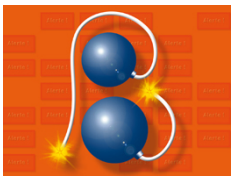

PJI
Master 1 informatique

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 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 résulats seront traités.

Chapter 1

Contexte et méthodes

Pour commencer, posons les bases de notre réflexion ainsi que des outils utilisés. Nous nous intéressons plus précisément, aux peptides non-ribosomiques (ou NRPs).

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).

1.1.1 NRPs

Peptides ribosomiques Pour se rafraichir la mémoire, un peptide est une molécule composée d'acides aminés reliés entre eux par des liaisons peptidiques. Dans la plupart des cas, ils résultent de la traduction d'un ARNm au sein du ribosome, produisant ainsi le matériel nécessaire à tout organisme pour subvenir à ses besoins, tel que les hormones, les enzymes, les neurotransmetteurs, ... Ceux-ci sont les peptides ribosomiques, mais ceux qui nous interessent pour l'étude sont les peptides non-ribosomiques.

Peptides non-ribosomiques Les peptides non-ribosomiques sont des molécules produites par des micro-organismes (bactéries et champignons) ayant un large domaine biologiques et pharmacologiques d'application (antitumeur, antibiotique, ...) . Ils sont composés d'une grande variété de monomères tels que des acides aminés, mais aussi des lipides ou des sucres. Des liaisons non-peptidiques peuvent être formées entre certains monomères, ce qui produit des peptides contenant des cycles et/ou des branchements. Dans le projet, nous étudierons entre autre la possibilité d'un lien entre la structure d'un peptide et son activité. Se sont des peptides indépendant du ribosome, qui sont produit par une complexe enzymatique les NRPS (non-ribosomal peptides synthetase). Ils ont la caractéristique d'être de petite taille et d'être fort diversifiés.

Clusters de monomères Ainsi, certains monomères auxquels on attribue certaines propriétés sont regroupés en cluster (ou famille). Voir le 'Rangement

en cluster de certains monomères utilisés pour l'étude' dans le glossaire. Dans notre étude, nous nous pencherons sur la questions de savoir si l'appartenance majoritaire ou partiel d'un peptide à un cluster a une influence ou non sur son activité.

1.1.2 Norine

Norine (NOnRibosomal peptides with INE) est la première base de données entièrement dédiée aux peptides non ribosomiques. Créée par l'équipe BONSAI, elle repertorie plus de 1700 NRPs et pas moins de 528 monomères. Accessible via http://bioinfo.lifl.fr/norine/ . Elle permet d'avoir la description complète d'un peptide, son nom, ses synonymes, sa composition monomérique et chimique, etc ... ; de même pour les monomères.

1.1.3 Problématique

L'étude précédente a permis de montrer la forte relation entre la décomposition monomérique d'un NRP et son activité biologique. Comme évoqué un peu précédement, nous allons analyser les liens qui peuvent exister entre la décomposition d'un peptide en monomère, en cluster de monomères et sa structure, avec son activité. Pour se faire, nous allons pousser plus loin la précédente étude, et essayer de trouver les meilleurs critères pour la prédestination de l'activité d'un peptide. Cela permettra, par la suite, de déterminer l'activité d'un peptide en analysant sa décomposition et structure.

1.2 Méthodes

TODO : schéma résumant les étapes effectuées

1.2.1 Récupération des données

Récupération des données Etant en possession de toutes les données sur les peptides de Norine, nous allons baser notre étude dessus. La première partie consiste donc à relever les informations sur les peptides que nous souhaitons traiter. Pour se faire nous prélevons uniquement la décomposition en monomères et sa structure pour voir les liens qu'il a. Pour des questions pratique, nous ne considérons que les peptides ayant une unique activité de façon à avoir des résultats cohérents, hormis dans le cas d'un peptide ayant comme capacité 'surfactant' en plus d'une autre, nous le gardons en ne conservant que la deuxième propriété.

Préparation des données Il sera nécessaire de ne garder que les peptides ayant une activité dépassant un certain seuil pour ne pas bruiter les données. Puis selon l'étude que l'on souhaite faire, il faudra faire le comptage de chaque monomère de la base de données pour chaque peptide. De même que le comptage de chaque cluster de monomère. Ou encore, le comptage de lien simple, double, aisni de uite jusque 5 .

1.2.2 Utilisation de méthodes d'apprentissage

Une fois les données récupérées,et préparées, 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 utile de s'en servir pour prédire l'activité d'un peptide ou non. Ainsi, via l'utilisation de Weka, nous appliquerons NaiveBayes, SMO et LibLinear. TODO : Revoir les descriptions

Naive Bayes Cette méthode repose sur le théorème de Bayes. Avec une phase d'entrainement et une phase de test. On fait de l'apprentissage probabiliste sur l'ensemble des peptides, ainsi lorsque l'on test à quelle activité appartient un nouveau peptide, on prend celle qui a la plus grande probabilité.

SMO Méthode d'apprentissage supervisé qui construit au fur et à mesure de la lecture des données une fonction objective via une descente de gradient.

LibLinear LibLinear reste dans la même idée.

1.2.3 Analyse des mesures de robustesse

Une fois le traitement des données fait, on analyse les résultats avec la précision, la sensibilité, la F-measure et le ROC.

Precision La précision est la probabilité que les données soient bien triées, une mesure de la qualité des méthodes d'apprentissage sur les données. Donc un taux élevé est attendu dans la recherche.

Sensibilité A l'inverse, la sensibilité est la probabilité que les données soient mal classées. Il vaut donc mieux qu'elle soit faible.

F-measure La F-measure allie la précision et la sensibilité, ainsi si elle approche 1 c'est un bon résulat et si elle s'approche de 0 c'est l'inverse.

AUC ou ROC La courbe ROC est la courbe représentant les faux-positifs (valeur rangée dans une classe alors qu'elle n'en fait pas partie) en fonction des vrai positifs (valeur bien rangée).

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

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 résulats 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 trouver des combinaisons de critères nous permettant d'apprendre avec plus de certitudes l'activité d'un NRP.

Pour connaitre 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

Rangement en cluster de certain monomères utilisé pour l'étude

Chromophores non siderophore
ChrD;ChrA;ChrAct

Siderophores
ChrI;ChrP;OH-Asp;D-OH-Asp;OH-His;Ac-OH-Orn;D-Fo-OH-Orn;D-OH-Orn;D-Ac-OH-Orn;D-OH-cOrn;OH-Orn;NAc-Fo-OH-Orn;Fo-OH-Orn;OH-cOrn

Peptaibols (antibiotiques)
NAc-Dpr;Ac-Ser;Ac-Ival;Ac-Val;Ac-Trp;Ac-Phe;Ac-Aib;NAc-Leu;Ivalol;Leuol;Valol;Ileol;Pheol;Trpol;Se
Leuol;Aib;Ival;4OH-Pro;Et-Nva

Hpg et derives (antibiotiques)
Hpg;D-OH-dHpg;Cl2-Hpg;NMe-Hpg;Cl-Hpg;OH-dHpg;D-Hpg

Sucres (antibiotiques)
2OMe-Rha;Ara;Aco;D-Gal;D-Ara;Ere;Glc;Oli;D-Glc;bD-Gal;U4oxo-Van;Rha;D-Man;4oxo-Van;Act;Lyx;Ria;Van

Références