Université de Provence Aix-Marseille I

Ecole Doctorale des Sciences de la Vie et de la Santé

Thèse de doctorat

Présentée en vue d'obtenir le grade de

DOCTEUR ES SCIENCES DE L'UNIVERSITE DE PROVENCE

Discipline : Bioinformatique

Par

Vincent Lombard

Structuration et exploration d'information génomique et fonctionnelle des enzymes actives sur les glucides

Soutenue devant le jury composé de

Dr Gilles Labesse	**Rapporteur**
Dr Gabrielle Potocki-Veronese	**Rapporteur**
Dr. Garry Paul Gippert	**Examinateur**
Dr Etienne Danchin	**Examinateur**
Dr Bernard Henrissat	**Co directeur de thèse**
Pr Pedro Maldonado Coutinho	**Co directeur de thèse**

Remerciements

Je tiens tout d'abord à remercier mes rapporteurs, Gilles Labesse et Gabrielle Potocki-Veronese ainsi que mes examinateurs Garry Paul Gippert et Etienne Danchin d'avoir accepté d'évaluer ce travail.

J'adresse toute ma gratitude à mes co-directeurs Bernard Henrissat et Pedro Maldonado Coutinho qui m'ont épaulé et fait confiance à tout moment.

Merci Bernard pour ton aide professionnelle et personnelle. Je suis toujours disponible pour garder ta maison.

Merci Pedro de m'avoir supporté à tes cotés (surtout mes sauts d'humeur).

Je vous remercie tous deux pour tant de simplicité et d'intelligence réunies.

Merci aux personnes du groupe de glycogenomique qu'ils soient permanents ou non-permanents.

Je remercie le directeur de l'AFMB Yves Bourne.

I would like to acknowledge the people from Novozyme who have funded my phD: I really enjoyed visiting you in Denmark and working with you.

L'équipe informatique du laboratoire (Eric, Denis) pour leur aide précieuse. Merci à Polo et Phillipe pour leurs nombreux services. Merci à mes nombreux partenaires sportifs (les 2 Phillipe L, Antoine, Bruno, Eric). Merci Karine et Loic pour nos formidables conversations.

Et plus généralement, je remercie très chaleureusement tous les membres du laboratoire AFMB avec qui j'ai passé trois belles années.

Mes remerciements se dirigent également à tout ceux qui ont suivi mon parcours et m'ont soutenue depuis toujours, toute ma famille en particulier mes parents, mes sœurs, ma nièce, mes amis d'adolescence (Fred, Lionel, Nico, Denis, Tom).

Ma dernière pensée s'adresse à Flo et au petit bout de chou qui écoutera peut être cette thèse d'une oreille discrète confortablement installé dans le ventre de sa maman à qui je fais de tendres baisés.

Table des Matières

Liste des Figures et des Tableaux

Liste des abréviations

ADN	Acide désoxyribonucléique
ATV	A tree viewer
BLAST	Basic local alignment search tool
CAZy	Carbohydrate-active enzymes
CBD	Carbohydrate-binding domain
CBM	Carbohydrate-binding module
CE	Carbohydrates estérase
CPU	Central processing unit
EC	Enzyme commission
EMP	Metabolic pathways database
FN3	Fibronectine de type III
GB	GenBank
GH	Glycosides hydrolases
GI	GenBank identifier
GP	Genpept
GPI	Glycosylphosphatidylinositol
GT	Glycosyltransférase
HCA	Hydrophobic cluster analysis
HMM	Hidden Markov model
HTML	Hypertext mark-up language
IUBMB	International Union of Biochemistry and Molecular Biology
NCBI	National Center for Biotechnology Information
PDB	Protein database
Pfam	Protein family database
PL	Polysaccharides lyase

PMD	Protein mutant database
PSSM	Position-specific scoring matrices
RG	Rhamnogalacturonanne
RG-I	Rhamnogalacturonanne de type I
RG-II	Rhamnogalacturonanne de type II
SLH	S-layer homology
SP	Swissprot
UNK	Unknown
XynU	Xylanase U

I. Introduction

I.1. Les sucres et enzymes agissant sur les sucres

I.1.1. Les glucides

Les glucides sont des constituants primordiaux de la vie. Ils jouent un rôle essentiel dans de nombreux processus biologiques à la fois au niveau structural et fonctionnel. Sous forme de saccharides et de glycoconjugués, ils constituent une partie substantielle de la biomasse produite sur terre et représentent une source potentielle d'énergie renouvelable de première importance.

I.1.1.1. Les monosaccharides

D'un point de vue chimique, les glucides les plus simples sont les monosaccharides dont la formule générale peut-être décrite comme $(CH2O)_n$, n pouvant varier de 3 à 10. Les différents atomes de carbone sont liés par des liaisons simples formant des chaines linéaires courtes. Dans un monosaccharide, un des atomes de carbone fait partie d'un groupe carbonyle. Si ce carbonyle est en position terminale, on a alors un aldéhyde, et le monosaccharide correspondant est dénommé aldose (**Figure 1a**). Si le carbonyle est en position interne de la chaine, on a une cétone et le monosaccharide qui en découle est nommé cétose (**Figure 1b**). Les atomes de carbone restants sont liés à des groupes hydroxyle (ou alcool). Chaque glucide peut-être décrit sous deux formes D ou L totalement symétriques selon l'orientation du carbone asymétrique de l'extrémité de la chaine opposée à celle contenant le groupe carbonyle (**Figure 1**).

En solution, les monosaccharides contenant au moins 5 atomes de carbone peuvent subir spontanément une cyclisation par condensation intramoléculaire entre une des fonctions alcool et le groupe carbonyle. On obtient ainsi des formes hémiacétal et hémicetal à partir d'un aldose ou d'un cétose, respectivement. Le carbone portant originalement le groupe carbonyle devient ainsi un nouveau centre chiral dans la molécule qui adopte une des deux configurations anomériques

différentes désignées comme les anomères α et β. Les différentes formes linéaires et cycliques sont en équilibre thermodynamique en solution.

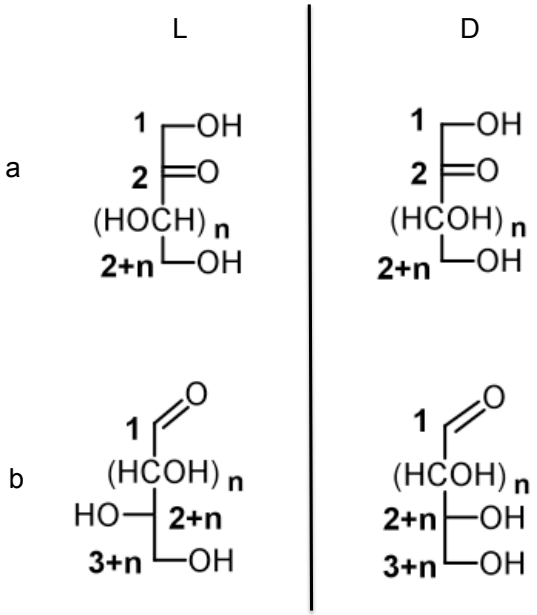


Figure 1:Structure linéaire simplifiée des monosaccharides. Représentation de : (a) L- et D-cétoses ; (b) L- et D-aldoses. La barre verticale représente un plan de symétrie.

I.1.1.2. Les glucides complexes

Les glucides complexes, ou glycanes, sont formés par la condensation d'unités monosaccharidiques unies de façon covalente par une liaison O-glycosidique, qui est formée lorsque l'hydroxyle d'une sucre réagit avec le carbone anomérique d'un autre sucre en libérant une molécule d'eau. Les liaisons glycosidiques résultantes sont chimiquement stables [1]. La complexité des structures glucidiques augmente considérablement avec le nombre et la nature des divers résidus de monosaccharides présents. Cette complexité liée à la grande diversité engendrée par les monosaccharides est due à :

- des longueurs de chaines carbonées linéaires variées
- de nombreux centres de chiralité possibles dans chaque chaine
- de nombreuses structures cycliques possibles (configuration α ou β, nombre d'atomes dans les cycles qui peut varier entre 5 et 6)

- de nombreuses modifications chimiques possibles sur les différents groupes hydroxyle

De plus, une grande diversité de structures est possible car les différents groupes hydroxyle présents dans un monosaccharide sont susceptibles de participer à de nombreux types de liaisons glycosidiques. Les structures contenant au maximum une vingtaine de résidus sont typiquement nommées oligosaccharides. Lorsqu'on dépasse la vingtaine de résidus, on parle communément de polysaccharides.

Un exemple simple de polysaccharide est la cellulose, l'un des principaux constituants de la paroi végétale et également produit par certaines bactéries [2, 3]. La cellulose est un polymère linéaire constitué de résidus de D-glucose liés par des liaisons glycosidiques du type β-1,4. Ce polymère comprend entre 2000-6000 résidus chez les bactéries et 13000-14000 résidus chez les plantes [4] (**Figure 2**).

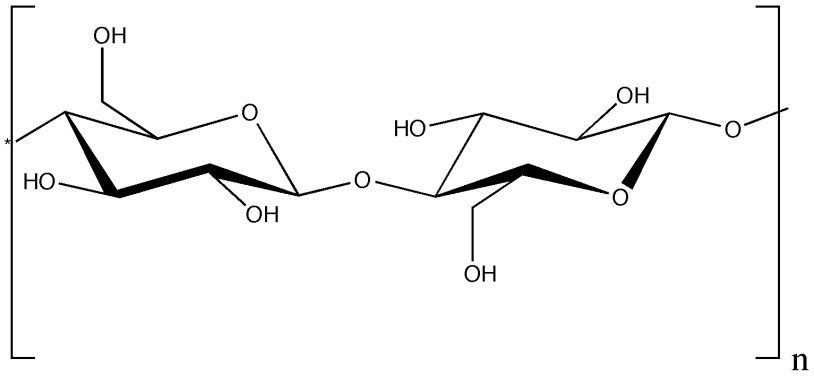


Figure 2: Structure de la cellulose. Spatialement, les chaines de cellulose crystalline présentent comme sous-unité structurale deux résidus de D-glucose dont l'orientation relative par rapport à l'axe formé par la chaine tourne de 180º. Ces sous-unités sont répétées n fois.

Les glucides s'associent souvent de façon covalente à d'autres classes de molécules comme les protéines ou les lipides. Les molécules résultantes sont des glycoconjugués qui peuvent être classés en différents sous-groupes tels que les glycoprotéines, les glycolipides, les protéoglycanes, *etc.* Les glycoconjugués sont des composants essentiels des membranes cellulaires, de la matrice extracellulaire et du milieu aqueux intra- ou extra-cellulaire. La composante glucidique des glycoconjugués, contribue généralement à leur stabilité, à leur solubilité ou à l'adoption d'une conformation particulière de l'aglycone.

On retiendra chez les glycanes que leur structure peut présenter une grande complexité, particulièrement lors de la création de structures ramifiées. Cette complexité leur permet une grande diversité de rôles biologiques [5]. Parmi ces rôles on trouve: le stockage d'énergie (polysaccharides de réserve : amidon, glycogène, fructane) ; la fonction structurale (e.g. peptidoglycane, cellulose, chitine, agarose, pectine etc.) ; ou encore l'adressage et la signalisation moléculaire (e.g. glycosylation des protéines et d'autres glycocojugués) [6].

I.1.2. Structures et mécanismes d'action des enzymes agissant sur les glucides

Un grand nombre de protéines agit sur les glucides soit comme support de reconnaissance et d'interactions, soit en changeant leur composition ou leurs propriétés. Parmi ces dernières, on trouve des enzymes communément nommées enzymes actives sur les glucides (en anglais *carbohydrate-active enzymes* ou CAZymes). L'action des CAZymes permet de favoriser le flux de carbone et d'énergie dans la cellule et dans l'environnement. Leur rôle est de synthétiser, de modifier ou de déconstruire les glycanes. La grande diversité de structures glucidiques entraine une grande diversité de reconnaissance et d'actions catalytiques et par conséquent une grande diversité enzymatique.

I.1.2.1. Les unités fonctionnelles des protéines : les modules

Pour pouvoir agir sur les glucides *in vivo*, les gènes des enzymes codent pour divers éléments structuraux ou fonctionnels de la séquence d'acides aminés résultante. La composante essentielle présente dans chaque enzyme est celle d'une unité catalytique capable de faciliter, du point de vue moléculaire, des réactions sur des substrats spécifiques en vue de l'obtention d'un ou plusieurs produits. Physiquement, l'enzyme doit non seulement reconnaître son substrat mais aussi, par juxtaposition de différents résidus d'acides aminés et éventuellement de certains cofacteurs, de créer un environnement qui facilite une réaction catalytique spécifique. À cette unité fonctionnelle correspond ainsi un segment polypeptidique nommé également « module ». Dans un même polypeptide, d'autres composantes peuvent être présentes pour permettre :

- l'adressage intra- ou extracellulaire ;
- la formation de complexes multienzymatiques via des interactions protéine-protéine ;
- la fixation sur le substrat ou sur un composant associé ;
- la réalisation d'activités catalytiques complémentaires.

L'ensemble des tâches réalisées par l'enzyme peut être associé à un ou plusieurs segments particuliers du polypeptide [7-9]. Les différents segments d'un polypeptide ont souvent des rôles associés permettant de les classer en différents groupes: (i) les signaux peptidiques liés à la sécrétion ou associés à des modifications ou des interactions particulières ; (ii) les segments non- ou peu-structurés de liaison (e.g. *linkers*) ou de fixation membranaire ou transmembranaire (e.g. SLH de l'anglais *S-layer homology*, régions transmembranaires) ; (iii) les modules avec un repliement autonome associés aux fonctions enzymatiques, de liaison au substrat ou d'interaction protéine-protéine, etc. Ces différents groupes de modules peuvent être classés en familles structurales au sein desquelles ils partageront des propriétés similaires aux niveaux moléculaire et fonctionnel.

Les enzymes actives sur les glucides, en particulier les formes extracellulaires, présentent fréquemment une structure plurimodulaire. La plurimodularité au sein d'une même protéine peut se traduire par la répétition d'un même module ou par l'association de plusieurs modules exhibant des fonctions complémentaires de différente nature. De nombreux exemples ont été étudiés et publiés dans la littérature. Chez *Clostridium thermocellum*, la proteine xylanase U (XynU) présente plusieurs fonctions complémentaires [10]. Elle comporte simultanément des modules avec des activités xylanase, acetylxylan estérase, de fixation au xylane et un domaine dockerine (un composant participant aux complexes multienzymatiques appelés cellulosomes via des interactions protein-proteine avec des cohésines [11]). On trouve souvent ce type d'arrangement combinant ces mêmes modules catalytiques et d'adhésion au substrat chez certaines bactéries (e.g. *Clostridium cellulovorans* [12], *Cellvibrio mixtus* [13]).

Les modules catalytiques peuvent également présenter un degré variable de promiscuité et ainsi catalyser plusieurs réactions. Ce phénomène est fréquent chez les substrats glucidiques présentant des similarités structurales. Par exemple, la β-mannanase chez *Caldocellum saccharolyticum* présente une activité endo-β-1,4-xylanase mais aussi une activité endo-β-1,4-glucanase dans le même module catalytique [14]. Chez *Trichoderma reesei* ou *Bacillus cereus*, le module catalytique de certaines cellulases présente simultanément des activités cellobiohydrolase et chitosanase [15]. Les exemples sont nombreux et ont tendance à se multiplier avec l'augmentation de la caractérisation expérimentale des enzymes. L'existence de protéines multi-modulaires, et la présence d'un même module dans plusieurs de ces protéines attestent que le module est une unité structurale et que cette unité se répète au sein d'un même organisme [16-18]. L'identification et les propriétés des modules jouent un rôle primordial dans la classification des CAZymes.

I.1.2.2. Une classification à différents niveaux hiérarchiques

I.1.2.2.1. La classification EC

En 1955, l'*International Union of Biochemistry and Molecular Biology* (IUBMB) a formé une commission dédiée à la nomenclature des enzymes: l'*Enzyme Commission* (EC), ayant pour objectif de créer une nomenclature unifiée des activités enzymatiques utilisable par la communauté scientifique [19]. À chaque activité décrite est associé un code de classification à quatre chiffres, chacun représentant un niveau hiérarchique de classification. Ainsi, le premier niveau se décompose en six classes d'activités enzymatiques connues : les oxydoréductases, les transférases, les hydrolases, les lyases, les isomérases et les ligases ou synthétases. Le deuxième niveau correspond à une sous-classe et informe sur le type de composé ou de groupement chimique impliqué dans la réaction. Le troisième niveau correspond au type de réaction. Enfin, le quatrième niveau correspond à l'identifiant d'une activité enzymatique individuelle. Par exemple, pour toutes les glycosidases la nomenclature EC 3.2.1.x est utilisée, dans laquelle x est un numéro valide donné par la classification. Le dernier chiffre est fréquemment substitué par un simple '-' si la fonction n'a pas encore été décrite dans la classification.

Les numéros EC décrivent chaque activité enzymatique en réduisant les ambiguïtés et en limitant la prolifération de noms triviaux. Cependant, avec l'accumulation de séquences et de données biochimiques, l'utilisation de la nomenclature EC présente des limitations dans sa capacité d'annoter et de distinguer les activités enzymatiques. Des exemples illustrant des limitations de la classification EC chez les CAZymes sont présentés ci dessous.

Dans certains cas, le même nombre EC 3.2.1.4 est donné aux endo-β-1,4-glucanases sans tenir compte du mécanisme catalytique (inversion ou rétention). Il serait raisonnable de diviser chacun de ces ensembles en deux groupes mécanistiques homogènes quand la description des nombres EC n'est pas suffisamment précise.

Dans d'autres cas, l'attribution de numéros ECs peut être ambigüe en raison de limitations expérimentales couplées avec des classes non discriminantes. Par exemple, selon la nature et la finesse des testes biochimiques réalisés on peut décrire une protéine comme étant une α-amylase (EC 3.2.1.1) ou une α-amylase maltohexaose-spécifique (EC 3.2.1.98). Les α-amylases produisant spécifiquement du maltohexaose sont en réalité un sous-ensemble des α-amylases, ce qui est fréquemment escamoté.

Enfin, de nombreuses enzymes ayant une caractérisation biochimique se retrouvent sans numéro EC ou avec un numéro EC incomplet (par exemple 3.2.1.-).

I.1.2.2.2. Vers une classification structurale

La diversité de propriétés et le comportement de certaines enzymes n'étant pas toujours définissables par une simple référence EC, une nouvelle approche de classification des enzymes agissant sur les glucides s'imposait. Ces raisons ont amené, dès 1989, Bernard Henrissat à chercher une alternative de classification basée sur la structure. Son étude sur les enzymes actives sur les glucides a débuté avec l'analyse de similarité de séquences, à partir de l'outil « *hydrophobic cluster analysis* » (HCA) [20] sur les séquences protéiques de cellulases [21]. Contrairement aux outils d'alignement classiques, l'analyse par HCA a permis de relier des séquences ayant des divergences importantes ce qui à l'époque s'est avéré indispensable. En effet, il était particulièrement laborieux d'établir des relations entre séquences compte tenu de leur faible quantité disponible et de la divergence parfois importante entre elles. Les séquences ayant une forte similarité et présentant des signatures communes entre elles ont ainsi été classées dans les mêmes familles.

A l'origine, une classification des cellulases en six familles a été créée et les informations associées ont été enregistrées sous un simple format tabulaire [21]. L'enrichissement de cette classification a présenté immédiatement quelques limitations car des séquences homologues aux cellulases pouvaient présenter d'autres activités. En 1991, la classification en familles par similarité de séquence s'élargit à l'ensemble des glycosides hydrolases (GHs) [22]. Les cellulases, originalement classées en 1989, ont ainsi été intégrées au sein de familles n'ayant pas nécessairement que de la cellulose pour substrat. Ce nouvel effort de classification a abouti au classement de 291 séquences connues pour leur capacité à hydrolyser des substrats glucidiques, et scindé en 35 familles de glycoside hydrolases distinctes [22]. Cette liste a évolué au fur et à mesure des années [23, 24] avec l'introduction de nouvelles familles. De nouvelles catégories d'enzyme ont été ajoutées par la suite telles que les glycosyltransférases (GTs) [25, 26], les polysaccharides lyases (PLs) [27], les carbohydrates estérases (CEs) [27] et, en complément, les modules d'adhésion aux glucides (CBMs) [28, 29]. Ces différentes catégories seront expliquées ultérieurement.

I.1.2.2.3. Les niveaux de la classification structurale des CAZymes

Cette nouvelle classification répartit les enzymes et les protéines présentant des relations structurales ou mécanistiques apparentées selon quatre niveaux hiérarchiques [23, 30] (**Figure 3**):

(i) la **« superfamille »** est basée uniquement sur la similitude de repliement.

(ii) le **« clan »**, regroupe des familles possédant un repliement et une machinerie catalytique identique. La machinerie se trouve positionnée dans les mêmes éléments de structure de tous les membres du clan [31, 32]. Il est important de distinguer la notion de clan de celle de superfamille qui ne groupe les protéines que sur un aspect structural [33].

(iii) la **« famille »**, regroupe des protéines ayant une similarité de séquence suffisante au niveau du module commun qui les caractérise. Le repliement d'une protéine étant dépendant de sa séquence [34], les modules d'une même famille possèdent donc tous la même architecture structurale. De plus, la plupart des modules au sein d'une famille partagent aussi des résidus catalytiques qui sont souvent strictement conservés [35]. Le mécanisme d'action enzymatique des glycosidases agissant par inversion ou rétention de la conformation est dicté par le positionnement spatial des résidus catalytiques. Ce mécanisme n'est identique que si les membres d'une même famille partagent le même repliement et les mêmes résidus fonctionnels [36]. De nombreuses familles sont polyspécifiques, ce qui ne permet pas toujours d'extrapoler la spécificité d'une nouvelle enzyme à partir de l'assignement d'une famille.

(iv) le quatrième niveau de hiérarchisation, nommé « **sous-famille »**, a été proposé afin d'essayer de palier au problème des familles polyspécifiques. En effet, les familles peuvent regrouper des enzymes ayant des activités et des spécificités différentes. Si cette diversité est importante, il devient alors difficile de prédire la fonction d'une enzyme à partir de son appartenance à une famille. La subdivision des familles en sous-familles a donc été proposée afin d'essayer d'améliorer ce niveau de prédiction, devenu de plus en plus recherché avec l'apparition des grands projets de génomique. Bien que ce niveau de hiérarchisation n'a pas encore été étendu publiquement à l'ensemble de la classification, quelques efforts ponctuels ont été réalisés dans certaines familles [37]. Plus récemment, une première approche automatique a été utilisée afin d'établir des sous-familles dans certaines familles de GHs [38]. Une approche similaire est développée dans cette thèse pour les familles de PLs [39].

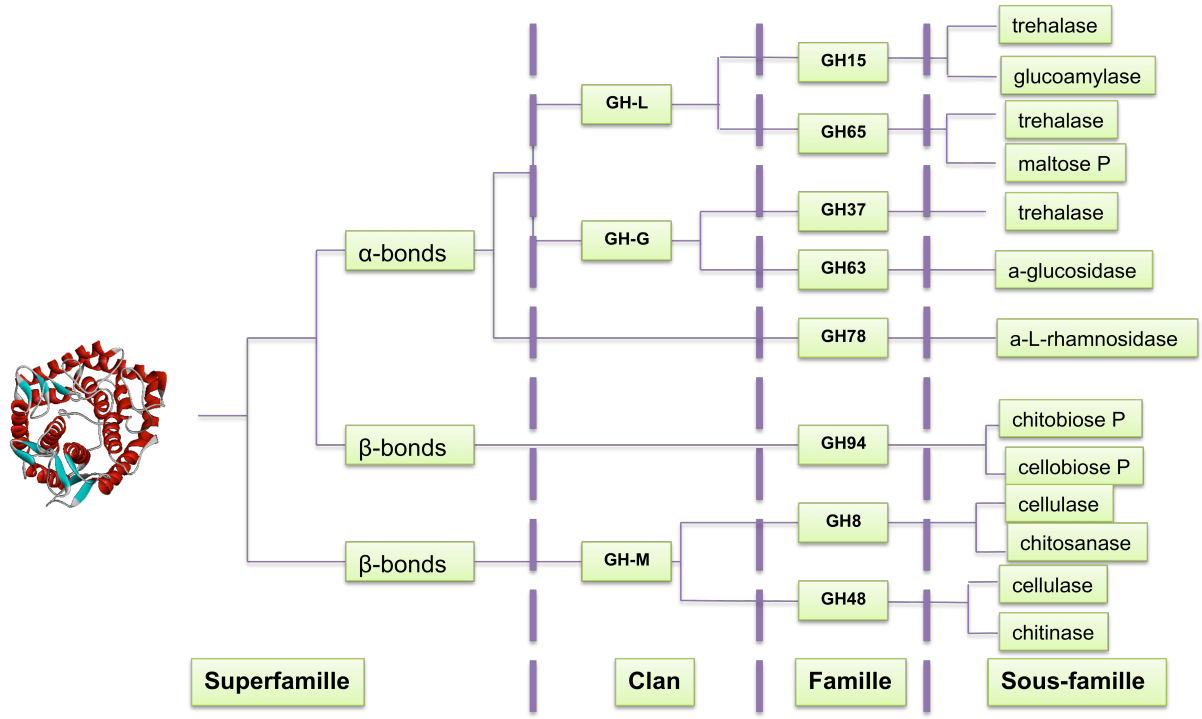


Figure 3: Exemple des différents niveaux hiérarchiques associés à la classification des « *Carbohydrate active enzymes* ». A partir d'un repliement tridimensionnel unique $(\alpha/\alpha)_6$ différents niveaux sont possibles: le clan, les familles et les sous-familles.

I.1.2.2.4. La nomenclature des familles de CAZy

Chaque catégorie est nommée par l'acronyme de la classe d'enzymes à laquelle ses membres appartiennent. Au sein de chaque catégorie, les familles sont rendues uniques par un numéro incrémenté à chaque nouvelle création (par exemple GH1 pour famille 1 des GHs). Ce système simple permet d'éviter une confusion entre le nom de la famille et les enzymes présentes dans cette même famille. Si des familles sont éliminées ou une classe enzymatique est changée, les anciennes désignations ne sont pas recyclées. De façon à respecter une cohérence avec la stratégie utilisée pour la création des familles, les sous-familles sont nommées avec des numéros qui suivent la famille (par exemple GH1_1 pour famille 1 de glycoside hydrolases sous-famille 1).

I.1.2.3. La base de données CAZy

La classification des enzymes participant à la dégradation, à la biosynthèse et à la modification des glucides est regroupée au sein de la base de données experte CAZy (accessible en ligne à l'adresse www.cazy.org) [40]. C'est en 1997 qu'une première version organisée de la base de données CAZy a été créée. Les données étaient classées en format tabulaire par groupe d'activités, regroupant pour chaque protéine des descriptions simples, leurs activités, des accessions de séquence et de structure, et une description modulaire. Ces tableaux ont permis de générer les pages HTML « *HyperText Mark-Up Language* » statiques de la base de données [27].

En Septembre 1998 lors de la mise en ligne de la première version, CAZy comptait environ 3500 protéines différentes et comportait déjà les différentes classes enzymatiques décrites auparavant. La croissance du nombre de séquences déposées rendait particulièrement fastidieux le travail manuel de curation des différents tableaux. Ainsi, avec plus de 5000 séquences protéiques différentes, la gestion de la modularité des CAZymes au format tabulaire n'était plus humainement possible du fait notamment de la redondance de l'information entre tableaux.

En 1999, une base de données relationnelle en MySQL (www.mysql.com) et son interface utilisateur en PHP (www.php.net) adaptée aux données ont vu le jour. Cette base de données a repris toute l'information existante en évitant la redondance d'informations et en facilitant la gestion des données. Cette nouvelle structure a permis à CAZy de croître, d'envisager l'intégration d'autres niveaux d'information et d'automatiser un nombre de taches lié aux données de séquence et de structure issues majoritairement de GenBank [41] de Swissprot [42] et de la PDB [43].

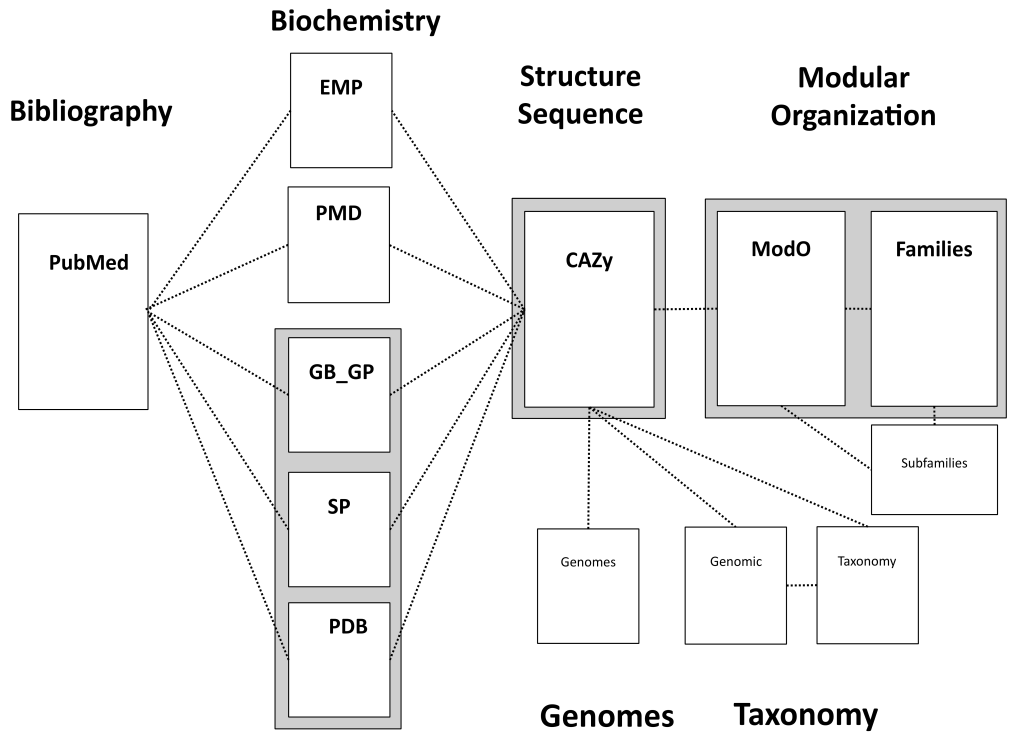


Figure 4: Schéma de la base de données CAZy en 2008. Les rectangles encadrés en gris correspondent aux tables originales issue du schéma relationnel de 1999. Les abréviations de certaines tables sont des noms de base de données : EMP [44] pour « Metabolic Pathways Database », PMD [45] pour « Protein Mutant Database », GB_GP [41] pour « GenBank » et « GenPept », SP [42] pour « SwissProt » et PDB [43] pour « Protein Data Bank ».

A partir de 2002, les champs d'information de la base de données relationnelle ont été élargis, tout en conservant son architecture d'origine, pour comporter des références bibliographiques, des données biochimiques provenant des bases de données EMP [44] et PMD [45], et des données taxonomiques. Cette structure a permis d'accommoder l'information de plus de 120000 séquences protéiques différentes (**Figure 4**). En 2008, suite à des difficultés grandissantes de gestion des données taxonomiques, génomiques et fonctionnelles, une refonte de CAZy s'est imposée. Cette restructuration constitue une partie des travaux décrits dans ce manuscrit.

I.1.2.4. Les différentes catégories d'activités dans la base de données CAZy

Les catégories d'enzymes étudiées et leurs modules associés dans la base de données CAZy sont décrites ci-dessous.

(i) Les glycosides hydrolases (GHs) comprennent majoritairement deux types d'activités mécanistiquement proches, les glycosidases (EC 3.2.1.x) et les transglycosidases (une partie des EC 2.4.1.x). Ces enzymes hydrolysent ou créent les liaisons glycosidiques soit entre glucides soit entre un glucide et un aglycone. Les GHs agissent selon deux modes distincts qui traduisent le changement de configuration du carbone participant dans la formation de la liaison glucidique entre le substrat et le produit [23, 24, 27, 46]. Ces deux modes ont pour résultat :

(a) la rétention de la configuration, où la stéréochimie du carbone anomèrique du produit est le même que celui du substrat (**Figure 5**).

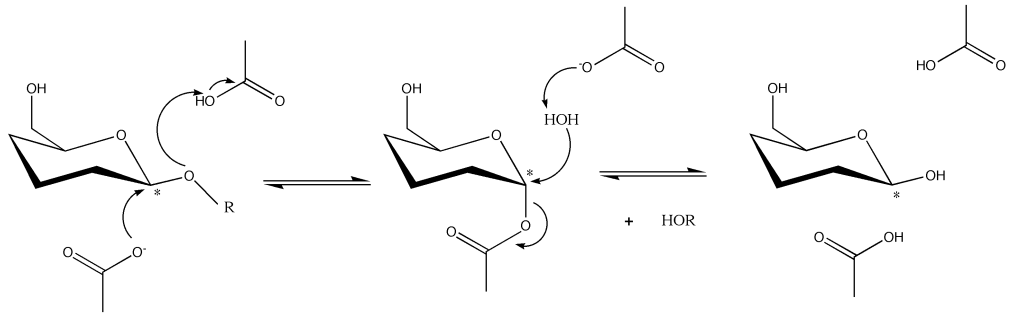


Figure 5: Schéma du mécanisme de rétention de configuration d'une β-glycosidase où (*) représente le carbone asymétrique [47].

(b) l'inversion de la configuration, où la stéréochimie du carbone anomèrique du produit est l'opposé de celui du substrat (**Figure 6**).

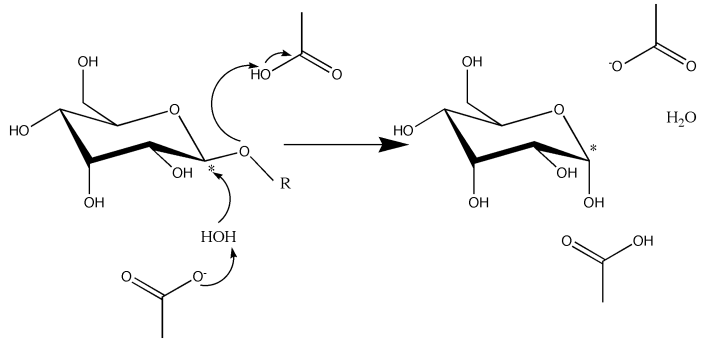


Figure 6: Schéma du mécanisme d'inversion de configuration d'une β-glycosidase où (*) représente le carbone asymétrique [47].

(ii) Les glycosyltransferases (GTs), correspondant aux EC 2.4.x.x, catalysent la formation des liaisons glycosidiques d'un donneur glucidique phospho-activé sur un accepteur glucidique ou non glucidique. Elles assurent en majorité l'assemblage des glycoconjugués et des sucres complexes. Comme chez les GHs, il est possible de diviser les GTs en fonction de leur mode d'action, celles à mécanisme d'inversion et celles à mécanisme de rétention de la conformation anomérique par rapport à la stéréochimie du sucre donneur [48].

(iii) Les polysaccharides lyases (PLs), correspondant au EC 4.2.2.x, clivent les liaisons glycosidiques par un mécanisme de β-élimination sans intervention de molécule d'eau [39, 49]. L'action des PLs aboutit à la formation d'une extrémité non réductrice insaturée et d'une nouvelle extrémité réductrice. Cette catégorie est à la base d'une partie des travaux décrits dans ce manuscrit.

(iv) Les carbohydrate estérases (CEs) catalysent la désacylation en O ou en N de sucres estérifiés. Deux grands groupes de substrats sont hydrolysés par les estérases: les esters d'acide uronique (ex : les esters méthyliques de la pectine) et ceux de glucides neutres estérifiés par des chaines d'acides aliphatiques courts (e.g. le xylane acétylé). L'action des CEs permet souvent de faciliter l'action d'autres enzymes comme les GHs et les PLs sur les polysaccharides estérifiés [50, 51].

(v) Les modules d'adhésion aux glucides (de l'anglais *Carbohydrate-Binding Modules* ou CBMs) sont les seuls éléments de la classifications de CAZy à ne pas avoir d'activité enzymatique [52, 53]. Ces modules sont souvent associés a des modules ayant une activité catalytique (GH, GT, CE, PL ainsi que d'autres modules non-couverts par CAZy). Leur rôle le plus reconnu est leur capacité de liaison ou d'adhésion aux saccharides afin de faciliter l'association des enzymes sur son substrat [52, 53]. Dans certaines familles on parle également de multispécificité des CBMs car ceux-ci peuvent avoir pour cible plusieurs substrats.

(vi) Enfin, d'autres modules aux rôles divers (l'adhésion cellulaire, l'assemblage du cellulosome et l'ancrage des protéines) sont également annotés mais restent non accessibles publiquement. Ils viennent s'ajouter à la base de données CAZy afin d'étendre la description de chaque peptide. Ces modules catalytiques ou non catalytiques peuvent se retrouver associés aux autres modules de CAZymes. Un exemple typique est celui des modules de fibronectine de type III (FN3) [54]. La plupart du temps, ces modules FN3 interconnectent un domaine catalytique et un CBM ce qui laisse présumer un rôle de « *linker* globulaire ». Cependant, d'autres fonctions ne sont pas à exclure telles que l'ancrage cellulaire ou la liaison aux sucres [55, 56].

Des régions particulières plus générales telles que les segments transmembranaires, les peptides signaux ou les « *linkers* » inter-modules sont aussi gérés au sein du système d'annotation de CAZy. Enfin, des régions non caractérisées mais présentant des homologies dans plusieurs séquences de CAZymes différentes sont répertoriées sous le terme de familles de modules « X ». Plus de 240 familles de modules « X » ont été définies jusqu'à présent et peuvent représenter des modules potentiels de CAZymes. Ces familles constituent une ressource de découverte importante car plusieurs familles de modules X définies dans le passé correspondent actuellement à des familles de modules CBM ou d'enzymes.

I.1.2.5. Les approches bioinformatiques de mise en évidence des modules

La structure modulaire d'une protéine peut-être mise en évidence par des approches structurales ou simplement par des approches bioinformatiques [22]. L'analyse de séquences en acides aminés de différentes protéines permet souvent l'identification de régions homologues dont les contours ou limites sont plus ou moins faciles à déterminer en fonction du degré de similitude. Idéalement, les limites d'un module dans une séquence polypeptidique sont déterminées par analyse des structures tridimensionnelles. En absence de données structurales de référence, l'identification de modules et de leurs limites se fait par analyse bioinformatique.

L'information des limites est ultérieurement propagée lors de l'annotation modulaire de séquences homologues [40].

Afin de faciliter l'analyse manuelle des curateurs, plusieurs approches bioinformatiques ont été utilisées au fil du temps permettant de mettre en évidence la modularité au sein des protéines actives sur les glucides. Ces approches ont varié en fonction du développement des analyses et de la spécialisation de la classification de familles de protéines. Les principales approches utilisées historiquement ont été :

(i) l'analyse des amas hydrophobes ou HCA [20], une approche basée sur la détection de segments structuraux constituant le cœur hydrophobe des protéines globulaires. L'intérêt de cette approche est que des similarités dans le repliement tridimensionnel peuvent être détectées entre des protéines possédant des identités de séquences très faibles (<20% d'identité). De plus, l'analyse HCA permet de discerner les bornes des modules globulaires.

(ii) les méthodes de recherche basées sur des alignements de séquence : typiquement par BLAST [57] ou FASTA [58]. Les séquences cibles identifiées sont classées selon un score de similarité ou selon la probabilité d'obtenir ce score par hasard (E-value). Ces deux approches permettent la détection rapide, dans un groupe de séquences, des protéines présentant des similarités avec la séquence d'intérêt.

(iii) les alignements multiples de séquence [59] : les logiciels ClustalW [60], MUSCLE [61] étant les principaux utilisés dans CAZy. Un alignement de séquences de protéines homologues doit révéler des caractéristiques communes conservées après divergence. Ces caractéristiques sont importantes pour la fonction ou pour la structure de chacune de ces protéines [62]. L'analyse d'alignements peut faciliter la définition de bonnes bornes modulaires et mettre en relief des régions peu conservées. Pour remplir ces deux critères, l'échantillon de séquences sélectionné pour l'alignement doit être de qualité, c'est à dire, que l'ensemble de séquences choisies soit divers mais suffisamment similaire. Les paramètres empiriques utilisés sont

difficilement quantifiables car ils résultent de nombreuses années de pratique dans la classification d'enzymes actives sur les glucides et varient de famille en famille [40].

(iv) l'utilisation des profils ou motifs de séquence correspondant à des modules ou à des régions protéiques. Différents outils permettent de comparer des séquences individuelles aux bibliothèques de motifs de familles. L'intérêt de ces profils réside dans leur capacité de synthétiser l'information de familles de modules et ainsi de détecter de nouveaux modules appartenant aux mêmes familles. Trois types de profils, reposant sur des méthodologies différentes sont utilisés : les profils ou PSSM « *Position-Specific Scoring Matrices* » (e.g. CD-Search [63]) résumant l'information de séquences alignées dans une matrice à 2 dimensions, les profils HMM [64] (e.g. Pfam [65]), structures composées d'états et de probabilités de transitions et enfin les patterns ou motifs (e.g. PROSITE [66]) résumant l'information à des positions jugées importantes dans un alignement.

Ces différents outils sont utilisés à présent pour annoter la modularité au sein des CAZymes. Il est important de remarquer que la nature apparemment « aléatoire » de la disposition de modules dans une protéine peut poser des problèmes aux outils bioinformatiques lors de l'annotation modulaire de séquences. En effet, la similarité de séquences partielles (correspondant à un ou plusieurs modules) ne garantit pas l'annotation modulaire d'une séquence protéique et engendre souvent des erreurs de résultat d'analyse [17]. Il est donc primordial de raisonner au niveau du module si l'on souhaite obtenir une analyse pertinente. La classification des enzymes se fait donc selon des propriétés structurales propres aux enzymes et à l'unité évolutive de leurs modules (**Figure 7**).

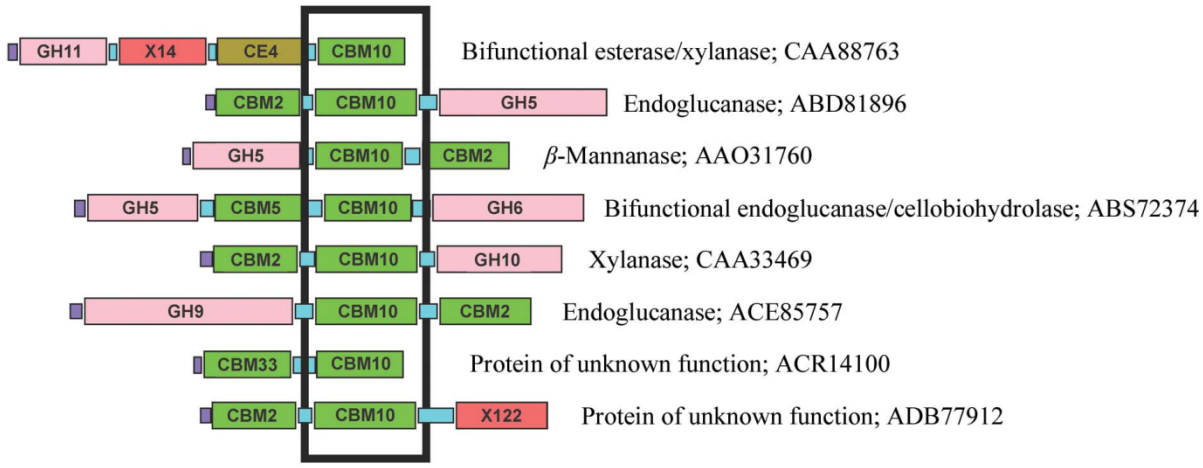


Figure 7: Exemple de variation de la modularité au sein de la famille CBM10.

I.1.2.6. La procédure de mise à jour

En 1999, date de création de la base de données relationnelle, une procédure de mise à jour automatisée a été mise en place. Le but de cette approche était d'alléger le processus d'ajout de nouvelles séquences par analyse comparative des séquences et modules de la base de données CAZy afin d'actualiser les différentes familles. Quotidiennement, ce sont plusieurs centaines voir souvent plusieurs milliers de nouvelles séquences de GenBank [41] contenues dans un fichier rendu disponible par le NCBI (« fichier du jour ») qui doivent être comparées aux entrées existantes dans CAZy. Le nombre de séquences à analyser est d'ailleurs en augmentation constante en raison de l'accélération du débit des projets de séquençage. Au début des années 2000, une procédure basée sur BLAST [57] permettait d'annoter ces séquences de manière semi-automatique. Cette procédure consistait à réaliser une comparaison par BLAST de chaque nouvelle séquence contre la bibliothèque des modules de CAZy et à reporter le résultat sous une forme graphique. L'interprétation de ces résultats était ensuite entièrement laissée à l'annotateur et pouvait s'avérer très subjective quant au seuil de *e-value* à partir duquel un module pouvait être défini ou non. En effet, un seuil de nature probabiliste s'avère variable selon le type, la longueur et la famille du module.

Bien qu'efficace, puisque utilisée avec succès pendant près de 10 ans, cette procédure reposant uniquement sur la similarité de séquences souffrait toutefois de limitations qui empêchaient une automatisation plus poussée. La procédure a évolué en 2008 avec l'introduction de HMMs (modèles de Markov cachés [64]) pour les familles de modules. De plus une combinaison d'analyse de *patterns* et d'identification de signaux peptiques et de segments transmembranaires est réalisée par Phobius [67, 68]. Le *pipeline* d'annotation semi-automatique développé en PERL (www.perl.org) a pour étapes :

- une recherche de séquences qui ont une corrélation avec les entrées de la base de données CAZy pour alléger le fichier original et regrouper toutes les CAZymes potentielles ;
- une comparaison des séquences avec BLAST sur des bibliothèques de modules de CAZy, les séquences présentant un score supérieur à 80% sont entrées automatiquement ;
- un deuxième tri est fait sur les séquences restantes contre les bibliothèques de BLAST et les profils HMMs correspondant à chaque famille et sous-famille. Si les bornes déterminées par les deux comparaisons sont similaires, la séquence est entrée automatiquement dans la base de données CAZy, sinon elle est laissée pour être analysée manuellement par le curateur. Celui-ci décide alors de conserver ou de rejeter la séquence [9, 22, 24].

Cette procédure est à la base de la mise en place de l'automatisation du *pipeline* d'analyse des génomes et métagénomes.

I.2. L'analyse des CAZymes dans le contexte des biocarburants

Les CAZymes ont de multiples applications biotechnologiques, car les glucides constituent des sources de matières premières majeures pour notre société. Les enzymes permettant de transformer les glucides sont utilisées par un large panel d'industries. On compte parmi celles-ci

les industries des détergents, du textile, des produits de pâtisserie et de panification, du bioethanol, de l'alcool, de l'alimentaire et du papier [69].

Mes travaux de thèse ont été financés par la société Novozymes A/S, Bagsvaerd, Danemark, qui est à présent le plus grand producteur mondial d'enzymes industrielles (47% de part de marché en 2010). La société Novozymes emploie 5300 personnes dans 30 pays et leur marché s'étend sur environ 700 produits dans le monde. Les principaux axes commerciaux de cette société sont les enzymes qui représentent 92% de leurs ventes, mais également les microorganismes et les ingrédients biopharmaceutiques. Cette firme est très impliquée sur le marché des biocarburants et recherche entre autres à l'échelle industrielle, de nouvelles enzymes qui optimiseront des cocktails enzymatiques adaptés à la dégradation de la biomasse végétale. Cette dégradation à pour but de produire des sucres fermentescibles afin d'obtenir de l'éthanol [70], un des constituants principaux des biocarburants.

I.2.1. Qu'est ce qu'un biocarburant ?

Les biocarburants sont des carburants obtenus à partir de la matière organique renouvelable issue de la biomasse par opposition aux carburants issus de ressources fossiles [71]. Actuellement, deux filières principales sont développées ; la première est le biodiesel, reposant sur l'huile d'origine végétale et ses dérivés, et la deuxième est le bioéthanol obtenu par fermentation à partir de l'amidon ou de la lignocellulose.

Au sein de la filière alcool (ou bioéthanol), on distingue les biocarburants de première et de deuxième génération. Les biocarburants de première génération utilisent comme matière première les glucides de réserve des plantes agricoles (céréales, canne à sucre, betterave, *etc*) [72]. Ces matières premières étant également utilisés pour l'alimentation humaine, la production de biocarburants entre en compétition avec la production alimentaire et engendre de nombreuses polémiques. En opposition, les biocarburants de seconde génération sont générés à partir de

plantes non alimentaires ou de résidus agricoles. Seuls les aspects de production de bioéthanol à partir de la lignocellulose seront traités dans ce manuscrit.

La paroi externe des plantes riches en glucides est centrale pour la valorisation de la lignocellulose. Il est ainsi concevable d'utiliser un grand nombre de sources de lignocellulose provenant des pailles, des tiges, des feuilles, des déchets verts [73] ou même des plantes dédiées à la croissance rapide (e.g. *Miscanthus* [74]). La production de biocarburants de deuxième génération nuit donc moins aux productions agricoles à visée alimentaire et permet de valoriser des matières premières peu exploitées. L'obtention d'éthanol par cette voie permettra de remplacer partiellement l'essence pour les transports automobiles, en contribuant à limiter la consommation de carburants fossiles.

I.2.2. *Trichoderma reesei* et la dégradation de la biomasse

La production de bioéthanol de deuxième génération repose à présent sur le développement de souches industrielles de champignons filamenteux capables de produire, à bas coût, des cocktails enzymatiques pour la saccharification de la matière première. Le champignon saprophyte mésophile *Trichoderma reesei* présente un grand intérêt industriel (**Figure 8**).

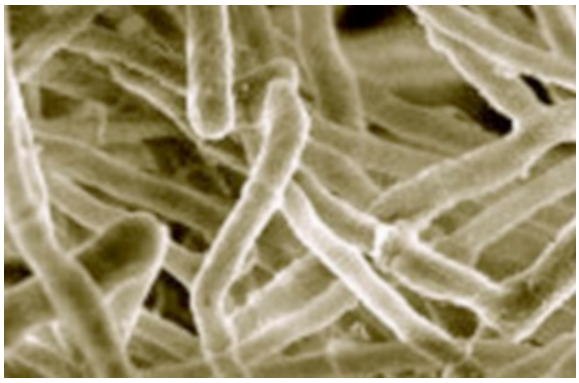


Figure 8 Image au microscope électronique de *Trichoderma reesei* (Photo: Irma Salovuori, VTT Biotechnology)

T. reesei a été découvert et analysé lors de la Seconde Guerre mondiale dans le Pacifique Sud car il dégradait les toiles de coton de l'armée américaine. Ce champignon est actuellement

considéré comme la référence pour la saccharification des polysaccharides dans le monde entier [75]. *T. reesei* est donc devenu l'organisme de choix pour la majorité des projets de recherche dans le domaine car il est capable de produire des cocktails complets de cellulases et hemicellulases à plus de 50 g/l [76].

L'étude du génome de *T. reesei* a révélé un nombre très limité d'activités enzymatiques permettant la digestion des composants de la paroi des plantes (cellulases, hemicellulases, pectinases) en comparaison à ce qui est trouvé habituellement chez d'autres champignons ayant les mêmes propriétés [77]. Le surprenant constat des limitations de cet organisme modèle suggère que sa sélection par l'homme a reposé surtout sur son extraordinaire capacité de production de protéines. Son cocktail enzymatique se prête alors à de nombreuses améliorations en vue d'une saccharification plus efficace pour la production de bioéthanol [77].

I.2.3. Les éléments de la paroi cellulaire des plantes

La paroi cellulaire des plantes est composée de lignocellulose constituée de glucides (incluant la cellulose, les hémicelluloses et la pectine) et de lignine, riche en composés phénoliques (**Figure 11**). La paroi végétale comprend une paroi primaire et une paroi secondaire qui se forment successivement [78]. La paroi primaire est présentée comme un réseau lâche de microfibrilles de cellulose englobées dans une matrice amorphe fortement hydratée de pectines et d'hémicelluloses. Par contre, la paroi secondaire est une structure inextensible et faiblement hydrophobe, constituée de cellulose et de lignine. La description de chacun des composants est décrite maintenant :

(i) La cellulose constitue la trame principale de la paroi primaire. Cet assemblage polysaccharidique est constitué d'un motif répété de deux résidus de β-D-glucopyranose reliés par une liaison osidique β-1,4 et dénommées cellobiose. L'existence de liaisons d'hydrogène inter-chaînes permet d'établir un réseau de nature cristallin résultant d'une accumulation linéaire de chaines. Ces réseaux cristallins sont à l'origine de microfibrilles de

cellulose. Cet assemblage lui confère diverses propriétés physiques, dont l'insolubilité dans l'eau [79].

(ii) L'hémicellulose joue un rôle fondamental dans le maintien de l'architecture des fibrilles de cellulose. Elle est formée d'un ensemble de polysaccharides branchés dont les monomères sont majoritairement constitués de résidus de D-xylopyranose, de D-mannopyranose, de D-galactopyranose et de L-arabinofuranose. Les composants des hémicelluloses les plus importants sont les xyloglucanes, les glucomannanes, et les arabinoxylanes (**Figure 9**). La classe la mieux étudiée correspond aux xyloglucanes car ils sont les composants majeurs de la paroi primaire. Ils sont constitués d'une chaîne principale formée de résidus de xylopyranose reliés par des liaisons β-1,4- rapellant la cellulose et de courtes chaînes latérales de D-xylose, D-galactose et L-fucose. La présence de chaînes latérales empêche la formation de microfibrilles de xyloglucane mais permet de contracter des liaisons hydrogène en surface avec la cellulose [80].

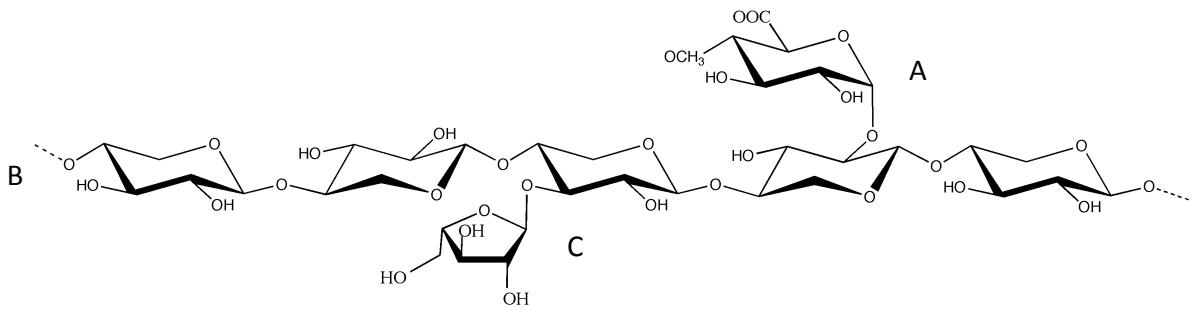


Figure 9: Représentation d'une molécule de xylane, un des composants de l'hémicellulose. La molécule de xylane a pour chaine principale des résidus liée en β-1,4 de xylopyranose (B) sur lesquels peuvent s'ajouter des résidus tel que l'arabinofuranose (C) et/ou l'acide (4-O-methyl)-glucuronopyranosidique (A).

(iii) La pectine résulte d'un ensemble de polymères complexes, présents en abondance dans la lamelle moyenne et la paroi primaire [81]. Le squelette de la chaine principale est riche en résidus d'acide D-galacturonique auquel s'ajoutent des chaines de rhamnogalacturonannes (RG) décomposées en RG de type I (RG-I) et II (RG-II).

Le squelette du RG-I comprend alternativement des résidus d'acide α-D-galacturonique et de l'α-L-rhamnopyranose (**Figure 10**A). Différents substituants polysaccharidiques neutres (e.g. arabinane, galactane, arabinogalactane) formant de larges chaînes latérales sont capables de se greffer à ce squelette osidique, via des liaisons au niveau du groupe 4-OH du L-rhamnopyranose.

La structure du RG-II, est de nature bien plus complexe mais semble cependant présente de manière universelle dans les plantes. Il comprend un squelette de structure variable, contenant des résidus d'acide D-galacturonique, de L-rhamnopyranose, de D- et L-galactopyranose, des D-glucopyranose ou encore de L-fucopyranose. Les substituants des chaines latérales sont de nature et de diversité tout aussi large, ce qui donne au final des structures ramifiées assez complexes identifiées comme les zones hérissées des pectines. (**Figure 10**B)

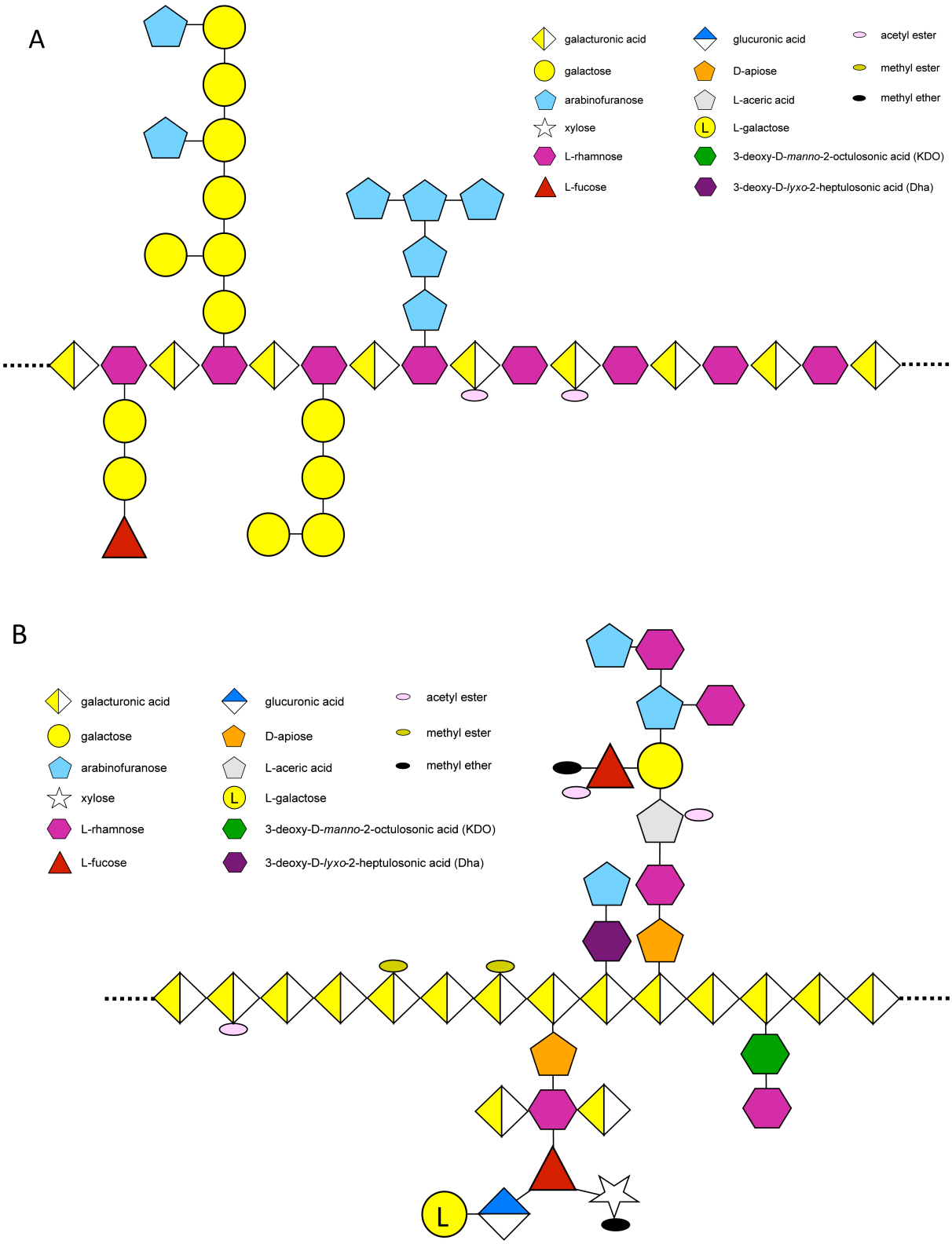


Figure 10: Schéma de la structure du (A) rhamnogalacturonanne de type I (RG-I) et du (B) rhamnogalacturonanne de type II (RG-II) On observe différents types de chaines latérales dont la composante chimique est différente de la chaine principale. Figure adaptée de Mohnen *et al.* [81]

(iv) La lignine a pour fonction principale d'apporter une rigidité, une imperméabilité à l'eau et une grande résistance à la décomposition de la paroi. Bien que la lignine forme un réseau tridimensionnel hydrophobe complexe, l'unité de base se résume essentiellement à une unité de phénylpropane. Cette unité de base est associée de façon covalente à la cellulose ou à l'hémicellulose via l'acide férulique. Les trois constituants de base formant la lignine sont les monolignols : l'alcool 4-coumarylique, l'alcool coniférylique et l'alcool sinapylique. La lignine est caractéristique des plantes vasculaires terrestres. Sa nature chimique lui confère un rôle de barrière de protection contre l'attaque microbienne car elle est extrêmement résistante à divers agents chimiques et à la dégradation biologique [82].

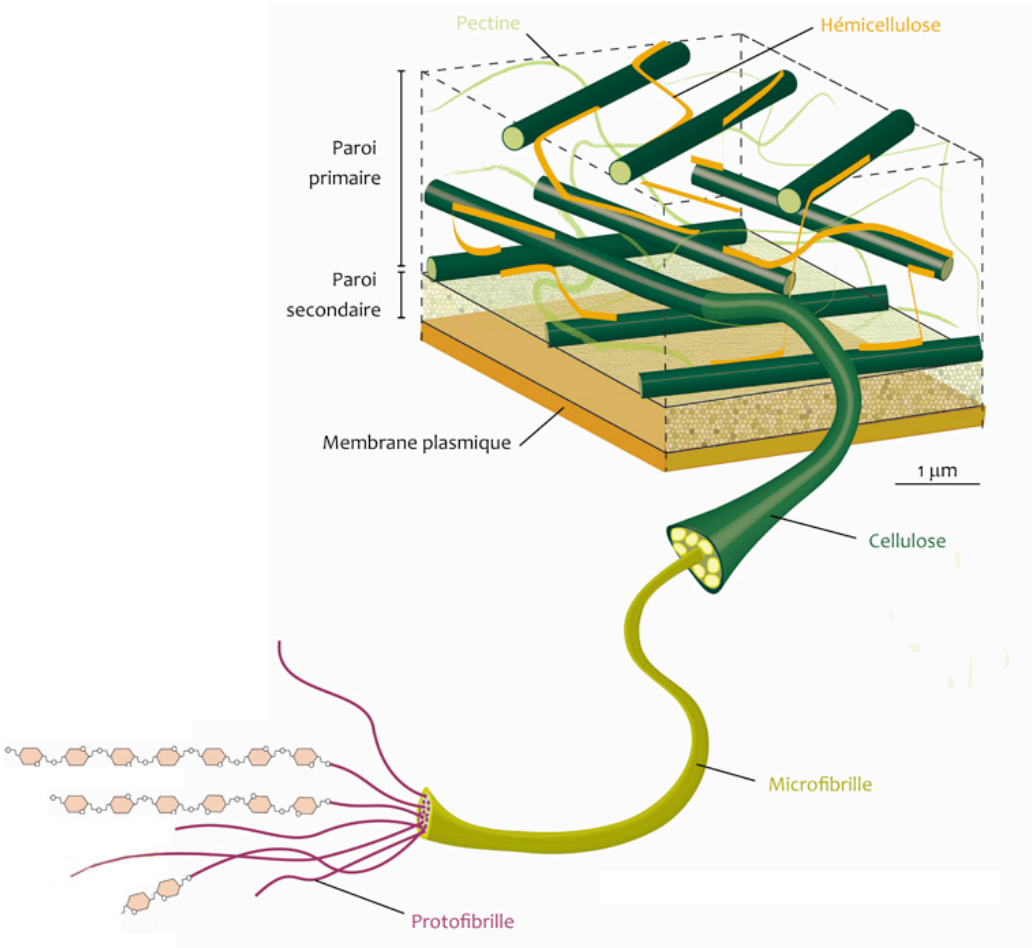


Figure 11: Schéma récapitulatif de la structure imbriquée des différents éléments de la paroi végétale

I.2.4. Recherche de nouvelles enzymes intervenant dans la dégradation de la biomasse

Comme nous l'avons vu dans le chapitre précédent, la paroi végétale constitue une source abondante de carbone renouvelable capable de fournir des sucres fermentescibles afin d'obtenir des biocarburants. La transformation de végétaux en sucres simples passe d'abord par différents types de prétraitement puis subit l'hydrolyse par des enzymes produits par des microorganismes industriels (typiquement des champignons filamenteux). Chaque étape à échelle industrielle de la déconstruction mécanique à la conversion en éthanol en passant par l'hydrolyse, doit prendre en considération plusieurs paramètres (coût, rendement, impact environnemental *etc.*). L'étape de transformation de la lignocellulose joue un rôle majeur dans la viabilité du procédé envisagé car son coût est estimé à la moitié du prix de revient de l'éthanol produit. Ce coût est essentiellement causé par le prix des cocktails enzymatiques [83] utilisés en quantité à cause de leur faible rendement. Des mélanges d'enzymes sont déjà commercialisés par différentes sociétés mais doivent être améliorés. Les recherches actuelles des industriels ont pour but de réduire la quantité d'enzymes à utiliser soit par l'ajout de nouvelles activités, soit par l'utilisation d'enzymes thermostables à haute température, soit par l'intégration d'enzymes agissant en synergie.

I.3. Objectifs de mes travaux

Mon sujet de recherche s'inscrit dans un objectif d'identification *in silico* de nouvelles enzymes ayant une activité sur la conversion de la biomasse avec pour emphase l'analyse de génomes fongiques. Tous ces travaux sont en lien direct avec la recherche de données de séquences génomiques et de données biochimiques présentes dans la base de données CAZy.

Dans la première partie, j'introduirai les besoins d'une nouvelle structure de base de données et d'une nouvelle interface utilisateur. Ces besoins étaient principalement liés à une meilleure gestion des génomes, des familles et sous-familles, des fonctions et de nombreuses autres informations. Le but de cette nouvelle implémentation a été d'améliorer la qualité des informations, simplifier la biocuration et d'optimiser la recherche de nouvelles enzymes.

Dans la deuxième partie seront traités des exemples d'utilisation de l'interface ainsi que l'intégration de nouveaux outils permettant l'exploration des données de CAZy. Tout d'abord l'analyse des familles de polysaccharide lyases et la création de sous-familles, dont l'homogénéité fonctionnelle a été révélée. Puis la description d'une stratégie d'identification de nouvelles protéines potentiellement impliquées dans la dégradation de la biomasse végétale par détection systématique de protéines modulaires portant des modules d'adhésion aux composants de la paroi végétale. Enfin, je décrirai les approches automatisées que j'ai implémentées, capables d'analyser de grands volumes de données (méta)génomiques pour en extraire le contenu en CAZymes. Chacune de ces parties sera présentée dans une section distincte pour finir sur des conclusions et des perspectives générales

II. Etudes

II.1. Nouvelle interface et bases de données CAZy

II.1.1. Historique de CAZy

La base de donnée CAZy actuelle est l'héritière de nombreux tableaux au format Excel de la classification structurale de modules enzymatiques et auxiliaires [27]. Suite aux difficultés de gestion manuelle du fait d'un grand nombre d'informations, une première version de la base de données relationnelle CAZy a été construite en 1999. Cette base a repris toute l'information existante sous forme tabulaire, en évitant la répétition et en facilitant la gestion des données. La base CAZy a évolué avec son temps selon les demandes des principaux utilisateurs qui étaient les membres de l'équipe de Glycogénomique. La structure mise en place a permis à CAZy de croître et d'intégrer différents niveaux d'information. Jusqu'à présent, chaque entrée de la base de données correspondait à une séquence protéique avec: (i) ses numéros d'accession dans les bases de données publiques (UniProt [42], GenBank [41], PDB [43]) ; (ii) le nom de l'organisme d'où elle provient lié à sa taxonomie disponible au NCBI ; et, si existante, (iii) de l'information sur sa biochimie, sa structure ainsi que sa bibliographie. Grâce à cette structure, CAZy a atteint un grand niveau de notoriété au sein de la communauté glycobiologique [84]. Avec l'avènement des techniques « omiques » (surtout génomique), la base de données a connu une forte croissance de données (**Figure 12**) engendrant à nouveau des difficultés de gestion. Par conséquent, il a fallu restructurer complètement la base de données et son interface utilisateur.

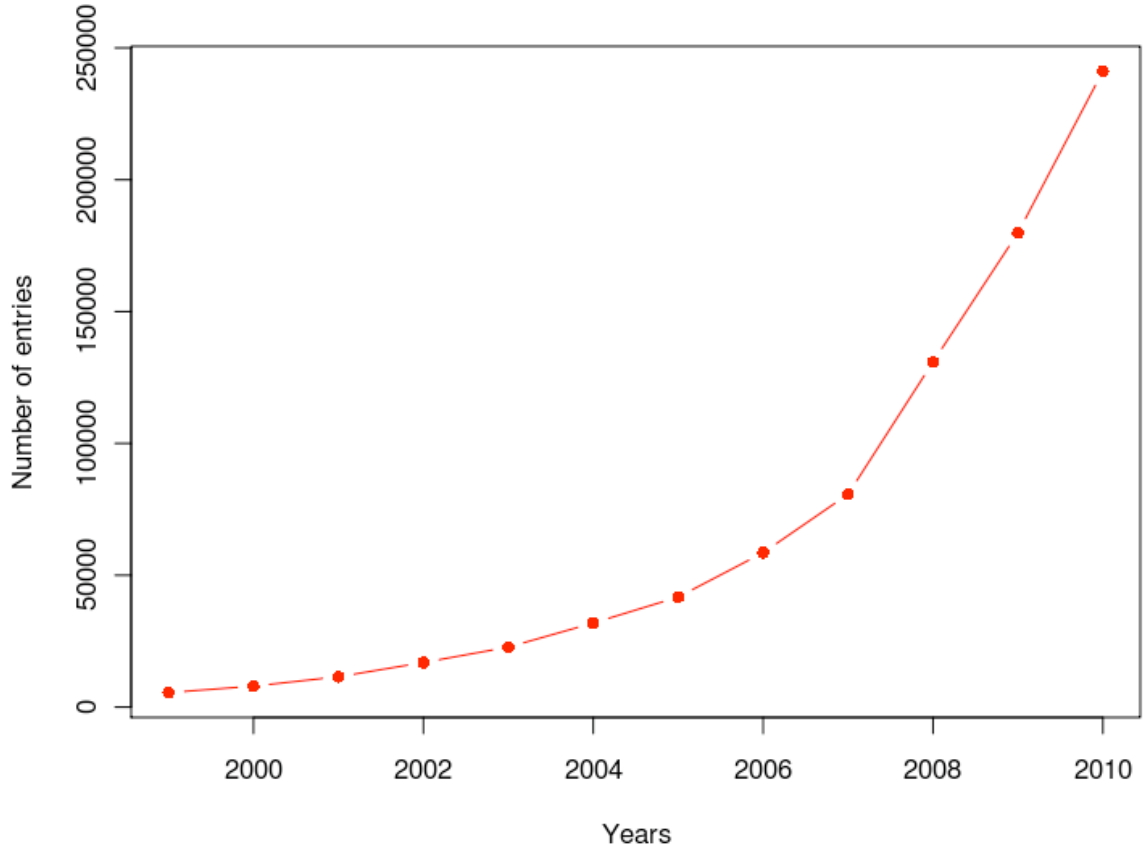


Figure 12: Évolution du nombre d'entrées dans la base de données CAZy de 1999 à 2010. Cette évolution est provoquée par l'augmentation du nombre de séquences dans chaque famille ainsi que par l'intégration de nouvelles familles.

II.1.2. Besoins d'une infrastructure de meilleure qualité

A force d'additions, de modifications d'informations et de tables sous MySQL, la base dans sa version de 1999 a évolué en accumulant un certain nombre de problèmes. Ils pouvaient être d'ordre structural liés aux faiblesses du schéma relationnel et de son interface mais également d'ordre qualitatif liés aux données.

II.1.2.1. Problèmes liés au schéma relationnel

Avant d'énumérer les différents problèmes liés au schéma relationnel, il est bon de se remémorer quelques principes élémentaires sur les bases de données. Une base de données relationnelle est un stock d'informations décomposées et organisées par des relations ou des

tables conformément au modèle de données. Les éléments fondamentaux d'une base de données sont les suivants :

- le schéma de la base de données est tout simplement des groupes d'objets qui sont apparentés et reliés entre eux ;
- les tables servent à stocker les informations auxquelles l'utilisateur doit accéder. C'est l'unité fondamentale de stockage physique des données dans la base ;
- les champs ou colonnes sont les informations contenues dans les tables ;
- les clés sont des valeurs de colonne d'une table qui permet d'établir des relations parent/enfant entre deux tables. Il existe deux types de clés : primaires et étrangères. Une clé primaire rend un champ unique dans une table. Elle sert généralement à joindre des tables apparentées ou à interdire la saisie d'enregistrements dupliqués. Une clé étrangère est la référence de la clé primaire d'une autre table. Elle est définie dans des tables enfant et assure qu'un enregistrement parent a été créé avant un enregistrement enfant et que l'enregistrement enfant sera supprimé avant l'enregistrement parent.

La première génération de la base de données CAZy a été construite afin d'intégrer un nombre de séquences qui pouvait atteindre la centaine en un mois. De nos jours, l'ajout de séquences peut dépasser les 8000 entrées par mois (**Figure 13**). La raison principale de la modification de la base de données a été la prise en considération de l'abondance des séquences obtenues grâce aux nouvelles techniques de séquençage. Cette abondance ne peut être canalisée que si la structuration du schéma relationnel est adaptée. En effet, la structure de la première génération possédait des tables trop spécifiques qui ont alourdi le schéma relationnel. Par exemple, les tables GB_GP, SP, PDB (**Figure 4**) étaient représentatives des bases de données correspondantes mais possédaient des colonnes avec des noms identiques. Dans la nouvelle structure toutes les informations de ces trois tables ont été regroupées en une seule nommée « annotation » en évitant ainsi toute spécificité et redondance (voir schéma relationnel en

Annexe D). De plus, ces tables ne possédaient pas de clés étrangères permettant des relations entre elles, ce qui a engendré la répétition de colonnes identiques au sein de différentes tables. L'ensemble de ces problèmes a surchargé la structure de la base de données et engendré un problème de rapidité d'accès aux données.

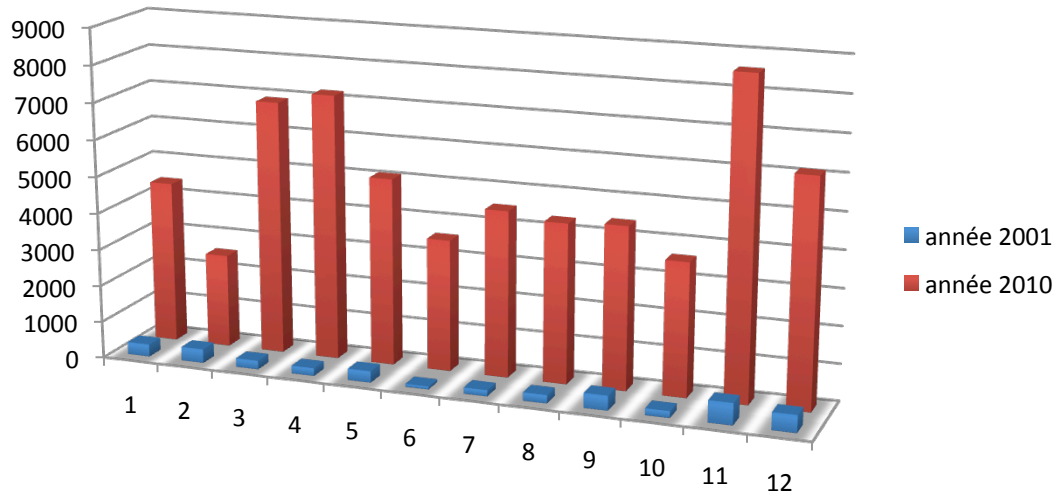


Figure 13: Nombre mensuel de CAZymes entrées dans la base de données en 2001 et 2010. L'arrivée de séquences dans la base de données suit majoritairement les fluctuations de déposition de séquences qui sont relâchées par GenBank.

II.1.2.2. Problèmes liés à l'interface

En 1999, date de création de la base de données relationnelle, une interface d'administration des données à également vu le jour pour que les utilisateurs principaux, curateurs et chercheurs puissent analyser, modifier et ajouter des données.

L'interface a été construite autour du logiciel phpMyAdmin (www.phpmyadmin.net) ayant des droits d'utilisation ou de modification ouvert à tout utilisateur ou « *open source* ». Il a donc été facile d'intégrer à cette interface générique un nombre de modules de programmation écrit en

PHP (www.php.net) aux fonctionnalités spécifiques à CAZy. Cette interface nommée phpCAZy est devenue moins performante avec le temps pour les raisons suivantes :

(i) Les versions de PhpMyAdmin se sont succédées au cours du temps alors que celles utilisées par CAZy ne pouvaient pas évoluer aussi régulièrement à cause d'une conversion fastidieuse et du manque de personnel dédié. Le développement s'est effectué au détriment des versions plus actuelles et donc plus optimisées de phpMyAdmin.

(ii) L'interface « phpCAZy » a été modifiée par de nombreux informaticiens intégrant différents projets sur des périodes plus ou moins longues. Ces modifications apportées ont rendu l'outil CAZy difficile à maintenir, au fur et a mesure des années.

(iii) L'outil phpMyAdmin n'a pas été créé par l'équipe de CAZy, rendant difficile la tâche de compréhension du programme et de ses subtilités.

(iv) La création de la nouvelle structure de la base de données a engendré des modifications importantes au sein du logiciel « phpCAZy » car trop spécifique à l'ancien schéma relationnel.

II.1.2.3. Problèmes d'ordre qualitatif

Certains des problèmes identifiés dans le modèle de la base de 1999 n'étaient pas intrinsèques à la structure de la base de données ou de l'interface, mais plutôt dus à l'accumulation d'erreurs de limites au sein des descriptions modulaires. Ces erreurs ont été générées par une gestion devenue difficile à cause d'un volume de données de plus en plus important. Malgré une annotation modulaire fine des séquences de CAZy par les curateurs, les bornes de certains modules sont parfois appelées à évoluer. Historiquement, les annotations modulaires étaient réalisées par homologie avec les modules des séquences préexistantes dans la base de données. Lorsqu'une nouvelle séquence présentait un degré de similarité faible par rapport aux séquences déjà connues, de petites imprécisions pouvait être introduites lorsqu'on établissait les bornes d'un nouveau module. Ces imprécisions pouvaient s'accumuler lors de l'intégration de nouvelles

séquences toujours plus divergentes par rapport aux séquences ayant servi originalement de modèle lors de la création de famille de modules. Ces données étaient cependant susceptibles d'être corrigées grâce à l'arrivée de nouvelles informations structurales. La difficulté majeure de cet exercice résidait en la propagation des bornes à l'ensemble des modules concernés car il s'agissait d'un travail fastidieux qui impliquait jusqu'alors de corriger manuellement chaque séquence individuellement. Afin de répondre à ce problème, un outil a été créé permettant de traiter un grand nombre de séquences de façon beaucoup plus commode et systématique [68]. Cet outil possède un énorme potentiel d'affinement des bornes et est d'une grande utilité pour gérer un grand volume de données. Il est donc impératif de conserver les outils de travail préexistants et de créer une infrastructure unifiée afin de faciliter la gestion des informations. Finalement, il est primordial d'établir une nouvelle procédure semi-automatique permettant une meilleure reconnaissance des bornes modulaires des CAZymes.

II.1.3. Nouvelle Structure de CAZy

En raison de l'ensemble des problèmes engendrés par la première génération de la base de données et de son interface, une refonte totale s'est avérée nécessaire. L'avantage de cette infrastructure est de faciliter la gestion de nouvelles capacités de recherches dans les domaines de la génomique et de la métagénomique.

II.1.3.1. Nouvelle Base de données

La création d'une nouvelle structure de base de données a pris en considération de nombreux objectifs :

- les nouvelles demandes des utilisateurs (*e.g.* une meilleur gestion des fonctions, de la taxonomie, de la modularité *etc.*) ;
- la conception de nouvelles thématiques de recherche ;
- la conservation des éléments positifs de la précédente version ;
- la gestion de grands volumes de séquences ;

- la résolution de divers problèmes identifiés précédemment.

Un schéma relationnel a été créé (voir Annexe D) en mettant l'accent sur une meilleure gestion des informations à différents niveaux et en privilégiant plusieurs points développées ci-dessous.

II.1.3.1.1. Rapidité d'accès aux données

La structure des tableaux a été optimisée pour que les interactions entre les tables soient plus logiques, en ajoutant des index sur les champs importants et en évitant la redondance afin de permettre la gestion d'un grand volume d'informations (**Figure 14**).

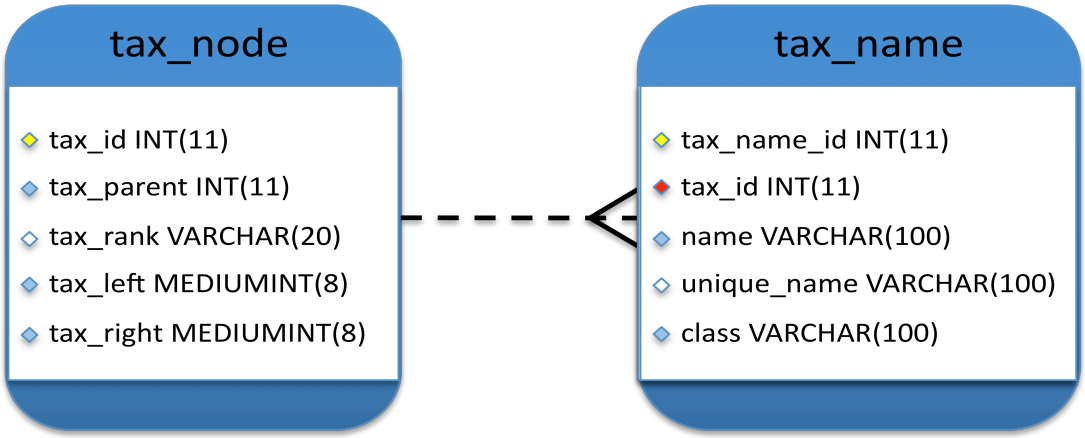


Figure 14: Exemple de structure et d'interaction entre des tables MySQL. Ces deux tables gèrent la taxonomie du NCBI dans la base de donnée CAZy. Le losange jaune représente la clé primaire de la table, le losange bleu clair les index et le losange rouge la clé étrangère. Un index est crée sur les colonnes tax_left et tax_right de la table tax_node de la base de données afin de rendre plus rapide l'accès de ces champs (voir schéma relationnel de la base de données en Annexe D).

Pour une gestion optimale de la taxonomie et un accès rapide aux informations, nous avons utilisé la méthode de « présentation intervallaire des arborescences » (de l'anglais « *preorder tree traversal algorithm* »). Cette technique constitue une manière performante de représenter les hiérarchies de type arborescentes en créant des intervalles adjacents et recouvrants. Elle est basée sur le principe où les feuilles (e.g. 'tax_left', 'tax_right' dans la table 'tax_node') de l'arbre situées au même niveau ont des intervalles adjacents et des nœuds (e.g. 'tax_parent' dans la table 'tax_node') englobant des feuilles ou d'autres nœuds ayant des intervalles recouvrant. Cette méthode permet l'interrogation et la recherche en une seule requête au sein d'une

arborescence (**Figure 15**). D'autre part, afin d'éviter la redondance des CAZymes, une table dédiée uniquement à la gestion des GIs (numéros d'identifiants de séquence protéique du NCBI) est mise à jour quotidiennement.

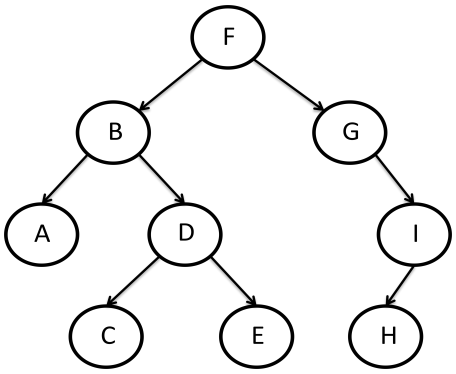


Figure 15: Exemple de 'Preorder traversal sequence'. La lecture de cet arbre commence par la racine (F) puis se poursuit de gauche a droite dans l'ordre suivant : B, A, D, C, E, G, I, H

II.1.3.1.2. Vocabulaire plus structuré

Un des points marquants de la version précédente de la base de données CAZy était le manque de structure au sein de la taxonomie. A chaque nouvelle protéine entrée dans CAZy, la lignée taxonomique était ajoutée entièrement sans tenir compte des mises à jour occasionnelles apportées par le NCBI (nouveaux organismes, souches, espèces). Dans la nouvelle version, la taxonomie des CAZymes utilise spécifiquement les identifiants taxonomique du NCBI. Cette information est mise à jour deux fois par semaine afin de permettre la gestion de son évolution quotidienne. De plus, dans CAZy, les protéines sont liées à des fonctions (évidences expérimentales). Ces fonctions sont attribuées lorsque des évidences expérimentales sont extraites de la littérature sous la forme de numéros EC [40] qui présentent pourtant des limitations (voir introduction). De ce fait, les fonctions ont été organisées en suivant l'arborescence de la classification EC modifiée afin d'être mieux adaptées aux cas des CAZymes (**Figure 16**). Une organisation hiérarchique des données fonctionnelles permet la résolution des ambigüités identifiées jusqu'à présent car elle limite les prédictions aux niveaux supérieurs si nécessaires.

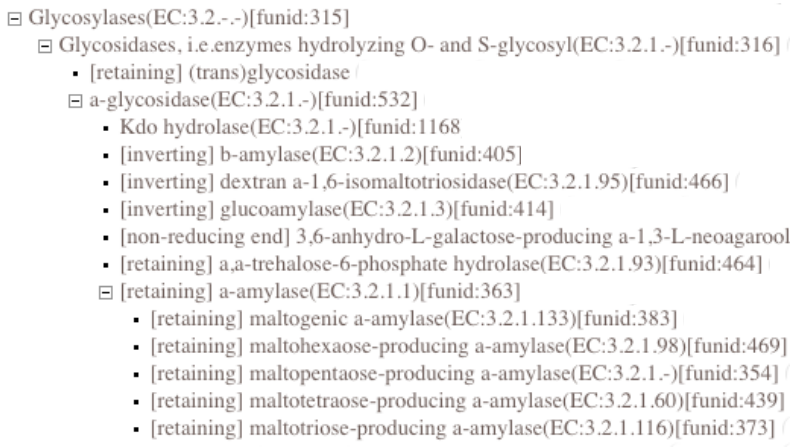


Figure 16: Exemple de classification d'enzymes (EC) hiérarchique utilisée pour les CAZymes. Chaque activité a une identité discriminée.

II.1.3.1.3. Diffusion multiutilisateur de l'information

Une meilleure gestion de la diffusion des données a été implémentée en créant un statut de partage dans la table 'genome'. Cette gestion différencie les génomes analysés en publics, semi-publics (statut de partage spécifique pour des collaborations), privés (uniquement interne à CAZy) ou automatiques (analyses des CAZymes généré automatiquement, i.e. sans intervention de curateur).

L'accès aux données est amélioré grâce à la création de comptes utilisateurs spécifiques avec un système de mots de passe. Les principaux utilisateurs sont les membres du groupe « Glycogénomique » de l'AFMB. Cependant, avec le temps et le succès de CAZy, un nombre grandissant d'utilisateurs a été autorisé à accéder aux informations. Ils sont a présent pris en considération car l'interrogation de la base de données ne se limite plus au groupe mais à l'ensemble du personnel autorisé du laboratoire voir également des collaborateurs externes venus travailler sur un sujet précis.

Au final, la structure de la base de données est résumée dans la **Figure 17** et son schéma relationnel est visible en Annexe D. Cette structure a été conçue pour traiter un large volume d'informations et être suffisamment robuste pour la prochaine décennie.

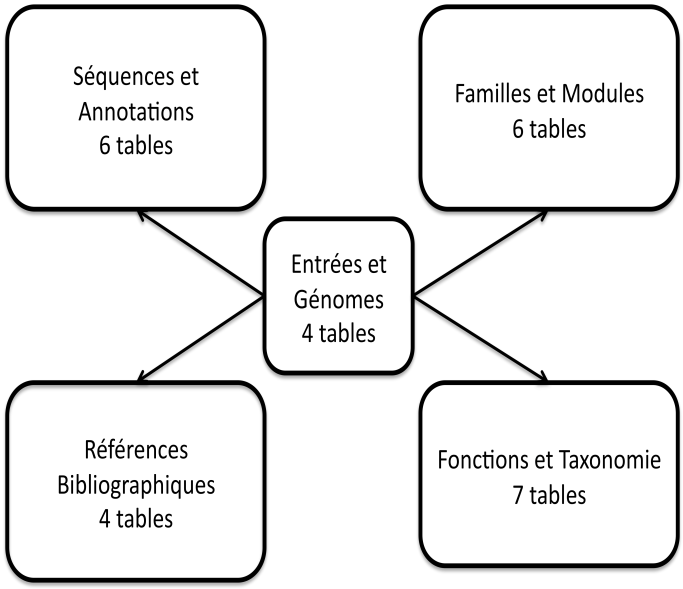


Figure 17: Schéma simplifié de la nouvelle structure de la base de données CAZy. L'ensemble est structuré autour d'un axe principale (les CAZymes) qui est relié à quatre grands groupes d'informations : les annotations, les fonctions, les références bibliographiques et les familles d'enzymes et de modules.

II.1.3.2. Nouvelle interface

La nouvelle structure de la base de données a nécessité l'implémentation d'une interface. Elle a été élaborée par l'équipe, à l'inverse de la réutilisation de logiciels « *open source* » comme au préalable. Les raisons de ce choix ont été les suivantes :

- le contrôle total du logiciel par l'équipe,
- la création d'une interface spécifique à la base de données CAZy,
- la possibilité d'éviter les mises à jour récurrentes, vectrices de possibles incompatibilités avec le système.

L'interface a été conçue en langage PERL (www.perl.org) car les principaux programmeurs impliqués en ont la maitrise. Les bibliothèques Bioperl [85] et DBIx::class (www.dbix-class.org) ont été intégrées aux bibliothèques PERL afin d'alléger la programmation. Bioperl permet la manipulation et le traitement de données biologiques issues de la génomique, la génétique ou de l'analyse du transcriptome. La bibliothèque Bioperl fournit des outils exploités dans l'interface de CAZy pour l'indexation, l'interrogation et l'extraction de données ainsi que des filtres pour un

très grand nombre de sorties de logiciels usuels en bioinformatique (BLAST, Clustalw, HMMer etc.). La bibliothèque DBIx::class permet de structurer la gestion des accès des logiciels aux bases de données MySQL, en apportant une certaine robustesse grâce à son interface orientée objet.

Une période de transition d'un mois en 2008 a été nécessaire entre la mise en place de la nouvelle structure de la base de données, le transfert des données et la création d'une ébauche fonctionnelle de la nouvelle interface. Le mot d'ordre de cette transition a été de transférer et construire au plus vite l'interface afin d'éviter l'interruption prolongée du service CAZy. En effet, avec un flux moyen d'environ un génome par jour additionné avec les fichiers du jour (daily release ftp : ftp://ftp.ncbi.nih.gov/genbank/daily-nc/) provenant du NCBI, il était urgent de créer cette interface pour que les curateurs ne cumulent pas trop de retard sur le travail de vérification et d'entrée d'informations dans la base de données. La période de transition a été délibérément choisie pendant l'été car le mois d'août connait une chute du volume de données libéré par le NCBI. Bien que très rudimentaire à ses débuts, l'interface a très vite évolué jusqu'à surpasser les fonctionnalités de la précédente. Un grand travail d'automatisation des analyses s'est ajouté aux améliorations afin de rendre le traitement de données le plus autonome possible pour alléger la tâche du curateur. La nouvelle interface facilite, également, la navigation au sein de la base afin d'explorer et de comprendre au mieux le rôle, la fonction et les interactions des CAZymes (**Figure 18**).

Figure 18: Exemple de page d'information d'une CAZyme de la nouvelle interface utilisateur. La partie (A-B) permet de naviguer au sein de la base de données. Le détail d'une entrée est représenté de la partie C à J. (A) zone de recherche au sein de la base, (B) menu, (C) titre descriptif, (D) fonction attribuée aux modules provenant de l'étude bibliographique, (E) séquence, (F) représentation graphique modulaire de l'entrée, (G) fonctions connues des familles et sous-familles de modules d'une entrée, (H) informations sur les base de données de référence à l'origine des séquences, (I) les autres entrées avec la même séquence, (J) références bibliographiques liées à l'entrée.

II.2. Applications du nouvel environnement de travail

II.2.1. Etude de la qualité des données des familles et des sous-familles : Classification des PLs

Afin de tester la nouvelle interface il était important de pouvoir traiter un volume significatif de données. Le choix s'est porté sur la catégorie des PLs car elle a pour particularité d'être la classe de CAZymes de plus petite taille et donc facilement gérable en terme de quantité de données. Une première approche d'analyse des PLs avait été préalablement effectuée [68]. Cette étude avait permis de mettre en évidence une partie des sous-familles des PLs. La poursuite de l'analyse a eu pour but de présenter les résultats de façon globale et cohérente. De nouvelles familles et sous-familles ont été implémentées en vérifiant leur robustesse et servant de test au nouvel environnement de travail. L'essentiel de l'analyse est présenté et publié dans un article [39] qui se trouve en Annexe A.

II.2.1.1. Vérification de la cohérence des familles de PLs

Les polysaccharide lyases (PLs) sont des enzymes (EC 4.2.2.-) qui clivent les liaisons glycosidiques des polysaccharides anioniques par un mécanisme de β-élimination [46]. La particularité des PLs, à l'instar des hydrolases, est de catalyser cette coupure sans intervention d'une molécule d'eau. Les PLs sont présentes chez tous les organismes, des virus aux archées, aux bactéries et aux eucaryotes comme les champignons ou les plantes. Le mécanisme catalytique employé par les PLs suit les trois étapes suivantes [46, 86]:

(i) L'abstraction du proton sur le carbone 5 du sucre de l'acide (ou ester) uronique par un groupement amine basique de la chaine latérale.

(ii) La stabilisation de la charge négative générée par délocalisation avec le groupement carbonyle en position 6.

(iii) L'élimination du groupement OR en position 4, facilitée par un transfert de proton d'un groupement acide pour conduire à la formation de l'acide (ou ester) hexenuronic à l'extrémité non réductrice nouvellement formée.

En fonction de la composition en monosaccharide du substrat et de sa conformation dans le site actif de la PL, le proton porté par le carbone 5 et le groupement éliminable en position 4 peuvent être en syn ou anti l'un de l'autre. Cela implique que l'élimination peut avoir lieu par un mécanisme concerté ou consécutif (**Figure 19**). Dans les deux cas, l'état de transition est stabilisé par la présence d'un cation bivalent (souvent Ca^{2+}) ou d'un acide aminé du site actif chargé positivement.

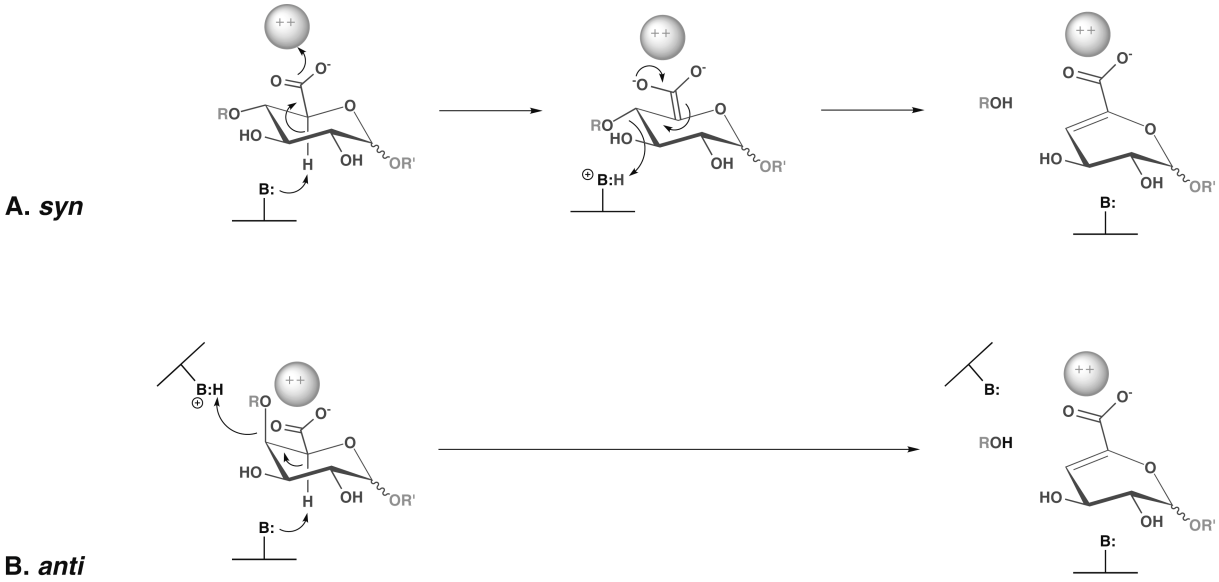


Figure 19: (A) Mécanisme d'élimination syn observé chez la chondroitine lyase, (B) Mécanisme d'élimination anti observé chez la pectate lyase. Dans les deux cas, les polysaccharides sont coupés pour générer un groupement acide 4-desoxy-hex-4-eneuronique à l'extrémité non réductrice. Les substrats gluco- ou galacto- conduisent essentiellement au même produit d'hydrolyse par perte du centre asymétrique en position C4

La catégorie des PLs a été créée en 1998 aux débuts de la base de données en cherchant les séquences homologues d'enzymes caractérisés [27]. Depuis leur création, de nouveaux modules ont enrichi les différentes familles de PLs (**Figure 20**) dont le nombre a augmenté pour atteindre plus d'une vingtaine actuellement. L'annotation modulaire des PLs a été réalisée selon des relations d'homologie établie contre les modules préexistants [40]. Du fait de leur ancienneté dans

la base de données, un travail d'analyse a été réalisé de manière à assurer de la cohérence entre les premières familles de PLs créées et celles nouvellement classées.

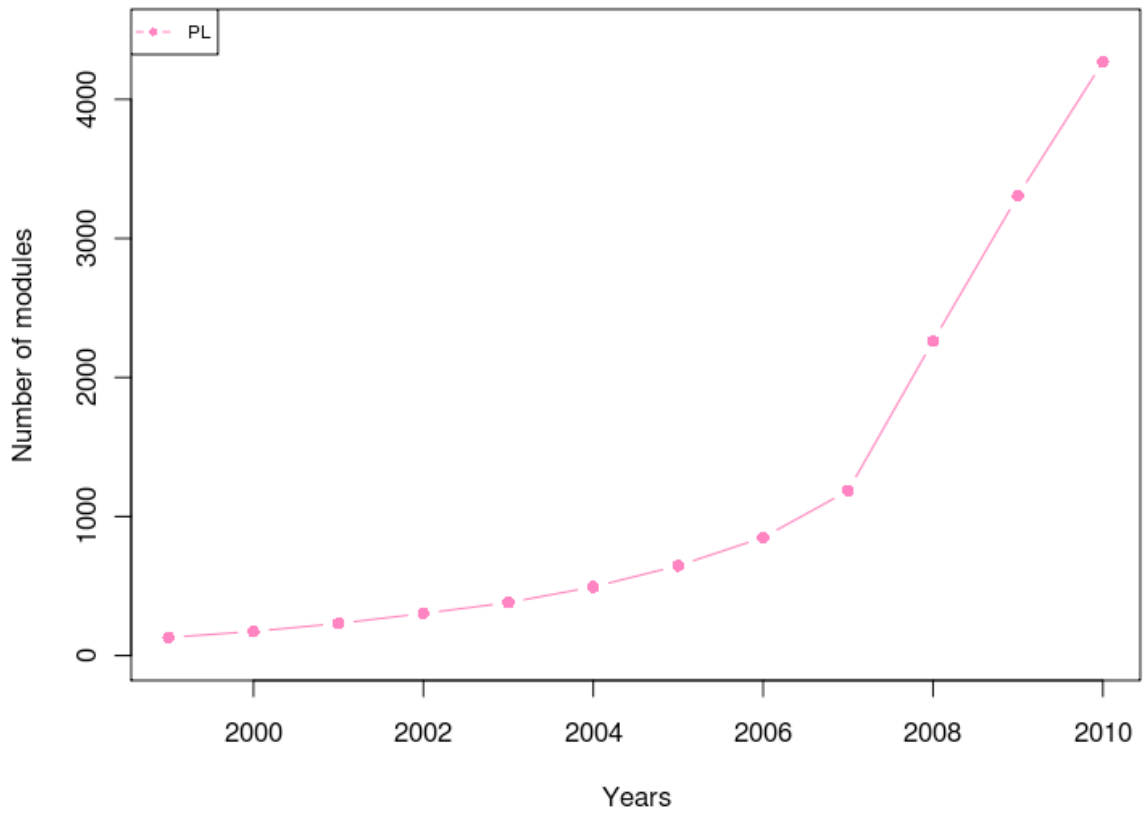


Figure 20: Évolution du nombre de modules des polysaccharide lyases (PLs) de 1999 à 2010.

Depuis la création de la classification CAZy, les familles de PLs ont connu une accumulation de dégénérescences des limites modulaires lors de l'insertion de nouvelles entrées par la procédure semi-automatique journalière, ce problème est décrit auparavant. Pour vérifier que les différentes familles de PL n'ont pas dégénéré, un travail en deux étapes a été réalisé:

(i) Tout d'abord un alignement multiple des séquences de chaque famille a été construit. La qualité générale de l'alignement a été estimée, en particulier par vérification de l'existence de positions conservées correspondant aux résidus catalytiques avérés ou potentiels.

(ii) Une deuxième étape consistait à contrôler la singularité des différentes familles en les comparant les unes aux autres avec l'outil PSI-BLAST qui permet de détecter des relations distantes [87]. En partant de séquences choisies aléatoirement au sein d'une famille, dix itérations de PSI-BLAST avec un seuil de E-value de 10^{-3} ont été réalisées. L'ensemble des séquences de la famille devrait être retrouvé en un nombre réduit d'itérations.

Les alignements de la première étape se sont révélés harmonieux, signe de leur bonne homogénéité et d'une bonne curation. Lors des analyses PSI-BLAST, l'ensemble des séquences au sein de chaque famille a été retrouvé dès la deuxième ou la troisième itération en fonction de leur taille, quelle que soit la séquence choisie comme point de départ. Ces études ont permis d'obtenir des ensembles cohérents.

II.2.1.2. Division des familles de PLs en sous-familles

La création des sous-familles à été réalisée en utilisant une procédure analogue à celle utilisée antérieurement au laboratoire pour l'étude de la famille GH13 [38] qui comprend une grande diversité d'enzymes aux structures similaires agissant sur l'amidon/glycogène et des composants apparentés. La division des familles de PL en sous-familles a été réalisé majoritairement en utilisant la CAZyBox, outil créé par Thomas Bernard doctorant au laboratoire de 2004 à 2008 au sein de l'équipe de « «Glycogénomique » [68]. Ce logiciel intègre l'ensemble des outils et des étapes nécessaires à la création des sous-familles. Tout d'abord, un important travail d'interfaçage avec la nouvelle structure de CAZy a été effectué afin de permettre un échange d'extraction et d'écriture de données entre l'outil et la base. Il fallait en particulier modifier les accès du logiciel à la base de données et structurer les informations sur les modules, les familles, les sous-familles en fonction du nouveau schéma relationnel.

La procédure de caractérisation des sous-familles consiste tout d'abord à extraire les séquences de l'ensemble des modules catalytiques de chaque famille pour éviter toute interférence de segments additionnels lors de l'analyse. Les domaines catalytiques isolés ont été soumis à un

alignement multiple effectué avec MUSCLE version 3.7 [61]. Une matrice de distances est ensuite générée en utilisant le modèle de substitution BLOSUM62 [88]. Cette matrice est utilisée par l'algorithme SECATOR [89] capable de créer des sous-groupes distincts de séquences lors de la reconstruction d'arbres phylogénétiques. La robustesse de l'arbre généré a été testée en même temps par une procédure Jackknife [90]. Cette procédure consiste à enlever de façon aléatoire des séquences de la famille tout en faisant varier les paramètres de l'algorithme de partitionnement en répétant cette étape un certain nombre de fois, typiquement 10000 fois. Les séquences fréquemment rencontrées dans le même sous-groupe (généralement ayant 80% d'identité) sont ensuite classées dans la même sous-famille. Au final, seules les sous-familles contenant au moins 5 membres ont été conservées. Les séquences non assignées à des sous-familles feront l'objet d'une nouvelle analyse ultérieure, lorsqu'un nombre suffisant de séquences sera intégré dans la base de données.

II.2.1.3. Résultats et discussion

II.2.1.3.1. La structure modulaire

Comme indiqué précédemment, les enzymes agissant sur les sucres sont souvent modulaires et les PLs n'échappent pas à cette règle (**Figure 21**). Les modules le plus souvent liés aux PLs sont les CBMs car certains facilitent l'interaction avec leurs substrats. D'autres arrangements sont également possibles comme ceux résultant d'une fusion avec des modules qui permettent des liaisons avec d'autres macromolécules. C'est le cas par exemple des modules de type SLH qui associent les protéines à la paroi cellulaire de certaines bactéries [91] ou les dockerines [11] qui jouent un rôle important dans l'assemblage des cellulosomes. Les PLs peuvent aussi être liées à d'autres modules ayant une activité catalytique comme les CEs ou ayant une fonction encore inconnue (X).

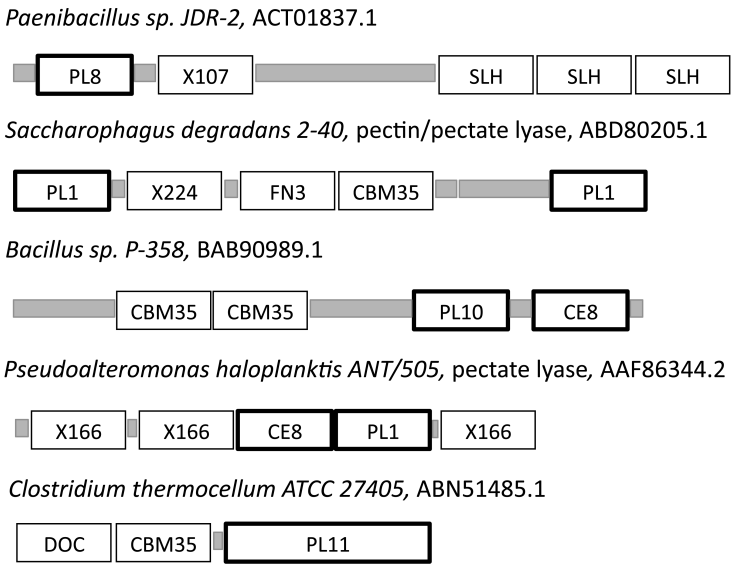


Figure 21: Exemples de modularité chez les polysaccharide lyases (PLs). Les modules de PLs peuvent être fusionnés avec des carbohydrate-binding modules (CBMs), des modules de carbohydrate estérase (CEs), des modules de dockerines (DOC), des domaines S-*layer homology* (SLH), des modules fibronectine de type 3 (FN3) et des modules conservés aux fonctions inconnus (X). Les régions non annotées sont représentées en gris. Chaque protéine est identifiée par un numéro GenBank.

II.2.1.3.2. Familles et repliements

En 1999, on pouvait compter dans CAZy une centaine de PLs arrangées en neuf familles [27]. Dès lors, le nombre de familles de modules de PL identifié a augmenté pour atteindre 22 familles en 2011. Si les familles ont connu un tel essor, c'est sans aucun doute grâce à l'apport continuel de caractérisations biochimiques de nouvelles enzymes et à l'analyse de leur structure tridimensionnelle. Les PLs forment ainsi la classe des CAZymes la mieux définie puisque 12% de ses membres ont été caractérisés contre 8% chez les GHs et seulement 3% chez les GTs et les CEs. En effet, seul le repliement tridimensionnel de deux familles (PL12 et PL17) reste à déterminer sur la vingtaine actuelle. Les différents repliements connus des PLs sont décrits dans la **Figure 22**. Un des plus intrigants est le repliement associé à la famille PL16 car la structure comporte trois chaines polypetidiques imbriquées pour former un triple feuillet de β-hélices. Les gènes de cette famille sont présents chez les phages de *Streptococcus* afin de faciliter la pénétration

de leur hôte et chez les bactéries de ce groupe qui les ont probablement acquis par transfert horizontal [92].

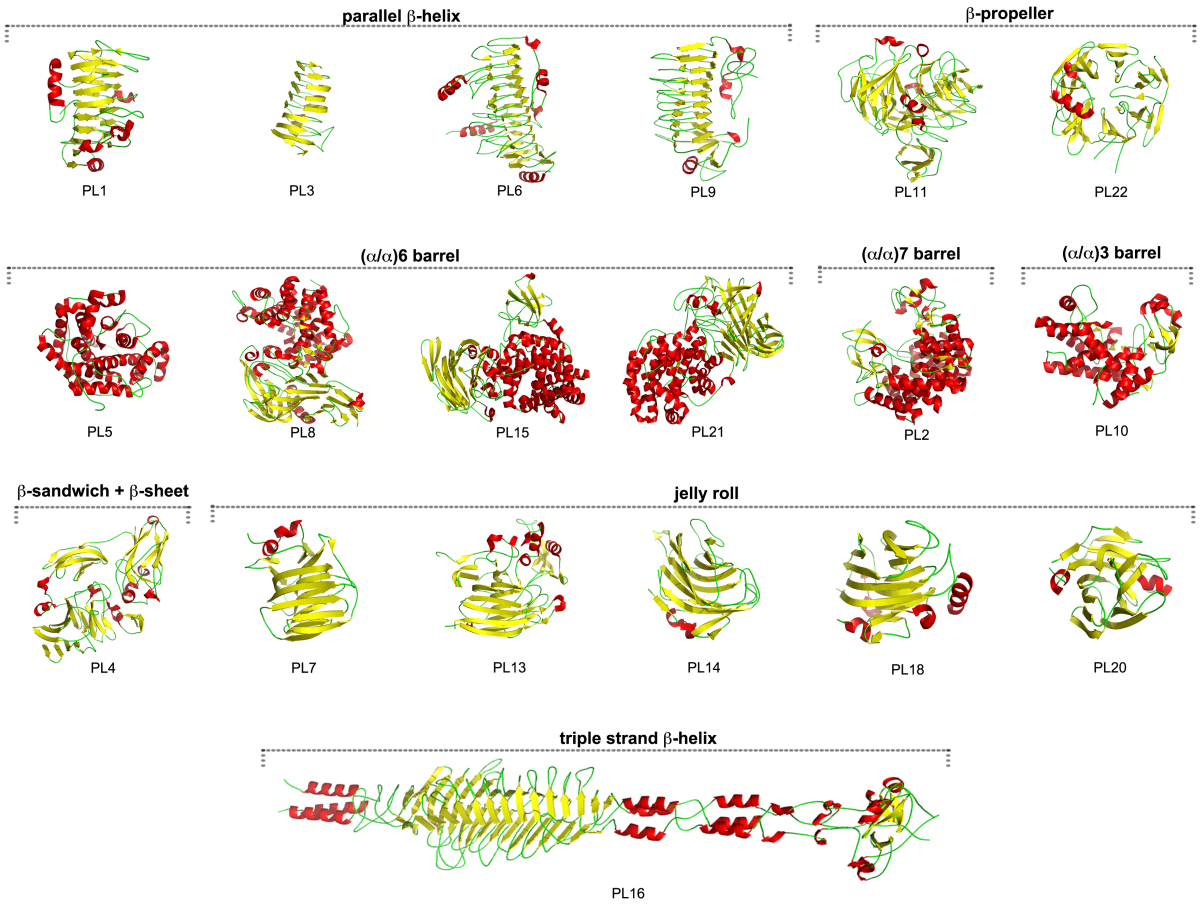


Figure 22: Repliements et structures représentatifs des différentes familles de polysaccharide lyases (PLs) présentes dans CAZy. Les éléments de structures secondaires sont coloriés en rouge pour les hélices α et en jaune pour les feuillets β. Ce panel de structures comprend les entrées PDB suivantes: PL1 (PDB code: 2QY1), PL3 (1EE6), PL6 (1OFM), PL9 (1RU4), PL11 (2ZUY), PL22 (3C5M), PL5 (1HV6), PL8 (1OJM), PL15 (3AO0), PL21 (2FUQ), PL2 (2V8K), PL10 (1GXN), PL4 (INKG), PL7 (1UAI), PL13 (3IKW), PL14 (3AON), PL18 (1J1T), PL20 (2ZZJ) and PL16 (2YW0).

L'abondance des repliements suggère que les PLs ont évolué de façon convergente à partir d'ancêtres de structures différentes. L'exemple le plus extrême au niveau de la convergence des PLs est sans doute entre les familles de pectate lyases PL1 et PL10 où des repliements totalement différents possèdent une même machinerie catalytique [93]. Nous noterons également que certaines PLs peuvent partager leur repliement avec d'autres CAZymes. Le cas le plus marquant est celui de la similarité structurale de plusieurs enzymes avec les repliements en β-hélice parallèle

qui se retrouve chez les pectate lyases de la famille PL1 mais également chez les polygalacturonases de la famille GH28 et les pectine methylesterases de la famille CE8 [94]. Les similarités structurales et l'intérêt pour la pectine comme substrat laissent présager que ces différentes familles partagent un ancêtre commun.

L'étude des familles de PLs révèle également la similarité structurale des différents substrats reconnus au sein de familles individuelles. La famille PL8 est polyspécifique car elle regroupe actuellement quatre types différents d'activité : hyaluronate lyase (EC 4.2.2.1), xanthan lyase (EC 4.2.2.12), chondroitine AC lyase (EC 4.2.2.5) et chondroitine ABC lyase (EC 4.2.2.20). Au niveau enzymatique, ces protéines utilisent une machinerie catalytique identique. Elles coupent la liaison C-O en position 4 du résidu d'acide glucuronique (ou de son epimère l'acide L-iduronique pour la chondroitine ABC lyase). La différence majeure entre ces substrats est la nature du résidu glucidique en position 4 de l'acide glucuronique (**Figure 23**).

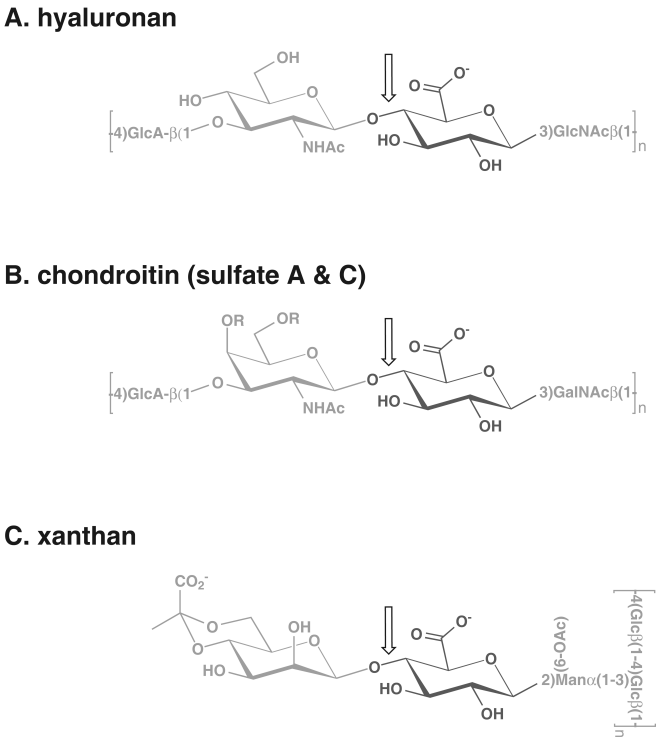


Figure 23: Structure de substrats de la famille PL8. A, hyaluronan; B, chondroitine (R = R' = H) et chondroitine sulphates A (R = SO3-, R' = H) et C (R = H, R' = SO3-); C, xanthan. Le résidu acide glucuronique est en gris foncé, et la liaison clivable est représentée par une flèche. Tous les monosaccharides ne sont pas dessinés.

II.2.1.3.3. Les sous-familles

Les séquences regroupées en familles n'ont pas une relation univoque avec un substrat spécifique car certaines familles peuvent avoir plusieurs activités enzymatiques. Nous avons évalué les sous-familles pour vérifier leur homogénéité fonctionnelle. En utilisant l'approche décrite précédemment, un total de 41 sous-familles a été créé à partir de 21 familles de PLs. Lors de leur création, ces sous-familles représentaient 72% de la totalité des séquences de PLs présentes dans la base de données CAZy. Les séquences non assignées à une sous-famille le seront probablement plus tard lorsqu'un nombre suffisant de séquences proches seront identifiées. Parmi toutes les sous-familles de PLs, seulement sept ne possèdent pas encore de membre caractérisé expérimentalement. On compte 38 sous-familles monospécifiques, soit 90% du total des sous-familles créées.

Malgré un nombre limité d'enzymes caractérisées dans certaines sous-familles, on constate une meilleure corrélation avec la spécificité du substrat par rapport aux familles. On compte cependant trois cas de sous-familles polyspécifiques : PL1_5, PL9_1 et PL14_3. La polyspécificité de PL1_5 et PL9_1 provient des enzymes présentant une action endo (EC 4.2.2.2) et/ou exo (EC 4.2.2.9). Sachant que ces deux activités enzymatiques ont le même substrat, la différence réside simplement dans le degré de polymérisation du produit. Dans la sous-famille PL14_3, la polyspecificité est associée à la présence de poly- et d'oligo-alginate lyases (EC 4.2.2.3 et 4.2.2.- respectivement). Ici encore, la différence est très subtile. Sachant que la liaison coupée est identique, la différence d'activité est due au degré de polymérisation du substrat.

II.2.1.4. Conclusion

L'intérêt principal de cette analyse était de tester simultanément la nouvelle interface CAZy et de vérifier, au sein des familles de PLs, la cohérence des groupes et sous-groupes proposés en utilisant des méthodes d'analyse intégrées à l'interface. Les critères pris en considération

comprenaient les biais liés aux différentes méthodes de partitionnement automatique (influence de l'échantillonnage, de la méthode elle même) et à l'intégration d'informations complémentaires (*e.g.* caractérisation biochimique, modularité). La division finale en familles et sous-familles reste dépendante d'une analyse humaine, dans laquelle s'inscrit l'intuition du chercheur et sa familiarité avec les protéines étudiées. Bien qu'intégrant une part de subjectivité, ces approches semi-automatiques permettent d'affiner les résultats produits par les méthodes entièrement automatiques plus simples et plus généralistes, expliqués par la suite. Les résultats sont très satisfaisants dans la plupart des cas.

II.2.2. Etude et exploration de l'information modulaire : les CBMs

Un autre sujet de recherche en application directe avec le « nouveau CAZy » a été consacré à la recherche de nouvelles enzymes qui dégradent la biomasse végétale. L'équipe « Glycogénomique » dirigé par Bernard Henrissat est en étroite collaboration sur ce sujet avec la société Novozymes A/S, Bagsvaerd, Danemark. Cette collaboration à pour but de favoriser le développement de la production de biocarburants de deuxième génération à partir de lignocellulose. La stratégie d'approche envisagée est la découverte *in silico* de nouvelles enzymes qui permettraient d'améliorer les cocktails enzymatiques actuellement utilisés dans la dégradation et la saccharification de la biomasse végétale (voir introduction).

La recherche *in silico* de nouvelles enzymes s'est essentiellement concentrée sur l'analyse et l'exploration des génomes fongiques et bactériens présents dans la base de données. En effet, bien que la dégradation de la cellulose puisse être réalisée par différentes espèces, seuls certains champignons et bactéries présentent un intérêt pour la production industrielle d'enzymes lignocellulolytiques [95].

L'avantage des champignons réside en leur grande capacité de production de protéines d'intérêt et de synthèse de sécrétomes directement exploitables comme cocktails industriels. L'étude s'est ainsi portée sur l'ensemble des champignons qu'ils soient saprophytes dépendants

de la matière organique végétale en décomposition, parasites d'un autre organisme ou symbiontes en étroite collaboration avec d'autres êtres vivants.

L'intérêt de l'analyse de bactéries cellulolytiques réside dans leur diversité ainsi que dans leur énorme potentiel d'évolution et d'adaptation qui leur confère une certaine richesse au niveau génomique. Au même titre que les champignons, cette grande diversité peut jouer un rôle prépondérant dans la découverte de nouvelles enzymes d'intérêt pour l'amélioration des cocktails enzymatiques.

II.2.2.1. Approches expérimentales

Un des objectifs de la nouvelle structure de la base de données a été de faciliter la gestion et l'analyse de données génomiques. La nécessité d'un tel effort résulte de l'énorme croissance du nombre de génomes banalisé par les nouvelles techniques de séquençage. En Février 2011 on comptait dans la base de données CAZy plus de 1500 génomes publics et plus de 400 privés provenant entre autres de différents consortia génomiques.

II.2.2.1.1. L'analyse des génomes

La plupart des génomes analysés sur le site CAZy (www.cazy.org) sont issus de la base de données publique GenBank [40] mais certains peuvent provenir de collaborations et, dans ce cas, sont annotés en interne comme des « génomes privées ». Afin de mieux comprendre la distribution des CAZymes dans les champignons, j'ai analysé plusieurs génomes fongiques parmi lesquels *Heterobasidion annosum*, *Agaricus bisporus*, *Schizophyllum commune* [96]. L'obtention des séquences de génomes fongiques comprend typiquement des étapes de séquençage, d'assemblage et de détermination des meilleurs modèles protéiques [97]. Au final, l'ensemble des modèles obtenus sera étudié par un consortium d'annotation composé de spécialistes de l'organisme à analyser, et d'experts en différents domaines complémentaires (sécrétome, glycogénomique, *etc.*).

Le groupe de « Glycogénomique » a souvent été sollicité pour son expertise dans le domaine des glucides et de leurs enzymes. L'étude des génomes fongiques au sein du groupe à débuté dans les années 2000 par les champignons ayant un rôle important dans la dégradation de la biomasse végétale comme *Phanerochaete chrysosporium* [98], *Trichoderma reesei* [77] et *Aspergillus niger* [99], mais également, ayant des styles de vie différents comme le saprophyte coprophyle *Podospora anserina* [100], le symbionte micorhyze *Laccaria bicolor* [101] et le phytopathogène *Nectria haematococca* [102].

L'examen de ces génomes par le groupe d'annotateurs de CAZy se fait par vérification manuelle. Cette étape précède l'analyse réalisée par le logiciel de détection semi-automatique des CAZymes sur les données au format FASTA constitué des meilleurs modèles protéiques fournis par les centres de séquençage. Ces modèles subissent une première comparaison contre une bibliothèque comprenant l'ensemble des séquences des CAZymes, sur toute leur longueur. Les séquences résultantes sont ensuite comparées aux bibliothèques de modules. Deux rapports sont créés pour chaque séquence identifiée comme CAZyme potentielle. Le premier rapport est issu de l'analyse classique d'annotation par homologie avec BLAST [57]. Le second provient de l'approche CAZyModO [68] qui combine simultanément des résultats BLAST avec ceux obtenus par HMMer (hmmer.janelia.org). En fonction de différents critères, ces résultats sont soit entrés automatiquement dans la base de données, soit laissés pour analyse à l'annotateur qui doit évaluer la séquence ainsi que les résultats associés pour éventuellement les entrer manuellement avec ou sans correction (**Figure 24**). Les entrées automatiques représentent typiquement 50-85% du total des séquences analysées, en fonction de la qualité (la présence de modèles fragmentaires fait chuter le taux d'entrées automatiques) et du degré de proximité avec des résultats déjà présents dans CAZy. Les informations sur les CAZomes (ensemble des CAZymes d'un génome) sont ensuite regroupées et envoyées aux consortia. L'étude des CAZymes identifiées au sein des génomes fongiques permet un profilage glycobiologique.

Par exemple, l'étude du CAZome de *T. reesei* a révélé un nombre très limité d'enzymes ayant des activités permettant la digestion des composants de la paroi des plantes (cellulases, hemicellulases, pectinases) en comparaison à ce qui est trouvé habituellement chez des champignons saprophytes [77].

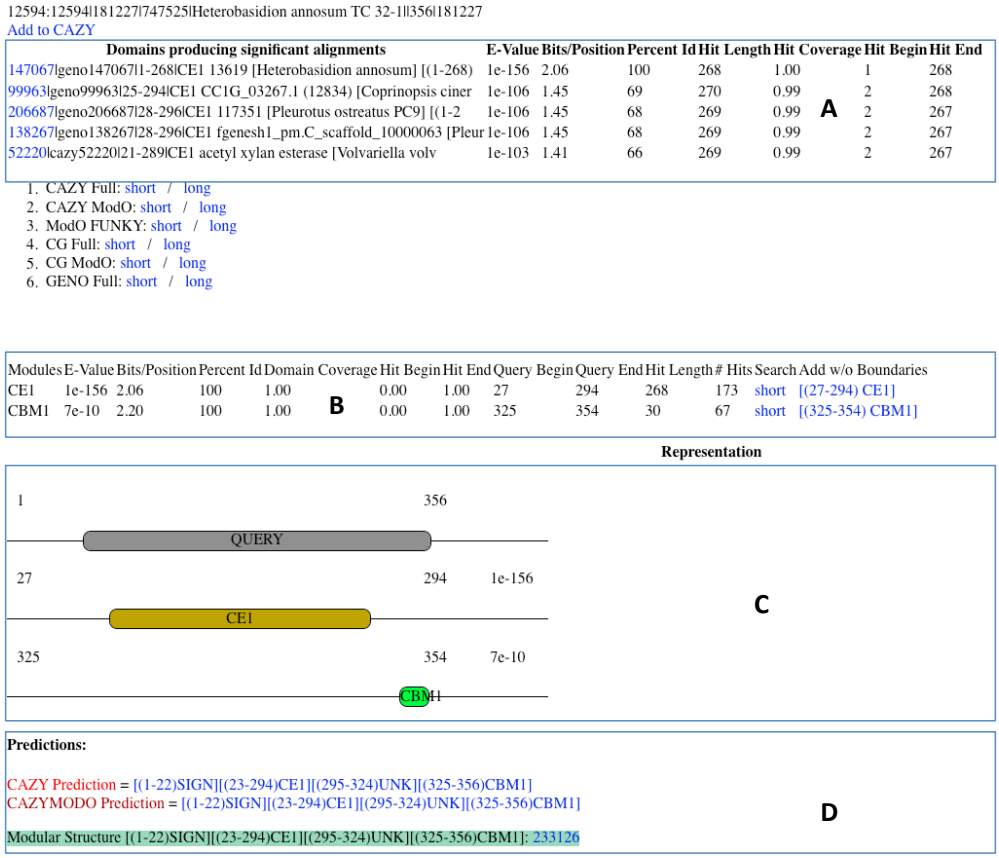


Figure 24: Exemple de rapport d'une CAZyme rentrée automatiquement dans la base de données CAZy provenant de l'analyse du génome fongique *Heterobasidion annosum*. Ce rapport comporte 4 parties : (A) le résultat des meilleurs corrélation de BLAST avec la bibliothèque de modules, (B) la liste des différents types de modules ayant un seuil suffisant, (C) le rapport graphique discernant les limites de chaque module identifié (D) les prédictions de modularité résultantes des différentes approches analytiques.

II.2.2.1.2. L'analyse des CBMs

L'analyse des génomes a permis la mise en évidence de nombreux CBMs parmi une multitude de nouveaux modules. L'identification des CBMs a débuté en 1986 lors de l'étude de l'action limitée de la protéolyse d'une cellobiohydrolase I chez *T. reesei* [103]. Cette étude a montré que parmi les enzymes hydrolytiques responsables de la dégradation de polysaccharides insolubles, certaines pouvaient être constituées d'une unité à activité catalytique et d'un domaine non

catalytique. La fonction de ces modules auxiliaires est l'adhésion sur les chaines glucidiques afin de rapprocher la région catalytique de son substrat. Les sucres insolubles reconnus par de telles enzymes sont multiples et incluent, entre autres, la cellulose, la chitine, le glucane, l'amidon/glycogène, l'inuline, le pullulane et le xylane. Les CBMs rendent ainsi plus efficace l'action des modules catalytiques grâce à leur fonction de proximité ou de ciblage [104, 105]. Ils jouent un rôle important dans la dégradation de la lignocellulose c'est pourquoi mes études se sont dirigées spécifiquement sur les modules de fixation de la cellulose.

II.2.2.1.3. Recherche de fusions avec des CBMs

La recherche *in silico* de nouvelles enzymes s'est portée sur l'identification de toutes les familles de CAZy (GH, PL, CE, X *etc.*) fusionnées avec des CBMs, et tout particulièrement sur les modules appartenant aux familles de X dont les fonctions restent encore à découvrir. Mon travail s'est concentré principalement sur l'étude de deux CBMs correspondant aux familles CBM1 et CBM2 connus pour leur interaction avec les polymères de la paroi végétale et la cellulose en particulier.

Les modules de la famille CBM1, appelés auparavant CBD1 [28, 29], sont essentiellement présents chez les champignons et sont connus pour leur affinité à la cellulose [106] bien qu'ils puissent aussi interagir avec la chitine [107]. Les modules CBM2, appelés auparavant CBD2, sont d'origine principalement bactérienne [108] et sont connus pour leur capacité d'adhésion à la cellulose, au xylane [28, 29] et à la chitine [109]. A l'inverse de la famille CBM1 chez les champignons, il existe chez les bactéries plusieurs familles de CBMs permettant la fixation à la cellulose autres que la famille CBM2. Parmi les plus abondantes, on trouve les familles CBM3, CBM6 et CBM10 [110]. Pour chaque famille de CBMs une analyse visant à identifier de nouvelles enzymes d'intérêt intervenant dans la dégradation de la lignocellulose a suivi les étapes suivantes :

(i) L'élaboration d'une liste de toutes les familles de modules en lien avec ce CBM.

(ii) L'identification des modules dont la famille est commune aux champignons et aux bactéries afin de vérifier leur importance dans la nature.

(iii) L'identification des familles absentes ou sous-représentées chez *Trichoderma reesei* qui est le producteur des cocktails enzymatique de référence.

(iv) L'étude des familles de modules X aux fonctions inconnues ou des régions non connues (UNK) liées aux CBMs. Les modules ou régions inconnues peuvent présenter un intérêt s'ils se retrouvent dans d'autres CAZymes et peuvent peut être jouer un rôle dans la dégradation de la lignocellulose.

Différents outils interfacés avec la base de données ont permis des approches complémentaires. Ces outils sont soit des applications récemment interfacées, soit des logiciels déjà présents lors de la première génération de la base de données CAZy, soit de nouveaux logiciels intégrés dans la nouvelle interface. Les différents outils et approches utilisés sont décrits ci-dessous.

(i) Cytoscape est un outil de visualisation, de modélisation et d'analyse moléculaire et génétique d'interactions au sein d'un réseau [111]. Il permet aux utilisateurs de déterminer et d'analyser les inter-connectivités au sein d'une liste de gènes ou de protéines. Cytoscape a été utilisé pour visualiser graphiquement le réseau de fusions inter-modulaires des CBMs dans la base de données. Cet outil a mis en évidence les corrélations entre modules et permis de regrouper en « *clusters* » les modules qui interagissent entre eux.

(ii) Flymod est un outil de création graphique de modules protéiques basé sur le programme fly (www.w3perl.com/fly) adapté à CAZy. Cet outil génère graphiquement une vue schématique et générale de chaque classe du CAZome de différents organismes et permet ainsi de comparer visuellement un organisme avec un autre.

(iii) CAZyBox [68] est une application résultant de l'intégration de l'ensemble des outils et des étapes nécessaires à la création des sous-familles dans un seul et même logiciel. La CAZyBox

permet de gagner un temps considérable lors de l'assignement en sous-familles et de tester aisément plusieurs hypothèses de subdivisions. Cette application intègre plusieurs outils Java d'analyse de séquences dont Jalview [112] pour visualiser les alignements et ATV [113] pour visualiser les arbres phylogéniques.

(iv) « Compare génome » est un nouveau logiciel interfacé dans la structure de CAZy permettant la comparaison de CAZomes entre différents organismes. Cet outil affiche au format tabulaire le décompte de familles et de classes différentes, et calcule les corrélations les plus probantes entre modules d'une liste de génomes sélectionnée par l'utilisateur.

II.2.2.2. Résultats et discussion

II.2.2.2.1. Interprétation graphique des combinaisons modulaires de CBMs

Une première approche a été d'analyser la catégorie des CBMs et leurs interactions avec d'autres modules en utilisant Cytoscape. La représentation graphique de Cytoscape a permis d'identifier des liens uniques établi via des fusions entre certaines classes de module. A titre d'exemple, nous avons restreint dans ce manuscrit notre analyse aux familles de modules CBM1 et CBM2 (**Figure 25**).

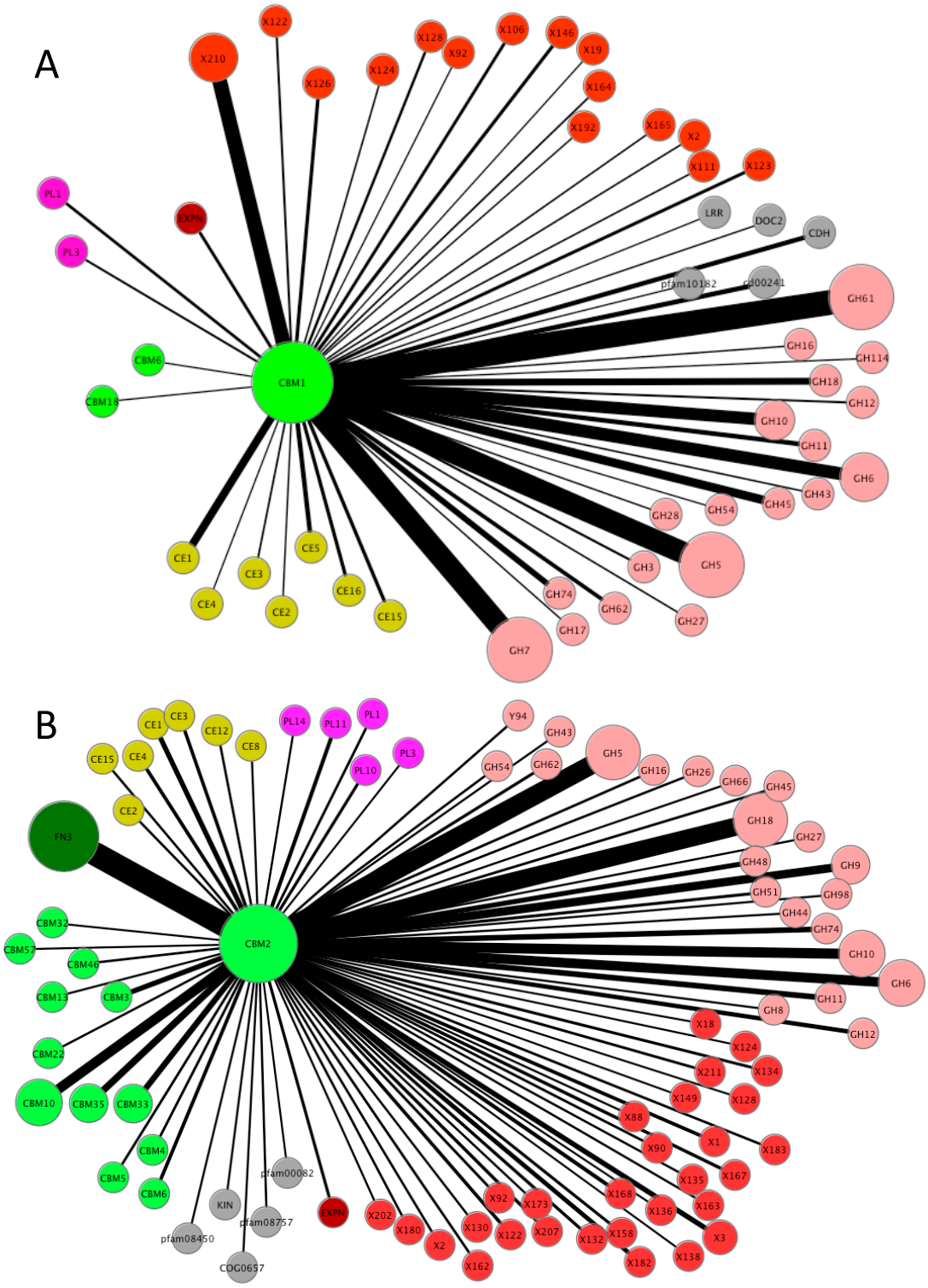


Figure 25: Schéma des familles fusionnées aux modules de la famille CBM1 (A) et CBM2 (B). Cette représentation graphique a été créée avec l'outil Cytoscape [111]. La grosseur des traits est représentative de l'abondance relative de cette combinaison. Les différentes classes de CAZY sont caractérisées par un jeu de couleur : X en rouge, CBMs en vert, CEs en marron, GHs en rose, PLs en magenta et d'autres domaines caractérisés en gris.

Les modules de la famille CBM1 connus pour leur capacité d'adhésion à la cellulose [106], se trouvent fusionnées à une trentaine de modules enzymatiques différents parmi lesquels on identifie 19 familles de GHs, 2 de PLs et 7 de CEs. Les enzymes les plus souvent associées aux modules CBM1 sont des membres des familles GH5 (cellulase, chitosanase, β-mannosidase etc.), GH61 (endoglucanase), GH7 (cellobiohydrolase, chitosanase, endoglucanase), impliqués entre autres dans la dégradation de la cellulose, du xylane et du xyloglucane. Il est intéressant d'observer également des fusions avec certains modules catalytiques comme les carbohydrate déshydrogénases (CDH) ou non catalytiques comme les expansines (EXPN) tous responsables d'activité sur la paroi végétale. Cette étude fait ressortir également un nombre varié de modules X internes à CAZy dont les plus abondants sont ceux de la famille X210 et dans une moindre mesure X123, X126 et X146.

Les modules de la famille CBM2 sont associés à une trentaine de types de modules enzymatiques parmi lesquels on identifie 21 familles de GHs, 5 de PLs et 7 de CEs. Les fusions avec d'autres modules non catalytiques sont plus nombreuses et plus diversifiées par rapport à la famille CBM1. Cette diversité pourrait résulter d'une forte capacité d'adaptation et de recombinaison des bactéries. Les modules enzymatiques associés aux modules CBM2 sont majoritairement les membres des familles GH5 (cellulase, chitosanase, β-mannosidase etc.), GH6 (endoglucanase, cellobiohydrolase), GH10 (endo-β-1,3-xylanase, endo-β-1,4-xylanase) et GH18 (chitinase, endo-β-N-acetylglucosaminidase). Curieusement, on observe une très grande diversité de familles de modules de CBMs fusionnées aux modules CBM2. Cette multiplicité peut être expliquée par un mécanisme de reconnaissance en plusieurs étapes. Tout d'abord, le module CBM2 se fixe à la surface cellulosique puis glisse sur la surface cristalline jusqu'à ce que les autres modules CBMs interagissent avec leur substrat d'adhésion pour permettre l'action catalytique du polypeptide [105]. La **Figure 25**B montre de nombreuses fusions avec les modules X et en particulier avec ceux des familles X3, X134, X182.

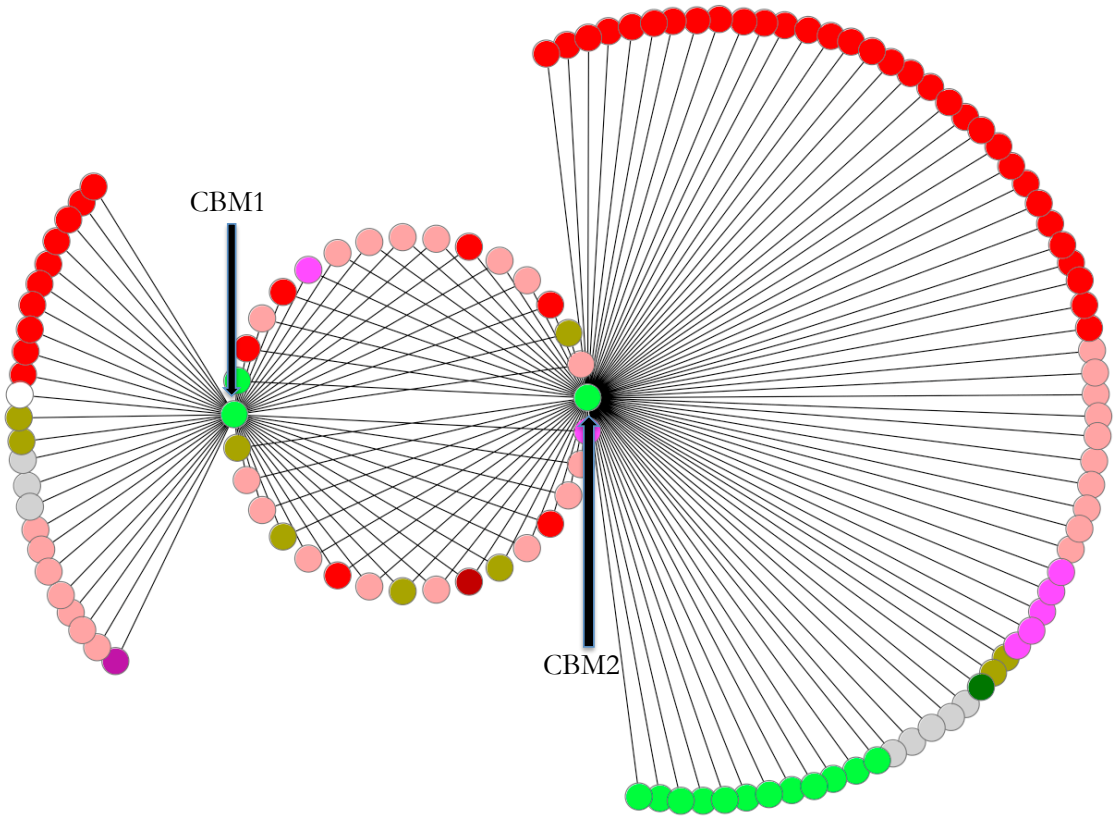


Figure 26: Schéma représentatif des modules liés aux familles de modules CBM1 et CBM2. Les modules situés sur le cercle extérieur sont liés soit au CBM1 soit au CBM2. Le cercle interne représente les modules communs au CBM1 et CBM2 (Avril 2011). Les différentes classes de CAZY sont caractérisées par un jeu de couleur : X en rouge, CBMs en vert, CEs en marron, GHs en rose, PLs en magenta et d'autres domaines caractérisés en gris.

En utilisant le même procédé la représentation des familles de module lies aux CBM1 et CBM2 est possible **Figure 26**. On obtient ainsi facilement la liste des modules communs entre ces deux familles. Finalement, les familles de X résultantes liées aux modules de CBMs et présentant un certain intérêt dans la recherche de nouvelles enzymes, sont regroupées dans le **Tableau 1** ci-dessous.

Tableau 1: Récapitulatif des familles de modules X fusionnées avec des modules CBM1 ou CBM2 (mai 2010). Les familles de X en rouge sont les familles communes avec les modules CBM1 et CBM2.

Module	Famille de module X
CBM1	X2, X19, X92, X106, X111, X122 X123, X124, X126, X128, X146, X164, X165, X192, X210
CBM2	X1, X2, X3, X18, X88, X90, X92 X122 X124 X128, X130, X132, X134, X135, X136, X138, X149, X158, X162, X163, X167, X168, X173, X180, X182, X183, X202, X207, X211

Les résultats de cette étude font ressortir des différences au niveau des combinaisons de modules fusionnés selon la famille analysée. A une échelle plus générale, toutes les combinaisons de modules portant un CBM ont été analysées (**Figure 27**). L'observation de la multiplicité des combinaisons existantes dans tous les organismes de la base de données révèle les modules spécifiques de chaque famille dans CAZy. Ces modules spécifiques se trouvent à l'extérieur du cercle principal de la **Figure 27**. La comparaison entre la famille CBM1 chez les champignons et les modules CBMs de fixation à la cellulose chez les bactéries montre un nombre total de modules spécifiques plus important chez ces derniers.

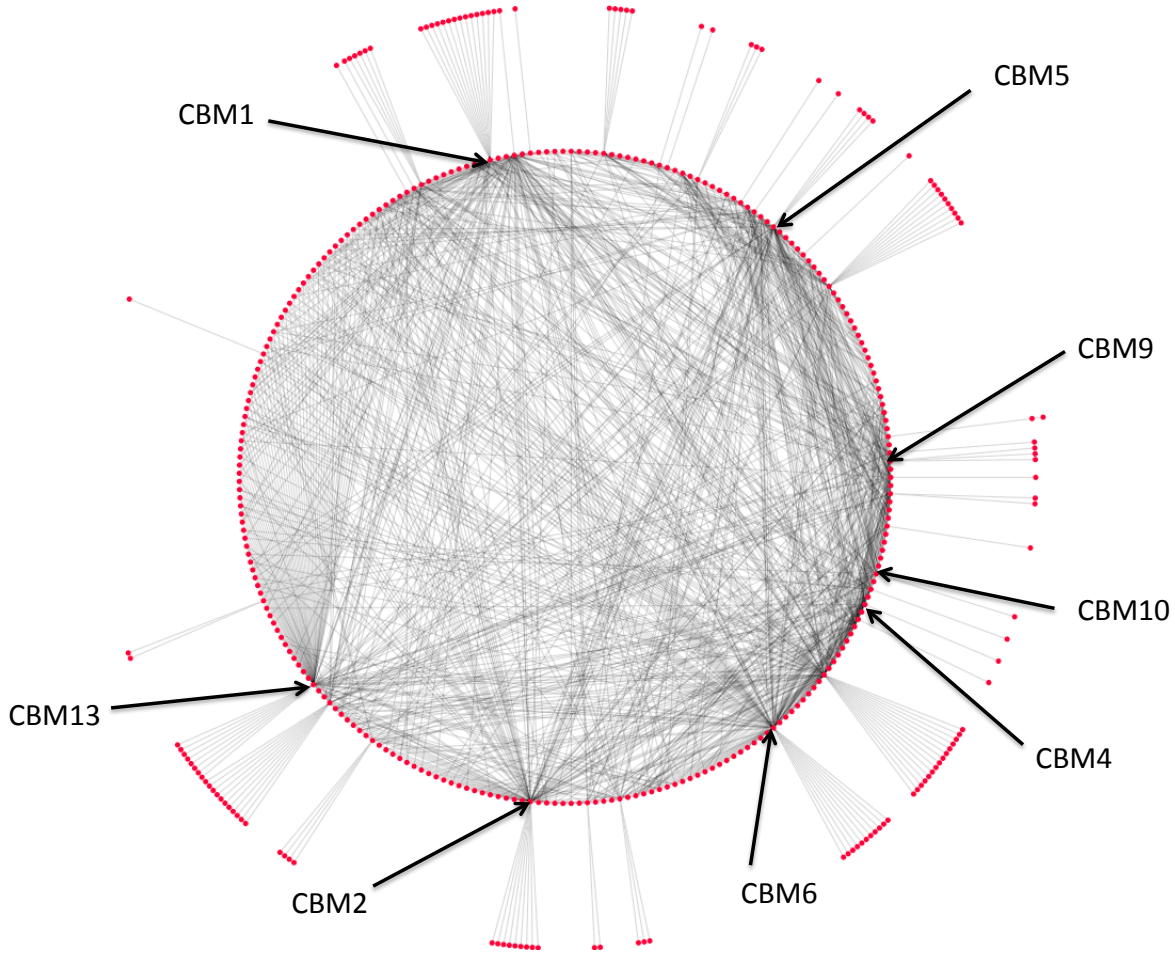


Figure 27: Schéma représentatif de toutes les combinaisons modulaires de CAZymes existantes dans tous les organismes de CAZy. A part la familles CBM1 qui est fongique toutes les autres familles sont bactériennes. Les traits noirs représentent les liaisons entre modules

II.2.2.2.2. Interprétation graphique des combinaisons modulaires au sein des génomes

L'approche basée sur l'affichage systématique de toutes les fusions identifiées peut être complétée par l'analyse de comparaison de génomes. L'outil « Flymod » permet la visualisation de la modularité des CAZymes présents dans une sélection de génomes de différents organismes. A titre d'exemple, on peut observer les ensembles de modules fusionnés aux CBM1 dans différents génomes fongiques (**Figure 28**) afin de localiser des régions X et UNK qui peuvent être intéressantes pour la recherche de nouvelles enzymes. On trouve logiquement une faible quantité d'activités enzymatiques fusionnées au CBM1 chez *Fusarium oxysparum* car c'est un pathogène qui cible la pectine et l'hémicellulose plutôt que la cellulose [97].

Inversement chez *P. anserina* on recense un nombre élevé de modules CBM1 car cet organisme coprophile intervient sur les restes végétaux non digérés riches en xylane et en cellulose cristalline [100].

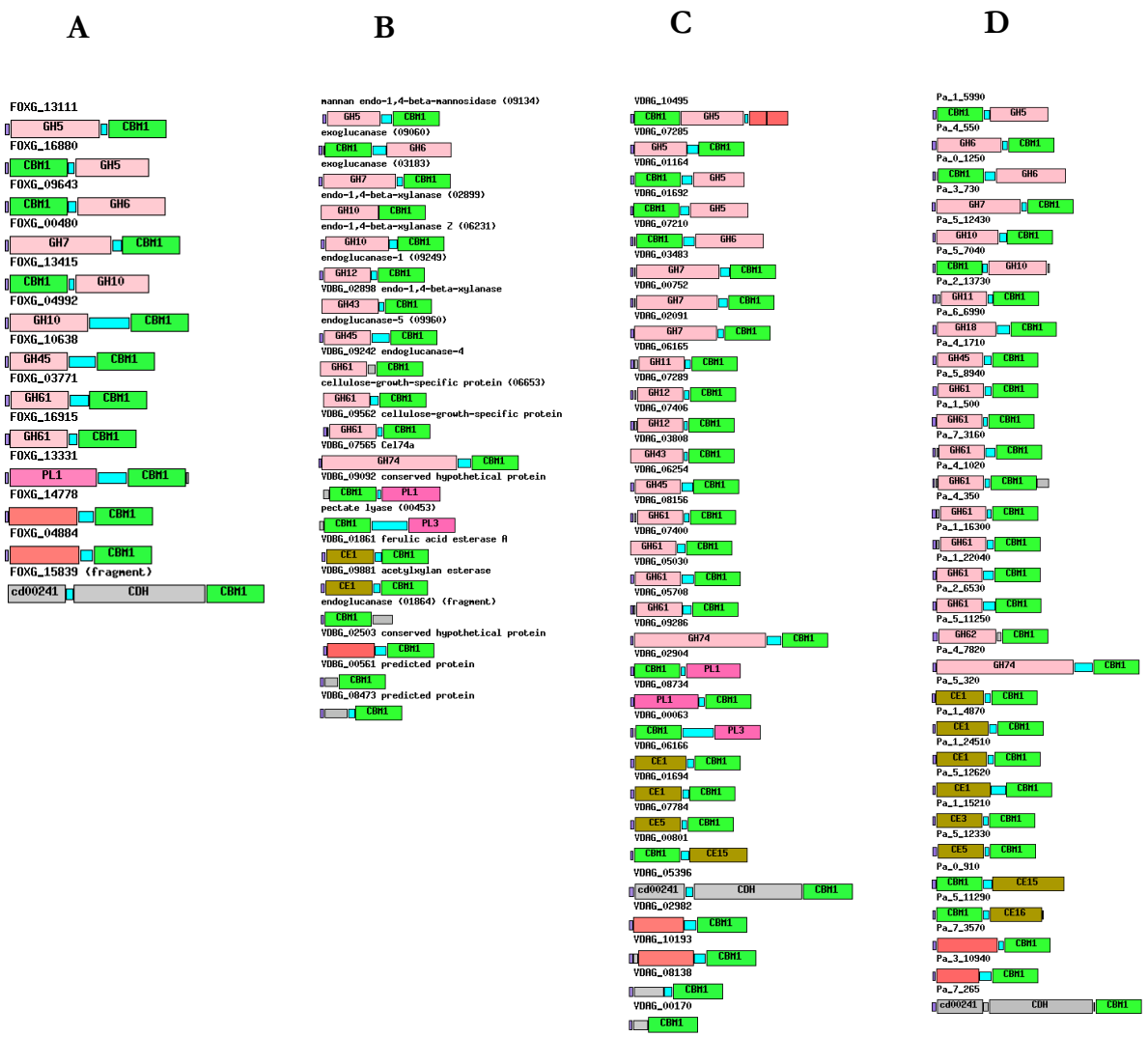


Figure 28: Représentation graphique des protéines de la famille CBM1 (rectangle vert) chez différents organismes. (A) *Fusarium oxysporum*, (B) *Verticillium albo-atrum*, (C) *Verticillium dahliae*, (D) *Podospora anserina*. Les couleurs des différentes descriptions modulaires sont : GHs (rose clair), PLs (rose foncées), CEs (marron), Xs (rouge), linkers (bleu clair), signaux peptidiques (violet), les autres régions (gris).

II.2.2.2.3. Les comparaisons de génomes

L'utilisation des outils graphiques Cytoscape et Flymod pour la visualisation de toutes les combinaisons de fusions permet de nombreuses interprétations. Ce type d'analyse peut servir d'ébauche à des études plus approfondies.

L'interface « Compare génome » permet une vue d'ensemble sous la forme tabulaire d'un certain nombre de génomes sélectionnés. Les vues résultantes permettent non seulement l'identification de la majorité des acteurs du métabolisme des glucides d'un organisme, mais

donnent aussi des pistes sur les grandes tendances de ce métabolisme en favorisant certains composés ou classes de composés [114]. Bien qu'utiles pour l'analyse de génomes uniques, ces approches révèlent toute leur puissance lors de la comparaison de données issues d'organismes ADDIN EN.CITE.DATA [115], en dévoilant les faiblesses et les atouts enzymatiques de **Tableau 2** est une représentation sous forme tabulaire de la **Figure 28** donnant l'information par organisme du nombre de familles fusionnées au CBM1.

Tableau 2: Table de comparaison de 4 génomes de champignons. La première ligne indique les modules liés au CBM1 sur différentes protéines. La table montre le décompte de ces modules dans 4 espèces de champignon différents. Le gradient de couleur permet de visualiser la pauvreté (vert) ou la richesse (rouge) des modules en lien avec les CBM1 selon les espèces.

CBM1	cd00241	CDH	CE1	CE15	CE16	CE3	CE5	GH10	GH11	GH12	GH18	GH43	GH45	GH5	GH6	GH61	GH62	GH7	GH74	PL1	PL3	X123	X126	X128	X192	X2
Fusarium oxysporum	0	1	0	0	0	0	0	2	0	0	0	0	1	2	1	2	0	1	0	1	0	0	0	2	0	0
Verticillium albo-atrum	0	0	2	0	0	0	0	2	0	1	0	1	1	1	1	3	0	1	1	1	1	0	1	0	0	0
Podospora anserina	1	1	4	1	1	1	1	2	1	0	1	0	1	1	2	8	1	1	1	0	0	1	1	0	0	0
Verticillium dahliae	1	1	2	1	0	0	1	0	1	2	0	1	1	4	1	4	0	3	1	2	1	0	1	1	1	1

L'avantage de ce type de représentation est de donner une vue condensée des données par famille et de permettre l'affichage de ces informations chez un grand nombre de génomes. Cet exemple fait ressortir un grand nombre de modules de la famille GH61 qui peuvent intervenir dans la dégradation de la cellulose [83] chez les champignons et surtout chez *P. anserina*. Cette vue permet aussi de montrer les modules communs chez toutes les espèces analysées (*e.g.* GH45, GH61) ainsi que les familles spécifiques à un seul organisme comme la famille GH62 chez *P. anserina*.

L'outil « compare génome » a permis de regrouper les informations de plusieurs génomes fongiques afin de comparer et d'identifier rapidement les familles qui sont absentes ou sous-représentées chez *T. reesei* (**Tableau 3**).

Tableau 3: Comparaison entre *T. reesei* et 56 autres champignons des familles fusionnées soit au CBM1 soit au CBM2 soit aux deux. Les cases grisées représentent les familles que l'on trouve liées à la fois dans la famille CBM1 et CBM2

domaine fusionné au CBM1	domaine fusionné au CBM2	nombre chez Trichoderma reesei	nombre moyen chez 56 champignons
GH3		13	12.9
GH5	GH5	11	14.2
GH6	GH6	1	1.4
GH7		2	3.2
	GH8	0	0.1
	GH9	0	0.3
GH10	GH10	1	3
GH11	GH11	4	2.5
GH12	GH12	2	2.7
GH16	GH16	16	15.9
GH17		4	4.3
GH18	GH18	20	13.7
	GH26	0	0.7
GH27	GH27	8	2.9
GH28		4	7.3
GH43	GH43	2	10.2
	GH44	0	0
GH45	GH45	1	1.1
	GH48	0	0
	GH51	0	1.9
GH54	GH54	2	0.6
GH61		3	11.3
GH62	GH62	1	1.1
	GH66	0	0
GH74	GH74	1	0.7
	GH98	0	0
GH114		0	0.9

Le **Tableau 3** met en évidence des familles de modules de GHs spécifiques des bactéries et des champignons mais également ceux communs aux deux groupes d'organismes. De plus, afin de comparer la composition en familles de GHs chez *T. reesei* une moyenne de ces familles de modules est calculée chez 56 espèces fongiques. Tout d'abord, on constate qu'il existe une certaine homogénéité entre le nombre de GHs fusionnées avec le CBM1 de *T. reesei* et le nombre moyen de GHs correspondant chez les autres champignons. Cependant quelques exceptions sont à noter. Par exemple, les familles GH18 et GH27 semblent très répandues chez *T. reesei.*

L'abondance relative des modules de la famille GH18 (majoritairement connue pour son activité chitinase) peut être expliquée par une capacité accrue chez *T. reesei* de dégradation de la paroi fongique riche en chitine. En effet plusieurs espèces de *Trichoderma* sont des mycoparasites symbiotiques des plantes et leur activité chitinolytique jouerait ainsi le rôle de protecteur contre l'agression d'autres champignons sur les plantes colonisées [116, 117]. Inversement les familles GH10 (ayant des activités endo-β-1,4-xylanase et endo-β-1,3-xylanase) et GH43 (ayant des activités β-xylosidase, arabinanase et arabinofuranosidase) sont moins représentées chez *T. reesei.* L'addition d'enzymes appartenant à ces familles pourrait représenter un intérêt pour l'amélioration des cocktails enzymatiques industriels produits par *T. reesei.*

Dans ce même tableau les familles de GHs spécifiques aux champignons ou aux bactéries sont également prises en considération. Les familles GH26 (β-1,3-xylanase, β-mannanase), GH51 (endoglucanase), GH114 (endo-α-1,4-polygalactosaminidase) présentent un grand intérêt car ils sont absents chez *T. reesei.* Ces familles peuvent être des acteurs importants dans la dégradation de la paroi végétale et des candidats jouant sur l'amélioration de *T. reesei* pour les biocarburants. Cette amélioration a pu être démontrée en ajoutant une β-mannanase de la famille GH26 de *P. anserina* au cocktail industriel issu de *T. reesei* [118].

Au final, cette étude peut être appliquée à toutes les classes de modules CAZy et en particulier aux familles de X inconnues afin de trouver de nouvelles enzymes d'intérêts.

II.2.2.2.4. Etudes phylogénétiques des familles de modules X

Une dernière approche plus fine a été accomplie grâce à l'outil « CAZyBox » en analysant les familles de CAZy et leurs arbres phylogéniques. Après avoir repéré tous les domaines X ou UNK liés aux CBMs (l'étude ne se restreint plus au CBM1 ou CBM2) impliqués dans la fixation à la cellulose, ces différentes séquences sont récupérées dans la base de données en utilisant la CAZyBox. Apres avoir été alignées, un arbre phylogénique est construit à partir d'une matrice de similarité (typiquement BLOSUM62) choisie. L'étude des différentes familles de modules X a

révélé pour certaines un possible intérêt. Cette approche est illustrée par l'étude représentative de la famille de modules inconnus X131. Cette famille se trouve essentiellement dans les génomes fongiques mais également dans un petit groupe de bactéries (**Figure 29**).

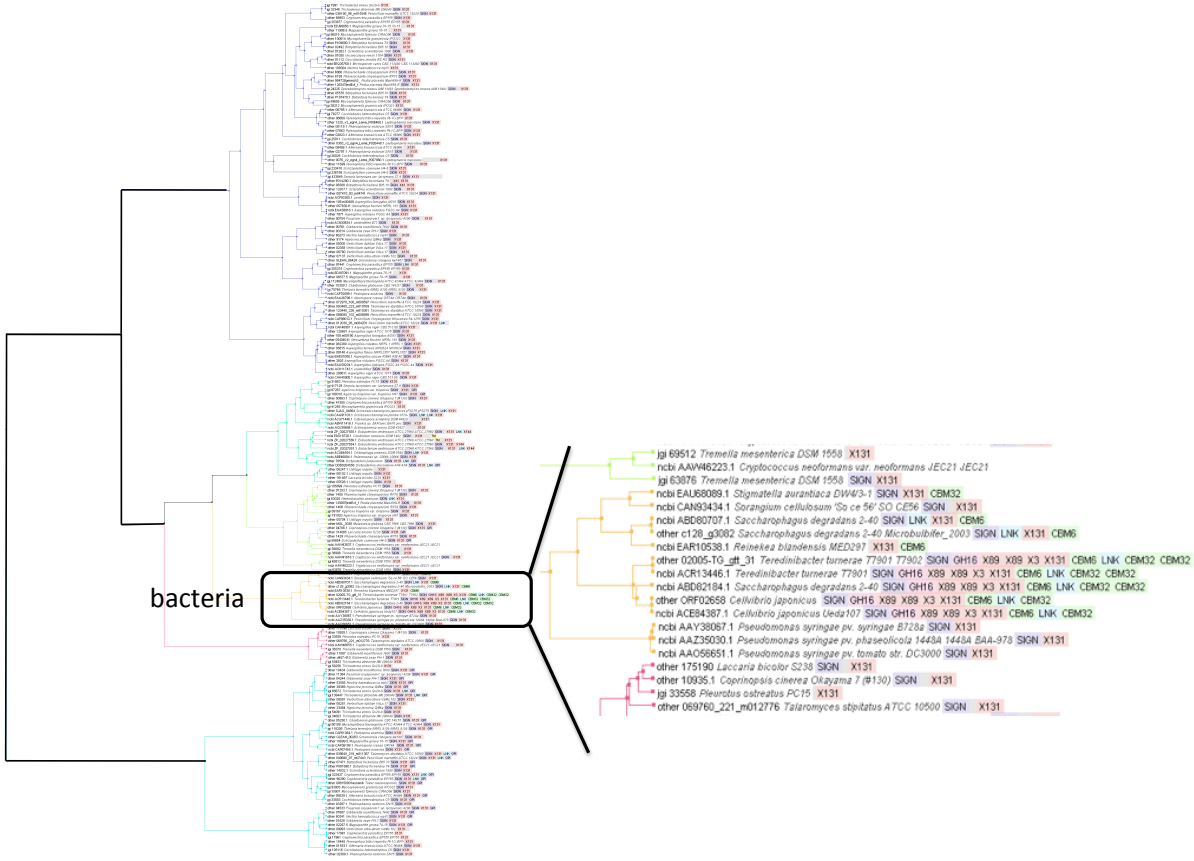


Figure 29: Arbre phylogénétique regroupant l'ensemble des séquences possédant le module X131. Toutes les séquences de cet arbre sont fongiques à l'exception des protéines contenues dans l'encadré noir qui sont bactériennes

Toutes les protéines possédant un module de la famille X131 sont secrétées car elles possèdent un signal peptidique. De nombreuses protéines fongiques portent une ancre glycosylphosphatidylinositol (GPI) qui permet l'accrochage à la paroi cellulaire [119]. De plus, les protéines bactériennes possèdent pour la plupart un module CBM6 qui permet l'adhésion au xylane et au β-glucane [104]. La famille X131 comprend plus de 200 membres et contient de nombreuses sous-familles distinctes. En approfondissant l'analyse de l'alignement et en analysant les régions conservées, on a pu identifier la machinerie catalytique et stéréochimique de coupure car des similarités lointaines avec des membres du clan GH-A ont été retrouvées. X131 est

finalement une famille possédant un domaine catalytique du clan GH-A qui regroupe les familles : GH1, GH2, GH5, GH10, GH17, GH26, GH30, GH35, GH39, GH42, GH44, GH50, GH51, GH53, GH59, GH72, GH79, GH86, GH113. Sachant que les membres de se clan dégradent les liaisons β-glycosidiques [120] et que X131 se lie à des modules de type CBM6 une réduction des substrats à analyser est ainsi possible. De plus le fait que beaucoup de séquences fongiques portent une ancre GPI laisse suggérer que ce module peut dégrader un composant de la paroi fongique ou des composants proches.

II.2.2.3. Conclusion

L'utilisation des nouveaux outils de CAZy a contribué à différentes stratégies d'analyse de génomes fongiques et bactériens en fonctions des familles de CBMs et a permis non seulement d'analyser les informations de la base de données mais aussi de tester la qualité de ces informations. La recherche de nouvelles enzymes a constitué un bon exercice pour tester la robustesse de la nouvelle structure ainsi que l'exactitude des données enzymatiques de CAZy. Enfin, l'approche utilisée, résumée dans la **Figure 30**, a permis de restreindre à 0,1% le champ de données des enzymes d'études. Cette analyse a suggéré de nouvelles pistes sur les activités impliquées dans la dégradation de la biomasse végétale grâce à la découverte de familles de module X fusionnées aux modules de CBMs. Les résultats obtenus *in silico* restent encore à être testés par la suite par Novozymes.

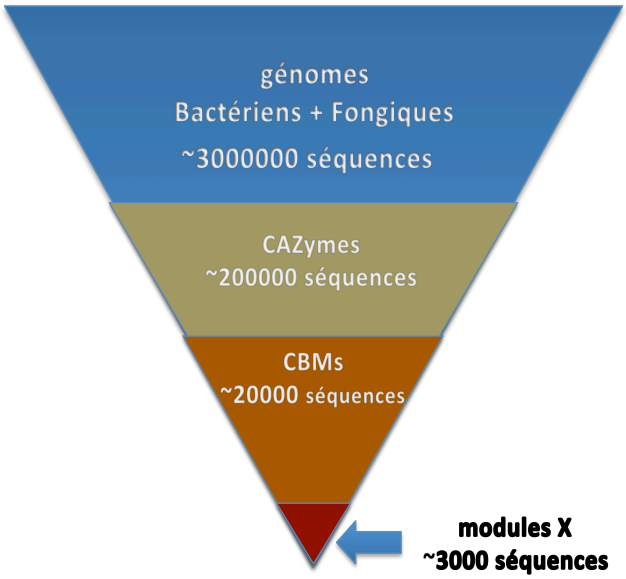


Figure 30: Récapitulatif de la stratégie de recherche de nouvelles enzymes. Le nombre de séquences est représentatif des informations de la base de données en 2011. A partir d'un grand nombre de séquences génomique, la recherche de nouvelles enzymes revient à chercher les CAZymes et plus précisément les séquences possédant des CBMs de fixation à la cellulose. L'analyse, par la suite, est restreinte à un nombre de séquence moins important lors de l'étude de modules inconnues (X) fusionnés à un CBM d'intérêt sur le même polypeptide.

II.2.3. Automatisation de l'analyse des métagénomes

II.2.3.1. Vers une nouvelle ère

L'arrivée de la génomique a totalement changé le panorama des CAZymes. En effet, les séquences associées à des génomes correspondent approximativement à 95% du contenu actuel de CAZy. Historiquement, les enzymes issues d'efforts génomiques n'étaient pas caractérisées ce qui de nos jours n'est plus le cas. En effet, certains projets génomiques sont maintenant initiés sur des organimes dont certaines enzymes ont été caractérisées. Une vue globale des CAZymes de chaque génome, couvrant l'essentiel des familles d'enzymes impliquées dans la biosynthèse, le remaniement et la dégradation des glycanes, est rendue possible grâce à la nouvelle interface (voir chapitre précédent). Plus récemment, un nombre croissant de séquences caractérisées a résulté d'efforts métagénomiques, alors qu'auparavant ces séquences provenaient essentiellement de génomes connus [121].

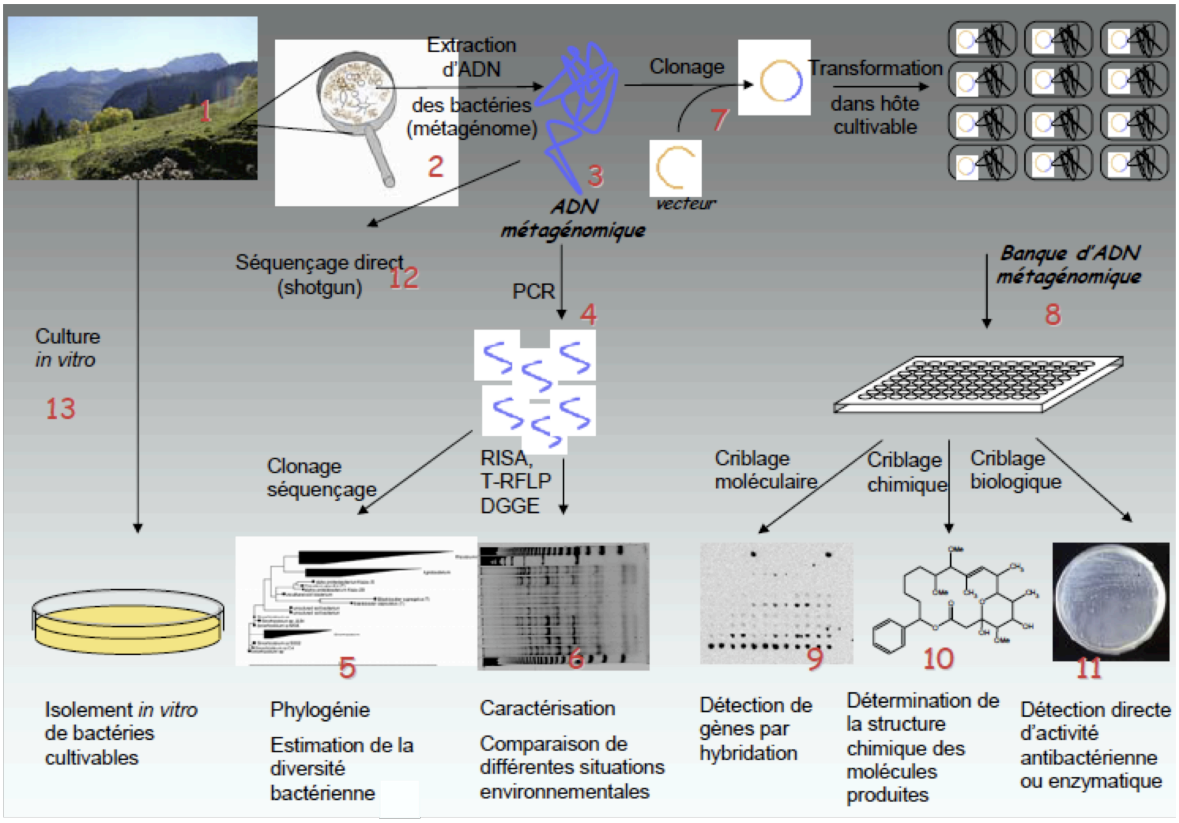


Figure 31: Schéma général des approches métagénomiques: A partir d'un échantillon prélevé dans l'environnement (1) (sol, eau, etc.) des traitements sont appliqués afin d'extraire et purifier l'ADN des microorganismes présents (2). Il existe plusieurs possibilités de traiter l'ADN « métagénomique » extrait (3). Une des méthodes parmi les plus utilisées est d'effectuer une étape d'amplification PCR (4), les produits ainsi générés pouvant être clonés et séquencés offrant par exemple la possibilité d'estimer la diversité bactérienne par génération d'arbres phylogénétiques (5). Une autre approche consiste à utiliser des méthodes de « *fingerprinting* » qui produisent des profils de bandes rendant compte de la diversité bactérienne au sein de l'échantillon analysé (6). La seconde approche consiste à cloner directement l'ADN métagénomique extrait (7) dans un hôte bactérien domestiqué comme E. coli permettant de générer des banques d'ADN métagénomique. Il existe au moins 3 méthodes pour cribler ces banques, de façon moléculaire par hybridation après que l'ADN ait été extrait et déposé sur membrane (9) mais aussi en analysant chimiquement le surnageant de culture de façon à identifier les composés produits suite à l'expression de l'insert (10). Enfin, une dernière méthode consiste à étaler les clones métagénomiques sur un milieu sélectif ne révélant ainsi que ceux dont l'expression des gènes de l'insert permet la croissance ou la révélation d'activités biochimiques d'intérêt (11). Dans ces deux dernières méthodes, l'expression des gènes métagénomiques est nécessaire requérant un changement d'hôtes d'expression de façon à cribler un nombre plus élevé d'inserts provenant de bactéries phylogénétiquements éloignées. La dernière approche d'exploitation de l'ADN métagénomique implique son séquençage direct par shotgun (12), celle-ci nécessite d'énormes capacités de séquençage pour fournir des résultats exploitables. Ces différentes approches de métagénomique doivent compléter l'approche traditionnelle de culture in vitro (13). La plupart des analyses métagénomique de l'équipe Glycogénomique ont été réalisées sur des échantillons soumis au séquençage direct par shotgun.

La métagénomique est une approche nouvelle et puissante, qui permet d'analyser les génomes de l'ensemble des microorganismes d'une niche écologique sans étape de culture [122]. Ces microorganismes sont souvent inconnus car non-cultivables ou non cultivés. Une étude métagénomique consiste en l'extraction d'une population microbienne de sa niche écologique, suivie de la purification et du séquençage de son ADN par les méthodes à haut débit (**Figure 31**). Deux types d'approches distinctes d'analyse métagénomique existent actuellement. La première est une approche orientée car elle permet de cribler un échantillon par rapport à une fonction particulière (*e.g.* activité enzymatique ou antibiotique) [123]. Cette analyse à l'avantage d'avoir pour résultat l'identification de gènes ou de groupes de gènes complets associés directement ou indirectement à la fonction désirée. Elle permet de détecter des nouveaux gènes ayant une fonction importante impliquée par exemple dans la virulence ou ayant un intérêt industriel. La seconde approche métagénomique est moins spécifique car l'échantillon biologique correspondant à un microenvironnement donné est tout simplement séquencé dans sa totalité sans tenir compte des fonctions potentielles. Pour des systèmes où le niveau de couverture est élevé, il est parfois possible de reconstituer le génome entier de certains microorganismes, ce qui peut apporter une vue plus riche de leur potentiel biologique. L'intérêt de cette approche réside dans la possibilité d'étudier des génomes complets des microorganismes particulièrement abondants dans un environnement donné [124].

Le groupe de « Glycogénomique » est de plus en plus sollicité pour effectuer l'analyse des données issues de projets métagénomiques. Une majorité de ces projets s'orientent sur la recherche et l'analyse de l'ensemble des séquences d'enzymes dégradant les glucides dans un échantillon biologique donné. Typiquement, ces projets visent soit une identification taxonomique de la communauté présente, soit une estimation du spectre et emphase fonctionnelle des activités enzymatiques identifiées ou estimées, soit les deux simultanément. Ces communautés ou métagénomes peuvent correspondre à des niches microbiennes localisées dans

la nature comme des échantillons collectés dans les océans, le sol ou au sein de macro organismes comme l'être humain [125]. La particularité des approches métagénomiques des CAZymes réside dans leur pouvoir à dévoiler l'étendue du métabolisme des sucres sans même donner de détail dû au manque de données biochimiques de référence pour beaucoup de familles.

II.2.3.2. Vers une l'automatisation des analyses CAZy

Le volume de données issu d'un échantillon métagénomique peut varier de quelques milliers de séquences à plusieurs millions et leurs longueurs peuvent être courtes ou assez longues pour couvrir la taille d'un gène. Du fait de la quantité importante d'informations, l'analyse ne peut donc plus être soumise au même traitement que les séquences des mises à jour du NCBI ou celui effectué pour un simple génome. La curation semi-automatique de séquences décrite dans les chapitres précédents devient un exercice pratiquement impossible. Pour surmonter ces difficultés, un *pipeline* automatique de reconnaissance des CAZymes a été créé. Afin de réduire le temps de calcul sur grand volume de données, il est impératif d'avoir accès à des serveurs possédant un nombre élevé de processeurs (CPU). Les calculs de l'analyse métagénomique décrite ci-dessous ont été effectués grâce à un *cluster* que le laboratoire AFMB héberge. Ce *cluster* comprend 32 processeurs à 4 cœurs groupés dans 16 serveurs parallélisés possédant chacun 24 Go de mémoire vive soit au total 128 CPUs. Chacun des cœurs peut traiter simultanément deux processus ("*hyperthreading*"). L'analyse de chaque échantillon est réalisée en deux étapes :

(i) Utilisation du programme BLASTX [57] sur le *cluster* afin de comparer chacune des séquences nucléotidiques du métagénome contre des bibliothèques non redondantes de séquences issues de CAZy en utilisant BLOSUM62 comme matrice de substitution. Comme l'échantillon étudié contient souvent des séquences bactériennes, l'analyse se fait en utilisant une bibliothèque composée exclusivement des CAZymes de procaryotes afin de rendre les recherches plus rapides. Seules les séquences présentant une *E-value* inférieure ou meilleure que 10^{-6} sont conservées.

(ii) Les séquences conservées identifiées lors de la première étape sont analysées par le programme FASTX [126] contre des librairies de modules issues de CAZy. FASTX est un sous programme du logiciel FASTA [58] capable de comparer des séquences d'ADN contre des bibliothèques de séquences protéiques, en examinant tous les cadres de lecture et permettant des permutations de cadre, maximisant la longueur des alignements résultants. Là encore seules les protéines ayant une *E-value* inférieure à 10^{-6} sont conservées. Cette approche permet de garder les CAZymes ayant une bonne similarité avec les protéines de la base de données et d'éviter les faux positifs dus aux structures modulaires variables des CAZymes.

Cette approche a été testée sur des données issues de différentes études publiées tel que le métagénome digestif du termite [127] et du kangourou nain, ou *w*allaby de Tammar (*Macropus eugenii*) [128]. L'analyse de ces deux études a abouti à des conclusions similaires. Par conséquent, seule l'analyse des données issues de l'étude du *Wallaby* de Tammar sera décrite ici (**Tableau 4**). Les données du projet *M. eugenii* comprenaient 53388 fragments d'ADN à analyser. Ces fragments d'une taille moyenne de 990 nucléotides ont été comparés contre CAZy en utilisant les étapes d'analyse bioinformatique précitées.

Tableau 4: Comparaison des résultats obtenus lors de l'étude du métagénome du *wallaby* de Tammar effectué dans l'article Pope *et al.* [128] et ceux obtenus par analyse automatique contre CAZy. Les différentes catégories de familles d'enzymes sont celles présentes dans l'article de référence. Ces catégories ont été maintenues pour comparaison. Une partie des résultats obtenus par le *pipeline* automatique a été confirmée manuellement (en rouge).

Catégorie		CAZy	Pope et al.
"Cellulases"			
	GH5	57	10
	GH6	0	0
	GH7	0	0
	GH9	12	0
	GH44	0	0
	GH45	0	0
	GH48	0	0
	Total	**69**	**10**
"Endohemicellulases"			
	GH8	7	1
	GH10	27	11
	GH11	0	0
	GH12	0	0
	GH26	21	5
	GH28	32	2
	GH53	14	9
	Total	**101**	**28**
"Debranching enzymes"			
	GH51	31	12
	GH54	0	0
	GH62	0	0
	GH67	9	5
	GH78	62	25
	Total	**102**	**42**
"Oligosaccharide-degrading"			
	GH1	84	61
	GH2	121	24
	GH3	146	72
	GH29	7	2
	GH35	14	3
	GH38	9	3
	GH39	7	1
	GH42	18	8
	GH43	127	10
	GH52	0	0
	Total	**533**	**184**

Catégorie	CAZy (cont.)	
Other CAZymes		
	GH4	5
	GH13	227
	GH15	1
	GH16	14
	GH18	14
	GH20	15
	GH23	29
	GH24	20
	GH25	45
	GH27	9
	GH30	10
	GH31	39
	GH32	30
	GH33	7
	GH36	35
	GH37	1
	GH55	4
	GH57	9
	GH59	1
	GH63	1
	GH64	1
	GH65	5
	GH73	34
	GH74	1
	GH76	1
	GH77	48
	GH81	1
	GH84	3
	GH88	8
	GH89	5
	GH92	38
	GH94	46
	GH95	10
	GH97	41
	GH98	1
	GH103	2
	GH104	1
	GH105	22
	GH106	11
	GH108	2
	GH109	9
	GH111	2
	GH112	1
	GH113	2
	GH114	1
	GH115	23
	GH117	1
	Total	**836**

Les résultats obtenus lors de l'analyse du métagénome du *wallaby* ont montré une grande différence entre le nombre de gènes identifiés par famille trouvé dans l'article de Pope *et al.* [128] et les résultats obtenus avec notre *pipeline* automatique. En effet, le nombre total de GHs identifiées en utilisant les bibliothèque de CAZy est approximativement quatre fois plus importante par rapport à l'analyse de Pope *et al.*. Afin de valider les résultats de l'analyse automatisée, la centaine de séquences des familles de CAZy attribuées à la catégorie « endohemicellulase » a été vérifié manuellement par BLASTX contre la bibliothèque de modules de CAZy (voir **Tableau 4**). Cette vérification a révélé l'exactitude à 100% des données du *pipeline* automatique. Afin d'interpréter les différences observées, il est important de comprendre la stratégie d'analyse originale d'identification des GHs de Pope *et al.* [128]. Leur analyse a consisté à rechercher parmi les séquences du métagénome des motifs de profiles HMM issus la base de données Pfam [65] correspondants aux familles de CAZy suivant l'approche utilisée pour étudier le métagénome de termite [127]. Pour les familles sans motif Pfam identifiés, une analyse BLAST avec des séquences d'enzymes caractérisées issues de CAZy a complété les résultats. L'analyse basée sur les motifs Pfam semble peu adaptée car basée sur des motifs ne couvrant pas forcement la totalité des familles dans CAZy. De plus, la recherche HMM peut donner des scores peu élevés lors de l'analyse de séquences courtes ou fragmentaires issue des données métagénomiques. Enfin, les analyses BLAST complémentaires permettant d'élargir la couverture des recherches à l'ensemble des familles de GHs, n'ont été réalisées que sur un échantillon de séquences caractérisées non représentatif de l'entièreté de ces familles. La qualité de l'approche automatique est intimement liée à l'analyse FASTX [126] couplée aux bibliothèques exhaustives issues de CAZy car elles permettent d'avoir des résultats significatifs plus nombreux.

Un *pipeline* réalisé sur un principe similaire d'authentification automatique de CAZymes basé sur des analyses BLAST et HMM (pour une annalyse modulaire plus fine) a été également testé sur des génomes en donnant des résultats concluant avec 98% d'entrées confirmées manuellement par les curateurs de CAZy comme correctement prédites.

III. Conclusions et Perspectives

Le travail effectué durant cette thèse a permis de développer le support logistique nécessaire à l'étude et à la recherche des CAZymes tout en améliorant la base de donnée CAZy et son interface. La base de données a connu une refonte complète en prenant en considération les problèmes de l'ancienne base et les demandes des utilisateurs de CAZy pour les projets de recherche à venir. L'interface de la base de données a été conçue et réalisée entièrement indépendamment d'outils « open source » tels que PhPMyadmin. L'interface permet maintenant de faire des recherches approfondies à différents niveaux comme la recherche de sous-familles ou l'étude et la comparaison de génomes. Elle a déjà prouvé son efficacité en permettant à l'équipe de Glycogénomique de participer à de nombreux projets. L'implémentation d'approches prédictives de spécificité fine des CAZymes à partir de la seule séquence est donc désormais envisageable, la plupart des outils et des canevas méthodologiques nécessaires à cette tâche étant dorénavant disponibles. Ces outils seront utilisés pour l'étude de nouveaux génomes fongiques ce qui pourra contribuer entre autres à améliorer les cocktails enzymatiques actuellement utilisés dans la dégradation et la saccharification de la biomasse végétale. Mon implication dans le développement de la production de biocarburants de deuxième génération à partir de lignocellulose est toujours d'actualité et devrait se poursuivre sur l'analyse de nouvelles familles de X découvertes lors d'études génomiques et métagénomiques récentes.

Cependant, l'augmentation exponentielle des données va augmenter la charge de travail manuel engendrant un effort considérable de la part du curateur pour vérifier les entrées. A court terme, il sera donc nécessaire de rendre les analyses génomiques automatiques afin de délester le travail de vérification des curateurs. Cette expertise sera moins fine et approfondie qu'une analyse semi-automatique mais elle permettra des vues générales qui seront suffisamment fiables pour la comparaison des génomes. Ces données automatiques devront être prises en considération dans

la base de données CAZy ce qui engendrera une nouvelle logistique avec la création de nouvelles tables dans le schéma relationnel. L'exercice d'automatisation de l'analyse des génomes permettra aussi d'aborder différemment les recherches automatiques de CAZymes issues de données métagénomiques.

En effet, la glycogénomique pourra jouer un rôle prépondérant en apportant une vision globale lors de l'étude des données issue de séquençage à haut débit d'échantillons prélevés dans l'environnement. Les particularités des vues métagénomiques des CAZymes vont dévoiler l'étendue du métabolisme des sucres dans différents environnements. Comment fournir ces détails ? Comment les intégrer avec des données post-génomiques ? Ces problématiques devraient constituer un des thèmes de recherche de l'équipe pour les prochaines années. Une approche intéressante serait de créer un outil d'étude et de comparaison des familles de modules des génomes bactériens en fonction de leur mode de vie. Cette approche permettrait de générer une carte d'identité spécifique pour chaque organisme qui apporterait une information précieuse lors de l'étude métagénomique mais également lors des comparaisons de génomes. Une approche supplémentaire consisterait en l'élaboration d'un schéma de corrélation de familles modulaires par organisme et par fonction. Cette méthode statistique donnerait des informations supplémentaires lors de l'étude de synténie entre des génomes de différentes espèces. Cette approche serait aussi applicable lors de l'étude des séquences d'ADN métagénomiques mais cela exigerait le stockage des séquences nucléotidiques dans la base de données CAZy.

Finalement, l'augmentation exponentielle des séquences dans les bases de données va engendrer la construction de procédures automatiques encore plus performantes afin de permettre au curateur d'interagir de façon différente et sporadique avec les informations. Ces procédures nécessiteront probablement à nouveau une restructuration importante de la base de données.

Références Bibliographiques

1 Wolfenden, R., Lu, X. D. and Young, G. (1998) Spontaneous hydrolysis of glycosides. J. Am. Chem. Soc. **120**, 6814-6815
2 Matthysse, A. G. (1983) Role of bacterial cellulose fibrils in *Agrobacterium tumefaciens* infection. J. Bacteriol. **154**, 906-915
3 Matthysse, A. G., Deora, R., Mishra, M. and Torres, A. G. (2008) Polysaccharides cellulose, poly-beta-1,6-N-acetyl-D-glucosamine, and colanic acid are required for optimal binding of *Escherichia coli O157* : H7 strains to alfalfa sprouts and K-12 strains to plastic but not for binding to epithelial cells. Appl. Environ. Microbiol. **74**, 2384-2390
4 Jonas, R. and Farah, L. F. (1998) Production and application of microbial cellulose. Polym. Degrad. Stabil. **59**, 101-106
5 Laine, R. A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. Glycobiology. **4**, 759-767
6 Varki, A. and N., S. (2009) Historical Background and Overview. In Essentials of Glycobiology, 2nd edition (Editors, T. C. o. G., ed.), Cold Spring Harbor (NY)
7 Riley, M. and Labedan, B. (1997) Protein evolution viewed through Escherichia coli protein sequences: Introducing the notion of a structural segment of homology, the module. J. Mol. Biol. **268**, 857-868
8 Holm, L. and Sander, C. (1994) Parser for protein-folding units. Proteins. **19**, 256-268
9 Bourne, Y. and Henrissat, B. (2001) Glycoside hydrolases and glycosyltransferases: families and functional modules. Curr. Opin. Struct. Biol. **11**, 593-600
10 Fernandes, A. C., Fontes, C., Gilbert, H. J., Hazlewood, G. P., Fernandes, T. H. and Ferreira, L. M. A. (1999) Homologous xylanases from *Clostridium thermocellum*: evidence for bi-functional activity, synergism between xylanase catalytic modules and the presence of xylan-binding domains in enzyme complexes. Biochem. J. **342**, 105-110
11 Fontes, C. and Gilbert, H. J. (2010) Cellulosomes: Highly Efficient Nanomachines Designed to Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. In Annual Review of Biochemistry, Vol 79. pp. 655-681
12 Kosugi, A., Murashima, K. and Doi, R. H. (2002) Xylanase and acetyl xylan esterase activities of XynA, a key subunit of the *Clostridium cellulovorans* cellulosome for xylan degradation. Appl. Environ. Microbiol. **68**, 6399-6402
13 Millwardsadler, S. J., Davidson, K., Hazlewood, G. P., Black, G. W., Gilbert, H. J. and Clarke, J. H. (1995) Novel cellulose-binding domains, NodB homologues and conserved modular architecture in xylanases from the aerobic soil bacteria Pseudomonas fluorescens subsp. cellulosa and Cellvibrio mixtus. Biochem. J. **312**, 39-48
14 Gibbs, M. D., Saul, D. J., Luthi, E. and Bergquist, P. L. (1992) The beta-mannanase from caldocellum-saccharolyticum is part of multidomain enzyme. Appl. Environ. Microbiol. **58**, 3864-3867
15 Xia, W. S., Liu, P. and Liu, J. (2008) Advance in chitosan hydrolysis by non-specific cellulases. Bioresour. Technol. **99**, 6751-6762
16 Bork, P. (1991) Shuffled domains in extracellular proteins. FEBS Lett. **286**, 47-54
17 Sivakumar, A., Wilton, C. and Holm, L. (2006) From sequences to a functional unit. Physiol. Genomics. **25**, 1-8
18 Bork, P. (1992) Mobile modules and motifs. Curr.Opin.Struct.Biol. **2**, 413-421

19 Webb, E. C. (1992) Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. the International Union of Biochemistry and Molecular Biology by Academic Press.
20 Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. P. (1987) Hydrophobic cluster-analysis - an efficient new way to compare and analyzed amino acid sequences. FEBS Lett. **224**, 149-155
21 Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. and Mornon, J. P. (1989) Cellulase families revealed by hydrophobic cluster-analysis. Gene. **81**, 83-95
22 Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **280**, 309-316
23 Henrissat, B. and Bairoch, A. (1996) Updating the sequence-based classification of glycosyl hydrolases. Biochem. J. **316**, 695-696
24 Henrissat, B. and Bairoch, A. (1993) New families in the classification of glycosyl hydrolases based on amin-acid-sequence similarities. Biochem. J. **293**, 781-788
25 Campbell, J. A., Davies, G. J., Bulone, V. and Henrissat, B. (1997) A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. Biochem. J. **326**, 929-939
26 Coutinho, P. M., Deleury, E., Davies, G. J. and Henrissat, B. (2003) An evolving hierarchical family classification for glycosyltransferases. J. Mol. Biol. **328**, 307-317
27 Coutinho, P. M. and Henrissat, B. (1999) Carbohydrate-active enzymes: An integrated database approach. In Recent Advances in Carbohydrate Bioengineering (Gilbert, H. J., Davies, G. J., Henrissat, B. and Svensson, B., eds.). pp. 3-12
28 Tomme, P., Warren, R. A., Miller, R. C., Jr. , Kilburn, D. G. and Gilkes, N. R. (1995) Cellulose-binding domains: classification and properties. In Enzymatic Degradation of Insoluble Polysaccharides (Saddler, J. N. P., M., eds., ed.). pp. 142-163, American Chemical Society, Washington
29 Coutinho, P. M. and Henrissat, B. (1999) The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach. In Genetics, Biochemistry and Ecology of Cellulose Degradation. (K. Ohmiya, K. H., K. Sakka, Y. Kobayashi, S. Karita & T. Kimura eds, ed.). pp. 15-23, Uni Publishers Co., Tokyo
30 Beguin, P. (1990) Molecular-Biology of cellulose degradation. Annu. Rev. Microbiol. **44**, 219-248
31 Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P. and Davies, G. (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc. Natl. Acad. Sci. U. S. A. **92**, 7090-7094
32 Jenkins, J., Leggio, L. L., Harris, G. and Pickersgill, R. (1995) Beta-glucosidase, beta-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold beta/alpha architecture and with two conserved glutamates near the carboxy-terminal ends of beta-strands four and seven. FEBS Lett. **362**, 281-285
33 Henrissat, B. and Davies, G. (1997) Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. **7**, 637-644
34 Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. Embo J. **5**, 823-826
35 Henrissat, B. and Davies, G. J. (2000) Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. Plant Physiol. **124**, 1515-1519
36 Gebler, J., Gilkes, N. R., Claeyssens, M., Wilson, D. B., Beguin, P., Wakarchuk, W. W., Kilburn, D. G., Miller, R. C., Warren, R. A. J. and Withers, S. G. (1992) Stereoselective

hydrolysis catalyzed by related beta-1,4-glucanases and beta-1,4-xylanases. J. Biol. Chem. **267**, 12559-12561

37 Marques, A. R., Coutinho, P. M., Videira, P., Fialho, A. M. and Sa-Correia, I. (2003) Sphingomonas paucimobilis beta-glucosidase Bgl1: a member of a new bacterial subfamily in glycoside hydrolase family 1. Biochem. J. **370**, 793-804

38 Stam, M. R., Blanc, E., Coutinho, P. M. and Henrissat, B. (2005) Evolutionary and mechanistic relationships between glycosidases acting on alpha- and beta-bonds. Carbohydr. Res. **340**, 2728-2734

39 Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M. and Henrissat, B. (2010) A hierarchical classification of polysaccharide lyases for glycogenomics. Biochem J. **432**, 437-444

40 Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. **37**, D233-D238

41 Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2010) GenBank. Nucleic Acids Res. **38**, D46-D51

42 Apweiler, R., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., Antunes, R., Barrell, D., Bely, B., Bingley, M., Binns, D., Bower, L., Browne, P., Chan, W. M., Dimmer, E., Eberhardt, R., Fedotov, A., Foulger, R., Garavelli, J., Huntley, R., Jacobsen, J., Kleen, M., Laiho, K., Leinonen, R., Legge, D., Lin, Q., Liu, W. D., Luo, J., Orchard, S., Patient, S., Poggioli, D., Pruess, M., Corbett, M., di Martino, G., Donnelly, M., van Rensburg, P., Bairoch, A., Bougueleret, L., Xenarios, I., Altairac, S., Auchincloss, A., Argoud-Puy, G., Axelsen, K., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Bollondi, L., Boutet, E., Quintaje, S. B., Breuza, L., Bridge, A., deCastro, E., Ciapina, L., Coral, D., Coudert, E., Cusin, I., Delbard, G., Doche, M., Dornevil, D., Roggli, P. D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gehant, S., Farriol-Mathis, N., Ferro, S., Gasteiger, E., Gateau, A., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hulo, N., James, J., Jimenez, S., Jungo, F., Kappler, T., Keller, G., Lachaize, C., Lane-Guermonprez, L., Langendijk-Genevaux, P., Lara, V., Lemercier, P., Lieberherr, D., Lima, T. D., Mangold, V., Martin, X., Masson, P., Moinat, M., Morgat, A., Mottaz, A., Paesano, S., Pedruzzi, I., Pilbout, S., Pillet, V., Poux, S., Pozzato, M., Redaschi, N., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stanley, E., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Yip, L. N., Zuletta, L., Wu, C., Arighi, C., Arminski, L., Barker, W., Chen, C. M., Chen, Y. X., Hu, Z. Z., Huang, H. Z., Mazumder, R., McGarvey, P., Natale, D. A., Nchoutmboube, J., Petrova, N., Subramanian, N., Suzek, B. E., Ugochukwu, U., Vasudevan, S., Vinayaka, C. R., Yeh, L. S., Zhang, J. and UniProt, C. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. **38**, D142-D148

43 Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. **112**, 535-542

44 Burgard, A. P. and Maranas, C. D. (2001) Review of the Enzymes and Metabolic Pathways (EMP) database. **3**, 193-194

45 Kawabata, T., Ota, M. and Nishikawa, K. (1999) The protein mutant database. **27**, 355-357

46 Yip, V. L. and Withers, S. G. (2006) Breakdown of oligosaccharides by the process of elimination. Curr. Opin. Chem. Biol. **10**, 147-155

47 McCarter, J. D. and Withers, S. G. (1994) Mechanisms of enzymatic glycoside hydrolysis. Curr. Opin. Struct. Biol. **4**, 885-892

48 Lairson, L. L., Henrissat, B., Davies, G. J. and Withers, S. G. (2008) Glycosyltransferases: Structures, functions, and mechanisms. Annu. Rev. Biochem. **77**, 521-555
49 Garron, M. L. and Cygler, M. (2010) Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. Glycobiology. **20**, 1547-1573
50 Christov, L. P. and Prior, B. A. (1993) Esterases of xylan-degrading microorganisms: production, properties, and significance. Enzyme Microb. Technol. **15**, 460-475
51 Jayani, R. S., Saxena, S. and Gupta, R. (2005) Microbial pectinolytic enzymes: A review. Process Biochem. **40**, 2931-2944
52 Boraston, A. B., Bolam, D. N., Gilbert, H. J. and Davies, G. J. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. **382**, 769-781
53 Guillen, D., Sanchez, S. and Rodriguez-Sanoja, R. (2010) Carbohydrate-binding domains: multiplicity of biological roles. Appl. Microbiol. Biotechnol. **85**, 1241-1249
54 Pankov, R. and Yamada, K. M. (2002) Fibronectin at a glance. J. Cell Sci. **115**, 3861-3863
55 Alahuhta, M., Xu, Q., Brunecky, R., Adney, W. S., Ding, S. Y., Himmel, M. E. and Lunin, V. V. (2010) Structure of a fibronectin type III-like module from Clostridium thermocellum. Acta Crystallogr. F-Struct. Biol. Cryst. Commun. **66**, 878-880
56 Ficko-Blean, E., Gregg, K. J., Adams, J. J., Hehemann, J. H., Czjzek, M., Smith, S. P. and Boraston, A. B. (2009) Portrait of an Enzyme, a Complete Structural Analysis of a Multimodular beta-N-Acetylglucosaminidase from Clostridium perfringens. J. Biol. Chem. **284**, 9876-9884
57 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. J. Mol. Biol. **215**, 403-410
58 Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. **85**, 2444-2448
59 Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F. and Barton, G. J. (2010) Visualization of multiple alignments, phylogenies and gene family evolution. **7**, S16-S25
60 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673-4680
61 Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792-1797
62 Griffiths-Jones, S. and Bateman, A. (2002) The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. Bioinformatics. **18**, 1243-1249
63 Marchler-Bauer, A. and Bryant, S. H. (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res. **32**, W327-W331
64 Eddy, S. R. (1998) Profile hidden Markov models. Bioinformatics. **14**, 755-763
65 Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. and Bateman, A. (2010) The Pfam protein families database. Nucleic Acids Res. **38**, D211-D222
66 Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A. and Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. **38**, D161-D166
67 Kall, L., Krogh, A. and Sonnhammer, E. L. L. (2004) A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. **338**, 1027-1036

68 Bernard, T. (2008) Annotation et prédiction de la spécificité de substrat des enzymes actives sur les sucres. p. 450, Université de la mediterranée Aix-Marseille II, Marseille
69 Kirk, O., Borchert, T. V. and Fuglsang, C. C. (2002) Industrial enzyme applications. Curr. Opin. Biotechnol. **13**, 345-351
70 Fahey, J. (2010) Are we getting closer to 'clean' ethanol? Forbes Magazine. **185**
71 Solomon, B. D. (2010) Biofuels and sustainability. In Ecological Economics Reviews. pp. 119-134
72 Giampietro, M., Ulgiati, S. and Pimentel, D. (1997) Feasibility of large-scale biofuel production - Does an enlargement of scale change the picture? Bioscience. **47**, 587-600
73 Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W. and Foust, T. D. (2007) Biomass recalcitrance: Engineering plants and enzymes for biofuels production. Science. **315**, 804-807
74 Lewandowski, I., Clifton-Brown, J. C., Scurlock, J. M. O. and Huisman, W. (2000) Miscanthus: European experience with a novel energy crop. Biomass Bioenerg. **19**, 209-227
75 Mandels, M. and Reese, E. T. (1957) Induction of cellulase in *Trichoderma viride* as influenced by carbon sources and metals. J. Bacteriol. **73**, 269–278
76 Ilmen, M., Saloheimo, A., Onnela, M. L. and Penttila, M. E. (1997) Regulation of cellulase gene expression in the filamentous fungus Trichoderma reesei. Appl. Environ. Microbiol. **63**, 1298-1306
77 Martinez, D., Berka, R. M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S. E., Chapman, J., Chertkov, O., Coutinho, P. M., Cullen, D., Danchin, E. G. J., Grigoriev, I. V., Harris, P., Jackson, M., Kubicek, C. P., Han, C. S., Ho, I., Larrondo, L. F., de Leon, A. L., Magnuson, J. K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A. A., Schmoll, M., Terry, A., Thayer, N., Westerholm-Parvinen, A., Schoch, C. L., Yao, J., Barbote, R., Nelson, M. A., Detter, C., Bruce, D., Kuske, C. R., Xie, G., Richardson, P., Rokhsar, D. S., Lucas, S. M., Rubin, E. M., Dunn-Coleman, N., Ward, M. and Brettin, T. S. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat. Biotechnol. **26**, 553-560
78 Fry, S. C. (2011) Cell wall polysaccharide composition and covalent crosslinking. In Plant Polysaccharides: Biosynthesis and Bioengineering (Ulvskov, P., ed.). pp. 1-42, Blackwell Publishing Ltd.
79 Young, A. R. and Rowell, M. R. (1986) Cellulose structure modification and hydrolysis. Wiley, New York
80 Fry, S. C. (1989) Cellulases, hemicelluloses and auxin-stimulated growth: a possible relationship. Physiol. Plant. **75**, 532-536
81 Mohnen, D. (2008) Pectin structure and biosynthesis. Curr. Opin. Plant Biol. **11**, 266-277
82 Harris, P. J. and Stone, B. A. (2008) Chemistry and molecular organization of plant cell walls. In Biomass recalcitrance (Himmel, D. M. E., ed.). pp. 61-93, Blackwell, Oxford
83 Harris, P. V., Welner, D., McFarland, K. C., Re, E., Poulsen, J. C. N., Brown, K., Salbo, R., Ding, H. S., Vlasenko, E., Merino, S., Xu, F., Cherry, J., Larsen, S. and Lo Leggio, L. (2010) Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family. Biochemistry. **49**, 3305-3316
84 Hart, G. W. and Copeland, R. J. (2010) Glycomics Hits the Big Time. Cell. **143**, 672-676
85 Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E.,

Wilkinson, M. D. and Birney, E. (2002) The bioperl toolkit: Perl modules for the life sciences. Genome Res. **12**, 1611-1618

86 Yip, V. L. Y., Varrot, A., Davies, G. J., Rajan, S. S., Yang, X. J., Thompson, J., Anderson, W. F. and Withers, S. G. (2004) An unusual mechanism of glycoside hydrolysis involving redox and elimination steps by a family 4 beta-glycosidase from Thermotoga maritima. J. Am. Chem. Soc. **126**, 8354-8355

87 Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389-3402

88 Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. **89**, 10915-10919

89 Wicker, N., Perrin, G. R., Thierry, J. C. and Poch, O. (2001) Secator: A program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol. **18**, 1435-1441

90 Farris, J. S., Albert, V. A., Kallersjo, M., Lipscomb, D. and Kluge, A. G. (1996) Parsimony jackknifing outperforms neighbor-joining. Cladistics-Int. J. Willi Hennig Soc. **12**, 99-124

91 Fouet, A. and Mesnage, S. (2002) Bacillus anthracis cell envelope components. In Anthrax. pp. 87-113

92 Mishra, P., Kumar, R. P., Ethayathulla, A. S., Singh, N., Sharma, S., Perbandt, M., Betzel, C., Kaur, P., Srinivasan, A., Bhakuni, V. and Singh, T. P. (2009) Polysaccharide binding sites in hyaluronate lyase - crystal structures of native phage-encoded hyaluronate lyase and its complexes with ascorbic acid and lactose. Febs J. **276**, 3392-3402

93 Charnock, S. J., Brown, I. E., Turkenburg, J. P., Black, G. W. and Davies, G. J. (2002) Convergent evolution sheds light on the anti-beta-elimination mechanism common to family 1 and 10 polysaccharide lyases. Proc. Natl. Acad. Sci. U. S. A. **99**, 12067-12072

94 Jenkins, J. and Pickersgill, R. (2001) The architecture of parallel beta-helices and related folds. Prog. Biophys. Mol. Biol. **77**, 111-175

95 Rubin, E. M. (2008) Genomics of cellulosic biofuels. Nature. **454**, 841-845

96 Ohm, R. A., de Jong, J. F., Lugones, L. G., Aerts, A., Kothe, E., Stajich, J. E., de Vries, R. P., Record, E., Levasseur, A., Baker, S. E., Bartholomew, K. A., Coutinho, P. M., Erdmann, S., Fowler, T. J., Gathman, A. C., Lombard, V., Henrissat, B., Knabe, N., Kues, U., Lilly, W. W., Lindquist, E., Lucas, S., Magnuson, J. K., Piumi, F., Raudaskoski, M., Salamov, A., Schmutz, J., Schwarze, F., vanKuyk, P. A., Horton, J. S., Grigoriev, I. V. and Wosten, H. A. B. (2010) Genome sequence of the model mushroom *Schizophyllum commune*. Nat. Biotechnol. **28**, 957-U910

97 Ma, L. J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M. J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W. B., Woloshuk, C., Xie, X. H., Xu, J. R., Antoniw, J., Baker, S. E., Bluhm, B. H., Breakspear, A., Brown, D. W., Butchko, R. A. E., Chapman, S., Coulson, R., Coutinho, P. M., Danchin, E. G. J., Diener, A., Gale, L. R., Gardiner, D. M., Goff, S., Hammond-Kosack, K. E., Hilburn, K., Hua-Van, A., Jonkers, W., Kazan, K., Kodira, C. D., Koehrsen, M., Kumar, L., Lee, Y. H., Li, L. D., Manners, J. M., Miranda-Saavedra, D., Mukherjee, M., Park, G., Park, J., Park, S. Y., Proctor, R. H., Regev, A., Ruiz-Roldan, M. C., Sain, D., Sakthikumar, S., Sykes, S., Schwartz, D. C., Turgeon, B. G., Wapinski, I., Yoder, O., Young, S., Zeng, Q. D., Zhou, S. G., Galagan, J., Cuomo, C. A., Kistler, H. C. and Rep, M. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. **464**, 367-373

98 Martinez, D., Larrondo, L. F., Putnam, N., Gelpke, M. D. S., Huang, K., Chapman, J., Helfenbein, K. G., Ramaiya, P., Detter, J. C., Larimer, F., Coutinho, P. M., Henrissat,

B., Berka, R., Cullen, D. and Rokhsar, D. (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium strain RP78*. Nat. Biotechnol. **22**, 695-700

99 Pel, H. J., de Winde, J. H., Archer, D. B., Dyer, P. S., Hofmann, G., Schaap, P. J., Turner, G., de Vries, R. P., Albang, R., Albermann, K., Andersen, M. R., Bendtsen, J. D., Benen, J. A. E., van den Berg, M., Breestraat, S., Caddick, M. X., Contreras, R., Cornell, M., Coutinho, P. M., Danchin, E. G. J., Debets, A. J. M., Dekker, P., van Dijck, P. W. M., van Dijk, A., Dijkhuizen, L., Driessen, A. J. M., d'Enfert, C., Geysens, S., Goosen, C., Groot, G. S. P., de Groot, P. W. J., Guillemette, T., Henrissat, B., Herweijer, M., van den Hombergh, J., van den Hondel, C., van der Heijden, R., van der Kaaij, R. M., Klis, F. M., Kools, H. J., Kubicek, C. P., van Kuyk, P. A., Lauber, J., Lu, X., van der Maarel, M., Meulenberg, R., Menke, H., Mortimer, M. A., Nielsen, J., Oliver, S. G., Olsthoorn, M., Pal, K., van Peij, N., Ram, A. F. J., Rinas, U., Roubos, J. A., Sagt, C. M. J., Schmoll, M., Sun, J. B., Ussery, D., Varga, J., Vervecken, W., de Vondervoort, P., Wedler, H., Wosten, H. A. B., Zeng, A. P., van Ooyen, A. J. J., Visser, J. and Stam, H. (2007) Genome sequencing and analysis of the versatile cell factory *Aspergillus niger CBS 513.88*. Nat. Biotechnol. **25**, 221-231

100 Espagne, E., Lespinet, O., Malagnac, F., Da Silva, C., Jaillon, O., Porcel, B. M., Couloux, A., Aury, J. M., Segurens, B., Poulain, J., Anthouard, V., Grossetete, S., Khalili, H., Coppin, E., Dequard-Chablat, M., Picard, M., Contamine, V., Arnaise, S., Bourdais, A., Berteaux-Lecellier, V., Gautheret, D., de Vries, R. P., Battaglia, E., Coutinho, P. M., Danchin, E. G. J., Henrissat, B., El Khoury, R., Sainsard-Chanet, A., Boivin, A., Pinan-Lucarree, B., Sellem, C. H., Debuchy, R., Wincker, P., Weissenbach, J. and Silar, P. (2008) The genome sequence of the model ascomycete fungus *Podospora anserina*. Genome Biol. **9**

101 Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E. G. J., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H. J., Wuyts, J., Blaudez, D., Buee, M., Brokstein, P., Canback, B., Cohen, D., Courty, P. E., Coutinho, P. M., Delaruelle, C., Detter, J. C., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feussner, I., Gay, G., Grimwood, J., Hoegger, P. J., Jain, P., Kilaru, S., Labbe, J., Lin, Y. C., Legue, V., Le Tacon, F., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-Le Secq, M. P., Peter, M., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kues, U., Lucas, S., Van de Peer, Y., Podila, G. K., Polle, A., Pukkila, P. J., Richardson, P. M., Rouze, P., Sanders, I. R., Stajich, J. E., Tunlid, A., Tuskan, G. and Grigoriev, I. V. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature. **452**, 88-U87

102 Coleman, J. J., Rounsley, S. D., Rodriguez-Carres, M., Kuo, A., Wasmann, C. C., Grimwood, J., Schmutz, J., Taga, M., White, G. J., Zhou, S. G., Schwartz, D. C., Freitag, M., Ma, L. J., Danchin, E. G. J., Henrissat, B., Coutinho, P. M., Nelson, D. R., Straney, D., Napoli, C. A., Barker, B. M., Gribskov, M., Rep, M., Kroken, S., Molnar, I., Rensing, C., Kennell, J. C., Zamora, J., Farman, M. L., Selker, E. U., Salamov, A., Shapiro, H., Pangilinan, J., Lindquist, E., Lamers, C., Grigoriev, I. V., Geiser, D. M., Covert, S. F., Temporini, E. and VanEtten, H. D. (2009) The Genome of *Nectria haematococca*: Contribution of Supernumerary Chromosomes to Gene Expansion. PLoS Genet. **5**

103 Vantilbeurgh, H., Tomme, P., Claeyssens, M., Bhikhabhai, R. and Pettersson, G. (1986) Limited proteolysis of the cellobiohydrolase I from *Trichoderma reesei*: Separation of functional domains. FEBS Lett. **204**, 223-227

104 Boraston, A. B., Lammerts van Bueren, A., Ficko-Blean, E. and Abbott, D. W. (2007) Carbohydrate-protein interactions: carbohydrate-binding modules. In Comprehensive Glycoscience From Chemistry to Systems Biology (Kamerling, J. P., ed.). pp. 661-696, Elsevier B.V.
105 Gilbert, H. J. (2010) The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. Plant Physiol. **153**, 444-455
106 Tomme, P., Vantilbeurgh, H., Pettersson, G., Vandamme, J., Vandekerckhove, J., Knowles, J., Teeri, T. and Claeyssens, M. (1988) Studies of the cellulolytic system of *Trichoderma reesei QM 9414*. Analysis of domain function in two cellobiohydrolases by limited proteolysis. Eur. J. Biochem. **170**, 575-581
107 Limon, M. C., Margolles-Clark, E., Benitez, T. and Penttila, M. (2001) Addition of substrate-binding domains increases substrate-binding capacity and specific activity of a chitinase from *Trichoderma harzianum*. FEMS Microbiol. Lett. **198**, 57-63
108 Gilkes, N. R., Warren, R. A. J., Miller, R. C. and Kilburn, D. G. (1988) Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis. J. Biol. Chem. **263**, 10401-10407
109 Kikkawa, Y., Tokuhisa, H., Shingai, H., Hiraishi, T., Houjou, H., Kanesato, M., Imanaka, T. and Tanaka, T. (2008) Interaction force of chitin-binding domains onto chitin surface. Biomacromolecules. **9**, 2126-2131
110 Tomme, P., Warren, R. A. J., Miller, R. C., Kilburn, D. G. and Gilkes, N. R. (1995) Cellulose hydrolysis by bacteria and fungi. In Enzymatic degradation of insoluble carbohydrates (Saddler, J. N. and Penner, M. H., eds.). pp. 142-163, American Chemical Society, Washington
111 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. **13**, 2498-2504
112 Clamp, M., Cuff, J., Searle, S. M. and Barton, G. J. (2004) The Jalview Java alignment editor. Bioinformatics. **20**, 426-427
113 Zmasek, C. M. and Eddy, S. R. (2001) ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics. **17**, 383-384
114 Henrissat, B. and Coutinho, P. M. (2001) Classification of glycoside hydrolases and glycosyltransferases from hyperthermophiles. In Hyperthermophilic Enzymes, Pt A. pp. 183-201
115 Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P. M., Jaillon, O., Montanini, B., Morin, E., Noel, B., Percudani, R., Porcel, B., Rubini, A., Amicucci, A., Amselem, J., Anthouard, V., Arcioni, S., Artiguenave, F., Aury, J. M., Ballario, P., Bolchi, A., Brenna, A., Brun, A., Buee, M., Cantarel, B., Chevalier, G., Couloux, A., Da Silva, C., Denoeud, F., Duplessis, S., Ghignone, S., Hilselberger, B., Iotti, M., Marcais, B., Mello, A., Miranda, M., Pacioni, G., Quesneville, H., Riccioni, C., Ruotolo, R., Splivallo, R., Stocchi, V., Tisserant, E., Viscomi, A. R., Zambonelli, A., Zampieri, E., Henrissat, B., Lebrun, M. H., Paolocci, F., Bonfante, P., Ottonello, S. and Wincker, P. (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature. **464**, 1033-1038
116 Schuster, A. and Schmoll, M. (2010) Biology and biotechnology of *Trichoderma*. Appl. Microbiol. Biotechnol. **87**, 787-799
117 Viterbo, A., Ramot, O., Chernin, L. and Chet, I. (2002) Significance of lytic enzymes from *Trichoderma spp.* in the biocontrol of fungal plant pathogens. Antonie Van Leeuwenhoek. **81**, 549-556
118 Couturier, M., Haon, M., Coutinho, P. M., Henrissat, B., Lesage-Meessen, L. and Berrin, J. G. (2011) *Podospora anserina* Hemicellulases Potentiate the *Trichoderma*

reesei Secretome for Saccharification of Lignocellulosic Biomass. Appl. Environ. Microbiol. **77**, 237-246
119 De Groot, P. W. J., Ram, A. F. and Klis, F. M. (2005) Features and functions of covalently linked proteins in fungal cell walls. Fungal Genet. Biol. **42**, 657-675
120 Durand, P., Lehn, P., Callebaut, I., Fabrega, S., Henrissat, B. and Mornon, J. P. (1997) Active-site motifs of lysosomal acid hydrolases: Invariant features of clan GH-A glycosyl hydrolases deduced from hydrophobic cluster analysis. Glycobiology. **7**, 277-284
121 Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., Cantarel, B. L., Coutinho, P. M., Henrissat, B., Leclerc, M., Dore, J., Monsan, P., Remaud-Simeon, M. and Potocki-Veronese, G. (2010) Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. Genome Res. **20**, 1605-1612
122 Handelsman, J. (2004) Metagenomics: Application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. **68**, 669-+
123 Schloss, P. D. and Handelsman, J. (2003) Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14**, 303-310
124 Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. **428**, 37-43
125 Langer, M., Gabor, E. M., Liebeton, K., Meurer, G., Niehaus, F., Schulze, R., Eck, J. and Lorenz, P. (2006) Metagenomics: An inexhaustible access to nature's diversity. Biotechnol. J. **1**, 815-821
126 Pearson, W. R., Wood, T., Zhang, Z. and Miller, W. (1997) Comparison of DNA sequences with protein sequences. **46**, 24-36
127 Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., Cayouette, M., McHardy, A. C., Djordjevic, G., Aboushadi, N., Sorek, R., Tringe, S. G., Podar, M., Martin, H. G., Kunin, V., Dalevi, D., Madejska, J., Kirton, E., Platt, D., Szeto, E., Salamov, A., Barry, K., Mikhailova, N., Kyrpides, N. C., Matson, E. G., Ottesen, E. A., Zhang, X. N., Hernandez, M., Murillo, C., Acosta, L. G., Rigoutsos, I., Tamayo, G., Green, B. D., Chang, C., Rubin, E. M., Mathur, E. J., Robertson, D. E., Hugenholtz, P. and Leadbetter, J. R. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature. **450**, 560-U517
128 Pope, P. B., Denman, S. E., Jones, M., Tringe, S. G., Barry, K., Malfatti, S. A., McHardy, A. C., Cheng, J. F., Hugenholtz, P., McSweeney, C. S. and Morrison, M. (2010) Adaptation to herbivory by the *Tammar wallaby* includes bacterial and glycoside hydrolase profiles different from other herbivores. Proc. Natl. Acad. Sci. U. S. A. **107**, 14793-14798

Annexes

A

Biochem. J. (2010);**432**:437-444

A hierarchical classification of polysaccharide lyases for glycogenomics.

Lombard V, Bernard T, Rancurel C, Brumer H, Coutinho PM, Henrissat B.

Abstract :Carbohydrate-active enzymes face large substrate diversity in a highly selective manner with only a limited number of available folds. They are therefore subjected to multiple divergent and convergent evolutionary events. This and their frequent modularity render their functional annotation in genomes difficult in a number of cases. A classification of polysaccharide lyases (the enzymes that cleave polysaccharides using an elimination instead of a hydrolytic mechanism) is presented thoroughly for the first time. Based on the analysis of a large panel of experimentally characterized polysaccharide lyases, we examined the correlation of various enzyme properties with the three levels of the classification: fold, families and subfamilies. The resulting hierarchical classification, which should help annotate relevant genes in genomic efforts, is available and constantly updated at the Carbohydrate-Active Enzymes Database (www.cazy.org).

PMID: 20925655

BJ www.biochemj.org

Biochem. J. (2010) **432**, 437–444 (Printed in Great Britain) doi:10.1042/BJ20101185

A hierarchical classification of polysaccharide lyases for glycogenomics

Vincent LOMBARD*, Thomas BERNARD*[1], Corinne RANCUREL*, Harry BRUMER†, Pedro M. COUTINHO* and Bernard HENRISSAT*[2]

*Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Université de la Méditerranée, Université de Provence, Case 932, 163 Avenue de Luminy, 13288 Marseille cedex 9, France, and †School of Biotechnology, Royal Institute of Technology (KTH), AlbaNova University Centre, 106 91 Stockholm, Sweden

Carbohydrate-active enzymes face huge substrate diversity in a highly selective manner using only a limited number of available folds. They are therefore subjected to multiple divergent and convergent evolutionary events. This and their frequent modularity render their functional annotation in genomes difficult in a number of cases. In the present paper, a classification of polysaccharide lyases (the enzymes that cleave polysaccharides using an elimination instead of a hydrolytic mechanism) is shown thoroughly for the first time. Based on the analysis of a large panel of experimentally characterized polysaccharide lyases, we examined the correlation of various enzyme properties with the three levels of the classification: fold, family and subfamily. The resulting hierarchical classification, which should help annotate relevant genes in genomic efforts, is available and constantly updated at the Carbohydrate-Active Enzymes Database (http://www.cazy.org).

Key words: catalytic mechanism, enzyme family, functional annotation, modular structure, polysaccharide lyase.

INTRODUCTION

PLs (polysaccharide lyases) are a group of enzymes (EC 4.2.2.-) that cleave uronic acid-containing polysaccharides via a β-elimination mechanism to generate an unsaturated hexenuronic acid residue and a new reducing end at the point of cleavage (Figure 1) [1]. PLs are ubiquitous in nature, having been identified in organisms ranging from bacteriophages and Archaea, to Eubacteria and higher eukaryotes, such as fungi, algae, plants and mammals [2]. For all of these organisms, PLs represent a complimentary mechanistic strategy to the GHs (glycoside hydrolases; EC 3.2.1.-) [3] for the breakdown of C-6 carboxylated polysaccharides (Figure 1), with the contrasting feature that PL-catalysed cleavage occurs without intervention of a water molecule. PLs are implicit in diverse biochemical processes, including biomass degradation, tissue matrix recycling and pathogenesis [2,4–9]. Moreover, the widespread use of polyuronic acids in the food and medical sectors makes PLs attractive as specific catalysts for the modification of substrates such as pectins, alginates and heparins in biotechnological applications [2,10–12].

The catalytic mechanism employed by PLs (Figure 2) can be broadly described as consisting of three events: (i) abstraction of the C-5 proton on the sugar ring of a uronic acid or ester by a basic amino acid side chain, (ii) stabilization of the resulting anion by charge delocalization into the C-6 carbonyl group, and (iii) lytic cleavage of the O-4:C-4 bond, facilitated by proton donation from a catalytic acid, to yield a hexenuronic acid (or ester) moiety at the newly formed non-reducing chain end [1,13]. Depending on the monosaccharide composition of the substrate and its conformation in the PL active site, the proton removed from C-5 and the departing oxygen on C-4 may lie either *syn* or *anti* to each other. This, in turn, imposes certain requirements on the position of active site groups and the possibilities for a stepwise or concerted elimination reaction (Figures 2A and 2B). Polysaccharide recognition in PLs is often dependent on the interaction of tightly held bivalent cations (often Ca^{2+}), or positively charged amino acid side chains, with uronic acid groups in the substrate. Such cations may play an additional role in stabilizing the transient anion in the reaction pathway. The extent to which these molecular events are concerted, as well as the nature and individual contributions of the catalytic groups, in the mechanisms of specific enzymes have not been fully clarified, although significant advances have been made in a few cases (see [14,15] and references therein). Detailed structural information on the catalytic modules of PLs has been previously published [16].

In common with GHs, PLs frequently have multi-modular structures, in which the catalytic module can be appended to a variable number of ancillary modules such as CBMs (carbohydrate-binding modules) [17,18], other catalytic modules or modules with other functions (see below). Interestingly, many non-catalytic modules borne by PLs may also be appended to GHs. Following a full dissection of their modular organization, we have grouped the PLs into amino acid sequence-based families to provide a framework for structural and mechanistic studies. In the present paper we describe a hierarchical classification of PLs including subfamilies, families and clans/superfamilies, and we discuss the value of these levels for genome mining and functional prediction. This classification is implemented in the CAZy (Carbohydrate-Active Enzymes) Database (http://www.cazy.org) [19].

EXPERIMENTAL

Included and excluded enzyme classes

For the purpose of this family classification, the scope of the term PL is restricted to those enzymes that operate according to the general mechanisms described in Figures 1 and 2, to produce

Abbreviations used: CBM, carbohydrate-binding module; CE, carbohydrate esterase; GH, glycoside hydrolase; GT, glycosyltransferase; LT, lytic transglycosylase; PL, polysaccharide lyase.

[1] Present address: Biométrie et Biologie Évolutive, UMR CNRS 5558, UCB Lyon 1, Bât. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, France

[2] To whom correspondence should be addressed (email bernard.henrissat@afmb.univ-mrs.fr).

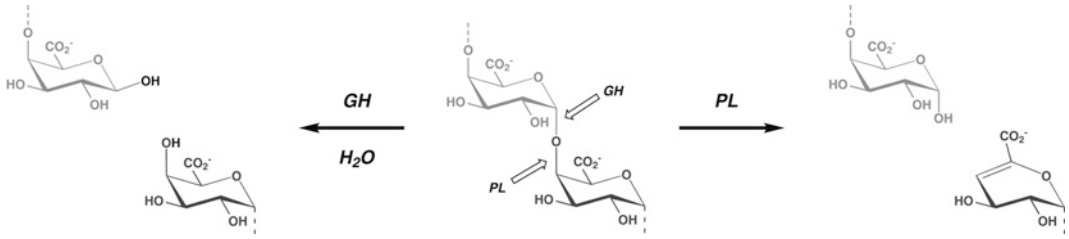


Figure 1 Comparison of the products of PL and GH, exemplified by polygalacturonate (pectate) cleavage

Both enzyme classes generate a new reducing chain end (light grey). GHs cleave the glycosidic bond (C-1′:O-4) by the addition of water, maintaining the 4-OH group at the new non-reducing chain end. PLs, in contrast, generate a hexeneuronic acid moiety (HexA, 4-deoxy-hex-4-eneuronic acid) at the new non-reducing end by eliminative cleavage of the O-4:C-4 bond.

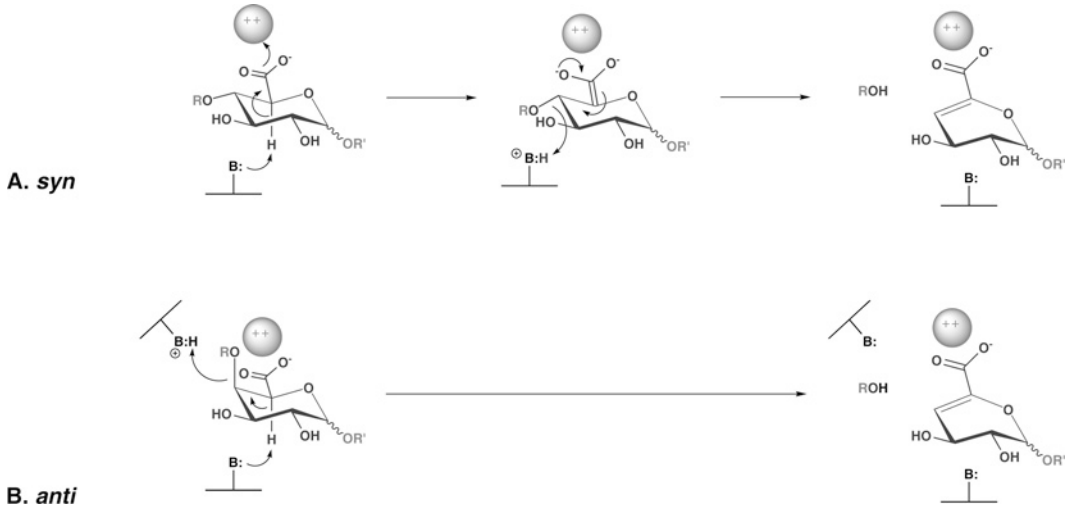


Figure 2 General mechanisms of PLs

(**A**) *syn*-Elimination, as in chondroitin lyase. (**B**) *anti*-Elimination, as in α-(1,4)-polygalacturonan (pectate) lyase. In both, polysaccharides are cleaved to produce a 4-deoxy-hex-4-eneuronic acid moiety at the newly formed non-reducing end of the chain; due to loss of the asymmetric center at C-4, gluco- or galacto-configured substrates yield essentially the same product (depending on the stereochemistry at C-1). As a prelude to chain scission, the C-5 proton adjacent to the carbonyl group is abstracted by a suitably poised basic amino acid sidechain (B:). Departure of the glycosidic oxygen is likely to be facilitated by proton donation from a catalytic acid (B:H). Co-ordinating and charge-stabilizing cations, Ca^{2+} or a positively charged amino-acid side chain, are also a common feature of PL active sites.

a terminal hexenuronic acid moiety by β-elimination. This is a clear distinction from the broader NC-IUBMB classification of carbon-oxygen lyases acting on polysaccharides into EC 4.2.2.- (http://www.chem.qmul.ac.uk/iubmb/enzyme/). In particular, the following enzymes are not included in PL classification described in CAZy, as they are structurally and mechanistically more similar to the GHs:

(i) exo-α-(1,4)-D-glucan lyases (EC 4.2.2.13) cleave malto-oligosaccharides to produce 1,5-anhydro-D-fructose without the intervention of a water molecule. These enzymes are structurally similar to GH31 α-glucoside hydrolases, with which they are currently classified. Analogous to other GH31 enzymes, the first step in the catalytic mechanism involves the formation of a covalent glycosyl-enzyme intermediate. However, in α-glucan lyases this intermediate decomposes through a *syn*-elimination mechanism, rather than hydrolysis [13,20].

(ii) LTs (lytic transglycosylases) cleave the β-(1,4)-glycosidic bond between the *N*-acetylmuramic acid and the *N*-acetylglucosamine residues of peptideglycan via a substrate participation mechanism, with no intervention of water, to yield a 1,6-anhydro sugar derivative [21]. LTs are structurally and mechanistically closely related to lysozymes and are currently classified in GH families GH23, GH102, GH103 and GH104 [22].

(iii) Levan and inulin fructotransferases (EC 4.2.2.16, EC 4.2.2.17 and EC 4.2.2.18) cleave fructo-oligosaccharides by intramolecular attack to yield various anhydro-fructodisaccharides. These enzymes are presently classified into GH91, along with a sequence-similar enzyme that hydrolyses the DFA III (α-D-fructofuranose β-D-fructofuranose 1,2′:2,3′-dianhydride) product of the EC 4.2.2.18 inulin fructotransferase [23,24]. As such, mechanistic commonality with GHs (and loosely with LTs) is predicted.

Family and subfamily groupings

The PL families were first built by searching sequence homologues of experimentally characterized enzymes. To avoid the creation of a large number of families, distant homologues were assigned to existing families. These families have been presented online in the CAZy database since its launch in 1998 with the occasional creation of novel families subsequent to the experimental characterization of PLs with no or insufficient similarity to known families. Within families, subfamilies have been defined by procedures similar to that described for the large GH family GH13, which is comprised of a diversity of starch-active enzymes of similar structure [25].

Briefly, in each family the sequences were edited to isolate the catalytic domains to avoid interference from the presence

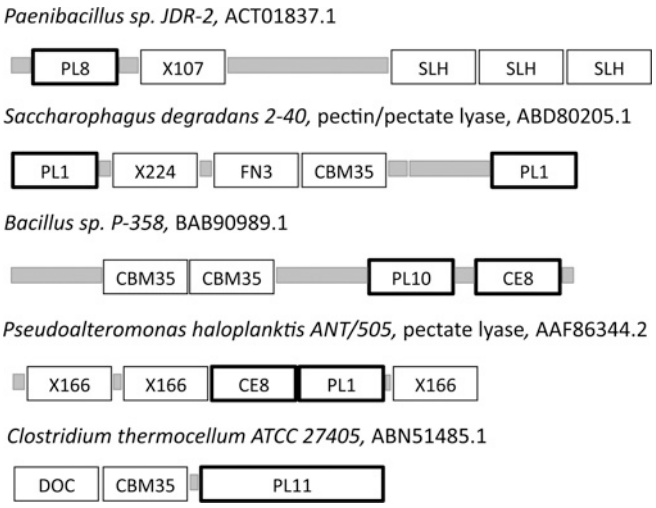


Figure 3 Examples of modular PLs

PL, polysaccharide lyase module; CBM, carbohydrate-binding module; CE, carbohydrate esterase module. Other modules include dockerins (DOC), S-layer homology domains (SLH), fibronectin type 3 domains (FN3) and conserved domains of unknown function (X). Unassigned regions are in grey. GenBank® accession numbers are indicated for each protein.

or absence of additional modules. The catalytic domains were then subject to a multiple sequence alignment using MUSCLE [26] and a distance matrix was created using the BLOSUM62 [27] substitution model. The distance matrix was then used as the input for an automatic analysis based on the SECATOR algorithm [28], which proposes the breakdown of the family into a number of subfamilies, based on a reconstructed phylogenetic tree. The robustness of the subfamilies was tested by a resampling approach whereby a proportion of the sequences were randomly removed from the sample. The clustering procedure was iterated typically 10000 times with random variations of the parameters of the automatic partitioning algorithm. The percentage of sequences removed from the sample was also picked randomly from 5 to 30% at each iteration. Sequences found in the same cluster over 80% of the time were assigned to the same subfamily. Finally, only subfamilies containing more than five members were retained in order to define significant subfamilies. Unassigned sequences will be subjected to a new round of analysis as more sequences become available.

RESULTS AND DISCUSSION

Modular structure of PLs

Carbohydrate-active enzymes are frequently composed of a modular structure, in which a catalytic module carries one or more ancillary modules [29]. PLs are no exception and there is a large variety of multi-modular PLs (Figure 3). Perhaps the most common situation is the occurrence of one or more CBMs in tandem with the catalytic PL module. However, other arrangements have been observed, such as the addition of domains that promote binding to other macromolecules, including SLH (S-layer homology) domains for cell attachment [30] or dockerin modules for cellulosome assembly [31]. Some PLs may even be arranged with an additional PL module or a complementary CE (carbohydrate esterase) module, as well as domains whose function is presently unknown (termed 'X' modules; Figure 3). The number of possible combinations of domains is in principle unlimited, and their presence poses a specific challenge for sequence-based family grouping and annotation. Whole genome annotations are particularly prone to false identification (and subsequent misleading functional annotation) due to spurious hits on ancillary modules common to two distinct proteins. Consequently, a systematic excision of the ancillary modules was performed prior to all sequence alignments of PLs, and indeed this approach is the principal *modus operandi* of the CAZy classification [19,32].

Families and folds

In April 1999 there were approx. 100 PL sequences arranged in nine families [33]. Since then, the number of PL sequences has increased approx. 20-fold, essentially due to whole genome-sequencing projects. Thanks to the biochemical characterization of many novel PLs, the number of PL families has progressively grown over the years to reach 21 in 2010. The corresponding 11 years of structural biology have vastly expanded knowledge of the three-dimensional structures of PLs (for a thorough review on three-dimensional structure–function relationships of PLs, see [16]), whereas only one of the initial nine families of PLs had a structural representative in 1999, the fold of only two (PL12 and PL17) out of the 21 current PL families remain to be determined (Figure 4).

PLs show a large variety of fold types, ranging from β-helices to α/α barrels (Figure 4). The abundance of PL folds indicates that PLs have been invented more than once during evolution, from totally different scaffolds. The most extreme example of the convergent evolution of PLs is perhaps with PL1 and PL10 pectate lyases, in which the different folds carry an identically poised catalytic machinery that performs the same reaction on the same substrate [34]. The plasticity of the active site of PLs to accommodate a variety of substrates is reminiscent of that of GHs [35]. Interestingly most of the PL folds are also found in GH families, an indication of possible common evolutionary origins between the two enzyme classes.

In addition to being well-characterized at the three-dimensional level, examination of the CAZy database (http://www.cazy.org) shows that more than 10% of the PLs in the database have been biochemically (kinetically) characterized, which is the highest proportion among all of the classes of carbohydrate-active enzymes described in CAZy [GHs, GTs (glycosyltransferases), PLs and CEs]. This wealth of biochemical data indicates that most PL families group enzymes with diverse substrate specificities (Table 1). This situation has been previously observed for other CAZyme classes, especially the GHs [32] and GTs [36]. One probable explanation for this phenomenon is that the number of available protein folds is considerably smaller than the number of carbohydrate structures and hence nature has adventitiously tuned existing scaffolds for exquisite substrate specificity.

Less immediately apparent, however, are the structural similarities among the various substrates processed by individual family members. As an example, family PL8 can be considered polyspecific, as it groups together three different enzyme activities: hyaluronate lyase (EC 4.2.2.1), xanthan lyase (EC 4.2.2.12) and chondroitin AC lyase (EC 4.2.2.5). Here, the common names of these polysaccharides belie the fact that these three types of enzymes act at the same position on the same sugar, i.e. they cleave the C-O bond at position 4 of unsubstituted glucuronic acid in the backbone (Figure 5). What differentiates the three substrates is the substituent attached to the 4-oxygen of the glucuronic acid, a situation similar to, for instance, GHs that exhibit aglycone specificity [37]. The three enzymes therefore can, and in the case of PL8 do, utilize an identical catalytic machinery to cleave their respective substrates.

Table 1 Activities in PL families and subfamilies

For the taxonomical range, A, Archea; B, Bacteria; E, Eukaryota; U, unclassified; V, virus. The characterized enymes were calculated as the number of characterized enzymes/total number of enzymes in each subfamily. An enzyme was considered characterized if we could identify published direct evidence for its activity.

Family	Taxonomical range	Subfamily	Known activities	Characterized enzymes
PL1	A,B,E,U			
		1	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	4/173
		2	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	1/53
		3	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	10/38
		4	Pectin lyase (EC 4.2.2.10)	18/40
		5	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	16/38
			Exo-polygalacturonate lyase (EC 4.2.2.9)	4/38
		6	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	19/66
		7	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	7/22
		8	Pectin lyase (EC 4.2.2.10)	4/22
PL2	A,B			
		1	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	4/21
		2	Exo-polygalacturonate lyase (EC 4.2.2.9)	2/23
PL3	B,E			
		1	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	3/137
		2	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	16/115
		3	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	1/8
PL4	B,E			
		1	Rhamnogalacturonan lyase (EC 4.2.2.-)	2/21
		2		0/15
		3	Rhamnogalacturonan lyase (EC 4.2.2.-)	1/7
		4	Rhamnogalacturonan lyase (EC 4.2.2.-)	1/6
PL5	B			
		1	Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	9/27
PL6	B		Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	1/21
			Chondroitin-sulfate-ABC endolyase (EC 4.2.2.4)	1/21
PL7	B,E			
		1	Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	1/20
		2		0/7
		3	Poly(α-L-guluronate) lyase (EC 4.2.2.11)	2/8
		4		0/5
		5	Poly(α-L-guluronate) lyase (EC 4.2.2.11)	4/17
PL8	A,B			
		1	Hyaluronate lyase (EC 4.2.2.1)	11/85
		2	Chondroitin-sulfate-ABC endolyase (EC 4.2.2.20)	3/30
		3	Chondroitin AC lyase (EC 4.2.2.5)	3/7
PL9	B,E			
		1	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	6/66
			Exo-polygalacturonate lyase (EC 4.2.2.9)	2/66
		2		0/18
PL10	B			
		1	Endo-polygalacturonate (pectate) lyase (EC 4.2.2.2)	5/47
PL11	A,B,E			
		1	Rhamnogalacturonan lyase (EC 4.2.2.-)	4/64
PL12	A,B			
		1		0/35
		2	Heparin-sulfate lyase (EC 4.2.2.8)	2/10
PL13	B		Heparin lyase (EC 4.2.2.7)	2/7
PL14	B,E,V			
		1	Glucuronate lyase (EC 4.2.2.-)	1/8
		2		0/6
		3	Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	1/13
			Exo-oligo-alginate lyase (EC 4.2.2.-)	1/13
PL15	B			
		1	Oligo-alginate lyase (EC 4.2.2.-)	3/5
PL16	B,V			
		1		0/14
		2	Hyaluronate lyase (EC 4.2.2.1)	2/8
		3	Hyaluronate lyase (EC 4.2.2.1)	1/10
PL17	B			
		1		0/14
		2	Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	1/15
PL18	B		Poly(β-D-mannuronate) lyase (alginate lyase) (EC 4.2.2.3)	3/6
			Poly(α-L-guluronate) lyase (EC 4.2.2.11)	3/6
PL20	B,E		Endo-β-1,4-glucuronan lyase (EC 4.2.2.14)	1/11
PL21	B		Heparin lyase (EC 4.2.2.7)	1/9
			Heparin-sulfate lyase (EC 4.2.2.8)	1/9
PL22	A,B		Oligogalacturonate lyase (EC 4.2.2.6)	1/47

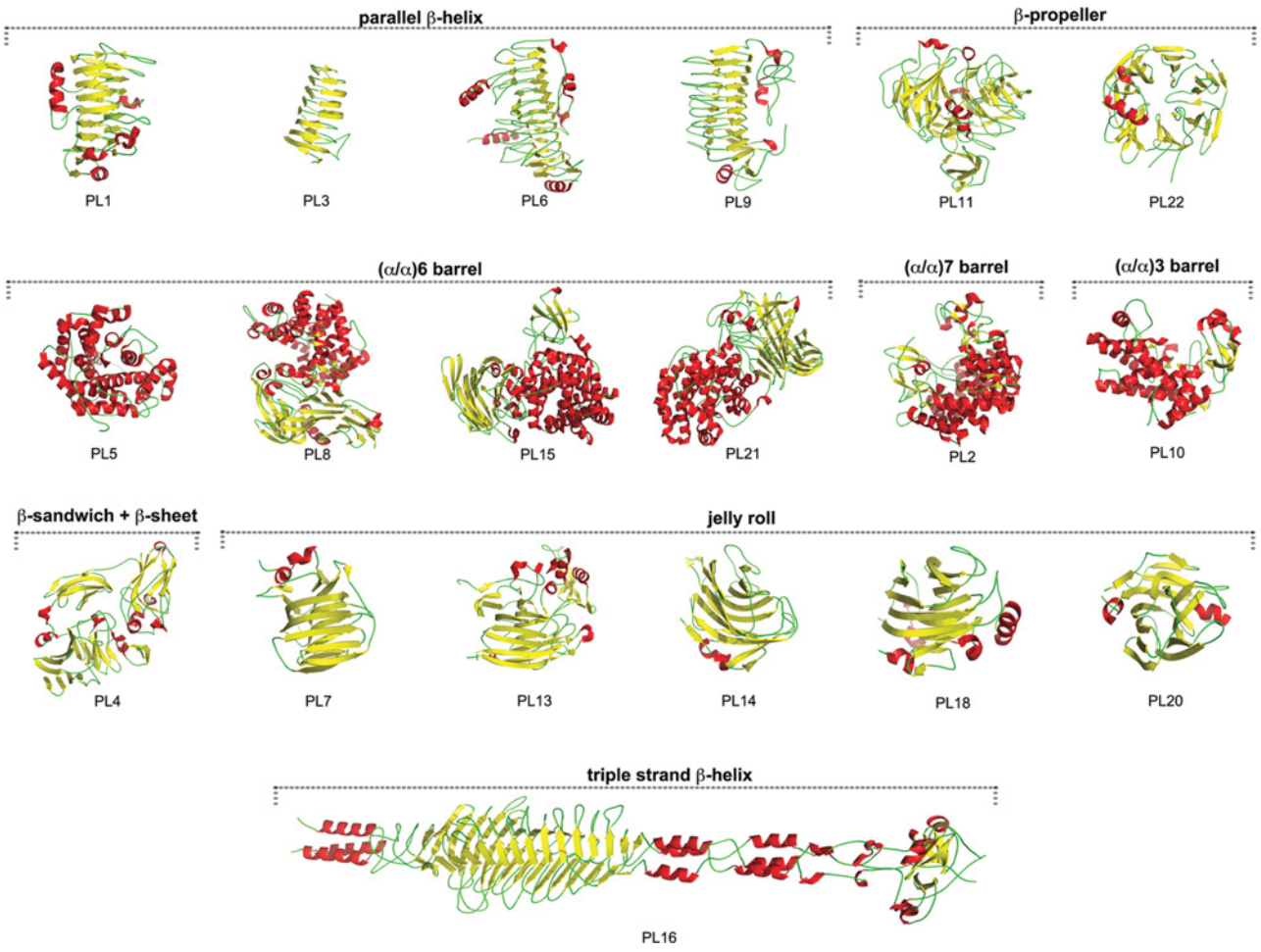


Figure 4 Folds and structures of PL families with known three-dimensional structures

Representative structures for each family are shown organized by fold. The following PDB entries are depicted: PL1 (PDB code: 2QY1), PL3 (PDB code: 1EE6), PL6 (PDB code: 1OFM), PL9 (PDB code: 1RU4), PL11 (PDB code: 2ZUY), PL22 (PDB code: 3C5M), PL5 (PDB code: 1HV6), PL8 (PDB code: 1QJM), PL15 (PDB code: 3AO0), PL21 (PDB code: 2FUQ), PL2 (PDB code: 2V8K), PL10 (PDB code: 1GXN), PL4 (PDB code: INKG), PL7 (PDB code: 1UAI), PL13 (PDB code: 3IKW), PL14 (PDB code: 3AON), PL18 (PDB code: 1J1T), PL20 (PDB code: 2ZZJ) and PL16 (PDB code: 2YW0). Within the β-propeller grouping, PL11 and PL22 members are composed of eight-bladed and seven-bladed β-propellers respectively.

Subfamilies

The functional prediction (i.e. substrate specificity) of thousands of putative carbohydrate-active enzymes derived from genome data is highly desirable, but requires a direct unequivocal relationship between sequence groupings and substrate specificity. Because the sequence-based families of PLs generally do not correlate with the fine substrate specificity, as described above, we have examined the definition of subfamilies to assess whether functional grouping and prediction could be improved. A similar approach was previously applied to the large polyspecific GH13 family of α-amylase-related enzymes, in which most of the sequence-derived subfamilies were indeed found to correspond to a single enzyme activity [25].

With the sequence data available to date, we were able to break down the 21 PL families into a total of 41 subfamilies covering 72 % of all sequences analysed (Table 1). The sequences that could not be assigned to subfamilies will most likely generate new subfamilies as more sequences become available in the future. The subfamilies are identified by an Arabic numeral following the family identifier; for instance, PL5_1 designates subfamily 1 within family PL5. As shown in Table 1, the vast majority of subfamilies have at least one representative that has been characterized with respect to substrate specificity; only seven subfamilies are lacking an experimentally characterized member. Depending on the subfamily, the cumulated biochemical characterization data varies from low (e.g. subfamilies PL3_1 and PL4_2) to high (e.g. PL1_5, PL1_6 and PL5_1). These variations can have a profound effect on any subsequent functional predictions based on subfamily membership, since reliability obviously depends (i) on the number of characterized enzymes per subfamily and (ii) on how detailed and reliably each characterization was performed.

We observe that of the 41 subfamilies identified here, 37 (90 %) appear to be monospecific, thus indicating that the subfamilies correlate with substrate specificity significantly better than at the family level. Only three subfamilies remained apparently polyspecific (i.e. grouping enzymes with different EC numbers): PL1_5, PL9_1 and PL14_3. These three subfamilies were further inspected to identify the origin of their polyspecificity. In the case of subfamilies PL1_5 and PL9_1, the apparent polyspecificity is due to the presence of both endo-acting (EC 4.2.2.2) and exo-acting (EC 4.2.2.9) polygalacturonate lyases. These two types of enzymes have exactly the same substrate specificity

Table 2 List of fully sequenced organisms with the highest number of PLs.

Organism	Reference	Number of PLs*	Description
Phytophthora infestans T30–4	[42]	67	Phytopathogenic oomycete
Phytophthora sojae	[43]	54	Phytopathogenic oomycete
Phytophthora ramorum	[43]	49	Phytopathogenic oomycete
Populus trichocarpa	[44]	39	Plant
Arabidopsis thaliana	[45]	34	Plant
Nectria haematococca mpVI	[46]	33	Phytopathogenic fungus
Saccharophagus degradans 2–40	[47]	32	Marine saprophytic bacterium
Meloidogyne incognita	[48]	30	Phytopathogenic nematode
Aspergillus oryzae RIB40	[49]	23	Phytopathogenic fungus
Aspergillus nidulans FGSC A4	[50]	21	Phytopathogenic fungus
Pedobacter heparinus DSM 2366	[51]	18	Soil bacterium
Oryza sativa Japonica Group	[52]	16	Plant
Dickeya zeae Ech1591	JGI-DOE CP001655	16	Phytopathogenic bacterium
Bacteroides thetaiotaomicron VPI-5482	[53]	16	Human gut bacterium
Cellvibrio japonicus Ueda107	[54]	14	Soil saprophytic bacterium
Streptomyces scabiei 87.22	Wellcome Trust Sanger Institute FN554889	13	Soil bacterium
Pectobacterium carotovorum subsp. *carotovorum* PC1	[55]	13	Phytopathogenic bacterium
Pectobacterium atrosepticum SCRI1043	[55]	13	Phytopathogenic bacterium
Dickeya dadantii Ech586 (*Erwinia chrysanthemi*)	JGI-DOE CP001836	13	Phytopathogenic bacterium
Actinosynnema mirum DSM 43827	[56]	13	Environmental bacterium isolated from grass

*Numbers derived from the CAZy database at http://www.cazy.org

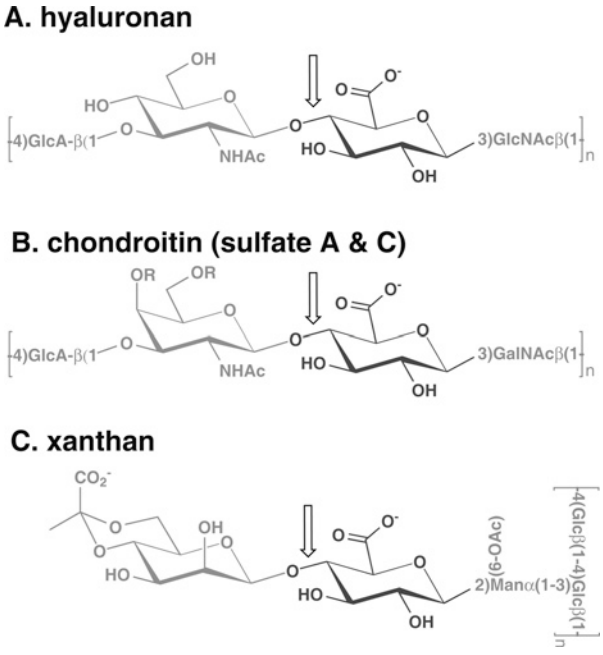


Figure 5 Structures of PL family 8 (PL8) substrates

(**A**) hyaluronan, (**B**) chondroitin (R = R′ = H) and chondroitin sulfates A (R = SO_3^-, R′ = H) and C (R = H, R′ = SO_3^-), and (**C**) xanthan. The common glucuronic acid residue is in dark grey, and the scissile bond is identified with an arrow. All monosaccharides not explicitly drawn are D-sugars in the pyranose ring form.

and differ only in the degree of polymerization of the released products. As with other types of carbohydrate-cleaving enzymes, the basis of endo- compared with exo-activity within a family is typically due to subtle details in the three-dimensional structure of the enzymes, and rigidly distinguishing the two activities can be tricky [38]. In the case of subfamily PL14_3, the apparent polyspecificity is associated with the presence of both poly- and oligo-alginate lyases (EC 4.2.2.3 and 4.2.2.- respectively). Here again, the difference is subtle: the bond cleaved is identical, and the difference in the definitions of the activities pertains to the degree of polymerization of the substrate. It may well be that such a difference is not biologically significant or, if it is, sequence data alone will never be able to sort one from the other.

Occurrence of PLs in genomes

We entered the genomic era approx. 15 years ago and the current pace of genome release is on the order of 1–2 per day. Next-generation sequencing will boost this flow of sequence data even further. Our analyses of more than 1300 genomes from diverse organisms, ranging from Archaea to higher plants and animals, show that the amount of PLs is usually low and consistently less than that of GHs (representing 3–5 % of the number of GHs). The most likely explanation for this observation is that the substrates of PLs, polysaccharides containing uronic acids, represent just a small proportion of all carbohydrate polymers. The organisms that have the largest number of PLs share a common focus: the plant cell wall. The genomes of both plants and micro-organisms that feed on living or dead plant tissue (phytopathogens or saprophytes respectively) typically encode large numbers of PLs (Table 2). The abundance of PLs in plants is due to the emergence of large multigene CAZy families [39] and the biological importance of pectin in plant development [40]. The pectic network contributes to the structural integrity of the plant cell wall and, as such, it is an obvious target for phytopathogens and symbionts (including bacteria, fungi, oomycetes and nematodes) to gain access via an arsenal of pectinolytic enzymes. And, because it is far more digestible than cellulose and lignin, pectin is also a delicacy for most saprophytic organisms, which draw nutrients from decaying plant material.

Recommendations for large-scale sequence annotation

Next-generation sequencing machines will deliver ever more sequences, whose utility largely depends on our ability to correlate them with molecular functions. The hierarchical classification

that we advocate here, based on fold, family and subfamily, provides a convenient way to produce the best possible functional assignments that take into account distance with experimentally characterized enzymes.

At the most general end of the spectrum, very distant similarity [such as that resulting from PSI-BLAST analyses or the use of degenerate HMMs (hidden Markov models)] should be used only to assign a protein to a folding class and not to a function. For example, Stam et al. [41] have shown that a PSI-BLAST search starting with a β-helical polygalacturonase (EC 3.2.1.15) of family GH28 retrieved β-helical pectate lyases of family PL1 and dextranases of family GH49 after only two iterations [41], despite the fact that these two enzymes employ distinctly different catalytic mechanisms. Although this may reflect ancient evolutionary events, the detection of such distant similarities is of little use when it comes to anticipating a molecular function.

The next level is the assignment to a family. This is typically reflected by significant BLAST scores over the entire length of the catalytic module (not the entire protein, which may contact multiply ancillary modules, see above). Here, similarity is sufficient to predict a global PL function, especially if the catalytic residues are conserved in the sequence under consideration. Even though the PL families are often polyfunctional, commonalities between the various substrates known to be cleaved by family members can guide experimental design to determine the actual specificity of novel enzymes.

Finally, the most fine-grained annotation is reached at the other end of the spectrum when a sequence can be assigned to one of the defined PL subfamilies. Two cases will arise. (i) The subfamily to which the sequence can be assigned contains one or several experimentally characterized members (the more the better). Here the function of the query protein can reasonably be assigned, for instance 'putative hyaluronate lyase'. (ii) The query protein belongs to a non-characterized subfamily, or does not belong to any defined subfamily. Here the precise substrate cannot be predicted with confidence, and the best possible annotation is simply 'putative polysaccharide lyase'.

One consequence of the above hierarchy is that functional predictions should be dynamic, varying as biochemical data accumulates in the various subfamilies. Additionally, we suggest that an EC number should only be assigned to the query protein and included in public databases when, and only when, the precise substrate specificity has been established experimentally, to avoid unchecked propagation of erroneous assignments. In general, we advocate a conservative approach to functional assignment based on sequence analysis, guided by the mantra that no annotation is better than a misleading annotation.

AUTHOR CONTRIBUTION

Vincent Lombard performed the analysis of PL sequences; Thomas Bernard, Corinne Rancurel and Pedro Coutinho constructed computer tools to analyse the PL sequences; Harry Brumer, Pedro Coutinho and Bernard Henrissat reviewed data; Bernard Henrissat and Pedro Coutinho designed research; Vincent Lombard, Harry Brumer and Bernard Henrissat wrote the paper.

FUNDING

This work was supported by Novozymes; the Swedish Research Council Formas, the Swedish Research Council, *Vetenskapsrådet* and the Wallenberg Wood Science Center (to H.B.).

REFERENCES

1 Yip, V. L. and Withers, S. G. (2006) Breakdown of oligosaccharides by the process of elimination. Curr. Opin. Chem. Biol. **10**, 147–155
2 Sutherland, I. W. (1995) Polysaccharide lyases. FEMS Microbiol. Rev. **16**, 323–347
3 Yip, V. L. Y. and Withers, S. G. (2004) Nature's many mechanisms for the degradation of oligosaccharides. Org. Biomol. Chem. **2**, 2707–2713
4 Abbott, D. W. and Boraston, A. B. (2008) Structural biology of pectin degradation by Enterobacteriaceae. Microbiol. Mol. Biol. Rev. **72**, 301–316
5 Gacesa, P. (1987) Alginate-modifying enzymes: a proposed unified mechanism of action for the lyases and epimerases. FEBS Lett. **212**, 199–202
6 Girish, K. S. and Kemparaju, K. (2007) The magic glue hyaluronan and its eraser hyaluronidase: a biological overview. Life Sci. **80**, 1921–1943
7 Herron, S. R., Benen, J. A. E., Scavetta, R. D., Visser, J. and Jurnak, F. (2000) Structure and function of pectic enzymes: Virulence factors of plant pathogens. Proc. Natl. Acad. Sci. U.S.A. **97**, 8762–8769
8 Vorwerk, S., Somerville, S. and Somerville, C. (2004) The role of plant cell wall polysaccharide composition in disease resistance. Trends Plant Sci. **9**, 203–209
9 Linhardt, R. J., Galliher, P. M. and Cooney, C. L. (1986) Polysaccharide lyases. Appl. Biochem. Biotechnol. **12**, 135–176
10 Wong, T. Y., Preston, L. A. and Schiller, N. L. (2000) Alginate lyase: review of major sources and enzyme characteristics, structure-function analysis, biological roles, and applications. Annu. Rev. Microbiol. **54**, 289–340
11 Hatch, R. A. and Schiller, N. L. (1998) Alginate lyase promotes diffusion of aminoglycosides through the extracellular polysaccharide of mucoid *Pseudomonas aeruginosa*. Antimicrob. Agents Chemother. **42**, 974–977
12 Cescutti, P., Scussolin, S., Herasimenka, Y., Impallomeni, G., Bicego, M. and Rizzo, R. (2006) First report of a lyase for cepacian, the polysaccharide produced by *Burkholderia cepacia* complex bacteria. Biochem. Biophys. Res. Commun. **339**, 821–826
13 Yip, V. L. Y., Varrot, A., Davies, G. J., Rajan, S. S., Yang, X. J., Thompson, J., Anderson, W. F. and Withers, S. G. (2004) An unusual mechanism of glycoside hydrolysis involving redox and elimination steps by a family 4 β-glycosidase from *Thermotoga maritima*. J. Am. Chem. Soc. **126**, 8354–8355
14 Rye, C. S., Matte, A., Cygler, M. and Withers, S. G. (2006) An atypical approach identifies TYR234 as the key base catalyst in chondroitin AC lyase. ChemBioChem **7**, 631–637
15 Shaya, D., Hahn, B. S., Bjerkan, T. M., Kim, W. S., Park, N. Y., Sim, J. S., Kim, Y. S. and Cygler, M. (2008) Composite active site of chondroitin lyase ABC accepting both epimers of uronic acid. Glycobiology **18**, 270–277
16 Garron, M. and Cygler, M. (2010) Structural and mechanistic classification of polysaccharide lyases. Glycobiology, doi:10.1093/glycob/cwq122
17 Boraston, A. B., Bolam, D. N., Gilbert, H. J. and Davies, G. J. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. **382**, 769–781
18 Guillen, D., Sanchez, S. and Rodriguez-Sanoja, R. (2010) Carbohydrate-binding domains: multiplicity of biological roles. Appl. Microbiol. Biotechnol. **85**, 1241–1249
19 Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. **37**, D233–D238
20 Lee, S. S., Yu, S. and Withers, S. G. (2003) Detailed dissection of a new mechanism for glycoside cleavage: α-1,4-glucan lyase. Biochemistry **42**, 13081–13090
21 Holtje, J. V., Mirelman, D., Sharon, N. and Schwarz, U. (1975) Novel type of murein transglycosylase in *Escherichia coli*. J. Bacteriol. **124**, 1067–1076
22 Blackburn, N. T. and Clarke, A. J. (2001) Identification of four families of peptidoglycan lytic transglycosylases. J. Mol. Evol. **52**, 78–84
23 Saito, K., Sumita, Y., Nagasaka, Y., Tomita, F. and Yokota, A. (2003) Molecular cloning of the gene encoding the di-D-fructofuranose 1,2′: 2,3′ dianhydride hydrolysis enzyme (DFA IIIase) from *Arthrobacter* sp H65–7. J. Biosci. Bioeng. **95**, 538–540
24 Sakurai, H., Yokota, A., Sumita, Y., Mori, Y., Matsui, H. and Tomita, F. (1997) Metabolism of DFA III by *Arthrobacter* sp. H65–7: purification and properties of a DFA III hydrolysis enzyme (DFA IIIase). Biosci. Biotechnol. Biochem. **61**, 989–993
25 Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M. and Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Eng. Des. Sel. **19**, 555–562
26 Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792–1797
27 Henikoff, S. and Henikoff, J. G. (1992) Amino-acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. **89**, 10915–10919
28 Wicker, N., Perrin, G. R., Thierry, J. C. and Poch, O. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol. **18**, 1435–1441
29 Gilbert, H. J. (2010) The biochemistry and structural biology of plant cell wall deconstruction. Plant Physiol. **153**, 444–455
30 Fouet, A. and Mesnage, S. (2002) *Bacillus anthracis* cell envelope components. Curr. Top. Microbiol. Immunol. **271**, 87–113

31 Fontes, C. M. and Gilbert, H. J. (2010) Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. Annu. Rev. Biochem. **79**, 655–681

32 Henrissat, B. and Davies, G. J. (2000) Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. Plant Physiol. **124**, 1515–1519

33 Coutinho, P. M. and Henrissat, B. (1999) Carbohydrate-active enzymes: an integrated database approach. In Recent Advances in Carbohydrate Bioengineering, pp. 3–12, The Royal Society of Chemistry, Cambridge

34 Charnock, S. J., Brown, I. E., Turkenburg, J. P., Black, G. W. and Davies, G. J. (2002) Convergent evolution sheds light on the anti-β-elimination mechanism common to family 1 and 10 polysaccharide lyases. Proc. Natl. Acad. Sci. U.S.A. **99**, 12067–12072

35 Davies, G. and Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases. Structure **3**, 853–859

36 Coutinho, P. M., Deleury, E., Davies, G. J. and Henrissat, B. (2003) An evolving hierarchical family classification for glycosyltransferases. J. Mol. Biol. **328**, 307–317

37 Czjzek, M., Cicek, M., Zamboni, V., Bevan, D. R., Henrissat, B. and Esen, A. (2000) The mechanism of substrate (aglycone) specificity in β-glucosidases is revealed by crystal structures of mutant maize β-glucosidase-DIMBOA, -DIMBOAGlc, and -dhurrin complexes. Proc. Natl. Acad. Sci. U.S.A. **97**, 13555–13560

38 Armand, S., Drouillard, S., Schulein, M., Henrissat, B. and Driguez, H. (1997) A bifunctionalized fluorogenic tetrasaccharide as a substrate to study cellulases. J. Biol. Chem. **272**, 2709–2713

39 Coutinho, P. M., Starn, M., Blanc, E. and Henrissat, B. (2003) Why are there so many carbohydrate-active enzyme-related genes in plants? Trends Plant Sci. **8**, 563–565

40 Caffall, K. H. and Mohnen, D. (2009) The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. Carbohydr. Res. **344**, 1879–1900

41 Stam, M. R., Blanc, E., Coutinho, P. M. and Henrissat, B. (2005) Evolutionary and mechanistic relationships between glycosidases acting on α- and β-bonds. Carbohydr. Res. **340**, 2728–2734

42 Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., Grabherr, M., Kodira, C. D., Raffaele, S., Torto-Alalibo, T. et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature **461**, 393–398

43 Tyler, B. M., Tripathy, S., Zhang, X. M., Dehal, P., Jiang, R. H. Y., Aerts, A., Arredondo, F. D., Baxter, L., Bensasson, D., Beynon, J. L. et al. (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science **313**, 1261–1266

44 Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**, 1596–1604

45 Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H. M., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R. et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301**, 653–657

46 Coleman, J. J., Rounsley, S. D., Rodriguez-Carres, M., Kuo, A., Wasmann, C. C., Grimwood, J., Schmutz, J., Taga, M., White, G. J., Zhou, S. G. et al. (2009) The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. **5**, e1000618

47 Weiner, R. M., Taylor, L. E., Henrissat, B., Hauser, L., Land, M., Coutinho, P. M., Rancurel, C., Saunders, E. H., Longmire, A. G., Zhang, H. T. et al. (2008) Complete genome sequence of the complex carbohydrate-degrading marine bacterium, *Saccharophagus degradans* strain 2-40(T). PLoS Genet. **4**, 13

48 Abad, P., Gouzy, J., Aury, J. M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C. et al. (2008) Genome sequence of the metazoan plant-specific nematode *Meloidogyne incognita*. Nat. Biotechnol. **26**, 909–915

49 Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K. I., Arima, T., Akita, O., Kashiwagi, Y. et al. (2005) Genome sequencing and analysis of *Aspergillus oryzae*. Nature **438**, 1157–1161

50 Galagan, J. E., Calvo, S. E., Cuomo, C., Ma, L. J., Wortman, J. R., Batzoglou, S., Lee, S. I., Basturkmen, M., Spevak, C. C., Clutterbuck, J. et al. (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A.fumigatus* and *A.oryzae*. Nature **438**, 1105–1115

51 Han, C., Spring, S., Lapidus, A., Glavina Del Rio, T., Tice, H., Copeland, A., Cheng, J.-F., Lucas, S., Chen, F., Nolan, M. et al. (2009) Complete genome sequence of *Pedobacter heparinus* type strain (HIM 762–3T). Stand. Genomic Sci. **1**, 54–62

52 Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science **296**, 92–100

53 Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. and Gordon, J. I. (2003) A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. Science **299**, 2074–2076

54 Deboy, R. T., Mongodin, E. F., Fouts, D. E., Tailford, L. E., Khouri, H., Emerson, J. B., Mohamoud, Y., Watkins, K., Henrissat, B., Gilbert, H. J. and Nelson, K. E. (2008) Insights into plant cell wall degradation from the genome sequence of the soil bacterium *Cellvibrio japonicus*. J. Bacteriol. **190**, 5455–5463

55 Bell, K. S., Sebaihia, M., Pritchard, L., Holden, M. T. G., Hyman, L. J., Holeva, M. C., Thomson, N. R., Bentley, S. D., Churcher, L. J. C., Mungall, K. et al. (2004) Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp *atroseptica* and characterization of virulence factors. Proc. Natl. Acad. Sci. U.S.A. **101**, 11105–11110

56 Land, M., Lapidus, A., Mayilraj, S., Chen, F., Copeland, A., Glavina Del Rio, T., Nolan, M., Lucas, S., Tice, H., Cheng, J.-F. et al. (2009) Complete genome sequence of *Actinosynnema mirum* type strain (101T). Stand. Genomic Sci. **1**, 46–53

Received 3 August 2010/1 October 2010; accepted 7 October 2010
Published as BJ Immediate Publication 7 October 2010, doi:10.1042/BJ20101185

B

Nat Biotechnol. (2010);**28**(9):957-963.

Genome sequence of the model mushroom Schizophyllum commune.

Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, de Vries RP, Record E, Levasseur A, Baker SE, Bartholomew KA, Coutinho PM, Erdmann S, Fowler TJ, Gathman AC, Lombard V, Henrissat B, Knabe N, Kües U, Lilly WW, Lindquist E, Lucas S, Magnuson JK, Piumi F, Raudaskoski M, Salamov A, Schmutz J, Schwarze FW, vanKuyk PA, Horton JS, Grigoriev IV, Wösten HA.

Abstract :Much remains to be learned about the biology of mushroom-forming fungi, which are an important source of food, secondary metabolites and industrial enzymes. The wood-degrading fungus Schizophyllum commune is both a genetically tractable model for studying mushroom development and a likely source of enzymes capable of efficient degradation of lignocellulosic biomass. Comparative analyses of its 38.5-megabase genome, which encodes 13,210 predicted genes, reveal the species's unique wood-degrading machinery. One-third of the 471 genes predicted to encode transcription factors are differentially expressed during sexual development of S. commune. Whereas inactivation of one of these, fst4, prevented mushroom formation, inactivation of another, fst3, resulted in more, albeit smaller, mushrooms than in the wild-type fungus. Antisense transcripts may also have a role in the formation of fruiting bodies. Better insight into the mechanisms underlying mushroom formation should affect commercial production of mushrooms and their industrial use for producing enzymes and pharmaceuticals.

PMID: 20622885

nature biotechnology

Genome sequence of the model mushroom *Schizophyllum commune*

Robin A Ohm[1], Jan F de Jong[1], Luis G Lugones[1], Andrea Aerts[2], Erika Kothe[3], Jason E Stajich[4], Ronald P de Vries[1,5], Eric Record[6,7], Anthony Levasseur[6,7], Scott E Baker[2,8], Kirk A Bartholomew[9], Pedro M Coutinho[10], Susann Erdmann[3], Thomas J Fowler[11], Allen C Gathman[12], Vincent Lombard[10], Bernard Henrissat[10], Nicole Knabe[3,18], Ursula Kües[13], Walt W Lilly[12], Erika Lindquist[2], Susan Lucas[2], Jon K Magnuson[8], François Piumi[6,7], Marjatta Raudaskoski[14], Asaf Salamov[2], Jeremy Schmutz[2], Francis W M R Schwarze[15], Patricia A vanKuyk[16], J Stephen Horton[17], Igor V Grigoriev[2] & Han A B Wösten[1]

Much remains to be learned about the biology of mushroom-forming fungi, which are an important source of food, secondary metabolites and industrial enzymes. The wood-degrading fungus *Schizophyllum commune* is both a genetically tractable model for studying mushroom development and a likely source of enzymes capable of efficient degradation of lignocellulosic biomass. Comparative analyses of its 38.5-megabase genome, which encodes 13,210 predicted genes, reveal the species's unique wood-degrading machinery. One-third of the 471 genes predicted to encode transcription factors are differentially expressed during sexual development of *S. commune*. Whereas inactivation of one of these, *fst4*, prevented mushroom formation, inactivation of another, *fst3*, resulted in more, albeit smaller, mushrooms than in the wild-type fungus. Antisense transcripts may also have a role in the formation of fruiting bodies. Better insight into the mechanisms underlying mushroom formation should affect commercial production of mushrooms and their industrial use for producing enzymes and pharmaceuticals.

The importance of mushroom-forming fungi in agriculture, human health and ecology underscores their biotechnological potential for a wide range of applications. The most conspicuous forms of these species, most of which are basidiomycetes, are their fleshy, spore-bearing fruiting bodies. Although these are primarily of economic value because of their use as food[1,2] (worldwide production of edible mushrooms amounts to ~2.5 million tons annually), mushrooms also produce anti-tumor and immunostimulatory molecules[1,2], as well as enzymes used for bioconversions[3]. Moreover, they have been identified as promising cell factories for the production of pharmaceutical proteins[4].

Despite their economic importance, relatively little is known about how mushroom-forming fungi obtain nutrients and how their fruiting bodies are formed. The vast majority of mushroom-forming fungi cannot be genetically modified, or even cultured under laboratory conditions. The basidiomycete *Schizophyllum commune*, which completes its life cycle in ~10 d, is a notable exception insofar as it can be cultured on defined media and there are a wealth of molecular tools to study its growth and development. It is the only mushroom-forming fungus for which genes have been inactivated by homologous recombination. The importance of *S. commune* as a model system is also exemplified by the fact that its recombinant DNA constructs will express in other mushroom-forming fungi[5]. In contrast, constructs that have been developed for ascomycetes are often not functional in mushroom-forming basidiomycetes.

S. commune is one of the most commonly found fungi and can be isolated from all continents, except for Antarctica. *S. commune* has been reported to be a pathogen of humans and trees, but it mainly adopts a saprobic lifestyle by causing white rot[6]. It is predominantly found on fallen branches and timber of deciduous trees. At least 150 genera of woody plants are substrates for *S. commune*, but it also colonizes soft-wood and grass silage[7]. The mushrooms of *S. commune* that form on these substrates are used as a food source in Africa and Asia.

In the life cycle of *S. commune*[8], meiospores germinate to form a sterile monokaryotic mycelium, in which each hyphal compartment

[1]Department of Microbiology and Kluyver Centre for Genomics of Industrial Fermentation, Utrecht University, Utrecht, The Netherlands. [2]Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [3]Department of Microbiology, Friedrich Schiller University, Jena, Germany. [4]Department of Plant Pathology and Microbiology, University of California, Riverside, California, USA. [5]CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands. [6]INRA, Biotechnologie des Champignons Filamenteux, Marseille, France. [7]Universités Aix-Marseille I & II, Marseille, France. [8]Chemical and Biological Process Development Group, Pacific Northwest National Laboratory, Richland, Washington, USA. [9]Biology Department, Sacred Heart University, Fairfield, Connecticut, USA. [10]Architecture et Fonction des Macromolecules Biologiques, Université Aix-Marseille I & II, Marseille, France. [11]Department of Biological Sciences, Southern Illinois University, Edwardsville, Illinois, USA. [12]Department of Biology, Southeast Missouri State University, Cape Girardeau, Missouri, USA. [13]Division of Molecular Wood Biotechnology and Technical Mycology, Büsgen-Institute, University of Göttingen, Göttingen, Germany. [14]Department of Biochemistry and Food Chemistry, University of Turku, Biocity A, Turku, Finland. [15]Wood Protection & Biotechnology, Empa, Swiss Federal Laboratories for Materials Testing and Research, St. Gallen, Switzerland. [16]Molecular Microbiology, Institute of Biology, Leiden University, Leiden, The Netherlands. [17]Department of Biological Sciences, Union College, Schenectady, New York, USA. [18]Present address: Dartmouth Medical School, Hanover, New Hampshire, USA. Correspondence should be addressed to J.S.H. (hortons@union.edu), I.V.G. (IVGrigoriev@lbl.gov) or H.A.B.W. (h.a.b.wosten@uu.nl).

Received 14 April; accepted 12 May; published online 11 July 2010; doi:10.1038/nbt.1643

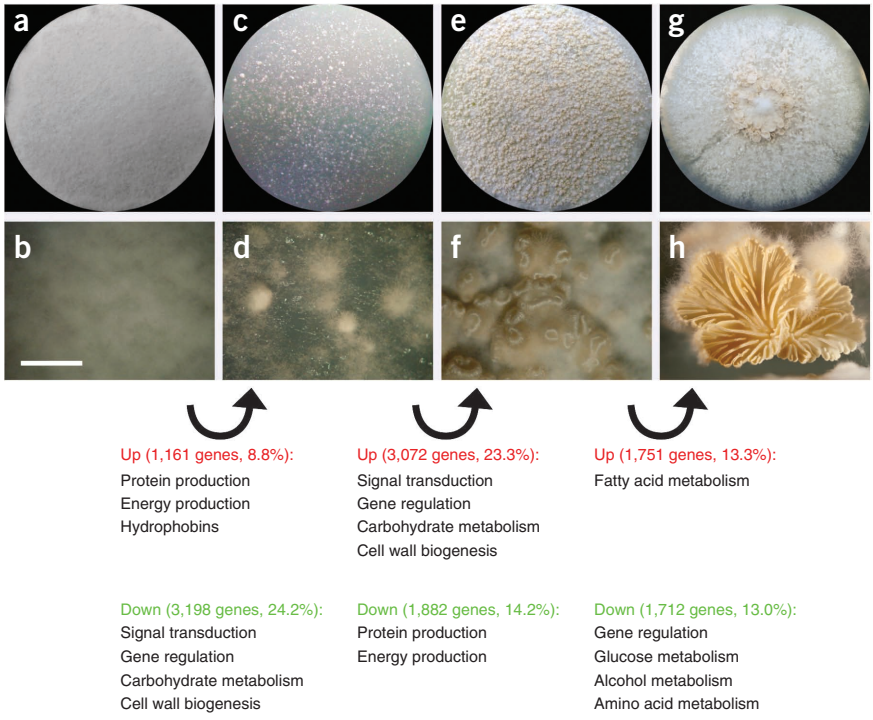


Figure 1 Development of *S. commune*. (**a**–**h**) Four-day-old (**a**–**f**) and 8-day-old (**g**,**h**) colonies grown from homogenates illustrate typical developmental stages in the life cycle of *S. commune*. A monokaryon generates sterile aerial hyphae that form a fluffy white layer on top of the vegetative mycelium (**a**,**b**). Aerial hyphae of a dikaryon interact with each other to form stage I aggregates (**c**,**d**), which, after a light stimulus, develop into stage II primordia (**e**,**f**). These primoridia further differentiate into sporulating mushrooms (**g**,**h**). Enrichment analysis shows that particular functional terms are over-represented in genes that are up- or downregulated during a developmental transition. These terms are indicated below the panels. **a**,**c**,**e**,**g** represent cultures grown in 9-cm Petri dishes, whereas **b**,**d**,**f**,**h** represent magnifications thereof. Scale bar, 1 cm (**h**), 2.5 mm (**b**,**d**) and 5 mm (**f**).

contains one nucleus. Initial growth of this mycelium occurs beneath the surface of the substrate, with formation of aerial hyphae a few days after germination (**Fig. 1a**,**b**). Monokaryons that encounter each other fuse, and a fertile dikaryon forms when the alleles of the mating-type loci *matA* and *matB* of the partners differ. A short exposure to light is essential for fruiting, whereas a high concentration of carbon dioxide and high temperatures (30–37 °C) are inhibitory. Mushroom formation is initiated with the aggregation of aerial dikaryotic hyphae. These aggregates (**Fig. 1c**,**d**) form fruiting-body primordia (**Fig. 1e**,**f**), which further develop into mature fruiting bodies (**Fig. 1g**,**h**). Karyogamy and meiosis occur in the basidia within the mature fruiting body, and the resulting basidiospores can give rise to new monokaryotic mycelia.

Here we report the genomic sequence of the monokaryotic *S. commune* strain H4-8 and illustrate the potential of this basidiomycete as a model system to study mushroom formation. Besides the importance of understanding the sexual reproduction of *S. commune* for the commercial production of mushrooms, insight into the basis of this species' capacity to degrade lignocellulose may inspire more effective strategies to degrade lignocellulosic feedstocks for biofuel production.

RESULTS

The genome of *S. commune*

Sequencing of the genomic DNA of *S. commune* strain H4-8 with 8.29× coverage (**Supplementary Table 1**) revealed a 38.5-megabase genome assembly with 11.2% repeat content (**Supplementary Results 1**). The assembly is contained on 36 scaffolds (**Supplementary Table 2**), which represent 14 chromosomes[9]. We predict 13,210 gene models, with 42% supported by expressed sequenced tags (ESTs) and 69% similar to proteins from other organisms (**Supplementary Tables 3** and **4**). Clustering of the proteins of *S. commune* with those of other sequenced fungi (a phylogenetic tree of the organisms used in the analysis is shown in **Supplementary Fig. 1**) identifies 7,055 groups containing at least one *S. commune* protein (**Supplementary Table 5**). Analysis of these clusters suggested that 39% of the *S. commune* proteins have orthologs in the Dikarya and are thus conserved in the Basidiomycota and Ascomycota (**Supplementary Table 6**). Notably, a similar percentage of proteins (36%) are unique to *S. commune*, as based on OrthoMCL analysis. Of these proteins, 46% have at least one inparalog (a gene resulting from a duplication within the genome) in *S. commune*. The uniqueness of the *S. commune* proteome is also illustrated by the over- and under-representation of protein family (PFAM) domains compared to other fungi (**Supplementary Results 2**) and the fact that only 43% of the predicted genes (5,703 out of the 13,210) could be annotated with a gene ontology (GO) term.

Global gene expression analysis

We used massively parallel signature sequencing (MPSS) to compare whole-genome expression at the four developmental stages, defined by monokaryons, stage I aggregates, stage II primordia and mature fruiting bodies (**Fig. 1**). The majority of genes are either expressed in all four stages (4,859 genes) or not expressed in any of them (5,308 genes) (**Fig. 2** and **Supplementary Table 7**). Of the 13,210 predicted genes, 59.8% are expressed in at least one developmental stage (**Supplementary Table 7**). Fewer of the unique *S. commune* genes meet this criterion, whereas a higher percentage was observed for genes that share orthologs with Agaricomycetes or more distant fungi (**Supplementary Table 6**). This suggests that *S. commune* genes lacking homology to any reported sequences are more stringently regulated than orthologs of genes reported for other species. This is consistent with the observation that genes that are apparently unique to *S. commune* are over-represented in the pool of genes that are differentially expressed during the four developmental stages studied (**Supplementary Tables 8** and **9**).

Antisense transcription is a widespread phenomenon in *S. commune* (**Fig. 2b**,**c**). Of the tags that could be related to a gene model, 18.7% originate from an antisense transcript; and 42.3% of the predicted genes have antisense expression during one or more of the four developmental stages studied (**Supplementary Tables 7** and **10**). Northern hybridization with strand-specific probes confirmed the existence of antisense transcripts of *sc4* (DOE JGI Protein ID 73533; data not shown). Whereas a relatively large number of genes expressed in the antisense direction are uniquely expressed in stage II (2,888 genes), relatively few genes are expressed in the antisense direction in all stages (1,195 genes) (**Fig. 2b**). Our data suggest that 4,302 genes are expressed

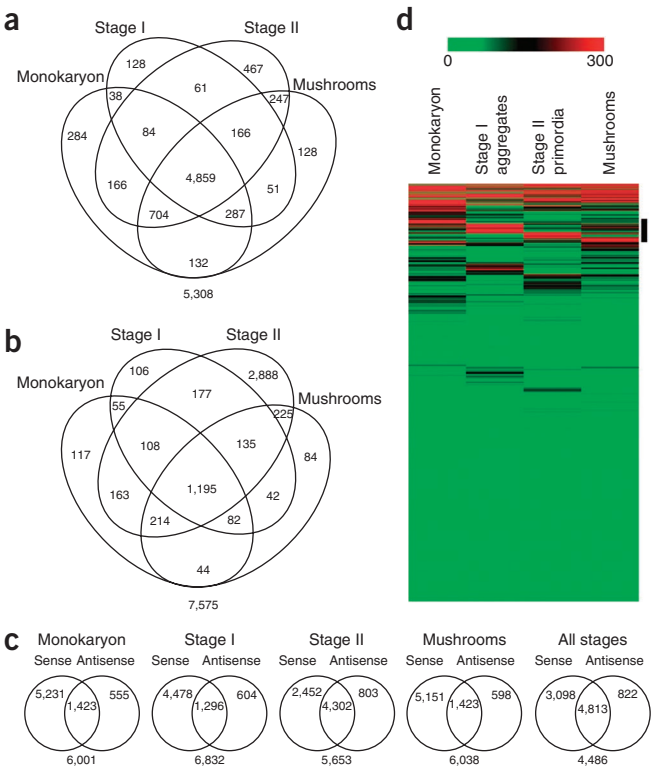


Figure 2 Gene expression in four developmental stages of *S. commune*. (**a**,**b**) The cutoff for expression is 4 tags per million (TPM). Venn diagrams show the overlap of genes expressed in the sense (**a**) and antisense (**b**) directions in the four developmental stages. For example, **a** shows that 61 genes are expressed in the sense direction in stage I and stage II, 4,859 genes are expressed in the sense direction in all stages, 132 genes are expressed in the sense direction in the monokaryon and mature fruiting bodies, and 5,308 genes are not expressed in the sense direction in any of the stages. (**c**) Venn diagrams of the overlap in genes that show sense and antisense expression in each developmental stage, and in all stages combined. (**d**) Heat map of expression of the *S. commune* genes in the four developmental stages. The bar at the top of the panel represents expression values between 0 and 300 TPM. Genes with expression values >300 TPM are also indicated in red. The bar on the right indicates a cluster of 366 highly expressed and differentially regulated genes. Annotation information for the genes in this cluster is given in **Supplementary Table 18**.

in both the sense and antisense directions during stage II (**Fig. 2c**). This overlap is larger for genes expressed during this phase of the life cycle than for the other developmental stages studied.

Fruiting-body development

We performed an enrichment analysis of functional annotation for the expression profiles of the developmental stages defined by monokaryons, stage I aggregates, stage II primordia and mature fruiting bodies. Functional terms involved in protein or energy production, or associated with hydrophobins, are over-represented in genes upregulated during formation of stage I aggregates (**Fig. 1** and **Supplementary Table 9**). Genes involved in signal transduction, regulation of gene expression, cell wall biogenesis and carbohydrate metabolism are enriched in the group of genes downregulated during the formation of stage I aggregates. These functional terms are enriched in the upregulated genes during formation of stage II primordia, whereas terms involved in protein and energy production are enriched in the downregulated genes (**Fig. 1** and **Supplementary Table 9**). Genes encoding transcription factors and genes involved in amino acid, glucose and alcohol metabolism are enriched in the group of genes downregulated during the formation of mature fruiting bodies.

As whole-genome expression was previously analyzed during mushroom formation in *Laccaria bicolor*[10], we next investigated whether the regulation of orthologous gene pairs of *L. bicolor* and *S. commune* might be correlated during fruiting. When we compared microarray expression profiles of free-living mycelium and mature fruiting bodies of *L. bicolor* to the MPSS expression profiles of monokaryotic mycelium and mature fruiting bodies of *S. commune*, we found that 6,751 expressed genes from *S. commune* had at least one expressed ortholog in *L. bicolor*. We determined the correlation of changes in expression of the functional annotation terms to which these orthologous pairs belong. There were 15 gene ontology terms, 2 KEGG terms, 4 KOG terms and 4 PFAM terms that showed a positive correlation in expression ($P < 0.01$; **Supplementary Table 11**). These terms include metabolic pathways (such as valine, leucine and isoleucine biosynthesis) and regulatory mechanisms (such as transcriptional regulation by transcription factors and signal transduction by G-protein α subunit). This indicates that regulation of these processes during mushroom formation is conserved in *S. commune* and *L. bicolor*.

Analysis of the matA and matB gene loci

Formation of a fertile dikaryon is regulated by the *matA* and *matB* mating-type loci. Proteins encoded in these loci activate signaling cascades (**Supplementary Results 3**) upstream of target genes. The target genes include those encoding enzymes and proteins that fulfill structural functions, such as hydrophobins (**Supplementary Results 4**), needed for the formation of fruiting bodies.

The *matA* locus of *S. commune* strain H4-8 appears to have more homeodomain genes than any fungal mating-type locus described thus far. This locus consists of two subloci, *Aα* and *Aβ*, which are separated by 550 kilobases (kb) on chromosome I of strain H4-8.

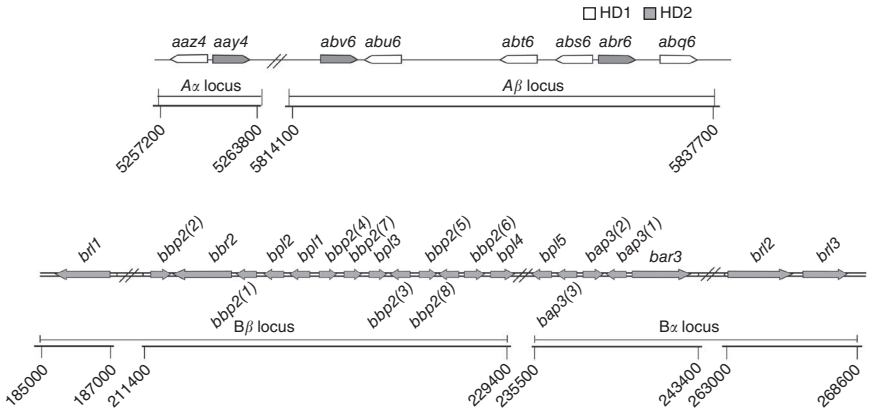


Figure 3 Distribution of genes encoding HD1 and HD2 homeodomain proteins in the *matA* locus and genes encoding pheromone receptors and pheromones in the *matB* locus of *S. commune* strain H4-8. The *matA* and *matB* loci are positioned on scaffolds 1 and 10, respectively. We identified an additional pheromone receptor gene, *brl4*, on scaffold 8.

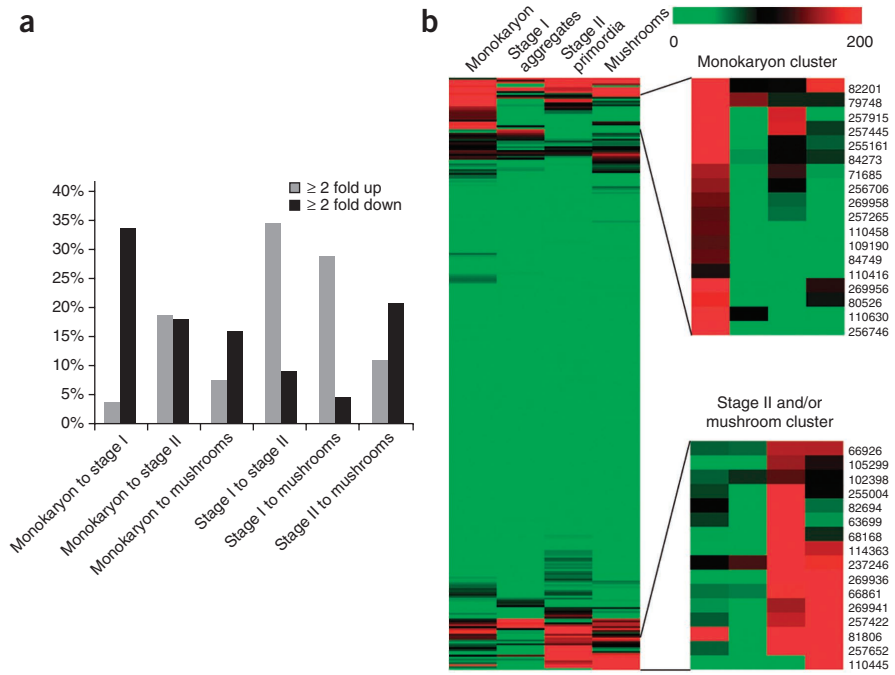


Figure 4 Expression of the 471 transcription factors in the genome of *S. commune*. (**a**) The histogram shows the percentage of transcription factor genes that are differentially expressed between stages of development. (**b**) The heat map shows a cluster containing predominantly monokaryon-specific transcription factors and a cluster containing predominantly stage II- and/or mushroom-specific transcription factors. These clusters are enlarged to the right of the heat map. The latter group contains two fungus-specific transcription factor genes, *fst3* and *fst4*.

Annotation revealed that the *Aα* locus of H4-8 contains two divergently transcribed genes, which encode the Y and Z homeodomain proteins of the HD2 and HD1 classes, respectively (**Fig. 3** and **Supplementary Table 12**). These two genes, *aay4* and *aaz4*, have been described previously[1]. A homeodomain gene has also been identified previously in the *Aβ* locus of H4-8 (ref. 11). Our genomic sequence revealed that this locus actually contains six predicted homeodomain genes: *abq6* (HD1), *abr6* (HD2), *abs6* (HD1), *abt6* (HD1, but lacking the nuclear localization signal), *abu6* (HD1) and *abv6* (HD2) (**Fig. 3** and **Supplementary Table 12**).

Annotation of the genomic sequence of *S. commune* reveals that the *matB* system contains more genes than previously envisioned. The *matB* locus comprises two linked loci, *Bα* and *Bβ*, which both encode pheromones and pheromone receptors[1] (**Fig. 3**). Previously, one pheromone receptor gene was identified in both *Bα3* and *Bβ2* of strain H4-8 (called *bar3* and *bbr2*, respectively)[12]. The genome sequence of *S. commune* reveals four additional genes with high sequence similarity to these pheromone receptor genes, which we call *B* receptor–like genes 1 to 4 (*brl1* to *brl4*; **Fig. 3**). Three of these genes are located near *bar3* and *bbr2* on scaffold 10, whereas one (*brl4*) is located on scaffold 8. MPSS analysis shows that the *brl* genes are expressed (**Supplementary Table 13**). In fact, of all receptor and receptor-like genes, *brl3* shows the highest expression under the conditions tested.

Three and eight pheromone genes have previously been identified at the *Bα3* and *Bβ2* loci, respectively[13]. We identified one additional pheromone gene, named *B* pheromone–like-5 (*bpl5*), at the *Bα3* locus. Moreover, four additional pheromone-like genes were detected at the *Bβ2* locus, called *bpl1* to *bpl4* (**Fig. 3**). Of these, only *bpl2* showed no expression in MPSS analysis (**Supplementary Table 13**). The *Bα* gene *bpl5* and three of the new *Bβ* pheromone-like genes show deviations from the consensus farnesylation signal, CAAX (where C is cysteine, A is aliphatic and X is any residue), with the variant motifs CASR, CTIA, CRLT and CQLT for Bpl5, Bpl1, Bpl2 and Bpl3, respectively. Previously, one of the pheromone genes (*bbp2(6)*) was shown to function with the deviant farnesylation signal CEVM[12]. This suggests that in *S. commune* only one amino acid residue in the consensus sequence of the farnesylation signal needs to be aliphatic.

Transcription factors

The genome of *S. commune* reveals genes encoding 471 putative transcription factors, of which 311 are expressed during at least one developmental stage (**Supplementary Table 14**). Of these genes, 56% are expressed in all developmental stages; 268 were expressed in the monokaryon, 200 during formation of stage I aggregates, 283 during formation of stage II aggregates and 253 during formation of mushrooms. We identified a cluster of monokaryon-specific transcription factors and a group of transcription factors upregulated in stage II primordia or in mature mushrooms, or both (**Fig. 4**). The latter group includes *fst3* (NCBI Protein ID: 257422) and *fst4* (NCBI Protein ID: 66861),

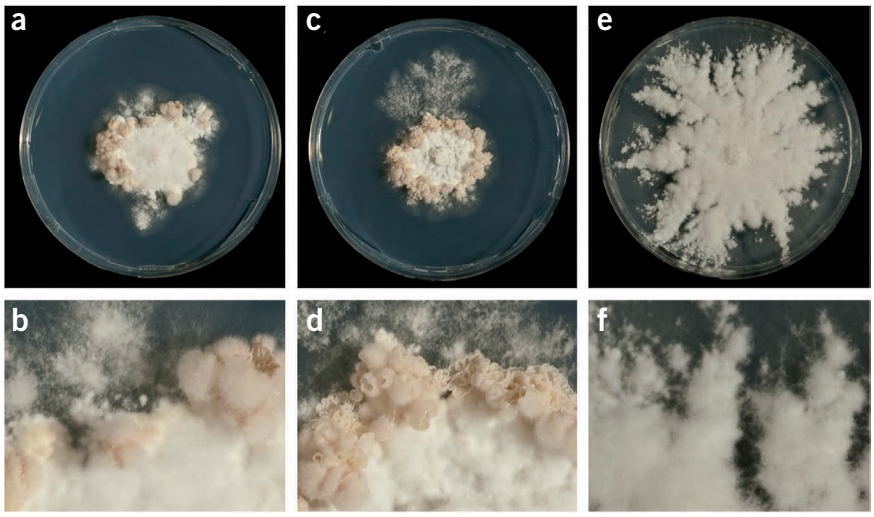


Figure 5 Transcription factors affecting fruiting body formation. (**a**,**b**) Wild-type dikaryon fruiting-body formation. (**c**–**f**) Fruiting-body formation in dikaryons in which *fst3* (**c**,**d**) or *fst4* (**e**,**f**) has been inactivated. Lower panels (**b**,**d**,**f**) show a magnification of part of the colonies shown in upper panels (**a**,**c**,**e**). Scale bar, 5 mm (**b**,**d**,**f**).

Table 1 Comparison of the number of FOLymes and CAZymes of *S. commune* with those of other fungi

	FOLymes											CAZymes			
Species	LO1	LO2	LO3	LDA1	LDA2	LDA3	LDA4	LDA5	LDA6	LDA7	LDA8	GH	GT	PL	CE
S. commune	2	0	1	1	0	2	1	0	4	4	1	240	75	16	30
C. cinerea	17	1	1	18	0	0	0	0	1	2	0	211	71	13	54
L. bicolor	9	1	0	4	0	0	0	0	3	2	0	163	88	7	20
P. placenta	2	0	0	3	0	0	0	0	0	1	2	124	51	4	13
P. chrysosporium	0	16	1	3	0	1	1	0	1	4	0	181	66	4	20
C. neoformans	0	0	0	0	0	0	0	0	0	0	0	75	64	3	8
U. maydis	0	0	0	0	0	1	0	1	1	1	0	101	64	1	19
S. cerevisiae	0	0	0	0	0	0	0	0	0	3	0	46	68	0	3
A. nidulans	1	0	1	0	1	0	0	0	1	0	0	250	91	21	32
N. crassa	5	0	2	1	0	0	0	1	1	1	1	173	76	4	23

LO1, laccases; LO2, peroxidases; LO3, cellobiose dehydrogenases; LDA1, aryl alcohol oxidases; LDA2, vanillyl-alcohol oxidases; LDA3, glyoxal oxidases; LDA4, pyranose oxidases; LDA5, galactose oxidases; LDA6, glucose oxidases; LDA7, benzoquinone reductases; LDA8, alcohol oxidases; GH, glycoside hydrolases; GT, glycosyl transferases; PL, polysaccharide lyases; CE, carbohydrate esterases.

which encode transcription factors that contain a fungus-specific Zn(II)2Cys6 zinc-finger DNA binding domain.

We inactivated the *fst3* and *fst4* genes via targeted gene deletions. The Δ*fst3* and Δ*fst4* monokaryons showed no phenotypic differences from the wild-type monokaryons. In contrast, the Δ*fst4* Δ*fst4* dikaryon did not fruit, but produced more aerial hyphae when compared to the wild type (**Fig. 5**). This suggests that Fst4 is crucial in the switch between the vegetative and reproductive phases of the *S. commune* life cycle. In contrast, the Δ*fst3* Δ*fst3* dikaryon formed more, albeit smaller, reproductive structures than those of the wild type (**Fig. 5**). As spatial and temporal regulation of fruiting-body formation and sporulation were not altered in the Δ*fst3* Δ*fst3* strain, we conclude that Fst3 inhibits the formation of clusters of mushrooms.

Wood degradation by *Schizophyllum commune*

As a white-rot fungus[6], *S. commune* degrades all woody cell wall components; in contrast, brown-rotters efficiently degrade cellulose but only modify lignin, leaving a polymeric residue. Lignin-degrading enzymes, which are commonly classified as FOLymes[14], comprise lignin oxidases (LO families) and lignin-degrading auxiliary enzymes that generate H_2O_2 for peroxidases (LDA families). The LO family consists of laccases (LO1), lignin peroxidases, manganese peroxidases, versatile peroxidases (LO2) and cellobiose dehydrogenases (CDHs; LO3). *S. commune* contains 16 FOLyme genes and 11 genes that encode enzymes distantly related to FOLyme enzymes (**Table 1** and **Supplementary Table 15**). The genome lacks genes encoding peroxidases of the LO2 family. However, it contains a CDH gene (LO3), two laccase genes (LO1) and 13 LDA genes, including four genes encoding glucose oxidases (LDA6) and benzoquinone reductases (LDA7) (**Table 1**).

S. commune appears to possess a more diverse assortment of FOLymes than the brown-rot fungus *Postia placenta* and the fungi that are known not to have ligninolytic activity (that is, *Ustilago maydis*, *Cryptococcus neoformans*, *Aspergillus nidulans*, *Neurospora crassa* and *Saccharomyces cerevisiae*; **Table 1**). In contrast, it has fewer FOLymes than either the coprophilic fungus *Coprinopsis cinerea* and the white-rot fungus *Phanerochaete chrysosporium*, which are predicted to possess 40 and 27 members, respectively[14].

Regarding polysaccharide degradation, *S. commune* has the most extensive machinery for degrading cellulose and hemicellulose of all of the basidiomycetes we examined. The Carbohydrate-Active Enzyme database (CAZy) identified 240 candidate glycoside hydrolases, 75 candidate glycosyl transferases, 16 candidate polysaccharide lyases and 30 candidate carbohydrate esterases encoded in the genome of *S. commune* (**Table 1** and **Supplementary Table 16**). Compared to the genomes of other basidiomycetes, *S. commune* has the highest number of glycoside hydrolases and polysaccharide lyases. *S. commune* is rich in genes encoding enzymes that degrade pectin, hemicellulose and cellulose (**Supplementary Table 17**). In fact, *S. commune* has genes in each family involved in the degradation of these plant cell wall polysaccharides. The *S. commune* genome is particularly rich in members of the glycosyl hydrolase families GH93 (hemicellulose degradation) and GH43 (hemicellulose and pectin degradation), and the lyase families PL1, PL3 and PL4 (pectin degradation) (**Supplementary Table 17**). The pectinolytic capacity of *S. commune* is further complemented by the presence of pectin hydrolases from families GH28, GH88 and GH105.

DISCUSSION

The phylum Basidiomycota contains roughly 30,000 described species, accounting for 37% of the true fungi[15]. The Basidiomycota comprises two class-level taxa (Wallemiomycetes and Entorrhizomycetes) and the subphyla Pucciniomycotina (rust), Ustilaginomycotina (smuts) and Agaricomycotina[16]. The Agaricomyotina include the mushroom- and puffball-forming fungi, crust fungi and jelly fungi. Genomic sequences are currently available for five members of the Agaricomycotina: *P. chrysosporium*[17], *L. bicolor*[10], *P. placenta*[18], *C. neoformans*[19] and *C. cinerea*[20]. Our 38.5-megabase assembly of the *S. commune* genome represents the first genomic sequence for a member of the family Schizophyllaceae. Thirty-six percent of the encoded proteins have no ortholog in other fungi. Only 43% of the predicted genes could be annotated with a gene ontology term, underscoring that much about the proteome of *S. commune* remains unknown. This percentage resembles that seen in other basidiomycetes: 30% in *L. bicolor*[10], 48% in *P. placenta*[18] and 49% in *P. chrysosporium*[17].

S. commune invades wood primarily by growing through the lumen of vessels, tracheids, fibers and xylem rays. Adjacent parenchymatic cells in the xylem tissue are invaded via simple and bordered pits. As a consequence of this approach to invasion, cellulose, hemicellulose or pectin can serve as the primary carbon source for *S. commune*. Indeed, the genome of *S. commune* probably encodes at least one gene in each family involved in the degradation of cellulose, hemicellulose and pectin. The large number of predicted pectinase genes is consistent with earlier studies describing *S. commune* as one of the best pectinase producers among the basidiomycetes[21]. *S. commune* also encodes carbohydrate-active enzymes that degrade other polymeric sugars, such as those acting on starch, mannans and inulins. Consistent with the wide variety of substrates that support its growth, *S. commune* has the most complete polysaccharide breakdown machinery of all basidiomycetes examined.

We know much less about how fungi degrade lignin than how they digest plant polysaccharides. Fungi are assumed to use FOLymes to degrade lignin[14]. Although members of the LO2 family of lignin oxidases are known to degrade lignin, it remains controversial whether laccases (LO1) and cellobiose dehydrogenases (CDHs; LO3) share this capacity. *S. commune* contains 16 genes encoding FOLymes. There are no members of the LO2 family, but the genome contains one CDH gene and two laccase genes. CDHs may participate in the degradation of cellulose, xylan and, possibly, lignin by generating hydroxyl radicals in a Fenton-type reaction. Laccases catalyze the one-electron oxidation of phenolic, aromatic amines and other electron-rich substrates with the concomitant reduction of O_2 to H_2O. They are classified as having either low or high redox potential[22], but it is not clear whether the two *S. commune* gene products belong to the high– or low–redox potential enzyme categories.

When the genomes of the white-rot fungi *S. commune* and *P. chrysosporium*[17] and the brown-rot fungus *P. placenta*[18] are compared, it is clear that *S. commune* has evolved its own set of FOLymes. *P. chrysosporium* lacks genes encoding laccases (LO1). It is thought to degrade lignin with the enzymes encoded by 16 isogenes of peroxidases (LO2), one CDH gene (LO3) and four genes of the multicopper oxidase superfamily. In contrast, *P. placenta* contains two laccase-encoding genes (LO1) but lacks members of the LO2 and LO3 families. As *S. commune* and *P. placenta* lack true LO2 FOLymes, one would expect a low number of LDAs that are responsible for H_2O_2 production for the peroxidases. This is not the case. *S. commune* contains more LDAs than *P. chrysosporium*. For instance, *S. commune* contains four glucose oxidase (LDA6) genes, whereas fungi seldom express more than one of these. In the absence of peroxidases of the LO2 family, it is expected that the glucose oxidases of *S. commune* serve another function. Glucose oxidases convert glucose into gluconic acid. This acid solubilizes inorganic phosphate and thus aids in the uptake of the nutrient[23].

The *matA* and *matB* mating-type loci of *S. commune* regulate the formation of a fertile dikaryon after the fusion of monokaryons that encounter one other. The genome sequence of this species now reveals that the mating type loci of *S. commune* contain the highest number of reported genes within such loci in the fungal kingdom. The *matB* locus comprises two linked loci, *Bα* and *Bβ*, which both encode pheromones and pheromone receptors[1]. Nine allelic specificities have been identified for both loci, resulting in 81 different mating types for *matB*. It was previously reported that the *Bα3* and *Bβ2* loci of H4-8 contain three and eight pheromone genes, respectively, and each contain one pheromone receptor gene[12,13]. We identified five additional pheromone genes and four additional pheromone receptor–like genes in the genome of H4-8. These newly identified receptor-like genes are present in a *matB* deletion strain, which has no pheromone response with any mate (T.J.F., unpublished data). This raises the question of whether the four receptor genes function in *matB*-regulated development. Expression of these genes, as discerned using MPSS, suggests that they do not represent pseudogenes.

The *matA* locus consists of two subloci, Aα and *Aβ*, of which 9 and 32 allelic specificities, respectively, are expected to occur in nature[1]. These loci are separated by 550 kb on chromosome I of strain H4-8. Such a large distance has not been found in other fungi that have a tetrapolar mating system. The functionally well-characterized *Aα* locus showed no substantial differences from the published descriptions[1]. It is composed of two genes encoding Y and Z homeodomain proteins of the HD2 and HD1 classes, respectively. The Y and Z proteins, as in other basidiomycetes, interact in non-self combinations to activate the A-pathway of sexual development[1,24]. Notably, a nuclear localization signal is present in Y but not in Z. This is consistent with non-self interaction of the two proteins taking place in the cytosol, followed by the translocation of the active protein complex into the nucleus[1].

The *Aβ* locus of *S. commune* has been studied much less than the *Aα* locus. Notably, *Aβ* reflects the highest degree of homeodomain-gene complexity for any fungal mating-type locus described to date. It contains four homeodomain genes of the HD1 class and two of the HD2 class. The *Aβ* locus of *S. commune* thus resembles that of *C. cinerea*, which consists of two pairs of functional HD1 and HD2 homeodomain genes (b and d)[25]. The large number of genes in *matAβ* would explain why recombination analyses predict as many as 32 mating specificities for this locus[26]. Overall, *S. commune* seems ideal for identifying the evolutionary pathways that have created high numbers of allelic specificities for enhancing outbreeding versus inbreeding rates.

As little is known about molecular processes that control formation of fruiting bodies in basidiomycetes, other than the role of the mating-type loci[8], we compared genome-wide expression profiles at four developmental stages. MPSS showed that relatively few genes were specifically expressed in the monokaryon (284 genes) and in stage I aggregates and the mature mushrooms (128 genes in both cases). Notably, 467 genes were specifically expressed in stage II primordia. This suggests that this stage represents a major developmental switch, an idea supported by the fact that genes involved in signal transduction and regulation of gene expression are enriched in the group of upregulated genes during formation of stage II primordia. A positive correlation of expression of these gene groups during mushroom formation in both *S. commune* and *L. bicolor* suggests that regulation of mushroom formation is a conserved process in the Agaricales.

Our analysis of gene expression in *S. commune* reveals a high frequency of antisense expression. About 20% of all sequenced mRNA tags originated from an antisense transcript, and >5,600 of the predicted genes showed antisense expression in one or more developmental stages. Antisense transcription was most pronounced in stage II primordia. At this stage, >4,300 genes were expressed in both the sense and antisense directions, and >800 genes were expressed in the antisense direction only. Previously, MPSS has revealed antisense transcripts in *Magnaporthe grisea*[27]. Little is known about the function of these transcripts in fungi. The circadian clock of *N. crassa* is entrained in part by the action of an antisense transcript derived from a locus encoding a component of the circadian clock[28], possibly through RNA interference. It is tempting to speculate that antisense transcripts also regulate mRNA levels in *S. commune*. Natural antisense transcripts in eukaryotes have also been implicated in other processes, such as translational regulation, alternative splicing and RNA editing[29]. The antisense transcripts of *S. commune* may likewise have such functions. In all these cases, the antisense transcripts could function in a developmental switch that occurs when stage II primordia are formed.

The apparently high conservation of gene regulation in the Agaricales led us to study the 471 genes predicted to encode transcriptional regulators. Of these, 268 were expressed in the monokaryon, whereas 200, 283 and 253 were expressed during formation of stage I aggregates, stage II primordia and mushrooms, respectively. The relatively high number of transcription factors expressed during formation of stage II primordia again points to a major switch that probably occurs during this developmental stage.

We identified a group of monokaryon-specific transcription factors and a group of transcription factors that are upregulated in stage II primordia or mature mushrooms, or in both. The *fst3* and *fst4* genes encode transcriptional regulators belonging to the latter group. Growth and development were not affected in monokaryotic strains

in which *fst3* or *fst4* were inactivated. Phenotypic differences were, however, observed in the dikaryon. The Δ*fst4* Δ*fst4* dikaryon did not fruit but produced more aerial hyphae than the wild type. In contrast, the Δ*fst3* Δ*fst3* dikaryon formed more, albeit smaller, fruiting bodies than the wild type. This suggests that Fst4 is involved in the switch between the vegetative and the reproductive phase, and that Fst3 inhibits formation of clusters of mushrooms. Inhibition of such clusters could be important in a natural environment to ensure that sufficient energy is available for full development of fruiting bodies. As *fst3* and *fst4* have homologs in other mushroom-forming fungi, it is tempting to speculate that they have similar functions in these organisms. This is supported by the observation that the homologs of *fst3* and *fst4* are upregulated in young fruiting bodies of *L. bicolor* compared to free-living mycelium[10]. In mature fruiting bodies of *L. bicolor*, the expression level of the homolog of *fst3* remains constant compared to young fruiting bodies, whereas the *fst4* homolog returns to the level expressed in the free-living mycelium.

In conclusion, the genomic sequence of *S. commune* will be an essential tool to unravel mechanisms by which mushroom-forming fungi degrade their natural substrates and form fruiting bodies. The large variety of genes that encode extracellular enzymes that act on polysaccharides probably explains why *S. commune* is so common in nature. Moreover, the genome sequence suggests that *S. commune* may have a unique mechanism to degrade lignin. Our MPSS data has provided leads on how mushroom formation is regulated, highlighting both the roles of certain transcription factors and the possible involvement of antisense transcription. Better understanding of the physiology and sexual reproduction of *S. commune* will probably have an impact on the commercial production of edible mushrooms and the use of mushrooms as cell factories.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Data availability and accession codes. *S. commune* assemblies, annotations and analyses are available through the interactive JGI Genome Portal at http://jgi.doe.gov/Scommune. Genome assemblies, together with predicted gene models and annotations, were also deposited at DDBJ/EMBL/GenBank under the project accession number ADMJ00000000. MPSS data have been deposited in NCBI's Gene Expression Omnibus with accession number GSE21265.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence Berkeley National Laboratory under contract no. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract no. DE-AC02-06NA25396. The work was also supported by the Dutch Technology Foundation STW, the Applied Science division of the Netherlands Organization for Scientific Research and the Technology Program of the Dutch Ministry of Economic Affairs.

AUTHOR CONTRIBUTIONS

J.S.H., H.A.B.W., T.J.F., W.W.L., A.C.G., E.K. and S.E.B. wrote the proposal; L.G.L., J.F.d.J., E.L. and R.A.O. isolated RNA and DNA and made libraries; E.L. and S.L. coordinated sequencing of the genome; J.S. assembled the genome; I.V.G., R.A.O., K.A.B., J.E.S., S.E.B., E.K. and H.A.B.W. coordinated the annotation process; I.V.G., A.A., A.S., J.E.S., T.J.F., E.R., A.L., F.P., U.K., J.K.M., J.F.d.J., R.P.d.V., P.M.C., V.L., B.H., W.W.L., A.C.G., P.A.v.K., K.A.B., J.S.H., E.K., S.E., N.K., R.A.O. and M.R. annotated genes; R.A.O., J.F.d.J., F.W.M.R.S. and L.G.L. performed experiments; R.A.O., J.F.d.J., F.W.M.R.S., L.G.L. and H.A.B.W. interpreted and designed experiments; R.A.O., A.A., J.E.S., E.R., J.F.d.J., R.P.d.V., F.W.M.R.S., K.A.B., E.K., J.S.H., I.V.G. and H.A.B.W. wrote the paper; R.A.O. and H.A.B.W. coordinated writing of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kothe, E. Mating-type genes for basidiomycete strain improvement in mushroom farming. *Appl. Microbiol. Biotechnol.* **56**, 602–612 (2001).
2. Kües, U. & Liu, Y. Fruiting body production in basidiomycetes. *Appl. Microbiol. Biotechnol.* **54**, 141–152 (2000).
3. Lomascolo, A., Stentelaire, C., Asther, M. & Lesage-Meessen, L. Basidiomycetes as new biotechnological tools to generate natural aromatic flavours for the food industry. *Trends Biotechnol.* **17**, 282–289 (1999).
4. Berends, E., Scholtmeijer, K., Wösten, H.A.B., Bosch, D. & Lugones, L.G. The use of mushroom-forming fungi for the production of N-glycosylated therapeutic proteins. *Trends Microbiol.* **17**, 439–443 (2009).
5. Alves, A.M. *et al.* Highly efficient production of laccase by the basidiomycete *Pycnoporus cinnabarinus*. *Appl. Environ. Microbiol.* **70**, 6379–6384 (2004).
6. Schmidt, O. & Liese, W. Variability of wood degrading enzymes of *Schizophyllum commune*. *Holzforschung* **34**, 67–72 (1980).
7. de Jong, J.F. *Aerial Hyphae of Schizophyllum commune: Their Function and Formation*. PhD thesis, Univ. Utrecht (2006).
8. Wösten, H.A.B. & Wessels, J.G.H. The emergence of fruiting bodies in basidiomycetes. in *The Mycota. Part I: Growth, Differentiation and Sexuality* (eds. Kües, U. & Fisher, R.) 393–414 (Springer, Berlin, 2006).
9. Asgeirsdottir, S.A., Schuren, F.H.J. & Wessels, J.G.H. Assignment of genes to pulse-field separated chromosomes of *Schizophyllum commune*. *Mycol. Res.* **98**, 689–693 (1994).
10. Martin, F. *et al.* The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* **452**, 88–92 (2008).
11. Shen, G.P. *et al.* The Aalpha6 locus: its relation to mating-type regulation of sexual development in *Schizophyllum commune*. *Curr. Genet.* **39**, 340–345 (2001).
12. Fowler, T.J., Mitton, M.F., Vaillancourt, L.J. & Raper, C.A. Changes in mate recognition through alterations of pheromones and receptors in the multisexual mushroom fungus *Schizophyllum commune*. *Genetics* **158**, 1491–1503 (2001).
13. Fowler, T.J., Mitton, M.F., Rees, E.I. & Raper, C.A. Crossing the boundary between the *Bα* and *Bβ* mating-type loci in *Schizophyllum commune*. *Fungal Genet. Biol.* **41**, 89–101 (2004).
14. Levasseur, A. *et al.* FOLy: an integrated database for the classification and functional annotation of fungal oxidoreductases potentially involved in the degradation of lignin and related aromatic compounds. *Fungal Genet. Biol.* **45**, 638–645 (2008).
15. Kirk, P.M., Cannon, P.F., David, J.C. & Stalpers, J.A.. *Ainsworth and Bisby's Dictionary of the Fungi* (CAB International, Wallingford, UK, 2001).
16. Hibbett, D.S. *et al.* A higher-level phylogenetic classification of the Fungi. *Mycol. Res.* **111**, 509–547 (2007).
17. Martinez, D. *et al.* Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.* **22**, 695–700 (2004).
18. Martinez, D. *et al.* Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. *Proc. Natl. Acad. Sci. USA* **106**, 1954–1959 (2009).
19. Loftus, B.J. *et al.* The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324 (2005).
20. Stajich, J.S. *et al.* Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc. Natl. Acad. Sci. USA* **107**, 11889–11894 (2010).
21. Xavier-Santos, S. *et al.* Screening for pectinolytic activity of wood-rotting basidiomycetes and characterization of the enzymes. *Folia Microbiol. (Praha)* **49**, 46–52 (2004).
22. Xu, F. *et al.* A study of a series of recombinant fungal laccases and bilirubin oxidase that exhibit significant differences in redox potential, substrate specificity, and stability. *Biochim. Biophys. Acta* **1292**, 303–311 (1996).
23. Xiao, C. *et al.* Isolation of phosphate-solubilizing fungi from phosphate mines and their effect on wheat seedling growth. *Appl. Biochem. Biotechnol.* **159**, 330–342 (2009).
24. Spit, A., Hyland, R.H., Mellor, E.J. & Casselton, L.A. A role for heterodimerization in nuclear localization of a homeodomain protein. *Proc. Natl. Acad. Sci. USA* **95**, 6228–6233 (1998).
25. Casselton, L.A. & Olesnicky, N.S. Molecular genetics of mating recognition in basidiomycete fungi. *Microbiol. Mol. Biol. Rev.* **62**, 55–70 (1998).
26. Raper, J. *Genetics of Sexuality of Higher Fungi* (The Roland Press, New York, 1966).
27. Gowda, M. *et al.* Deep and comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE, and oligoarray methods. *BMC Genomics* **7**, 310 (2006).
28. Kramer, C., Loros, J.J., Dunlap, J.C. & Crosthwaite, S.K. Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature* **421**, 948–952 (2003).
29. Lavorgna, G. *et al.* In search of antisense. *Trends Biochem. Sci.* **29**, 88–94 (2004).

ONLINE METHODS

Strains and culture conditions. *S. commune* was routinely grown at 25 °C on minimal medium (MM) with 1% (wt/vol) glucose and with or without 1.5% (wt/vol) agar[30]. Liquid cultures were shaken at 225 r.p.m. Glucose was replaced with 4% (wt/vol) glycerol for cultures used in the isolation of genomic DNA. All *S. commune* strains used were isogenic to strain 1-40 (ref. 31). Strain H4-8 (*matA43 matB41*; FGSC no. 9210) was used for sequencing. EST libraries were generated from H4-8 and from a dikaryon that resulted from a cross between H4-8 and strain H4-8b (*matA4 matB43*)[32]. Strains 4-39 (*matA41 matB41*; CBS 341.81) and 4-40 (*matA43 matB43*; CBS 340.81) were used for MPSS. These strains show a more synchronized fruiting compared to a cross between H4-8 and H4-8b. Partial sequencing of the haploid genome revealed that strains 4-40 and 4-39 have minor sequence differences (<0.2%) with strain H4-8 (data not shown).

Isolation of genomic DNA, genome sequencing and assembly. Genomic DNA of *S. commune* was isolated as described[30] and sequenced using a whole-genome shotgun strategy. All data were generated by paired-end sequencing of cloned inserts with six different insert sizes using Sanger technology on ABI3730xl sequencers. The data were assembled using the whole-genome shotgun assembler Arachne (http://www.broad.mit.edu/wga/).

EST library construction and sequencing. Cultures were inoculated on MM plates with 1% (wt/vol) glucose using mycelial plugs as an inoculum. Strain H4-8 was grown for 4 d in the light, whereas the dikaryon H4-8 × H4-8.3 was grown for 4 d in the dark and 8 d in the light. Mycelia of the dikaryotic stages were combined and RNA was isolated as described[30]. The poly(A)$^+$ RNA fraction was obtained using the Absolutely mRNA Purification kit and manufacturer's instructions (Stratagene). cDNA synthesis and cloning followed the SuperScript plasmid system procedure with Gateway technology for cDNA synthesis and cloning (Invitrogen). For the monokaryon, two size ranges of cDNA were cut out of the gel to generate two cDNA libraries (JGI library codes CBXY for the range 0.6 kb–2 kb and CBXX for the range >2 kb). For the dikaryon, cDNA was used in the range >2 kb, resulting in library CBXZ. The cDNA inserts were directionally ligated into vector pCMVsport6 (Invitrogen) and introduced into ElectroMAX T1 DH10B cells (Invitrogen). Plasmid DNA for sequencing was produced by rolling-circle amplification (Templiphi, GE Healthcare). Subclone inserts were sequenced from both ends using Big Dye terminator chemistry and ABI 3730 instruments (Applied Biosystems).

Annotation methods. Gene models in the genome of *S. commune* were predicted using Fgenesh[33], Fgenesh+[33], Genewise[34] and Augustus[35]. Fgenesh was trained for *S. commune* with a sensitivity of 72% and a specificity of 74%. Augustus *ab initio* gene predictions were generated with parameters based on *C. cinerea* gene models[20]. In addition, about 31,000 *S. commune* ESTs were clustered into nearly 9,000 groups. These groups were either directly mapped to the genomic sequence with a threshold of 80% coverage and 95% identity, included as putative full-length genes, or used to extend predicted gene models into full-length genes by adding 5′ and/or 3′ UTRs. Because multiple gene models were generated for each locus, a single representative model at each locus was computationally selected on the basis of EST support and similarity to protein sequences in the NCBI nonredundant database. This resulted in a final set of 13,210 predicted genes, of which 1,314 genes have been manually curated. In 66 cases, models were created or coordinates were changed.

All predicted gene models were functionally annotated by homology to annotated genes from the NCBI nonredundant set and classified according to Gene Ontology[36], eukaryotic orthologous groups (KOGs)[37], KEGG metabolic pathways[38] and Protein Family (PFAM) domains[39].

Repeat content. RepeatModeler 1.0.3 (http://www.repeatmasker.org/RepeatModeler.html) was used to generate *de novo* repeat sequence predictions for *S. commune*. Repeats were classified by comparison to the RepBase database (http://www.girinst.org/repbase/index.html). RepeatModeler produced 76 families of repeats used as a search library in RepeatMasker (http://www.repeatmasker.org/).

Orthologs of *S. commune* proteins in the fungal kingdom. Proteins of *S. commune* were assigned to orthologous groups with OrthoMCL version 2.0 (ref. 40) with an inflation value of 1.5. Members of such groups were assigned as orthologs (in the case of proteins from another species) or inparalogs (in the case of proteins from *S. commune*). Orthologs were determined in *C. cinerea*[20], *L. bicolor*[10], *P. placenta*[18], *P. chrysosporium*[17], *C. neoformans*[19], *U. maydis*[41], *S. cerevisiae*[42], *A. nidulans*[43] and *N. crassa*[44]. All-versus-all BLASTP analysis was performed using NCBI standalone BLAST version 2.2.20, with an *E* value of 10^{-5} as a cutoff. Custom scripts were used to further analyze the orthologous groups resulting from the OrthoMCL analysis. The evolutionary conservation for each orthologous group was expressed as the taxon this orthologous group was most specifically confined to (see **Supplementary Fig. 1**).

Representation analysis. FuncAssociate 2.0 (ref. 45) was used to study over- and under-representation of taxon-specific genes and of functional-annotation terms in sets of differentially regulated genes. Default settings were used, with a *P* value of 0.05 or 0.01 as the cutoff.

Protein families. The PFAM database version 24.0 (ref. 39) was used to identify PFAM protein families. Custom scripts in Python were written to group genes on basis of their PFAM domains. Differences in the number of predicted proteins belonging to a PFAM family across the fungal domains were determined using Student's *t*-test. When Agaricales were compared to the rest of the Dikarya, or when *S. commune* was compared to the Agaricales, only groups with a minimum of five members in at least one of the fungi were analyzed. When *S. commune* was compared to the rest of the Dikarya, only groups with a minimum of five members in at least four of the fungi were analyzed. In all cases, a *P* value of 0.05 was used as a cutoff. Similar results were obtained using the nonparametric Mann-Whitney *U*-test.

CAZy annotation. Annotation of carbohydrate-related enzymes was performed using the CAZy annotation pipeline[46]. Ambiguous family attributions were processed manually along with all identified models that presented defects (such as deletions, insertions or splicing problems). Each protein was also compared to a library of experimentally characterized proteins found in CAZy to provide a functional description.

FOLy annotation. Lignin oxidative enzymes (FOLymes)[14] were identified by BLASTP analysis of the *S. commune* gene models against a library of FOLy modules using an *e* value <0.1. The resulting 68 protein models were analyzed manually using the BLASTP results as well as multiple-sequence alignments and functional inference based on phylogeny[47]. Basically, a protein was identified as a FOLyme when it showed a similarity score above 50% with sequences of biochemically characterized enzymes. When the similarity score was <50% the proteins were scored as a FOLyme-related protein.

MPSS expression analysis. Total RNA was isolated from the monokaryotic strain 4-40 and from the dikaryon resulting from a cross between 4-40 and 4-39. A 7-day-old colony grown on solid MM at 30 °C in the dark was homogenized in 200 ml MM using a Waring blender for 1 min at low speed. Two milliliters of the homogenized mycelium was spread out over a polycarbonate membrane placed on top of solidified MM. Vegetative monokaryotic mycelium was grown for 4 d in the light. The dikaryon was grown for 2 and 4 d in the light to isolate mycelium with stage I aggregates and stage II primordia, respectively. Mature mushrooms 3 d old were picked from dikaryotic cultures that had grown for 8 d in the light. RNA was isolated as described[30]. MPSS was performed essentially as described[48] except that after DpnII digestion MmeI was used to generate 20-bp tags. Tags were sequenced using the Clonal Single Molecule Array technique (Illumina). Between 4.2 and 7.6 million tags of 20 bp were obtained for each of the stages. Programs were developed in the programming language Python to analyze the data. Tag counts were normalized to tags per million (TPM). Those with a maximum of <4 TPM in all developmental stages were removed from the data set. This data set consisted of a total of 40,791 unique tags. Of these tags, 61.7% and 58.6% could be mapped to the genome sequence and the predicted transcripts, respectively, using a perfect match as the criterion. The mapped tags accounted for 71.4% and 70.8%

doi:10.1038/nbt.1643

of the total number of tags, respectively. For comparison, 97.4% of the ESTs from *S. commune* strain H4-8 could be mapped to the assembly. Unmapped tags can be explained by sequencing errors in either tag or genomic DNA. Moreover, RNA editing may have altered the transcript sequencing to produce tags that do not match the genome perfectly. It may also be that the assigned untranslated region is incomplete or that the DpnII restriction site that defines the 5′ end of the tag is too close to the poly(A) tail of the mRNA. TPM values of tags originating from the same transcript were summed to assess their expression levels. A transcript is defined as the predicted coding sequence extended with 400-bp flanking regions at both sides.

Comparison of gene expression in *L. bicolor* and *S. commune*. Whole-genome expression analysis of *L. bicolor*[10] and *S. commune* was done essentially as described[49]. For *L. bicolor*, the microarray values from replicates were averaged. Expression values of genes were increased by 1, and the ratio between monokaryon and mushrooms (for *S. commune*), and between free-living mycelium and mature fruiting bodies (for *L. bicolor*), was log-transformed. All expressed genes from *S. commune* that had at least one expressed ortholog in *L. bicolor* were taken into account, resulting in a total of 6,751 orthologous pairs. These pairs were classified on the basis of functional-annotation terms. Correlation of changes in expression of these gene classes was expressed as the Pearson correlation coefficient. Only gene ontology terms with 10–200 pairs were used in the analysis. In the case of PFAM domains, a minimum of ten ortholog pairs were used.

Deletion of transcription factors *fst3* and *fst4*. The transcription factor genes *fst3* (NCBI Protein ID: 257422) and *fst4* (NCBI Protein ID: 66861) were deleted using the vector pDelcas[32]. Transformation of *S. commune* strain H4-8 was done as described[30]. Regeneration medium contained no antibiotic, whereas selection plates contained 20 μg ml^{-1} nourseothricin. Deletion of the target gene was confirmed by PCR. Compatible monokaryons with a gene deletion were selected from spores originating from a cross of the mutant strains with wild-type strain H4-8.3.

30. van Peer, A.F., de Bekker, C., Vinck, A., Wösten, H.A.B. & Lugones, L.G. Phleomycin increases transformation efficiency and promotes single integrations in *Schizophyllum commune*. *Appl. Environ. Microbiol.* **75**, 1243–1247 (2009).
31. Raper, J.R., Krongelb, G.S. & Baxter, M.G. The number and distribution of incompatibility factors in *Schizophyllum*. *Am. Nat.* **92**, 221–232 (1958).
32. Ohm, R.A. *et al.* An efficient gene deletion procedure for the mushroom-forming basidiomycete *Schizophyllum commune*. *World J. Microbiol. Biotechnol.* 10.1007/s11274–010–0356–0.
33. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
34. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).
35. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215–225 (2003).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
37. Koonin, E.V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).
38. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
39. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
40. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
41. Kämper, J. *et al.* Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444**, 97–101 (2006).
42. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
43. Galagan, J.E. *et al.* Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* **438**, 1105–1115 (2005).
44. Galagan, J.E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–868 (2003).
45. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. & Roth, F.P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).
46. Cantarel, B.L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
47. Gouret, P. *et al.* FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC Bioinformatics* **6**, 198 (2005).
48. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
49. McCarroll, S.A. *et al.* Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.* **36**, 197–204 (2004).

C

Nucleic Acids Res. (2009);**37**:D233-238.

The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B.

Abstract :The Carbohydrate-Active Enzyme (CAZy) database is a knowledge-based resource specialized in the enzymes that build and breakdown complex carbohydrates and glycoconjugates. As of September 2008, the database describes the present knowledge on 113 glycoside hydrolase, 91 glycosyltransferase, 19 polysaccharide lyase, 15 carbohydrate esterase and 52 carbohydrate-binding module families. These families are created based on experimentally characterized proteins and are populated by sequences from public databases with significant similarity. Protein biochemical information is continuously curated based on the available literature and structural information. Over 6400 proteins have assigned EC numbers and 700 proteins have a PDB structure. The classification (i) reflects the structural features of these enzymes better than their sole substrate specificity, (ii) helps to reveal the evolutionary relationships between these enzymes and (iii) provides a convenient framework to understand mechanistic properties. This resource has been available for over 10 years to the scientific community, contributing to information dissemination and providing a transversal nomenclature to glycobiologists. More recently, this resource has been used to improve the quality of functional predictions of a number genome projects by providing expert annotation. The CAZy resource resides at URL: http://www.cazy.org/.

PMID: 18838391

Published online 05 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D233–D238***
doi:10.1093/nar/gkn663

The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics

Brandi L. Cantarel, Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard and Bernard Henrissat*

Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS, Universités Aix-Marseille I & II, 163 Avenue de Luminy, 13288 Marseille, France

Received September 15, 2008; Accepted September 19, 2008

ABSTRACT

The Carbohydrate-Active Enzyme (CAZy) database is a knowledge-based resource specialized in the enzymes that build and breakdown complex carbohydrates and glycoconjugates. As of September 2008, the database describes the present knowledge on 113 glycoside hydrolase, 91 glycosyltransferase, 19 polysaccharide lyase, 15 carbohydrate esterase and 52 carbohydrate-binding module families. These families are created based on experimentally characterized proteins and are populated by sequences from public databases with significant similarity. Protein biochemical information is continuously curated based on the available literature and structural information. Over 6400 proteins have assigned EC numbers and 700 proteins have a PDB structure. The classification (i) reflects the structural features of these enzymes better than their sole substrate specificity, (ii) helps to reveal the evolutionary relationships between these enzymes and (iii) provides a convenient framework to understand mechanistic properties. This resource has been available for over 10 years to the scientific community, contributing to information dissemination and providing a transversal nomenclature to glycobiologists. More recently, this resource has been used to improve the quality of functional predictions of a number genome projects by providing expert annotation. The CAZy resource resides at URL: http://www.cazy.org/.

INTRODUCTION

Due to the extreme variety of monosaccharide structures, to the variety intersugar linkages and to the fact that virtually all types of molecules can be glycosylated (from sugars themselves, to proteins, lipids, nucleic acids, antibiotics, etc.), the large variety of enzymes acting on these glycoconjugates, oligo- and polysaccharides probably constitute one of the most structurally diverse set of substrates on Earth. Collectively designated as Carbohydrate-Active enZymes (CAZymes), these enzymes build and breakdown complex carbohydrates and glycoconjugates for a large body of biological roles (collectively studied under the term of Glycobiology). Therefore, CAZymes have to perform their function usually with high specificity. Because carbohydrate diversity (1) exceeds by far the number of protein folds, CAZymes have evolved from a limited number of progenitors by acquiring novel specificities at substrate and product level. Such a dizzying array of substrates and enzymes makes CAZymes a particularly challenging subject for experimental characterization and for functional annotation in genomes.

Nearly 20 years ago, the first foundation for a family classification of CAZymes was seen in an effort that classified cellulases into several distinct families based on amino-acid sequence similarity (2). Soon after, the family classification system based on protein sequence and structure similarities, was extended to all known glycoside hydrolases (2–4), and subsequently extended to all CAZymes involved in the synthesis, degradation and modification of glycoconjugates. The classification of CAZymes has been made available on the web since September 1998. Because based on amino-acid sequence similarities, these classifications correlate with enzyme mechanisms and protein fold more than enzyme specificity. Consequently, these families are used to conservatively classify proteins of uncharacterized function whose only known feature is sequence similarity to an experimentally characterized enzyme, avoiding overprediction of enzyme activities.

At present, CAZy covers approximately 300 protein families in the following classes of enzyme activities:

(1) Glycoside hydrolases (GHs), including glycosidases and transglycosidases (3–5). These enzymes constitute 113 protein families that are responsible for

*To whom correspondence should be addressed. Tel: +33 4 91 82 55 87; Fax: +33 491 26 67 20; Email: Bernard.Henrissat@afmb.univ-mrs.fr
Correspondence may also be addressed to Pedro M. Coutinho. Email: Pedro.Coutinho@afmb.univ-mrs.fr

the hydrolysis and/or transglycosylation of glycosidic bonds. GH-coding genes are abundant and present in the vast majority of genomes corresponding to almost half—presently about 47%—of the enzymes classified in CAZy. Because of their widespread importance for biotechnological and biomedical applications, GHs constitute so far the best biochemically characterized set of enzymes present in the CAZy database.

(2) Glycosyltransferases (GTs). These are the enzymes responsible for the biosynthesis of glycosidic bonds from phospho-activated sugar donors (6–8). They form over 90 sequence-based families and present in virtually every single organism and represent about 41% of CAZy at present.

(3) Polysaccharide lyases (PLs) cleave the glycosidic bonds of uronic acid-containing polysaccharides by a β-elimination mechanism (6). They are presently found in 19 families in CAZy (7), corresponding to only about 1.5% of CAZy content. Many PLs have biotechnological and biomedical applications and, despite their small overall number, they are among the CAZymes with the highest proportion of biochemically characterized examples present in the database.

(4) Carbohydrate esterases (CEs). They remove ester-based modifications present in mono-, oligo- and polysaccharides and thereby facilitate the action of GHs on complex polysaccharides. Presently described in 15 families (7), CEs represent roughly 5% of CAZy entries. As the specificity barrier between carbohydrate esterases and other esterase activities is low, it is likely that the sequence-based classification incorporates some enzymes that may act on non-carbohydrate esters.

(5) Carbohydrate-binding modules (CBMs). These are autonomously folding and functioning protein fragments that have no enzymatic activity *per se* but are known to potentiate the activity of many enzyme activities described above by targeting to and promoting a prolonged interaction with the substrate. CBMs are most often associated to the other carbohydrate-active enzyme catalytic modules in the same polypeptide and can target different substrate forms depending on different structural characteristics (9,10). However, occasionally they can be present in isolated or tandem forms not coupled with an enzyme. Roughly 7% of CAZy entries contain at least one CBM module. CBMs are presently classified in 52 families in CAZy (7).

In addition to protein families that are well curated by the CAZy database, CAZymes are known to contain domains not acting on carbohydrates, including other enzymes—such as proteases, myosin motors or phosphatases, etc.—and a variety of protein–protein or protein–cell wall binding domains—cohesins, SLHs, TPR, etc.

The CAZy family classification system covers all taxonomic groups, and provides the ground for common nomenclature for CAZymes across different glycobiologists (11,12) generally specialized only in some specific groups of organisms. Day-to-day inspection of new enzyme characterizations reported in the literature regularly led and continues to lead to the definition of new enzyme families. Significantly, the CAZy families, originally created following hydrophobic cluster analysis in the 1990s from very limited number of sequences available (2–6) and later complemented by BLAST- and HMMer-based sequence similarity approaches, are globally surviving the challenge of time in spite of a hundred-fold increase in the number of sequences.

DATABASE CONTENT

The CAZy database contains information from (i) sequence annotations from publicly available sources, namely the NCBI, including taxonomical, sequence and reference information, (ii) family classification and (iii) known functional information. This data allow the exploration of an enzyme (CAZyme), all CAZymes in an organism or a CAZy protein family. The addition of new family members and the incorporation of biochemical information extracted from the literature are updated regularly, following careful inspection. Newly released three-dimensional (3D) structures and genomes are analyzed as they are released by public databases. Daily update releases from GenBank form the bulk of sequence additions to the database (8) are complemented by weekly PDB releases (13). Presently only genome released through these GenBank releases are analyzed regularly, whereas other genomes protein predictions are analyzed upon request as part of collaborative efforts (*vide infra*).

Another feature of CAZy is that the number of families, the family-associated information and content are continuously updated. When new families are created, old previously released genomes and sequence in public databases are reanalyzed to take the additional new family into account to ensure completeness in sequence description. Internally, curators include and maintain all referenced biochemical and other characterization data from the literature and the analysis of full sets of protein sequences present in a single genome. Because of this continuous effort of data addition, new families are frequently added and reflect the advances in experimental characterization of CAZymes. New families are exclusively created based on the availability of at least one biochemically-characterized member for which a sequence is available and the information published in peer-reviewed scientific literature. This sequence then serves as a seed for the family that is gradually extended with sequences that share statistically significant similarity.

Only functional assignments based on experimental data are included in the CAZy database by the association of EC numbers to protein sequences. Therefore inferred functional assignments are not included. Experimental data are ideally a direct enzyme analysis, but also could include indirect evidence such as gene knockout experiments with extensive characterization. Because there is a shortage of EC numbers, relative to the number of functions characterized experimentally, some incomplete EC numbers such as 3.2.1.-, 2.4.1.-, 2.4.2.- and 2.4.99.- are

also included in the database. In addition, as the described functions in CAZy are only of enzymatic nature, additional and complementary binding and inhibitory functions known to be associated with several CAZy proteins will be curated and explored in the near future.

SEMI-AUTOMATIC MODULAR ASSIGNMENT

Carbohydrate-active enzymes, can exhibit a modular structure (Figure 1), where a module can be defined as a structural and functional unit (7,14). Each family in CAZy is dependent on the definition of a common segment in each full sequence that ultimately contains the catalytic or binding module. The definition of the limits within the sequence of the composing modules depends on available information derived from a combination of different approaches:

(1) protein 3D structures,
(2) reported deletion studies and
(3) protein-sequence analysis and comparisons.

Different sequence comparison tools are used to define enzyme families, particularly gapped BLAST (9) and HMMER (10) using hidden markov models (HMMs) made from each family. All the sequences corresponding to the catalytic and binding of carbohydrate-active enzymes are excised from the full protein sequence and grouped in a BLAST library. Positive hits against this 'high quality' library, are entered into the database by trained curators following manual check on a daily basis with a small number of sequences with high identity (>85%) ungapped alignments to previously examined sequences being entered automatically.

A new layer dealing with the analysis of whole protein sets issued from genomes has been introduced recently. Modular annotation has been in fact applied to genome data released by the NCBI, with over 750 genomes analyzed. Approximately 1–3% of the proteins encoded by a typical genome correspond to CAZymes (10,11). In addition to publicly released sequences, annotation of proteins in recently sequenced genomes prior to full release are regularly performed by the CAZy team in collaboration with scientists from all over the globe.

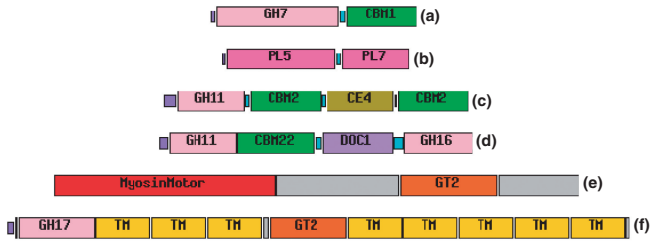


Figure 1. Examples of modular carbohydrate-active enzymes. (**a**) Cellobiohydrolase I from *Hypocrea jecorina* (SP P00725); (**b**) alginate lyase from *Sphingomonas* sp. A1 (GB BAB03312.1); (**c**) xylanase from *Cellulomonas fimi* (GB CAA54145.1); (**d**) xylanase D/lichenase from *Ruminococcus flavefaciens* (GB CAB51934.1); (**e**) chitin synthase from *Emericella nidulans* (GB BAA21714.1); (**f**) cyclicβ-1-3-glucan synthase from *Bradyrhizobium japonicum* (GB AAC62210.1).

MANUAL FUNCTIONAL ANALYSIS

All too often, functional annotation methods employed during whole genome annotation are erroneous and lack consistent language (12,15). While sequence similarity to genes annotated by GO or best BLAST hits can be a good-starting point to assignment to pathways or possible general functions, such as serine/theonine kinase, many automatic functional assignments are unfortunately much more specific. This is particularly true in the case of CAZymes, since related families of the latter group together enzymes of widely differing specificity.

The CAZy database employs practices that aim to eliminate the problems with automatic annotation. Biochemical characterization of new proteins from the literature is used to create new protein families, to annotate their referring entries and to update family descriptions (6). These biochemical assignments are also employed to help the manual curator estimate the likely general functions and add descriptions that indicate which enzymatically characterized proteins are related to new sequences. Inclusion of reference data compiled by communities centered on model organisms is considered for the future. Bibliographic references are included in CAZy by a specific layer that includes over 16 000 different bibliographic references. These references were extracted automatically from individual accessions using ProFal (16) and about one-third was entered manually.

When functional predictions are made, they arise from manual curation by examination of closely related sequences and when biochemical information is not available, such as the case for many genome projects, very general functional tags are used to convey general functions of a family. Recently, we have begun further breaking down families into subfamilies in the hope of grouping proteins by specificity using sequence similarity. This would allow us to give more insights into possible functions. This new classification can also give insights into conserved active sites and active site specificity, when comparing biochemically characterized enzymes. Currently subfamily assignments are available publicly only for GH13 (14), GH1, GH2 and GH5 (released with this publication). This effort will be continued in the future with many more subfamilies being incorporated into the CAZy knowledge base in the future. Subfamilies identify subgroups of sequences that are more homogeneous in their functional properties. Most identified subfamilies are monospecific. If polyspecific, the functional variability is low and typically limited to two or three EC activities. There, often the known subfamily functions often share a substrate or product. Furthermore, rational enzyme engineering may be used to switch the functions for several cases (data not shown). Subfamilies also open the door for further enzymatic characterization—a few subfamilies as still no known activity—or for the identification of meaningful targets for structural characterization.

LARGE-SCALE ANALYSIS AND COLLABORATION

Internal CAZy tools, such as our semi-automatic modular assignment presently allow the analysis of a larger number

of sequences than a few years ago, making it possible to perform large-scale analyses, such as the annotation of CAZyme systems in genomes and metagenomic investigations of the breakdown of complex carbohydrates. A typical genome analysis begins with the assignment of protein models to one or several CAZy families (depending on the number of CAZy modules present within the sequence). This family assignment is then followed by the prediction of general functional classes using a manual examination of alignments to closely related sequences, taking care to identify the retention of active-site residues. Once a genome is categorized by family and functional classes, gene content analysis is utilized to give insights into how newly sequenced organisms might be similar or different from closely related species. Differences in genome content, i.e. relative family size, might reflect the relative diversity or complexity of the inherent biological processes (17) and therefore, the biology of the compared species. For example, differences suggesting a more pronounced pectin metabolism in 'dicot' Arabidopsis versus 'monocot' rice have been noted (17) as well as expected differences in cell-wall metabolism between short-lived annual Arabidopsis versus long-lived poplar tree have been suggested (18). With the advent of a variety of post-genomic techniques, a new vision of the CAZymes as significant components of carbohydrate-based systems now emerges. Examples include: *N*- and *O*-glycosylation of proteins, starch metabolism, biosynthesis of the cell-wall and its subcomponents. Geisler-Lee *et al.* (19) have combined bioinformatics and transcriptome analysis of various poplar and Arabidopsis tissues and organs and have shown that CAZyme transcripts are particularly abundant in wood tissues.

NEW FEATURES

In addition to a website facelift, the new CAZy website comes with a host of new features. Primarily, we are now offering users the ability to search the CAZy site for information by GenBank protein accession number, family or organism rather than navigate long static pages as prior to 12/31/2008 (Figure 2). To the new site we are also including, pages to describe new releases, new genomes and other new features. In addition, tools developed in the lab are available for interactive use.

FUTURE TRENDS

The CAZy database is a fluid database always changing and growing as additional data becomes available.

Figure 2. (**A**) Once a search is performed, such as for a protein accession (P00275), the resulting page indicates the modular families that compose that protein. (**B**) Upon clicking the resulting links provided in A, users are directed to a page about the family and gives a listing of all annotated members.

In the last 2 years, the number of sequences in CAZY has nearly doubled and the number of available genomes is over 750. We believe this trend will continue in the coming years. Unfortunately, while sequencing is forever more rapid, progress in structural information and biochemical characterization is much slower. The number of biochemical data has grown only by 8% over the last 2 years (Figure 3). This means that the gap is widening between available sequences and biochemically characterized enzymes, making better methods for high-throughput biochemical characterization advantageous.

As started previously, we are actively pursuing the classification of subfamilies within each family. This further level of classification is important for instance to identify key residues or motifs important to define specificity. Finally, we hope to offer soon a page to submit sequences for a sequence similarity search and keyword search on our website.

AVAILABILITY ON THE WEB

The CAZy database is available at www.cazy.org. Information about selected families is available through the website and at www.cazypedia.org. Software from the group is available at www.cazy.org/tools.

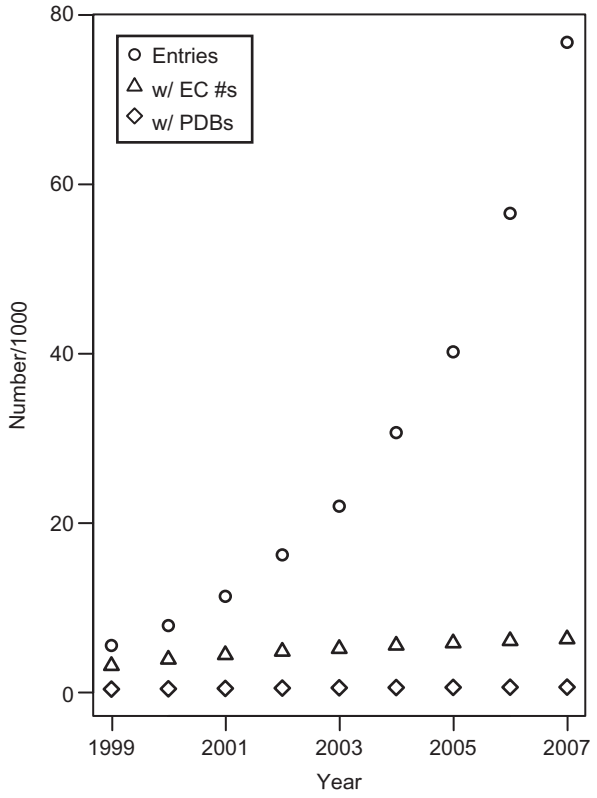


Figure 3. The number of protein containing CAZy modules were noted in December of the years 1999–2007. Within this set (Open circle), the number of enzymatically characterized proteins (triangle) and those with solved structures (open diamond) were also counted. In December 2007, <10% of proteins in CAZy were characterized enzymatically and <1% had a solved structure. In 8 years, the number of sequences has increased 14-fold, while the number of enzymatic and structural characterization has mearly doubled. Therefore, the porportion of proteins with functional and stuctural information is decreasing rapidly unless high throughput functional efforts are made in this category of enzymes.

FUNDING

The authors wish to thank the Departement des Sciences de la Vie of CNRS for a 2-year funding grant to B.L.C. and Novozymes for a contract supporting V.L.

Conflict of interest statement. P.M.C. is affiliated to Université de Provence (Aix-Marseille I) and B.H. and C.R. are members of CNRS.

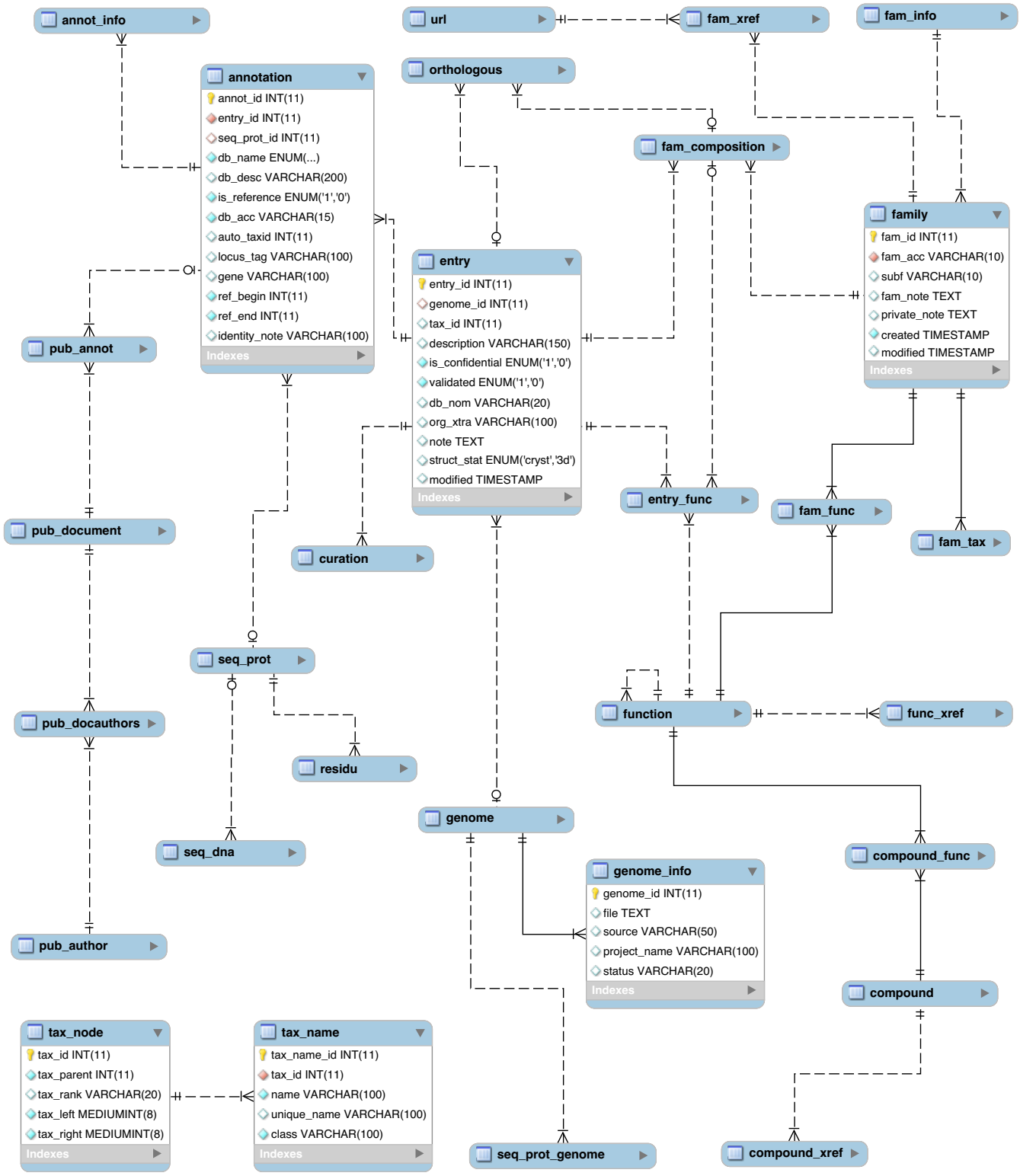
REFERENCES

1. Laine,R.A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields $1.05 \times 10(12)$ structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology*, **4**, 759–767.
2. Henrissat,B., Claeyssens,M., Tomme,P., Lemesle,L. and Mornon,J.P. (1989) Cellulase families revealed by hydrophobic cluster analysis. *Gene*, **81**, 83–95.
3. Henrissat,B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, **280 (Pt 2)**, 309–316.
4. Henrissat,B. and Bairoch,A. (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, **293 (Pt 3)**, 781–788.
5. Henrissat,B. and Bairoch,A. (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem. J.*, **316 (Pt 2)**, 695–696.
6. Yip,V.L. and Withers,S.G. (2006) Breakdown of oligosaccharides by the process of elimination. *Curr. Opin. Chem. Biol.*, **10**, 147–155.
7. Coutinho,P.M. and Henrissat,B. (1999) Carbohydrate-active enzymes: an integrated database approach. In Gilbert,H.J., Davies,G., Henrissat,H. and Svensson,B. (eds), *Recent Advances in Carbohydrate Bioengineering*. The Royal Society of Chemistry, Cambridge, pp. 3–12.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Eddy,S.R. (1995) Multiple alignment using hidden Markov models. In *Proc. Intl Conf. Intel. Syst. Molec. Biol. ISMB*, **3**, 114–120.
11. Davies,G.J., Gloster,T.M. and Henrissat,B. (2005) Recent structural insights into the expanding world of carbohydrate-active enzymes. *Curr. Opin. Struct. Biol.*, **15**, 637–645.
12. Doerks,T., Bairoch,A. and Bork,P. (1998) Protein annotation: detective work for function prediction. *Trends Genet.*, **14**, 248–250.
13. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al.* (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
14. Stam,M.R., Danchin,E.G., Rancurel,C., Coutinho,P.M. and Henrissat,B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel.*, **19**, 555–562.
15. Gilks,W.R., Audit,B., De Angelis,D., Tsoka,S. and Ouzounis,C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics (Oxford, England)*, **18**, 1641–1649.
16. Couto,F.M., Silva,J.M. and Coutinho,P.M. (2003) ProFAL: PROtein Functional Annotation through Literature. In Pimentel,E., Brisaboa,N.R. and Gomez, J. (eds), In *Proceedings of the 8th*

Conference on Software Engineering and Databases, Alicante, Spain, pp. 747–756.
17. Yokoyama,R., Rose,J.K. and Nishitani,K. (2004) A surprising diversity and abundance of xyloglucan endotransglucosylase/hydrolases in rice. Classification and expression analysis. *Plant Physiol.*, **134**, 1088–1099.
18. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, NY)*, **313**, 1596–1604.
19. Geisler-Lee,J., Geisler,M., Coutinho,P.M., Segerman,B., Nishikubo,N., Takahashi,J., Aspeborg,H., Djerbi,S., Master,E., Andersson-Gunneras,S. *et al.* (2006) Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.*, **140**, 946–962.

D

Schéma relationnel de la base de données CAZy


annot_info
url
fam_xref
fam_info
orthologous
annotation
annot_id INT(11)
entry_id INT(11)
seq_prot_id INT(11)
db_name ENUM(...)
db_desc VARCHAR(200)
is_reference ENUM('1','0')
db_acc VARCHAR(15)
auto_taxid INT(11)
locus_tag VARCHAR(100)
gene VARCHAR(100)
ref_begin INT(11)
ref_end INT(11)
identity_note VARCHAR(100)
Indexes
fam_composition
family
fam_id INT(11)
fam_acc VARCHAR(10)
subf VARCHAR(10)
fam_note TEXT
private_note TEXT
created TIMESTAMP
modified TIMESTAMP
Indexes
entry
entry_id INT(11)
genome_id INT(11)
tax_id INT(11)
description VARCHAR(150)
is_confidential ENUM('1','0')
validated ENUM('1','0')
db_nom VARCHAR(20)
org_xtra VARCHAR(100)
note TEXT
struct_stat ENUM('cryst','3d')
modified TIMESTAMP
Indexes
pub_annot
pub_document
curation
entry_func
fam_func
fam_tax
seq_prot
pub_docauthors
residu
function
func_xref
genome
seq_dna
compound_func
genome_info
genome_id INT(11)
file TEXT
source VARCHAR(50)
project_name VARCHAR(100)
status VARCHAR(20)
Indexes
pub_author
compound
tax_node
tax_id INT(11)
tax_parent INT(11)
tax_rank VARCHAR(20)
tax_left MEDIUMINT(8)
tax_right MEDIUMINT(8)
Indexes
tax_name
tax_name_id INT(11)
tax_id INT(11)
name VARCHAR(100)
unique_name VARCHAR(100)
class VARCHAR(100)
Indexes
seq_prot_genome
compound_xref

Structuration et exploration d'information génomique et fonctionnelle des enzymes actives sur les glucides.

Les glucides sont très répendus dans la nature et sont impliqués dans une multitude de phénomènes biologiques. Sous forme de saccharides et de glycoconjugués, ils constituent une partie substantielle de la biomasse produite sur terre et représentent une source potentielle d'énergie renouvelable de première importance. La diversité des glucides complexes est créée et contrôlée par un panel d'activités enzymatiques qui interviennent dans leur assemblage, dégradation et modification. L'étude structurale et fonctionnelle des enzymes actives sur les glucides (CAZymes) est à la base de multiples efforts de recherche appliquée en biotechnologie. L'industrie recherche actuellement des enzymes avec des activités et des spécificités encore plus performantes. L'activité de recherche de ces nouvelles enzymes est grandement facilitée par l'accumulation de séquences biologiques dans les bases de données, provenant notamment des études génomiques.

Mon sujet de recherche s'inscrit dans un objectif de développement d'outils pour la classification et l'identification de nouvelles enzymes impliqués dans la conversion de la biomasse. Tous ces travaux sont en lien direct avec la mise en place d'une nouvelle infrastructure de la base de données CAZy et l'analyse de données génomiques, métagénomiques et biochimiques. La refonte complète de la structure de la base de données préexistantes et de son interface a été ainsi réalisée. Cet effort a été validé par l'analyse des familles de polysaccharide lyases et la création de sous-familles, dont l'homogénéité fonctionnelle a été révélée. De plus, la détection systématique de protéines modulaires portant des modules d'adhésion aux composants de la paroi végétale a permis l'identification de nouvelles protéines potentiellement impliquées dans la dégradation de la biomasse végétale. Enfin, j'ai implémenté des approches automatisées capables d'analyser de grands volumes de données (méta)génomiques pour en extraire le contenu en CAZymes.

Structuration and exploration of genomic information and functional enzymes acting on carbohydrate-active enzymes.

Carbohydrates are widely distributed in nature, where they are involved in a multitude of important biological events. Saccharides and glycoconjugates constitute the main component of the biomass produced on earth, therefore they represent a plentiful source of renewable energy. The diversity of complex carbohydrates is created and controlled by a panel of enzyme activities involved in their assembly, degradation and modification. The structural and functional study of Carbohydrate Active enZymes on (CAZymes) has been the basis for many applied research efforts in biotechnology. For exemple, the biotechnology industry is currently searching enzymes with enhanced activities and specificities. The identification of new enzymes is potentially facilitated by the large-scale accumulation of gene sequences, particularly from current genomic studies.

This thesis aimed at developing tools for the classification and identification of new enzymes involved in biomass degradation. To this end, a new structure of the CAZy database was developed and applied to mining genomic, metagenomic and biochemical data. A complete reorganisation of the structure of the existing database and its interface has been achieved. In this effort the analysis of all known families of polysaccharide lyases has been validated and subfamilies were created, which revealed functional homogeneity. In addition, the systematic identification of modular proteins containing plant cell wall binding modules allowed the identification of new proteins potentially targeting plant biomass. Finally, I show that it is indeed possible to analyze large volumes of (meta)genomic data by automated methods in order to understand their CAZyme contents.