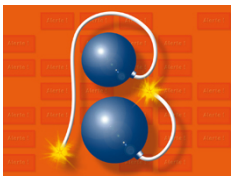

PJI
Master 1 informatique

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

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 donc d'analyser des peptides (petites protéines) dont nous connaissons déjà l'activité (antibiotique, anticancéreux, toxine, ...) et de voir quelles informations sur elle auraient pu nous faire prédire son activité. Par exemple, ... TODO (chercher un exmeple pour illustrer)

Le but est de pouvoir par la suite, connaissant ces critères, prédire l'activité d'un peptide. Ainsi, nous pourrons trouver de nouvelles molécules avec des capacités thérapeutiques intéressantes, qui pourrait être synthétisées plus facilement. Trouver de nouveaux médicaments efficaces donc.

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

Dans ce rapport, vous verrez donc le cahier des charges présantant le contexte et les différentes étapes à effectuer. Suivi de la mise en oeuvre, expliquant comment nous avons procédé. Et enfin les résulats seront traités.

Chapter 1

Cahier des charges et contexte

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 conférents des propriétés différentes. Un monomère ... TODO

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 Cahier des charges

1.2.1 récupération des données

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

1.2.2 Utilisation de weka

Une fois les données récupérées, nous appliquons différentes méthodes d'apprentissage dessus. Ainsi, nous pourrons convenir si les critères choisis apporte plus ou moins d'informations, et s'il est util 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 occurence 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'occurence 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 utilitées 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 fqire, nous avons utilisé 3 méthodes d'apprentissage comme cité précédement : 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 lancant 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

TODO : tableau comparatif conclusion tirée de celui-ci

Conclusion

Glossaire

Références