, et de comparer les résultats pour voir s'ils sont meilleurs et si nous avons affinée la recherche.

Pour ce faire, nous avons analysé les données de Norine (une base de données produites par l'équipe BONSAI).

Dans ce rapport, vous aurez une présentation plus détaillée du contexte et les différentes étapes à effectuer. Suivi de la mise en oeuvre, expliquant comment nous avons procédé. Et enfin les résultats seront traités.

# Chapter 1

## Contexte et méthodes

Pour commencé posons les bases de notre réflexion ainsi que des outils utilisés.

### 1.1 Contexte

Comme dit précédemment le projet s'est déroulé au sein de l'équipe BONSAI, dans la branche portant sur les peptides non-ribosomiques (ou NRPs). Mais avant d'entrer dans la partie technique expliquons ce qu'est un NRP, ainsi que Norine.

#### 1.1.1 NRPs

Tout d'abord un peptide, aussi appelé polymère, est constitués par des monomères . Relié entre eux par différentes liaisons, lui conférants des propriétés différentes. Un monomère ... TODO

**monomères**

**cluster**

#### 1.1.2 Norine

Norine est la première base de données entièrement dédiée aux peptides non ribosomiques. TODO

#### 1.1.3 Problématique

Détermination des meilleurs critères. Pourquoi? TODO

### 1.2 Méthodes

#### 1.2.1 Récupération des données

Selon les différents critères nécessaires, nous relevons les données ..

### 1.2.2 Utilisation de méthodes d'apprentissage

Une fois les données récupérées, nous appliquons différentes méthodes d'apprentissage dessus. Ainsi, nous pourrions convenir si les critères choisis apporte plus ou moins d'informations, et s'il est utile de s'en servir pour prédire l'activité d'un peptide ou non.

#### **Naive Bayes**

Explication du principe de bayes Exemple d'application: avec données, résultat, graph ...

#### **SMO**

idem

#### **LibLinear**

idem

### 1.2.3 Analyse des mesures de robustesse

Permet d'avoir la fiabilité de la prédiction

#### **Precision**

#### **Sensibilité**

#### **F-measure**

#### **AUC ou ROC**

## Chapter 2

# Mise en oeuvre

### 2.1 Récupération depuis Norine

**Les différents critères** Surfactant, une activité, lien , composition

### 2.2 Traitement des données

**Comptage en monomère** Prélèvement préalable des monomères. Traitement de l'attribut composition de chaque peptide Comptage pour chaque peptide de chaque occurrence des monomères Ajout au données

**Comptage en cluster** Prérequis : un fichier répertoriant les clusters de monomères. cad un groupe de monomères auquel on assure qu'ils ont transmette une propriétés Comptage pour chaque peptide du nombre d'occurrence des élément d'un cluster pour chaque cluster. Ajout aux données

**Comptage en lien** Traitement de l'attribut lien de chaque peptides Comptage de lien (de 1 à 5) pour chaque peptides. Ajout aux donnée

**Recupération au dessus d'un seuil** Comptage du nombre de peptide ayant une certaine activité pour chaque activité. Si le nombre de peptide possédant l'activité est inférieur au seuil, nous les enlevons de la base de données.

### 2.3 Choix des critères

Comme vu précédemment, il est possible de faire différents traitements sur les données. Ces traitements permettent de préparer les peptides pour voir si un critère apporte des connaissance sur ceuc-ci ou non. Donc un critère dans le cadre de ce projet, peut être la quantité et l'apparition de monomères, de même pour les clusters, ou encore le possession d'un certain nombre de lien.

Une fois que les données seront traitées par le programme, sur différents critères, nous pourrons comparer leur utilités en analysant la fiabilité de chaque prédiction

## 2.4 Lancement des méthodes d'apprentissage

Une fois les données prêtes, il ne reste plus qu'à les analyser et voir si nous en tirons beaucoup d'information. Pour ce faire, nous avons utilisé 3 méthodes d'apprentissage comme cité précédemment : le naïveBayes, le SMO et le LibLinear.

Pour nous faire nous avons utiliser un wrapper weka permettant de continuer à travailler sur python. Cependant, il a fallu installer une extension pour le LibLinear qui n'est pas implémenté de base. Pour se faire une librairie est ajoutée permettant de l'utiliser également.

ce programme de lancement se décompose en fonction, chacune lançant une méthode, prenant en entrée le fichier contenant les données et donnant en sortie les mesures de robustesse à analyser.

## 2.5 mesures de robustesse

Nous obtenons ainsi un ensemble de mesures de robustesse qu'il faut comparer pour voir quel critère ou quel combinaison de critère est la plus efficace pour obtenir des informations sur un peptide.

## Chapter 3

# Résultats

### 3.1 Résultats obtenus

TODO tableau des Qualités, graphes ...

Conclusion : quelles sont les meilleures combinaisons ...

### 3.2 Comparaison avec les résultats précédents

TODO tableau et/ou graphe comparatif

Conclusion : Avons-nous avancé depuis l'étude précédente? Quelles sont les meilleurs critère ou combinaison de critères pour déterminer l'activité d'un peptide ?

## Conclusion



Pour rappel, notre problématique était de chercher des meilleurs critères pour déterminer l'activité d'un peptide connaissant sa décomposition en monomères et sa structure. En comparant, les résultats de l'étude effectuée sur les NRPs de Norine quelques années auparavant et les résultats que nous avons obtenus en combinant différentes connaissances sur ces NRPs, nous avons trouvé des combinaisons de critères nous permettant d'apprendre avec plus de certitudes l'activité d'un NRP.

Pour connaître l'activité d'un NRP avec une certitude de ... minimum, il faut combiner ....

Nous avons donc pu aller au bout de ce projet, en trouvant une combinaison meilleur, ne reste plus qu'à analyser de nouveau NRP pour, qui sait, trouver un nouvel antibiotique ...

## Glossaire

## Références