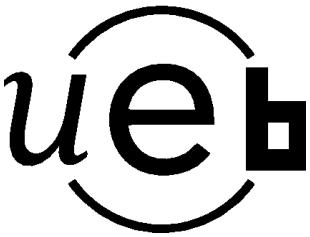
Classification et caractérisation de familles enzymatiques à l'aide de méthodes formelles

Gaëlle Garet

N^o d'ordre : 00000 ANNÉE 2014



THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Gaëlle GARET

préparée à l'unité de recherche Inria/Irisa – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
Composante universitaire : ISTIC

Classification et caractérisation de familles enzymatiques à l'aide de méthodes formelles

Thèse à soutenir à Rennes
le 16 décembre 2014

devant le jury composé de :

Jean-Christophe JANODET
Professeur à l'Université d'Evry-Val-d'Essonne / *Rapporteur*
Amedeo NAPOLI
Directeur de recherche au Loria, Nancy / *Rapporteur*
Colin DE LA HIGUERA
Professeur à l'Université de Nantes / *Examinateur*
Olivier RIDOUX
Professeur à l'Université de Rennes 1 / *Examinateur*
Mirjam CZJZEK
Directrice de recherche CNRS, Roscoff / *Examinatrice*
Jacques NICOLAS
Directeur de recherche à Inria, Rennes / *Directeur de thèse*
François COSTE
Chargé de recherche à Inria, Rennes / *Co-directeur de thèse*

Ainsi en était-il depuis toujours. Plus les hommes accumulaient des connaissances, plus ils prenaient la mesure de leur ignorance.
Dan Brown, Le Symbole perdu

Tu me dis, j'oublie. Tu m'enseignes, je me souviens. Tu m'impliques, j'apprends.
Benjamin Franklin

Remerciements

Je remercie tout d'abord la région Bretagne et Inria qui ont permis de financer ce projet de thèse.

Merci à Jean-Christophe Janodet et Amedeo Napoli qui ont accepté de rapporter cette thèse et à Olivier Ridoux, Colin De La Higuera et Mirjam Czjzek pour leur participation au jury.

J'aimerai aussi dire un grand merci à mes deux directeurs de thèse : Jacques Nicolas et François Coste, qui m'ont toujours apporté leur soutien tant dans le domaine scientifique que personnel. Jacques, merci pour ta patience, ta gentillesse et ton soutien à toute épreuve. Un grand merci surtout pour tout ce que tu m'a apporté au niveau scientifique, entre autre pour formaliser les problèmes rencontrés, et pour ta pile de bibliographie amassée pendant des années :) François, merci de m'avoir permis de faire mon stage de master qui m'a conduite ici, merci pour les discussions, constructives ou non, qui m'ont aidé à mieux comprendre les grammaires et les concepts qu'elles impliquent. Merci aussi de m'avoir donné l'occasion de découvrir le monde durant ces trois ans aux travers des voyages, je me souviendrais longtemps des voyages, Washington à l'occasion d'ICGI, Cordoba, les visites, le parc des condors ;)

Merci à tous ceux qui ont collaborés de près ou de loin à mes travaux, je pense notamment aux biologistes de la station de Roscoff qui m'ont toujours accueillis avec beaucoup de patience (et il en faut pour expliquer les concepts biologiques à une informaticienne !). En vrac, Myrjam, Gurvan, Thierry, Agnès, Catherine et tout ceux que j'oublie sur le moment.
Merci aussi à ceux qui m'ont accueilli dans leur labo en Argentine ou en Allemagne lors de mes différents séjours.

Un grand merci au groupe Symbiose (Dyliss, Genscale et Genouest), surtout ne changez rien ! ! Malgré les multiples départs et les arrivées, l'ambiance est toujours restée la même : excellente, ces 3 ans n'auraient pas été les mêmes sans chacun d'entre vous. Je me rappellerai des séminaires au vert, des midi activités, notamment les tournois de ping pong, l'escalade, le badminton (si si des fois j'y allais), les jeudi soirs amarillys, les afters de folie ou les samedi matins au marché, le voyage en Roumanie et j'en passe ... Tellement de bons souvenirs !
Merci à mathilde de m'avoir supportée dans le bureau pendant ces 3 ans (enfin plutot 2 ;)), les tours de smarties, les aprems entre filles, ...
Merci à ceux qui m'ont permi de squatter leur bureau sur la fin (Julie, Charles et Ayme-

rick) et à Claudia qui m'a supportée pendant la fin de la rédaction. Merci à Catherine, Coline ou encore Anaïs pour les pauses potins quand la rédaction devenait difficile. Un merci particulier à ceux qui ont toujours été là à n'importe quelle heure du jour ou de la nuit, mathilde, vincent, sylvain : le stage de M2, les discussions mc do, les conférences, les séjours à Roscoff, ...

Merci à tout ceux qui m'ont fait découvrir la vulgarisation scientifique et qui nous ont aidé dans l'organisation de Sciences en Cour[t]s, une aventure à la fois scientifique et humaine qui a permis de penser à autre chose de temps en temps. Notamment les orgas, parmi eux Sylvain, Coraline, Charles, Marie, Kévin, Nico, ... mais aussi tous les réals ! Et je souhaite bon courage aux nouveaux (entre autre Clovis, Fanny et Paulin) qui ont accepté de reprendre le flambeau pour continuer l'aventure !

Merci à mes amis de Quimper qui m'ont écouté pendant un certain nombre d'heure parler de cette thèse sans rien comprendre 'Je pense notamment à Camille, Elise, Céline, JR, Maureen, Kane, ... Merci à vous d'être là depuis nos années déjantées du lycée.

Et merci à toute ma famille qui m'a soutenue et encouragée dans cette aventure qu'est la thèse (même quand je n'étais pas toujours facile à vivre !), maman, Nico, Pierre, Audrey, Caro, ...

Enfin, merci à tout ceux que je n'ai pas cité mais qui se reconnaîtront.

Table des matières

Introduction

Prédire l'activité d'une enzyme à partir de sa séquence en acides aminés est une tâche d'une importance majeure pour la compréhension des réactions biochimiques avec de nombreuses retombées dans le domaine des biotechnologies. Cette thèse considère le problème de l'apprentissage de signatures caractéristiques de familles d'enzymes. Les enzymes sont des protéines particulières et on dispose déjà d'un certain nombre d'outils pour caractériser des familles de protéines et découvrir de nouveaux membres à ces familles. Le problème général est celui de l'inférence de modèles à partir d'un ensemble de séquences partageant une fonction commune. Parmi les modèles les plus utilisés, on trouve les modèles de Markov cachés ou encore les langages réguliers. Dans cette thèse nous nous intéressons à des modèles algébriques plus expressifs qui sont capables de capturer des interactions entre éléments éloignés sur la séquence de la protéine.

Dans ce but, nous nous sommes intéressés dans un premier temps à l'inférence de grammaires hors-contexte. Il s'agit d'un problème pour lequel il existe un certain nombre d'algorithmes heuristiques pour le traitement automatique des langues. Tous sont basés sur le principe de substituabilité de Harris pour lequel le travail de A. Clark [CE07] offre un cadre théorique et des résultats d'apprenabilité en introduisant la classe des langages hors-contexte substituables. Nous avons étendu ce principe en introduisant de nouvelles classes de langages et de nouveaux critères de généralisation basés sur des classes de substituabilité locales et/ou contextuelles. Ces nouveaux critères permettent ainsi de traiter des ensembles de séquences de protéines pour lesquels les exemples sont moins nombreux et les séquences beaucoup plus longues que pour les langues naturelles.

Afin de pouvoir utiliser ce modèle en pratique sur des ensembles de données réels, nous proposons un algorithme d'apprentissage efficace permettant de réduire la grammaire tout au long de l'apprentissage jusqu'à l'obtention d'une grammaire canonique non redondante du langage substituable cible. Nous avons ainsi obtenu de bons résultats sur des données réelles avec une haute spécificité et une bonne sensibilité grâce à une généralisation basée sur la substituabilité locale.

En pratique, la disponibilité des familles d'enzymes n'est pas immédiate. Il existe des groupes de séquences phylogénétiquement liées, appelés superfamilles, qui possèdent des motifs communs pour les repérer. Au sein de ces superfamilles, il existe un certain nombre de familles avec des activités diverses déterminées expérimentalement. Le coût et la difficulté des expérimentations ne permettent pas actuellement d'avoir des séquences d'apprentissage pour chaque famille existante.

Nous avons donc abordé le problème de la prédiction de l'appartenance d'une séquence à une famille. Nous présentons un classifieur basé sur l'identification de blocs de sous-séquences communes entre des séquences de familles connues et inconnues appartenant à la même superfamille. Nous nous appuyons pour cela sur la recherche de concepts formels construits sur le produit des blocs et des séquences. Contrairement à la plupart des classifieurs dans ce domaine, nous avons introduit les classes (familles) comme des objets à part entière. Nous traitons le problème non supervisé de la détection de nouvelles familles dans les séquences non étiquetées comme un problème d'optimisation en minimisant le

nombre de nouvelles familles tout en maximisant le support d'une nouvelle famille en terme de blocs caractéristiques. Nous avons utilisé ce classifieur sur des données provenant du génome d'une algue brune récemment séquencée, *Ectocarpus siliculosus*, relativement éloignée des espèces habituellement étudiées.

La dernière étape de l'étude est la création de grammaires hors-contexte pour chaque famille repérée au sein d'une superfamille. Ces caractérisations expressives permettent potentiellement de repérer des interactions intéressantes au sein des protéines.

Dans un premier temps, le chapitre 1 est consacré à l'introduction des concepts biologiques nécessaires à la compréhension de cette thèse. Il présentera les notions d'enzymes et de familles de séquences, ainsi que certaines méthodes utilisées pour caractériser un ensemble de séquences protéiques. Nous verrons ainsi les limites des approches utilisées actuellement.

puis, le chapitre 2 introduit les concepts mathématiques et informatiques utilisés pour modéliser les familles d'enzymes. Nous présentons dans ce but la théorie des langages et l'inférence grammaticale, et insistons sur l'état de l'art utilisant des langages réguliers.

Dans le chapitre 3, nous cherchons à obtenir des grammaires de la classe des grammaires hors-contexte, plus expressives que les grammaires régulières classiquement utilisées. Dans ce but, nous introduisons les concepts de substituabilités locale et contextuelle comme critères sur lesquels s'appuiera la généralisation des séquences d'apprentissage.

Le chapitre 4 est consacré à l'apprentissage supervisé et non supervisé des propriétés des différentes sous-familles présentes dans l'échantillon d'apprentissage grâce à l'analyse de concepts formels. Il sera ainsi possible de discriminer des ensembles de séquences présentant des activités inconnues. Cela permettra d'introduire une phase de sélection dans la méthode présentée au chapitre 2.

Enfin, le chapitre 5 présentera les perspectives et conclusions de ce travail.

Résumé

Cette thèse propose une nouvelle approche de découverte de signatures de familles (et superfamilles) d'enzymes. Dans un premier temps, étant donné un échantillon aligné de séquences appartenant à une même famille, cette approche infère des grammaires algébriques caractérisant cette famille. Pour ce faire, de nouveaux principes de généralisation et de nouvelles classes de langages ont été introduites sur la base de la substituabilité locale. Un algorithme a également été développé à cet effet qui produit une grammaire réduite, conservant la structuration des exemples, d'un langage substituable. Dans un second temps, ce manuscrit présente une méthode de classification des séquences d'une superfamille en familles à l'aide d'une analyse de concepts formels basée sur l'alignement des séquences qui permet la détection de nouvelles familles et la découverte des motifs fonctionnels pour améliorer les signatures précédentes.

Mots clés

bioinformatique, enzyme, famille, inférence grammaticale, grammaire algébrique, substituabilité, analyse de concepts formels

Abstract

This thesis proposes a new approach to discover signatures of families (and superfamilies) enzymes. At first, given a sample of aligned sequences belonging to the same family, this approach infers context-free grammars characteristic of this family. To do this, new principles of generalization and new classes have been introduced based on substitutability. An algorithm has also been developed for this purpose, which produces a reduced grammar able to retain the structure of examples. In a second step, this manuscript presents a method for classification of a superfamily sequences into families with a formal concept analysis based on alignement sequences allowing detection of new families and the discovery of patterns to improve functional previous signatures.

Keywords

bioinformatics, enzyme, family, grammatical inference, context-free grammar, substitutability, formal concept analysis