Etude structurale et fonctionnelle des polysaccharidases de Rhodopirellula baltica

Jérôme Dabin

THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Glycobiologie marine
Ecole doctorale Inter///Bio

Présentée par

M. Jérôme DABIN

Pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Etude structurale et fonctionnelle des polysaccharidases de *Rhodopirellula baltica*

Soutenue le 28 novembre 2008

devant le jury composé de :

M. Le Professeur	Bernard KLOAREG	Directeur de thèse
M. Le Docteur	Gurvan MICHEL	Co-directeur de thèse
M. Le Professeur	Steven BALL	Rapporteur
M. Le Professeur	Pédro COUTINHO	Rapporteur
Mme La Professeure	Alessandra CARBONE	Examinateur
M. Le Professeur	Franck-Oliver GLÖCKNER	Examinateur
Mme La Professeure	Claudine MAYER	Examinateur

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à ME : 01 42 34 68 41
Tél. pour les étudiants de MF à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

Car on ne réussit jamais seul…

Que de personnes à remercier pour ces années passées dans ce laboratoire !
Des personnes que j'ai croisées, certaines qui sont parties avant moi, et d'autres qui sont arrivées depuis …

Je tiens tout d'abord à remercier Mme Catherine BOYEN, directrice de l'UMR 7139 *Végétaux Marins et Biomolécules* ainsi que Mr Bernard KLOAREG, directeur de la Station Biologique de Roscoff et directeur de ma thèse, pour m'avoir accueilli au sein de leur laboratoire.

Je souhaiterais également livrer des remerciements chaleureux à toutes les personnes qui m'ont aidé à aller de l'avant et ont contribué à ce que ces années soient heureuses :

Mes premiers remerciements vont à mon équipe qui a été un vrai soutien, tant logistique qu'humain et m'a permis de survivre (surtout vers la fin !) :

- Merci tout d'abord à toi Gurvan, de m'avoir fait confiance et de m'avoir montré la dure voie du chercheur… Et je te rassure, prendre un physico-chimiste théoricien pour bosser sur une bactérie marine n'aura manqué ni d'impact, ni d'ambition !
- Mirjam et Tristan, vous avez été de vrais parents-poule pour moi, votre aide et vos encouragements ont fait une vraie différence dans ma lutte contre l'adversité;
- Murielle, sans qui le laboratoire cesserait de tourner ! Ton aide et ton amitié au cours de ma thèse m'auront apportés réconfort et (surtout ?) bons plans musicaux !
- Jan-Hendrick et Etienne, mes colistiers *Cheese Cake Boys* !
- Alexandra, Justina, Agnès, et toutes les personnes qui font ou ont fait de cette équipe une petite famille ;
- Babsi, pour sa bonne humeur légendaire, et qui m'a remonté le moral quand ses manips n'avançaient plus (!)

Beaucoup d'autres personnes ont traversé ma vie durant ma thèse, je souhaiterais donc leur témoigner toute mon affection :

- François, qui m'aura tant apporté, soutenu et nourri au grain… Ta chaleur humaine m'a plus d'une fois réconforté ! Ne cesses jamais d'être un Bouletto™ ;
- Alex, bien sûr, dont je crois avoir égayé le post-doc et dont j'espère garder l'amitié encore longtemps. Nous avons partagé tant d'épingles et tant de côtes de bœuf ensemble…
- Cécile, dont la vision crêpée et « Bob-like » restera à jamais gravée dans ma mémoire. A nos marmottes !

- Audrey, aussi sérieuse que drôle (c'est dire !), qui m'aura fait découvrir les Orgasmes Givrés™ et les très fameux *petits encas Banane-Roquefort* !
- Sabine, ma copaillassière (je sais, ce n'est pas très beau), qui a dû lutter contre ma naturelle propension à l'annexion territoriale et dont l'énergie est si communicative ;
- Marion, Manon, Vincent, Maud, PO, Steph, Diane, Gigi, Simon, Sarah, … et tous les thésards de la station, qui font de cet endroit un vrai petit cocon !
- Ludo (qu'il ne faut jamais sous-estimer !), Gaëlle (qui est la seule personne que je connaisse… qui connaisse absolument tout le monde !) et toute l'UMR 7139 bien sûr !
- La Station Biologique, microcosme marin où il fait si bon vivre ;

C'est un peu grâce à vous tous que j'ai compris la différence entre une algue…
« ha ?!? Mais il y en a plusieurs espèces ?!? » (sic)

Je souhaiterais également remercier toutes les personnes qui m'ont aidé pour la plupart depuis des années, parfois à distance, toujours avec gentillesse :

- Mes parents, Yves et Nicole, qui sont mes éternels soutiens, et qui m'ont donné bien plus que je ne pourrais jamais leur rendre ;
- Claire, ma petite sœur préférée… Et dire que tu n'es toujours pas venu à Roscoff ! Mais tu vas me reprocher d'être parti sans te laisser mon adresse !
- Mamie, Mamy, Papi, Papy et bien sûr toute ma famille ;
- Sissy et Toto (et Lina !), Bern, Luc, Philippe, Dav', Nono, … tant d'histoires avec chacun d'entre vous… Je les raconterais dans le Tome II !

Je souhaiterai enfin apporter un remerciement sincère à Mr Jean Delettré, directeur de l'école doctorale Inter///Bio au début de ma thèse, qui a cru en moi et grâce à qui j'ai pu avoir à écrire ces quelques pages.

Merci à vous!

Sommaire

Table des illustrations

Chapitre I

-

Introduction

I - Prologue : « Le projet Apollo de la biologie »

Cette année a célébré le cinquante-cinquième anniversaire de la découverte de la structure de l'acide désoxyribonucléique (ADN) par Watson et Crick (Watson and Crick, 1953), en même temps que le septième anniversaire de la publication annonçant le séquençage complet du génome humain (Venter JC *et al.*, 2001). Entre ces deux dates, 50 années d'évolution des mentalités sur l'utilité des constituants cellulaires, et de l'ADN en particulier, ainsi que des techniques de séquençage, ont été nécessaires pour aboutir à l'une des plus grandes révolutions scientifiques du vingtième siècle.

L'histoire de l'ADN commence à la fin du XIX^{ème} siècle, avec l'isolement de la « nucléine » des noyaux de globules blancs par Friedrich Miescher (Dahm, 2008). Le rôle précis de cette substance étrange, qui ne fait partie ni des protides, ni des glucides, ni des lipides, et dont ont savait juste qu'elle était riche en phosphore, est longtemps resté assez obscur pour les biologistes. On lui assignait dans les premiers temps suivant sa découverte, au mieux, un rôle de structuration du compartiment cellulaire, à la manière des polysaccharides. Les protéines, découvertes quasiment un siècle plus tôt, étaient alors considérées comme les composants cellulaires principaux, assurant les rôles fondamentaux dans le fonctionnement cellulaire, à commencer par celui de support de l'hérédité. La première démonstration du rôle de l'ADN dans le stockage de l'information génétique a eu lieu au cours des années 1940 en s'appuyant sur des expériences de transfert de plasmides entre deux souches de *Streptococcus pneumoniae* (voir Avery *et al.*, 1995). Il a cependant fallu attendre une autre dizaine d'années pour finir de convaincre, non sans mal, la communauté des biologistes. Cette nouvelle fonction de « support de l'information génétique » a permis de donner une base moléculaire au concept théorique de gène développé par Gregor Mendel, au cours de ses expériences sur l'hérédité (Mendel 1865). Malgré cette découverte majeure, la vision que beaucoup de biologistes avaient de l'ADN est restée très simplificatrice, ce dernier n'étant considéré que comme une sorte de collection de gènes, finalement assez passive dans la cellule.

L'influence certaine de la communauté des biochimistes, associée à la maîtrise de nombreuses techniques d'étude des protéines, continua de donner à la biochimie des protéines un grand avantage. Les difficultés techniques de leur manipulation constituaient en effet le principal frein à l'étude des acides nucléiques. A l'inverse, depuis le milieu des années 1950, la détermination de la séquence de l'insuline par Sanger (Sanger *et al.*, 1955) avait ouvert la voie du séquençage des protéines. Les premières séquences d'ADN,

obtenues à la fin des années 1960 après des efforts laborieux, ont été réalisées indirectement en séquencant, par des digestions enzymatiques ménagées, une copie ARN de la molécule d'ADN (voir Tableau I-1).

Année	Protéine	ARN	ADN	Nombre de résidus
1935	Insuline			1
1945	Insuline			2
1947	Gramicidine S			5
1949	Insuline			9
1955	Insuline			51
1960	Ribonucléase			120
1965		$ARNt_{Ala}$		75
1967		ARNr 5S		120
1968			Bactériophage λ	12
1978			Bphage ΦX174	5 386
1981			Mitochondrie	16 569
1982			Bactériophage λ	48 502

Tableau I-1 : Historique des travaux de séquençage jusqu'en 1982.
D'après le «Laboratory of Molecular Biology» (Cambridge) http://www2.mrc-lmb.cam.ac.uk/ archive/Sanger58.html

C'est en 1977 qu'ont été décrites, pour la première fois dans la littérature, et par deux équipes différentes, deux méthodes efficaces de séquençage d'acides nucléiques. La première, publiée par Maxam et Gilbert (Maxam and Gilbert, 1977), permet le séquençage de l'ADN par une série de réactions de coupure de l'ADN cloné dans un bactériophage, à l'aide de réactifs chimiques tels que l'hydrazine et la pipéridine. La seconde, publiée six mois plus tard par Sanger (Sanger *et al.*, 1977), s'appuie quant à elle sur la synthèse in vitro de l'ADN en présence d'une ADN polymérase, des 4 désoxyribonucléotides et des 4 didésoxyribonucléotides, synthétisés spécialement pour le séquençage de l'ADN, qui provoquent l'arrêt de la réaction de polymérisation. Une fois ces 4 didésoxyribonucléotides commercialement disponibles, c'est la méthode de Sanger qui a permis le réel développement du séquençage de l'ADN. C'est également au cours de ces années que les premières enzymes de restriction ont été découvertes, permettant une manipulation beaucoup plus facile de l'ADN. Néanmoins, le décalage entre la maîtrise des techniques liées à la manipulation de l'ADN et de celles des protéines s'est poursuivi jusqu'au début des années 1980. Durant cette période, le séquençage des protéines est resté nettement plus fréquent que celui de leur gène. Il est possible de dater le moment où la séquence des protéines va commencer à être déduite de la séquence nucléotidique de leur gène par la

publication de la séquence du gène de la β-galactosidase (Kalnins *et al.*, 1983) qui démontrera la qualité des séquences obtenables par cette méthode.

Les premières séquences d'ADN publiées après la mise au point des méthodes modernes de séquençage étaient de petite taille et correspondaient biologiquement à la séquence d'un gène et de ses régions régulatrices proches. L'idée de séquencer des génomes complets s'est cependant très vite répandue avec les avancées technologiques dans ce domaine. En effet, moins d'un an après la publication de la méthode de séquençage par Sanger, la première séquence complète du génome du bactériophage ΦX174 était publiée (Sanger *et al.*, 1978). Ce génome viral a, il est vrai, une petite taille (5386 pb) mais il a confirmé l'idée selon laquelle la connaissance de l'information génétique complète d'un organisme était non seulement accessible mais aussi intéressante scientifiquement.

C'est à partir de ces années de développement concomitant des techniques de séquençage et de l'exploitation des séquences, que la génétique moderne a pris son essor. Les années 1990 ont en particulier été celles de l'avènement de la biologie moléculaire. C'est avec l'accélération du développement des techniques ainsi que l'arrivée des premiers séquenceurs automatiques qu'est née l'idée du séquençage complet du génome humain[1]. Le premier projet a été lancé aux USA, dès 1986, par le département de l'énergie (Department of Energy ; DOE), suivi de peu en 1987 par l'institution de financement du secteur biomédical (National Health Institute ; NIH). En 1993, ces deux projets ont fusionné et un consortium international de laboratoires publics et privés a été créé afin de mener à bien le *Human Genome Project* (HGP). Dix ans plus tard, l'ampleur de ce projet lui vaudra d'être comparé à la conquête spatiale effrénée des années 1950, devenant le « projet Apollo de la biologie ».

Cette entreprise aura été un grand succès. Méthodologique tout d'abord, puisque les développements technologiques autour du HGP ont devancé de loin toutes les prévisions, en particulier grâce à l'implication de l'industrie. De nombreuses séquences de génomes ont de ce fait été publiées, bien avant celle de l'être humain (Tableau I-2). La gamma-protéobactérie *Haemophilus influenzae* a ainsi été le premier organisme dont le génome ait été séquencé (Fleischmann *et al.*, 1995). Le Tableau I-2 présente les premiers génomes séquencés dans chaque grand phylum.

[1] Le mot « génome » est issu de la contraction des mots *gene* et *chromosome*

Année	Type	Espèce	Taille (Mpb)	Nombre de gènes
1995	Bactérie	*Haemophilus influenzae*	1,8	1 650
1996	Archée	*Methanococcus jannaschii*	1,66	1750
1996	Levure	*Saccharomyces cerevisiae*	12	6000
1998	Ver	*Caenorhabditis elegans*	97	18 000
2000	Plante	*Arabidopsis thaliana*	115	25 500
2000	Insecte	*Drosophila melanogaster*	120	14 000
2001 - - 2004	Homme	*Homo sapiens*	3 000	50 000

Tableau I-2 : Historique du séquençage.
Historique des premiers génomes séquencés de plusieurs grands phyla de l'arbre de la vie. *H. influenzae* (Fleischmann *et al.*, 1995) ; *S. cerevisae* (Goffeau *et al.*, 1996) ; *M. janaschii* (Bult *et al.*, 1996) ; *C. elegans* (Consortium, 1998); *A. thaliana* (Consortium, 2000) ; *D. melanogaster* (Adams MD *et al.*, 2000) ; *H. sapiens* (Venter JC *et al.*, 2001).

Depuis la fin des années 1990 et jusqu'à aujourd'hui, de nombreuses initiatives de séquençage ont vu le jour et la liste des génomes publiés n'a ainsi pas cessé de croître pour atteindre le nombre de 8826 génomes séquencés et répertoriés dans la banque de données du NCBI (National Center for BioInformatics) en 2008 (Figure I-1).

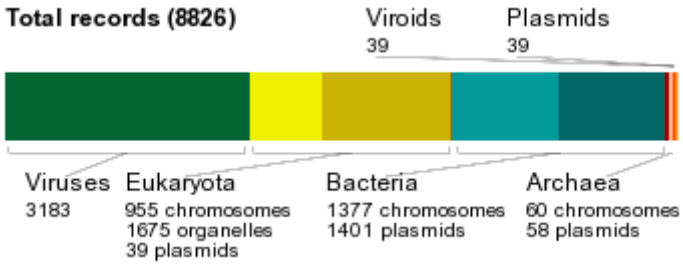


Figure I-1 : Total des génomes publiés en 2008.
Figure extraite du site du NCBI.

Ensuite, les progrès technologiques ont eu pour conséquence directe la publication du génome humain quatre ans avant la date prévue, même s'il aura fallu attendre 2004 pour en avoir une version finalisée et à haute résolution (Consortium, 2004). La coordination des différentes équipes internationales ainsi que l'interconnexion de plusieurs disciplines scientifiques ont abouti à la création d'une grande discipline d'interface qui a été appelée « **génomique** ». La génomique est donc plus une interface entre de nombreuses disciplines tournant autour de l'analyse des génomes qu'une technique d'étude en soit. Sur ce modèle, de nombreuses disciplines ont émergé dès le début des années 1990, s'intéressant à un

aspect particulier de l'analyse des différents génomes. Ainsi, si la génomique a été caractérisée par la création de banques de données et la génération d'un nombre important de séquences issues de génomes divers, la « post-génomique » relie l'ensemble des disciplines liées à l'exploitation de l'incroyable quantité d'information qui en découle. La liste suivante se propose d'en définir les plus générales et de souligner les formidables informations que l'on peut tirer de l'ensemble des données générées par la génomique.

La **génomique fonctionnelle** tout d'abord, repose sur l'analyse du fonctionnement des gènes et des autres composantes du génome. En effet, puisque le génome s'exprime au niveau d'un organisme, la connaissance des gènes, de leur produit, de leur régulation et de leurs interactions avec un environnement donné est indispensable pour prédire le fonctionnement de cet organisme dans cet environnement. Tandis que la génétique moléculaire classique s'intéresse tout au plus à quelques séquences à la fois, la génomique fonctionnelle opère, en parallèle, sur plusieurs centaines ou milliers de séquences d'ADN et de protéines, fournies par les projets de séquençage. Les différents outils bioinformatiques et les méthodes biologiques automatisées ont en particulier permis l'essor de cette discipline. Parmi ces techniques automatisées se trouve la famille des « omiques ». Si elles ne sont pas systématiquement impliquées dans des études de génomique, elles sont cependant nées de la volonté d'avoir un traitement le plus automatisé possible afin de permettre une analyse à grande échelle (sous-entendu sur plusieurs cibles *simultanément*). Par exemple on peut citer l'analyse de voies métaboliques ou de processus cellulaires complexes et ce à plusieurs niveaux d'échelle. La **transcriptomique** va ainsi s'intéresser aux phénomènes de transcription en analysant les ARN messagers produits, par exemple, après modification du contexte environnemental d'un organisme, afin d'étudier l'activation ou la répression de ses gènes (Gomase and Tagore, 2008). La **protéomique** est l'étude de l'ensemble des protéines d'un organisme à un temps donné dans des conditions expérimentales données (Pandey and Mann, 2000; Carrette *et al.*, 2006). La **métabolomique**, enfin, est la dernière-née de ces techniques. Son projet d'étude est la compréhension des flux de métabolites entre des populations de protéines (typiquement les cascades de transformations d'un substrat dans une voie métabolique) suite, encore une fois, à une modification environnementale (Styczynski *et al.*, 2007; May *et al.*, 2008).

La **génomique comparative** est un autre versant de la génomique fonctionnelle. Elle se propose d'étudier, cette fois, la fonction des gènes en comparant les différents génomes séquencés (Galperin and Kolker, 2006; Médigue and Moszer, 2007). Annoter un génome consiste à réaliser une succession de comparaisons de gènes ou de leur produit, avec des données bibliographiques disponibles dans d'autres organismes. La génomique comparative

cherche à déceler les originalités d'un génome ou, au contraire, à mettre en évidence les gènes, soit ubiquitaires au sein d'un environnement, soit indispensables à la survie d'une communauté, voire indispensables à la vie elle-même. Par exemple, l'ensemble des gènes impliqués dans les mécanismes de sécrétion a été identifié en analysant les génomes de plusieurs espèces de bactéries du genre *Bacillus*, permettant de définir son « sécrétome » (Tjalsma *et al.*, 2004). De la même manière, des études ont identifié un certain nombre de gènes présents dans l'ensemble des génomes séquencés (tous phyla confondus) allant dans le sens de la définition des gènes indispensables à la vie et suggérant l'idée d'un « génome minimum » (Barbazuk *et al.*, 2000; Check, 2002; Mizoguchi *et al.*, 2007; Gibson *et al.*, 2008).

La **métagénomique**, ou encore génomique environnementale, est l'étude des métagénomes, terme qui renvoie à l'ensemble des séquences d'ADN extraites de communautés multi-espèces prélevées à partir d'un environnement particulier. Ces communautés d'espèces sont le plus souvent composées d'organismes non-cultivables[2]. Il en résulte que le clonage des ADN issus de ces organismes non-cultivables en vue de leur séquençage doit être effectué directement à partir d'échantillons prélevés dans l'environnement. Plusieurs lignées d'organismes unicellulaires ont ainsi pu être identifiées par des approches de phylogénétique moléculaire (Schmidt *et al.*, 1991; Rodríguez-Valera, 2004; Schloss and Handelsman, 2005).

Enfin, la **génomique structurale** est l'étude de la structure tridimensionnelle des protéines issues des données de génomique. Cette discipline s'est proposé d'agir en synergie avec la génomique fonctionnelle. Il est en effet établi depuis de nombreuses années que structure et fonction sont liées (Fitch, 2000) et, dans le but d'identifier la fonction d'une grande quantité de protéines, une approche structurale apparaissait comme très puissante. Elle a bénéficié d'un fort soutien de la part de la communauté scientifique ainsi que des organismes de recherche. Ainsi, au début des années 2000, de grands projets de génomique structurale (PGS) ont vu le jour. Le projet *Protein Structure Initiative* (PSI) aux USA est à ce titre le plus important (*Nature Methods Editorial*, 2008), mais d'autres, au Canada, au Japon et en Europe, notamment, ont également été créés. Dans le cadre de la poursuite des travaux de génomique, ces PGS se sont focalisés sur la résolution de structures à haut débit par radiocristallographie aux rayons X. Le but premier de ce type de

[2] Il a été estimé que seul 1% des procaryotes de la plupart des environnements peuvent être cultivés (Amann *et al.*, 1990).

projets était véritablement de résoudre une structure pour chaque repliement possible, afin de permettre, par des méthodes de modélisation, la prédiction de la fonction de protéines inconnues à partir de protéines caractérisées. Le remplissage de l'espace des repliements devait également faciliter le phasage par remplacement moléculaire lors d'études cristallographiques. Les PGS ont produit pendant des années et produisent encore un grand nombre de structures (jusqu'à 200 par an pour PSI-2). A l'heure actuelle, la moitié des nouvelles structures déposées dans la banque de structures protéiques (Protein Data Bank ; PDB) a été résolue par un PGS. Le Tableau I-3 est extrait de Fox (Fox *et al.*, 2008) et résume la production de structures réalisées par les principaux PGS dans le monde depuis leur création d'après le site de *TargetDB* (http://targetdb.pdb.org/).

		Total	Asie[a]	Europe[b]	PSI[c]	PSI PC[d]	Autres PGS[e]
Sélection	PSI-1	66 490			64 262	19 611	
	PSI-2	87 653			79 582	75 815	
	Total	154 143	5 923	2 641	143 844	95 426	1 990
Clonage	PSI-1	50 117			48 454	14 472	
	PSI-2	56 815			49 326	46 848	
	Total	106 932	5 876	1 768	97 780	61 320	1 508
Purification	PSI-1	7 260			6 737	3 941	
	PSI-2	19 608			13 742	13 090	
	Total	26 868	5 218	599	20 479	17 031	572
Structures publiées	PSI-1	1 055			938	691	
	PSI-2	4 266			1 507	1 390	
	Total	5 321	2 680	103	2 445	2 081	93
Travail stoppé[f]	PSI-1	11 848			11 806	6 631	
	PSI-2	15 993			15 702	15 419	
	Total	27 841	36	45	27 508	22 050	252

Tableau I-3 : Résumé des résultats de génomique structurale.

Données tirées de *TargetDB* (http://targetdb.pdb.org/) (obtenues en décembre 2007) couvrant les périodes de temps PSI-1 (jusqu'à septembre 2005) et PSI-2 (d'octobre 2005 à maintenant). Tableau extrait de Fox (Fox *et al.*, 2008).

[a] RIKEN *Structural Genomics/Proteomics Initiative*, Japon.

[b] Les PGS d'Europe incluent *Bacterial Targets* à Information Genomique et Structurale-Centre National de la Recherche Scientifique, France (BIGS), *Israel Structural Proteomics Center* (ISPC), *Marseilles Structural Genomics Program*, France (MSGP), *Oxford Protein Production Facility*, England (OPPF), *Structural Genomics Consortium*, England-Canada-Sweden (SGC), *Structural Proteomics in Europe*, England (SPINE), *Mycobacterium Tuberculosis Structural Proteomics Project*, Germany (XMTB), et Paris-Sud *Yeast Structural Genomics*, France (YSG).

[c] Les centres PSI incluent : *Accelerated Technologies Center for Gene to 3D Structure* (ATCG3D), *Berkeley Structural Genomics Center* (BSGC), *Center for Eukaryotic Structural Genomics* (CESG), *Center for High-Throughput Structural Biology* (CHTSB), *Center for Structure of Membrane Proteins* (CSMP), *Integrated Centers for Structure and Function Innovation* (ISFI), J*oint Center for Structural Genomics* (JCSG), *Midwest Center for Structural Genomics* (MCSG), N*orthEast Structural Genomics Consortium* (NESG), *New York Consortium on Membrane* Protein Structure (NYCOMPS), *New York Structural Genomics Research Consortium* (NYSGXRC), *Southeast Collaboratory for Structural Genomics* (SECSG), *Structural Genomics of Pathogenic Protozoa Consortium* (SGPP), et *Mycobacterium Tuberculosis Structural Genomics Consortium* (TBSGC).

[d] PSI Production Centers. Les centres de production du PSI-2 sont : *Joint Center for Structural Genomics*, *Midwest Center for Structural Genomics*, *Northeast Structural Genomics Consortium*, et *New York Structural Genomics Research Consortium*.

[e] Les autres PGS repertories dans *TargetDB* sont : *Montreal-Kingston Bacterial Structural Genomics Initiative*, Canada (BSGI), et *Structure 2 Function Project*, USA (S2F).

[f] L'arrêt du traitement d'une cible sélectionnée peut survenir après des échecs expérimentaux, des avancées significatives d'autres programmes, ou une redéfinition de la sélection.

Le principal accomplissement de ces projets aura été l'extraordinaire développement méthodologique qui a permis d'automatiser, presque entièrement, les processus de production de protéines recombinantes, leur purification et leur cristallisation, de la collecte des données cristallographiques au traitement des données. Cependant, leur mission de découvrir de nouveaux repliements s'est heurtée au fait que si la structure d'une protéine est fortement corrélée à sa fonction, déduire l'une de l'autre est loin d'être trivial.

En effet, la fonction d'une protéine réside très précisément dans d'infimes détails structuraux (de Souza, 2007), rendant difficile les inférences fonctionnelles entre structures *a priori* voisines. La raison d'être des PGS a donc, au fil des ans, été redéfinie pour correspondre à des buts plus accessibles et surtout pour permettre une interaction forte avec les projets de génomique fonctionnelle. Une pléthore de revues traitent de ce sujet et sont actualisées chaque année (Eisenstein, 2007; Blow, 2008; *Nature Methods Editorial*, 2008; Fox *et al.*, 2008).

II - De l'annotation des génomes

La génomique ne se contente pas de fournir à la communauté scientifique des séquences brutes de génomes. Une de ses attributions principales, après le séquençage, est de fournir un génome qui présente non seulement une séquence fiable mais qui soit également annoté correctement. Le processus d'annotation est ainsi une problématique centrale dans la publication d'un génome et la qualité de l'étape d'annotation est primordiale pour exploiter, dans le futur, le génome de l'organisme en question. Sont développées, ici, les règles fondamentales de l'annotation fonctionnelle d'un génome, ainsi que les conditions nécessaires pour la mener à bien.

II.A - Notions préliminaires

II.A.1 - Notions d'homologie

L'homologie entre deux séquences est le lien d'ancestralité qui les relie. Ainsi, deux gènes seront dits **homologues** s'ils ont évolué à partir d'un ancêtre commun (Fitch, 2000). Des homologues peuvent dériver de cet ancêtre commun par deux types de mécanismes évolutifs: soit ils sont issus d'un évènement de spéciation entre deux organismes, et ils seront alors qualifiés de gènes **orthologues** ; soit ils sont issus d'un évènement de duplication à l'intérieur d'un même génome, et ils seront alors qualifiés de gènes **paralogues** (Figure I-2). Il est à noter au passage que la duplication génique est actuellement considérée comme un des mécanismes majeurs à l'origine de la diversité des fonctions biologiques. En effet, la copie d'un gène ne subit pas forcément la même pression de sélection que le gène parent (Ohno *et al.*, 1968; Galperin and Koonin, 1999) . La notion d'homologie de gènes est ainsi une notion qualitative qui n'est pas mesurable. Elle présente en outre la caractéristique d'être transitive: on peut déduire que deux gènes A et B sont homologues si les gènes A et C, et B et C, respectivement, sont homologues.

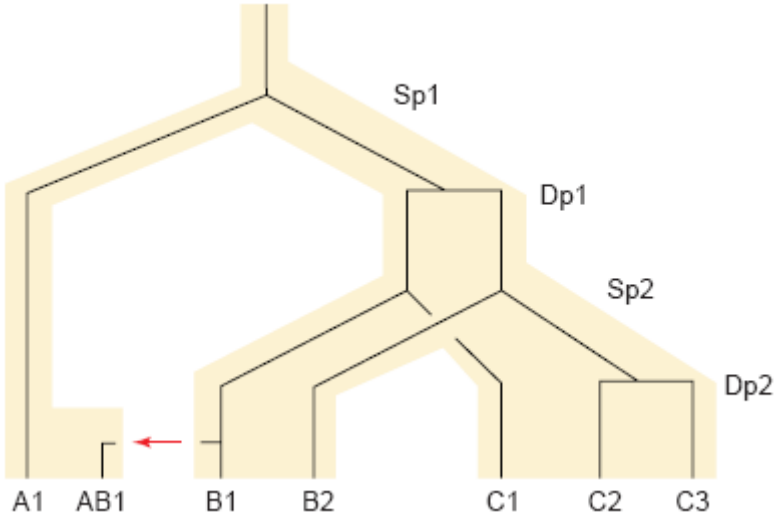


Figure I-2 : Présentation des phénomènes d'homologies.
Sp décrit une spéciation. *Dp* décrit une duplication de gène. La flèche rouge décrit un transfert horizontal.Figure extraite de Fitch (2000).

L'annotation d'un gène consiste à exploiter ses relations d'homologie avec des séquences de gènes proches et caractérisés. S'il est possible de prouver que deux séquences sont effectivement homologues, c'est-à-dire issues d'une séquence ancestrale, on peut alors en déduire, par principe, que leur fonction a de fortes chances d'être similaire mais pas forcément identique. En effet, la transmission de la fonction d'une séquence

ancestrale à ses descendantes n'est pas systématique, étant donnés les évènements de mutation ayant permis la divergence des séquences filles. Ainsi, deux séquences dites homologues peuvent ne pas avoir conservé la même fonction que leur ancêtre et *a fortiori* ne pas présenter la même fonction entre elles (Fitch, 2000).

Le calcul de la similitude entre deux séquences (exprimée en pourcentage) est l'outil utilisé afin d'estimer la validité d'un lien d'homologie. La similitude est le taux de conservation des résidus entre deux séquences (en prenant en compte les phénomènes de substitution et ceux d'insertion ou de délétion). La similitude globale entre deux séquences s'exprime en terme d'identité qui est le taux de conservation stricte des résidus et en terme de similitude (ou similarité) qui est le taux de substitution de résidus similaires.

II.A.2 - Notion de modularité

Les protéines ne présentent pas toujours un unique domaine structural sur l'entièreté de leur séquence. Il arrive en effet qu'elles soient consitituées d'une succession de modules le long de leur structure primaire. Un module se définit comme un fragment de séquence protéique présentant un repliement autonome et donc une fonction propre. Les modules d'une protéine peuvent être liés entre eux par des relations d'homologies ou au contraire ne présenter aucune similitude. Ce type de situation peut compliquer l'annotation d'un gène car les fonctions des modules constituant la protéine peuvent être très différentes. Un exemple, dans le cadre du métabolisme d'un polysaccharide sulfaté, pourrait être une enzyme présentant un module hydrolysant le polymère neutre, lié à un module sulfatase hydrolysant quant à lui les groupements sulfates du polymère chargé. L'activité globale de l'enzyme est bien la dégradation du polymère sulfaté, mais cette fonction n'émerge que de l'addition des activités de ses modules.

II.B - La bioinformatique au service de l'annotation

L'inférence fonctionnelle sur une séquence protéique inconnue commence par la recherche d'homologues (souvent orthologues) à cette séquence. Si la similitude entre cette séquence et son homologue est suffisante, c'est-à-dire supérieure à 25% (qui est considérée comme la limite en-dessous de laquelle il n'est plus possible de parler d'homologie) (Rost, 1999), alors il est possible de transférer la fonction de l'homologue vers la protéine étudiée. Plus la similitude sera grande, plus grande sera la confiance dans l'annotation. Cette analyse

de similitude utilise des algorithmes de comparaison de séquences, implémentés dans différents logiciels (le plus souvent accessibles en ligne), qui réalisent des alignements à partir de séquences stockées dans plusieurs banques de données, spécialisées ou non. Bien que ce soit le gène qui reçoive l'annotation, la recherche d'homologie est le plus souvent réalisée avec son produit, la séquence protéique. Les résultats d'une recherche à partir des séquences nucléotidiques sont en effet beaucoup plus difficiles à interpréter car l'utilisation d'un alphabet à 4 lettres aboutit à l'obtention de similitudes dues au hasard plus fréquemment qu'avec un alphabet à 20 lettres.

II.B.1 - Les alignements de séquences

Il existe plusieurs types d'alignements de séquences, qui reposent néanmoins sur des principes fondateurs. Les alignements par paires sont réalisés soit de façon globale (en alignant les deux séquences sur leur longueur entière), soit de façon locale (sur des segments de ces séquences). Les alignements multiples se proposent eux d'aligner des ensembles de séquences à travers la recherche de profils communs.

D'une manière générale, les alignements sont réalisés en recherchant une amorce de similitude entre les séquences, c'est-à-dire une portion de séquence commune présentant une similitude entre les séquences alignées. L'estimation de la similitude est réalisée en utilisant des matrices de scores de substitution définissant la probabilité qu'un acide aminé puisse être remplacé par un autre. Il existe plusieurs types de matrices, présentant chacune ses spécificités, mais deux sont particulièrement utilisées en biologie de par leur efficacité à rendre compte des phénomènes de substitutions d'acides aminés. Ainsi, les matrices de type BLOSUM (BLOcks Sustitution Matrix) (Henikoff and Henikoff, 1992) reposent sur les substitutions observées dans un alignement multiple de séquences protéiques, réalisé sans insertion, ni délétion. Les matrices de type PAM (Point Accepted Mutation) (Dayhoff *et al.*, 1978) reposent quant à elles sur des probabilités de mutation entre acides aminés, ajustées à un niveau de distance évolutive donnée. Il existe en effet, pour chacune de ces matrices, des versions permettant d'aligner des séquences plus ou moins similaires (de 30% à 90% de similitude).

Afin de rendre compte plus précisément des évènements évolutifs qui permettent de passer d'une séquence à une autre, les alignements cherchent également à inclure les phénomènes d'insertion et de délétion, couramment nommés les *indels*, qui se sont produits au cours de la divergence des séquences. Les *indels* sont des phénomènes qui peuvent fortement bruiter les alignements, et leur profusion est un signe que des évènements

évolutifs majeurs se sont produits entre deux séquences. Par conséquent, les algorithmes de calcul cherchent le plus souvent le meilleur alignement générant le moins possible ces *indels*. Leur quantité dans les alignements de séquences peut ainsi devenir un bon indicateur du degré d'homologie.

II.B.1.1 Alignement global contre alignement local

Dans le cadre de la recherche d'homologues avec une banque de données, un facteur critique (particulièrement à notre époque où ces banques prennent des proportions gigantesques) est le temps de calcul des différents alignements. Les algorithmes de programmation modernes permettent de réduire significativement ce temps de calcul, tout en prenant en compte les insertions et les délétions dans l'alignement. Ils reposent sur le principe de ne conserver que certains évènements de comparaison selon des critères prédéfinis. Afin de réduire encore les temps de calcul, des programmes générant des amorces de comparaison par des méthodes heuristiques ont été développés. On retiendra ainsi les programmes FASTA (A fast approximation to Smith-Waterman) (Pearson, 1990) et BLAST (Basic Local Alignement Search Tool) (Altschul *et al.*, 1990), qui génèrent chacun une liste de séquences présentant une certaine similitude avec la séquence d'entrée, et les trient selon un score reflétant la probabilité d'obtenir ces différentes séquences au hasard.

II.B.1.2 Les alignements multiples

Les programmes BLAST et FASTA permettent d'identifier dans une banque de données une liste de séquences similaires à une séquence d'entrée. Néanmoins, il est difficile d'aligner plus de deux séquences dans un temps raisonnable par ces méthodes de calcul. Or il est souvent très informatif d'aligner plus de deux séquences, pour déterminer par exemple les motifs caractéristiques d'une famille ou mettre en évidence des *indels* entre différentes séquences.

Des programmes d'alignement multiple ont été créés dans ce but. Certains d'entre eux sont assez utilisés, comme CLUSTAL W (Thompson *et al.*, 1994) et MULTALIN (Corpet, 1988), basés tous deux sur une succession d'alignements deux à deux de séquences qui sont affinés entre eux, ou encore MAFFT (Katoh *et al.*, 2002), qui s'appuie sur une analyse de transformée de Fourier rapide (FFT) de motifs similaires dans les séquences.

II.B.1.3 A la recherche de profils communs

Les alignements multiples de séquences issues d'une même famille permettent de mettre en évidence certaines positions conservées d'acides aminés. Ces alignements peuvent servir à créer des profils de séquences caractéristiques de ces familles. Deux types de méthodes de génération de profils ont été développés : les profils PSSM (Position-Specific Scoring Matrices) qui résument l'information de l'alignement dans une matrice et les profils HMM (Hidden Markov Models), qui sont des structures plus complexes composées d'états et de transitions, et qui basent leur recherche de similitude sur des distributions de probabilités. Ces profils se sont révélés être des outils extrêmement puissants pour déterminer des familles ou des domaines de protéines pouvant passer sous le seuil de détection des alignements classiques. Des banques de données spécialisées ont d'ailleurs été créées afin pourvoir prédire l'appartenance d'une protéine à une famille, sur la base de ces profils.

II.B.2 - Centralisation de l'information et banques de données

Des banques de données ont très vite été créées avec l'émergence des techniques de séquençage des protéines. Le volume de séquences n'ayant cessé de croître, ces banques se sont enrichies au fil des ans et certaines se sont spécialisées tandis que d'autres sont restées plus générales, pour permettre la centralisation de l'information. Ces banques de données peuvent être consultées en utilisant des systèmes d'interrogations croisées tel que SRS (Sequence Retrieval System) (Etzold *et al.*, 1996). Ces systèmes permettent de centraliser l'accès à l'information à partir d'une séquence et de croiser des données provenant de différentes banques. La section qui suit ne prétend pas répertorier l'intégralité de ces banques de données mais présenter à la fois une vue d'ensemble des plus utilisées dans le cadre de l'annotation d'un génome, et des plus pertinentes pour la suite de ce manuscrit. En effet, il existe à l'heure actuelle plus de 500 banques de données de divers types (information extraite de DBNet (Discala *et al.*, 2000), chacune essayant d'apporter un plus dans la recherche d'information.

II.B.2.1 Les banques de données généralistes

II.B.2.1.1 Banques de données des séquences nucléiques

Il existe trois grandes banques de données de séquences nucléiques, publiques et généralistes : la *GenBank* aux Etats-Unis (Benson *et al.*, 2008), l'*EMBL* en Europe (European Molecular Biology Laboratory. Sterk *et al.*, 2007) et la *DDBJ* au Japon (DNA Data Bank of Japan. Sugawara *et al.*, 2008). Grâce une forte coopération internationale depuis 1982, ces banques s'échangent leurs informations tous les jours. La *GenBank* est cependant le centre névralgique de cette interconnexion. Créée et maintenue par le National Center for Biotechnology Information (NCBI, USA), elle est issue de la fusion de 260 000 organismes de recherche (Benson *et al.*, 2008). En octobre 2008, cette banque de données contient, entre autres, les séquences de 706 génomes bactériens, 52 génomes archéens et 22 génomes eucaryotes. De plus, 1200 génomes de bactéries, 38 génomes d'archées et 336 génomes d'eucaryotes sont en cours de séquençage (source : NCBI).

II.B.2.1.2 Banques de données des séquences protéiques

L'ensemble des séquences protéiques, issues soit de soumissions directes soit de la traduction automatique des séquences d'acides nucléiques disponibles dans les banques de données précédemment citées, est accessible dans la banque publique GenPept créée et maintenue par le NCBI. Cette banque contient la collection de séquences protéiques la plus importante au monde. Elle présente cependant l'énorme inconvénient d'être mal annotée, car les annotations que l'on y trouve ne sont réalisées qu'avec un processus automatisé. En revanche, la banque UniProt KnowledgeBase (UniProtKB) (The UniProt Consortium, 2008), née de la fusion en 2002 des trois principales banques de données accessibles pour l'étude des protéines (SwissProt, TrEMBL et PIR), n'est composée, dans sa partie SwissProt, que de séquences protéiques à haute valeur ajoutée. Elle contient en effet des séquences présentant un niveau minimal de redondance et dont les annotations sont réputées très fiables, car révisées à la main par des experts à partir de la bibliographie. SwissProt propose en outre de nombreux liens avec plusieurs autres banques de données. Elle est maintenue par le Swiss Institute of Bioinformatics (SIB) en collaboration avec l'European Institute of Bioinformatics (EBI). La banque de données TrEMBL (Translation of EMBL) contient les traductions automatiques de l'ensemble des séquences nucléiques codantes de l'EMBL, non intégrées dans la banque SwissProt. Cette banque est également maintenue par le SIB et

l'EBI. Enfin, la banque PIR (Protein Information Ressource) rassemble l'ensemble des données protéiques contenues dans les principales banques mondiales. Elle est maintenue par le Georgetown University Medical Center, aux Etats-Unis.

II.B.2.2 Les banques de données spécialisées

La banque **ENZYME** regroupe des activités enzymatiques, classées selon une nomenclature systématique. Elle est maintenue par l'IUBMB (International Union of Biochemistry and Molecular Biology) (Tipton and Boyce, 2000) qui l'a créée en 1961 pour normaliser les descriptions des activités enzymatiques. Cette nomenclature classe les enzymes selon les réactions qu'elles catalysent. A l'heure actuelle, la nomenclature des numéros EC (Enzyme Commission) comporte 6 classes de réactions : les oxydoréductases, les transférases, les hydrolases, les lyases, les isomérases, et les ligases. Un code à 4 nombres permet de décrire chaque activité enzymatique de manière unique. Le premier nombre est compris entre 1 et 6, et correspond à la classe enzymatique de l'enzyme. Le deuxième nombre correspond à sa sous-classe et renseigne sur le type de liaison impliqué dans la réaction. Le troisième nombre indique le type d'atome de la liaison directement impliqué dans la réaction. Et enfin, le quatrième nombre correspond au substrat catalysé. Par exemple, le numéro EC 3.2.1.83 nous indique que l'enzyme appartient à la classe des hydrolases (3.-.-.-) et la sous-classe des glycosylases (3.2.-.-), qu'elle hydrolyse des liaisons *O*- ou *S*-glycosidiques (3.2.1.-), et enfin, qu'elle est active sur le κ-carraghénane (3.2.1.83). Il s'agit donc d'une glycoside hydrolase active sur le κ-carraghénane, soit une κ-carraghénase.

La banque **Carbohydrate-Active enZymes (CAZy)** (Henrissat, 1998) décrit les familles de modules (catalytiques ou non) d'enzymes qui dégradent, modifient ou créent des liaisons glycosidiques, en se basant sur les relations de structure existant entre ces modules. La banque classe les différentes familles en cinq grandes classes. Les *glycosyl transferases* (GT) transfèrent des sucres activés sur des groupements accepteurs ; les *glycoside hydrolases* (GH) hydrolysent les liaisons glycosidiques ou catalysent les réactions de transglycosylation ; les *polysaccharide lyases* (PL) clivent par β-élimination les liaisons glycosidiques impliquant des acides uroniques, libérant des sucres insaturés ; les *carbohydrate esterases* (CE) catalysent l'hydrolyse des liaisons esters qui constituent des décorations sur des polysaccharides ; et enfin, les *carbohydrate binding modules* (CBM) ne sont pas catalytiques mais peuvent se fixer avec une grande affinité sur certains motifs polysaccharidiques.

La **Protein Data Bank** (PDB) (Berman *et al.*, 2003) est une banque de structures tridimensionnelles. Elle rassemble toutes les structures publiées, quelle que soit leur nature (protéines, ADN, complexes obtenus par radiocristallographie ou RMN) et leur attribue un identificateur unique à quatre caractères composé de chiffres et de lettres (ex : *2BSP*). L'essor des projets de génomique structurale a fortement contribué à son succès. En effet, si elle contenait un peu plus de 15 000 entrées en 2001, elle est composée de plus de 50 000 structures en octobre 2008. De nombreuses banques de données sont dérivées de la PDB comme SCOP (Andreeva *et al.*, 2004) ou encore CATH (Greene *et al.*, 2007). Elles exploitent les informations structurales des protéines pour permettre leur classification en domaines structuraux.

PROSITE (Hulo *et al.*, 2008) est la première banque de données de familles de protéines et de domaines, créée en 1988 au Swiss Institute of Bioinformatics (SIB). PROSITE consiste en une collection de motifs assez courts (10 à 20 acides aminés) décrits par des profils et liés à une documentation sur la famille de protéines ou le domaine qu'ils permettent de détecter.

Pfam (Finn *et al.*, 2008) est une banque de données de domaines protéiques issus d'alignements multiples de séquences et de la génération de profils HMM. Une famille Pfam contient des annotations sur la fonction, des références bibliographiques, des liens vers d'autres banques de données et deux types d'alignements multiples : un alignement contenant seulement les membres représentatifs de la famille, utilisé pour calculer le profil HMM, et un alignement complet avec tous les membres de la famille trouvés avec le profil dans les banques Swiss-Prot et TrEMBL.

Le consortium **InterPro** (Mulder and Apweiler, 2008) a été créé en 1998 sous l'impulsion des banques de données pour permettre un recoupement des informations pertinentes issues des analyses de chaque banque. Ce serveur regroupe ainsi les données des quatre principales banques de données : PROSITE, Pfam, PRINTS (Attwood *et al.*, 2003) et ProDom (Servant *et al.*, 2002), ainsi que les informations issues de 12 autres banques.

Dans une volonté de recouper encore les informations, d'autres banques de données ont vu le jour avec, cette fois, le but de ne pas se contenter de décrire une famille mais d'établir les relations existant entre les protéines au sein des voies métaboliques. Des banques de données enzymatiques telle que BRENDA (Barthelmes *et al.*, 2007) permettent justement de décrire les activités enzymatiques et peuvent renvoyer sur les enzymes

agissant en amont ou en aval d'une voie métabolique. La banque KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa *et al.*, 2008) se propose quant à elle de décrire les grandes voies métaboliques et de positionner (par un système de graphiques très intuitifs) une enzyme ou un groupe d'enzymes donné dans la carte métabolique d'une cellule.

II.B.3 - Perspectives d'utilisation par l'annotateur

Il apparaît, au vu du grand nombre de données accessibles et de la multitude de banques de données s'interconnectant, qu'il est aisé de se perdre sur le chemin de l'information. Je n'ai de plus présenté qu'un nombre extrêmement restreint de banques de données et chaque annotateur peut avoir ses préférences et ses habitudes pour ses analyses d'homologies.

L'explosion du nombre de séquences ces dernières années et leur arrivée dans ces différentes structures est ainsi finalement à double tranchant. Cela permet de fiabiliser les comparaisons de séquences, en particulier en permettant de créer des familles dont la définition est de plus en plus stable. La caractérisation biochimique d'une enzyme dans une famille devient alors un début de piste très intéressant et valorisable pour comprendre les fonctions des autres membres de la famille. Cependant, l'arrivée massive de séquences a également tendance à noyer l'information pertinente au milieu d'un grand nombre de séquences non annotées. Ainsi, les banques de données et les outils pour chercher l'information pertinente doivent se perfectionner sans cesse pour ne pas se trouver saturés par les requêtes qui se font, elles aussi, de plus en plus nombreuses.

II.C - Méthodologies de l'annotation des génomes

L'annotation d'un génome n'est pas, loin s'en faut, une étape triviale. Il est nécessaire de suivre, pour chaque génome séquencé, une succession d'étapes qui permettront à terme son exploitation par la communauté scientifique : tout d'abord il faut réaliser une **annotation structurale** des gènes, suivie de leur **annotation fonctionnelle**. Pour procéder à ces deux niveaux d'annotation, les annotateurs ont à leur disposition les outils précédemment cités. Plusieurs techniques croisent les différentes informations disponibles dans les banques de données pour fournir une annotation la plus précise possible. Ces différentes techniques, leurs avantages, ainsi que leurs limitations, sont proposées dans la section qui suit.

II.C.1 - Que signifie « annoter un génome » ?

L'expression « annotation de génome » renvoie à l'ensemble des étapes suivant le séquençage d'un génome et permettant, à partir de sa séquence, de proposer une fonction pour chacun de ses gènes à un certain degré de confiance. L'ensemble du processus est basé sur l'inférence fonctionnelle à partir des homologues présents dans les banques de données. Les étapes critiques de l'annotation seront donc 1/ la prédiction des régions codantes et 2/ la qualité de l'inférence fonctionnelle, qui dépend elle-même du contenu des banques et des outils de recherche.

II.C.1.1 Annotation structurale

L'annotation structurale consiste à définir la position précise des gènes, ainsi que leur cadre de lecture. Ceci est réalisé en recherchant les cadres de lecture ouverts (ORF, Open Reading Frames) dans la séquence génomique. Le terme ORF fait référence à une portion de génome permettant de générer un ARN messager mature, qui, la plupart du temps, sera traduit en protéine. Un ORF commence donc par un codon *initiateur* et se termine par un codon *stop*. Dans les génomes eucaryotes, les ORFs sont, de plus, interrompus par des séquences non codantes appelées *introns* (Belshaw and Bensasson, 2006). Une fois la position d'un ORF déterminée, il est nécessaire de prédire où seront situées ses régions non codantes. Les difficultés de l'annotation structurale des génomes sont donc d'une part de prédire la position des ORFs, en veillant à ne pas introduire d'ORFs artéfactuels, et d'autre part de définir la position des régions régulatrices de ces ORFs (promoteurs, introns pour les eucaryotes, …).

Sur les dix dernières années, les méthodes d'identification des ORFs des séquences génomiques se sont basées sur trois techniques fondamentales (Brent, 2005) : le *séquençage de clones d'ADNc* sélectionnés aléatoirement, suivi de l'alignement de ces séquençages sur la séquence génomique ; la *recherche de structures d'ORFs* déjà répertoriées dans les banques de données et connues pour générer des protéines ; et enfin, la *recherche de novo d'ORFs*, c'est-à-dire sans référence à des séquences d'ADNc connues.

La difficulté de cette première étape dépend essentiellement de l'organisme dont est issu le génome. Ces méthodes de prédiction d'ORFs sont particulièrement efficaces sur les génomes procaryotes (dépourvues d'introns), mais le sont beaucoup moins sur les génomes eucaryotes dont les structures fines des régions régulatrices de l'ADN ne sont pas toujours connues (Brent, 2008).

II.C.1.2 Annotation fonctionnelle

Trois approches conceptuelles différentes sont utilisées pour procéder à l'annotation fonctionnelle d'un génome : *manuellement*, l'annotation étant réalisée « à la main » par une équipe d'annotateurs ; *automatiquement*, l'annotation étant réalisée cette fois de manière autonome par des logiciels ; et enfin *semi-automatiquement*, approche combinant les deux précédentes.

L'**annotation manuelle** consiste, comme son nom le suggère, en une annotation où les décisions d'inférence fonctionnelle sont réalisées par des équipes de biologistes annotateurs. Les équipes procèdent en effectuant des recherches d'homologies pour chaque gène dans les différentes banques de données disponibles. Les premiers génomes séquencés ont en particulier été annotés de la sorte, souvent en impliquant plusieurs équipes d'annotation. Il n'était en effet pas rare que plusieurs laboratoires allient leurs efforts pour produire une annotation de qualité. Ce processus était assez fastidieux, et ce d'autant plus que les outils d'annotation étaient pour l'essentiel en création. De nos jours, le recours à une annotation entièrement manuelle est de moins en moins fréquent en raison de la lenteur de ce processus, surtout vis-à-vis du taux d'accroissement du nombre de nouvelles séquences dans les banques de données.

L**'annotation automatique** consiste à l'automatisation entière du processus d'annotation. Des logiciels réalisent les recherches d'homologies sur les banques de données et procèdent aux inférences fonctionnelles. Ils croisent pour cela une série de mots clefs, sortis par les banques de données autour des séquences recherchées, avec des seuils de similitude. Cette méthode est dénommée « reciprocal best BLAST hit » (RBH) : la fonction du meilleur score (ou « hit ») d'alignement par BLAST est transférée si les deux séquences sont chacune le meilleur score de l'autre. La grande limite de ce type d'annotation réside dans la faible souplesse des croisements de données. En effet, les seuils de similitudes et les informations nécessaires à une annotation de qualité sont définis au niveau du génome, mais les écarts à la norme sont spécifiques à quasiment chaque famille de gènes. Il en résulte que le RBH peut générer des erreurs dans les cas non triviaux (qui peuvent constituer une part importante des annotations). Plusieurs logiciels d'annotation automatisée ont vu le jour ces dernières années (EnsEMBL, …) et des améliorations sont proposées régulièrement (voir par exemple (Fulton *et al.*, 2006) et (Moreno-Hagelsieb and Latimer, 2008).

L'**annotation semi-automatique** est un compromis entre les deux méthodes présentées. Un logiciel propose, pour chaque gène, une collection d'informations recueillies

dans les banques de données et l'annotateur décide si l'inférence fonctionnelle est valable ou non. Différents systèmes basés sur ce type d'annotation ont été créés ces dernières années. Certains, comme dans le projet HAMAP (Gattiker *et al.*, 2003), exploitent les résultats issus des analyses des séquences. D'autres présentent à l'annotateur un premier résultat d'annotation automatique complété d'une synthèse plus ou moins exhaustive des analyses de différentes banques de données (recherche de domaines, de peptides signaux, d'hélices transmembranaires, de signatures, …), le tout au sein d'une interface graphique. Un exemple de ce type d'approche est le projet GeneDB, développé par le Wellcome trust Sanger Institute de Cambridge (Hertz-Fowler *et al.*, 2004). Cette méthode d'annotation est très puissante du fait que l'annotateur a en quelques clics accès à une grande quantité d'informations, qui seraient fastidieuses à rassembler manuellement. Cela permet en outre de limiter les risques d'erreurs d'annotation puisque l'annotateur a accès à toutes les données essentielles du gène. Les inférences fonctionnelles sont donc révisées avec un esprit critique.

II.C.2 - Les améliorations du processus d'annotation

Le développement des banques de données a permis des avancées majeures dans l'exploitation de l'information contenue dans les séquences. Ces dernières années, différentes méthodologies utilisant les techniques de la génomique comparée ont été mises à profit dans les annotations, en particulier pour les séquences géniques présentant une faible homologie dans les banques de données. Il est ainsi possible de citer la création des groupes de domaines orthologues conservés (COG) qui associent des analyses phylogéniques aux comparaisons de séquences et permettent de définir de manière univoque le contenu des produits des gènes (Tatusov *et al.*, 1997). Je citerai encore les comparaisons inter-génomiques qui peuvent donner des informations contextuelles cruciales sur une séquence, étant donné que la conservation de la synténie d'un gène entre deux génomes peut être un gage de la conservation de sa fonction (Dandekar *et al.*, 1998; Wolf *et al.*, 2001).

Il est également vite apparu au travers des annotations successives des génomes qu'une annotation basée uniquement sur la fonction moléculaire des gènes n'était pas suffisante. Nombre d'exemples prouvent qu'une bonne similitude entre deux protéines ne garantit pas une conservation de fonction biochimique (cas de la lactate déshydrogénase, utilisée chez les vertébrés comme protéine de structure du cristallin – (Piatigorsky, 2003), pas plus qu'une conservation de la fonction biologique (cas des gènes homéotiques *Hox* qui

participent à la mise en place des nerfs crâniens chez les vertébrés, et à la mise en place des appendices chez les invertébrés (Hueber and Lohmann, 2008). Dans ce sens, la nomenclature *Gene Ontology* (GO) a été créée par le *Gene Ontology consortium* (Ashburner *et al.*, 2000). Dans ce système, la notion de fonction biologique d'un gène peut être décrite sur plusieurs niveaux. Le premier est sa fonction dans la biologie de la cellule (« Cell Biological Function ») : ce niveau renseigne sur le processus biologique dans lequel s'exprime le gène (respiration, division, …). Le second niveau est sa fonction moléculaire (« Molecular Function »): ce niveau renseigne sur l'activité biochimique concrètement réalisée par le produit du gène (glycoside hydrolase, kinase, protéase, récepteur, …). Le troisième et dernier niveau est la localisation cellulaire de la fonction du gène (« Cellular Component ») ; ce niveau renseigne sur le compartiment subcellulaire dans lequel agit le produit du gène (cytosol, Golgi, périplasme, …). L'EBI a dans ce sens développé le projet « Gene Ontology Annotation » (GOA) (Camon *et al.*, 2003) afin d'intégrer dans ses banques de données les termes de vocabulaire définis par le consortium GO.

II.C.3 - Un système imparfait…

Des erreurs se produisent, malgré tous les garde-fous implémentés dans les procédures d'annotation et les croisements de plus en plus faciles entre les banques de données. Leurs origines peuvent être extrêmement variées. Certaines proviennent de la surinterprétation de la fonction d'un gène, d'autres d'une inférence fonctionnelle fausse. D'une manière générale, et assez intuitivement, plus un gène est divergent par rapport à ses homologues (si il en a), ou plus il y a de protéines non caractérisées dans les familles de protéines orthologues, plus la probabilité de générer des erreurs d'annotation est grande.

II.C.3.1 Les biais des banques de données

Afin de comprendre en quoi le système actuel de stockage de séquences et de croisement de données est constitutivement biaisé, il importe de rappeler quelques chiffres. La banque de données nucléiques du NCBI GenBank contient à l'heure actuelle 76 millions de séquences (dont 15 millions ajoutées depuis 2007) (Figure I-3). Ce nombre comprend les séquences provenant d'environ 800 génomes (tous phyla confondus) et 1500 génomes en cours de séquençage seront publiés dans les années à venir. Et ce d'autant plus rapidement que GenBank intègre également les ébauches de génomes.

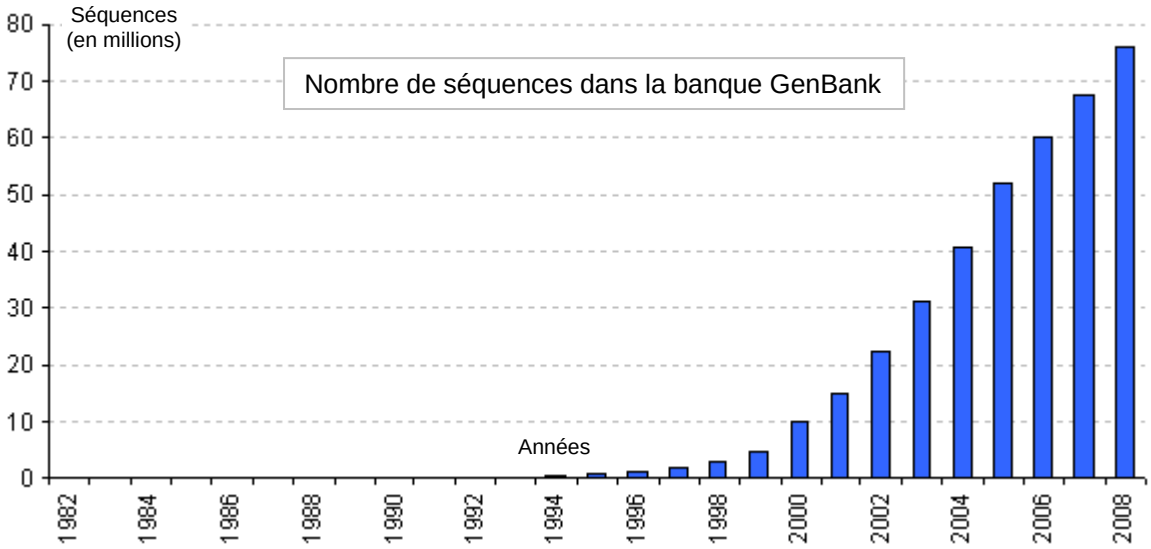


Figure I-3 : Nombre de séquences dans la banque de données GenBank.
Depuis sa création (données extraites du site www.ncbi.nlm.nih.gov)

Dans le même temps, la banque de données protéique SwissProt, qui contient les définitions des enzymes caractérisées expérimentalement et révisées manuellement contient 400 000 entrées.

Le décalage d'échelle (d'un facteur 200) entre le nombre de séquences disponibles dans les banques de données et le nombre de séquences expérimentalement caractérisées et référencées est assez frappant. Il traduit déjà le fait que le séquençage des acides nucléiques a rattrapé son retard par rapport à celui des protéines. De fait, les infrastructures existent maintenant pour séquencer des génomes rapidement et efficacement. Les centres de séquençage créés au cours des années 1990, et dont le but premier était de participer au projet de séquençage du génome humain, se sont en effet transformés, après avoir rempli leur objectif, en centres techniques de séquençage à haut débit. L'accélération des techniques a également permis de baisser le prix du séquençage par paire de bases (aujourd'hui autour de 0,5 $ / bp[3]) et les laboratoires peuvent maintenant obtenir le génome d'un organisme modèle à faible coût.

La croissance exponentielle du nombre de séquences déposées dans les banques de données est également due au développement de la métagénomique. En effet, ces dernières années, de grands programmes de séquençage d'environnements ont été lancés.

[3] Le projet du *1000$ genome* aux USA qui propose de séquencer pour 1000$ le génome de particuliers est assez emblématique de la facilité de l'accès au séquençage de nos jours.

A ce titre, le programme *Sorcerer II*[4] Global Ocean Sampling, également connu sous le nom de programme de séquençage de la mer des Sargasses, est de loin le plus vaste programme de ce type jamais réalisé : en trois expéditions, il a permis de générer plus de 5 millions de séquences (pour un total de 6 milliards de paires de bases, soit douze fois le génome humain) réparties sur 2000 espèces bactériennes (Gross *et al.*, 2007; Nicholls, 2007). Le gigantisme de ce type de projet est effectivement une source d'information formidable pour la catégorisation du vivant, la découverte d'espèces, le remplissage de familles, …. Cependant, il apparaît également évident qu'il n'est pas possible de procéder à une annotation révisée à la main de ce volume de données sans pareil, et *a fortiori* encore moins envisageable de caractériser ces séquences expérimentalement. Il est possible de tirer plusieurs observations de cet état de fait.

Tout d'abord, il apparaît que ce qui était au départ une avancée scientifique majeure s'est transformé lors de sa publication dans les banques de données en bruit de fond pour l'essentiel des comparaisons de séquences. En effet, toute prédiction de fonction aussi fiable soit elle reste une prédiction, et les exemples ne manquent pas où la protéine présente une fonction différente de son annotation. En l'occurrence, l'annotation automatique réalisée sur ces séquences est la procédure la moins (globalement) fiable du fait des biais présentés précédemment. Les génomes en cours de séquençage, et n'en étant encore qu'au stade d'ébauche, étant également mis à disposition dans les banques de données, le volume des entrées à faible valeur ajoutée a donc explosé ces dernières années. Cela justifie d'autant plus le recours à des banques spécialisées telle SwissProt dont les entrées sont beaucoup plus fiables.

Cependant, lorsque l'on s'intéresse plus en détail aux génomes séquencés d'une part, et aux séquences caractérisées d'autre part, il apparaît que d'autres biais peuvent survenir. En effet, l'analyse des statistiques de la banque SwissProt révèle que 15% seulement des séquences qu'elle regroupe ont été caractérisées expérimentalement, quand 65% sont issues d'une inférence par homologie. Dans le cas précis de la banque SwissProt, ces inférences révisées manuellement sont probablement souvent justifiées, mais encore une fois, une inférence ne remplacera jamais une caractérisation. De plus, sur les 400000 séquences de la banque, représentant 11500 espèces, 25% sont issues de 20 espèces, et parmi les plus connues : *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Arabisopsis*

[4] Pour l'anecdote, le nom de ce programme (*Sorcerer II*) provient du nom du yacht personnel de J. Craig Venter, reconverti en bateau scientifique pour l'occasion (Gross, 2007). C'est ce même J.C. Venter qui a créé Celera, la compagnie privée ayant très fortement contribué au programme Human Genome Project. Cette société a également synthétisé chimiquement un génome entier pour créer une espèce bactérienne artificielle.

thaliana, *Saccharomyces cerevisiae*, *Escherichia coli*, …. Lors du recours à cette banque de données qui fait partie de celles à haute valeur ajoutée, les comparaisons de similitude se réalisent donc essentiellement sur un échantillon réduit d'organismes (et les mêmes pour tous les génomes) provenant d'environnements pour la plupart d'entre eux voisins (homme, souris, rat, taureau, levure du boulanger, plantes, bactéries fécales et alimentaires, …), qui plus est sur des séquences elles-mêmes majoritairement issues d'une inférence fonctionnelle. Il s'agit donc typiquement d'une situation où une erreur d'annotation peut se propager sur des séquences, et peut au final s'imposer dans les annotations. On peut également se poser la question de la pertinence d'annoter des génomes couvrant de larges pans du vivant à partir d'un échantillon très restreint d'organismes. Le cas des sulfatases de *Rhodopirellula baltica* peut servir d'exemple pour étayer la question. En 2003, lors de la publication du génome de *R. baltica* (Glöckner *et al.*, 2003), les sulfatases à formylglycine ne forment qu'une petite famille d'enzymes homologues de quelques représentants. Ces enzymes sont finalement assez peu étudiées et il est rare d'en trouver plus d'une dizaine dans les organismes qui en possèdent. Leur fonction étant peu triviale à trouver, elles sont toutes annotées « arylsulfatases », en raison du type de substrat de référence utilisé pour caractériser leur activité (un arylsulfate). *R. baltica* possède plus de 100 de ces enzymes. A elle toute seule, elle décuple le nombre d'enzymes connues dans sa famille. D'autres bactéries marines se sont révélées depuis posséder également un grand nombre de sulfatases (Gurvan Michel, communication personnelle). Il est important de soulever ici que toutes ces enzymes de bactéries marines ont donc été annotées à partir d'un panel très restreint d'enzymes caractérisées provenant d'organismes terrestres, dont l'utilisation est probablement radicalement différente.

Il apparaît donc que le système actuel n'est pas parfait. Par exemple, des mesures du taux d'erreurs générées par la banque de données du projet GO (dans laquelle les annotations sont pourtant révisées à la main par des experts) ont montré qu'il allait de 15% à environ 50%, selon le niveau de révision (respectivement sans et avec recherche d'homologue par similitude) (Jones *et al.*, 2007).

Par ailleurs, maintenir une banque de données demande de grands moyens et une mise à jour quotidienne. Cela demande aussi de la crédibilité, qui est acquise par la publication de mises à jour de version par exemple. Cependant, l'intégration de nouvelles informations dans les banques de données n'est pas soumise à la règle de révision par les pairs qui prévaut en science. Des tentatives de réorganiser l'ensemble des banques sont actuellement à l'essai (Wren and Bateman, 2008).

Pour toutes ces raisons, l'utilisation des banques de données doit donc être réalisée avec une grande circonspection, et une réelle volonté de confrontation des informations.

II.C.3.2 Les erreurs d'annotation

Malheureusement, d'autres erreurs peuvent survenir au cours des processus d'annotation. Les annotateurs, qu'ils soient humains ou robotisés, ne sont pas infaillibles et ils peuvent générer des erreurs. Un certain nombre de ces erreurs ont été référencées par (Galperin and Koonin, 1998)

Parmi les premières erreurs, on trouve celles de prédiction des régions codantes. En effet, il n'existe toujours pas à l'heure actuelle d'algorithme de prédiction fiable des ORFs eucaryotes (Brent, 2008). La prédiction des ORFs procaryotes semble moins problématique, même si des erreurs peuvent survenir. Ceci a pour conséquence de gonfler artificiellement le volume des gènes dans les banques, rajoutant encore du bruit dans les recherches d'homologie et ralentissant les calculs.

On observe également des erreurs dans la description même des annotations, souvent en raison d'un mauvais choix de la méthode automatique d'annotation ou suite à des erreurs d'interprétation faites par l'annotateur. Ce type d'erreur n'est ni rare, ni mineur. Un exemple frappant est la comparaison des annotations du génome de *Mycoplasma genitalium* par trois équipes d'annotateurs indépendantes qui ont mis en évidence que 8% des prédictions présentaient des désaccords importants (Brenner, 1999). Ce type d'erreur peut aussi résulter d'une forte divergence de la séquence annotée par rapport au contenu des banques. Ainsi, des alignements partiels sur des régions non significatives peuvent aboutir à une inférence abusive (Devos and Valencia, 2001). Enfin, la structure multimodulaire d'une enzyme peut être la source de ces erreurs d'annotation.

Dans le cas présenté **Figure I-4**, la fonction A est sûre, elle a été caractérisée. L'inférence fonctionnelle est donc naturelle avec son proche homologue de la séquence 2. La séquence 2 est, elle, bimodulaire et son second module (module B) est un module non catalytique (fixation à un substrat par exemple). Cependant, la séquence 3 voit l'un de ses modules présenter une forte similitude avec le module B de la séquence 2. Une inférence fonctionnelle indue a ici lieu et la séquence 3 se retrouve posséder la fonction A. Ce type d'erreur peut être très dommageable car il peut attribuer des fonctions totalement fausses.

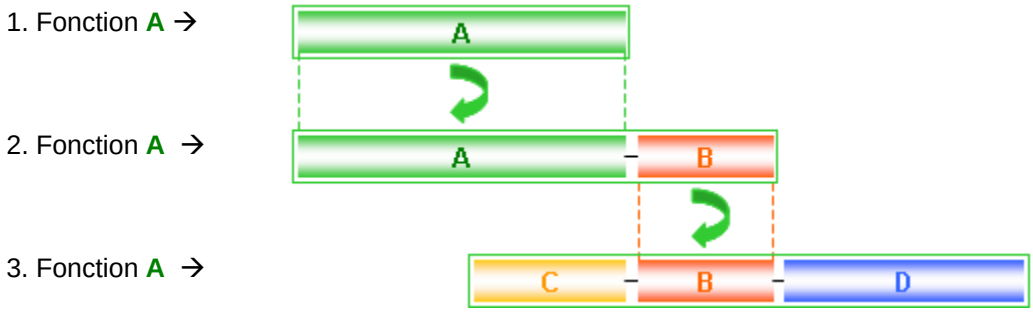


Figure I-4 : Modules et inférence fonctionnelle
Présentation d'une inférence fonctionnelle abusive issue d'une mauvaise interprétation d'un alignement.

Enfin un dernier cas ne relève pas totalement de l'erreur mais plutôt de l'imprécision. Il arrive assez fréquemment qu'une série de fonctions sortant d'un alignement apparaisse comme variée, alors qu'il s'agit d'une unique fonction présentant plusieurs descripteurs. Ceci peut faire perdre un temps précieux à l'annotateur et traduit un manque de nomenclature à l'échelle internationale.

II.C.4 - … en quête de rédemption ?

Devant la réelle amélioration des conditions d'annotation moderne, des projets de réannotation de génomes séquencés au cours des années 1990 ont commencé à émerger (Boneca *et al.*, 2003; Iliopoulos *et al.*, 2003). En effet, non seulement les outils de l'époque n'en étaient qu'à leurs balbutiements (quand ils existaient) mais en plus, les banques étaient infiniment moins fournies que de nos jours (voir Figure I-3). La réannotation des génomes permet de tester la performance de nouvelles méthodes mais aussi d'utiliser les dernières données pour trouver de nouvelles fonctions manquées dans les analyses précédentes.

III - Ecologie microbienne marine

Le séquençage du génome de la planctomycète marine *Rhodopirellula baltica* (Glöckner *et al.*, 2003) s'inscrit dans un projet de métagénomique porté par le laboratoire de Biologie Marine du Max Planck Institute (LBM-MPI) de Brême (Allemagne). Un des enjeux de ce programme est de comprendre la structure et les interactions des populations

microbiennes aquatiques dans l'environnement marin que constitue la mer Baltique. A ce titre, ce programme s'intéresse en particulier à la compréhension des rôles spécifiques des différents organismes qui agissent en tant que médiateurs des flux d'éléments, par exemple durant la reminéralisation du carbone organique (Llobet-Brossa *et al.*, 1998; Pernthaler and Amann, 2005; Pedros-Alio, 2006; Riemann *et al.*, 2008).

III.A - Les planctomycètes : des acteurs majeurs des écosystèmes marins

Au cours des quinze dernières années, les progrès de la métagénomique, c'est-à-dire ceux concernant les méthodes d'étude de séquences prélevées dans les échantillons environnementaux, ont permis de déterminer la composition et la diversité des assemblages microbiens dans nombre d'environnements marins et d'eaux douces.

C'est au cours de ces études qu'il a pu être mis en évidence que les planctomycètes étaient des bactéries abondantes dans de nombreux habitats marins et terrestres (Fuerst *et al.*, 1997; Neef *et al.*, 1998). Ce phylum présente des caractéristiques uniques qui ne sont rencontrées dans aucun autre phylum bactérien à l'heure actuelle.

III.A.1 - Un phylum bactérien très divergent

Le phylum *Planctomycetes,* dans le règne des *Bacteria,* comprend, à l'heure actuelle, sept genres (*Planctomyces, Blastopirellula*, *Jettenia*, *Gemmata*, *Isosphaera*, *Pirellula*, et *Rhodopirellula*). Ce phylum, bien que taxonomiquement équivalent aux vastes phyla que sont les *Proteobacteria* ou les *Firmicutes*, est cependant plus petit. Il n'est en effet composé que d'une seule classe (les *Planctomycetacia*), d'un seul ordre (*Planctomycetales*), et d'une seule famille, les *Planctomycetaceae*.

Le phylum *Planctomycetes* est indépendant et monophylétique au sein des *Bacteria*. Sa position phylogénétique ne semble cependant pas totalement stabilisée même si les dernières analyses phylogéniques semblent indiquer une parenté avec le phylum des *Chlamydiae*. Plusieurs hypothèses ont été formulées pour expliquer cet état de fait, notamment que ces bactéries pourraient être à évolution rapide. Il est également évoqué qu'elles puissent se situer à un embranchement profond du règne des *Bacteria* (Teeling *et al.*, 2004).

De plus, bien que considérées comme étant des bactéries Gram négative, leur paroi cellulaire ne contient pas de muréine, le peptidoglycane présent dans la paroi de l'ensemble des bactéries (excepté les *Mollicutes*) (Vollmer *et al.*, 2008). A la place, elles possèdent une structure protéique, glycosylée et hautement réticulée, très riche en cystéines et prolines (Liesack *et al.*, 1986).

La forme cellulaire des *Planctomycetes* indique que ces bactéries présentent une polarisation de leurs cellules. Le genre *Rhodopirellula* (signifiant *petite poire rouge*) se caractérise par des cellules en forme de poire, avec un pole (le petit bout de la poire) d'où émerge une structure mucilagineuse qui semble intervenir directement dans les phénomènes d'adhésion aux surfaces, et un autre (le grand bout de la poire) d'où émergent, par bourgeonnement, les cellules-filles flagellées, comme montré dans la Figure I-5 (Lindsay *et al.*, 2001).

Les *Planctomycetes* présentent enfin, ce qui est assez surprenant pour des bactéries, une compartimentation cellulaire dans laquelle une membrane sépare le paryphoplasme périphérique (sorte de très grand périplasme) qui ne contient aucun ribosome, du riboplasme intérieur (également appelé pirellulosome). A l'intérieur du riboplasme, se trouve un nucléoïde fibrillaire condensé. Curieusement, le genre *Gemmata* présente une membrane supplémentaire séparant le nucléoïde du reste du riboplasme (Lindsay *et al.*, 2001). La présence de ces membranes dont les phospholipides ont une composition en acides gras inhabituelle, est plutôt surprenante chez des cellules procaryotes et rappelle l'organisation cellulaire des eucaryotes (Figure I-5).

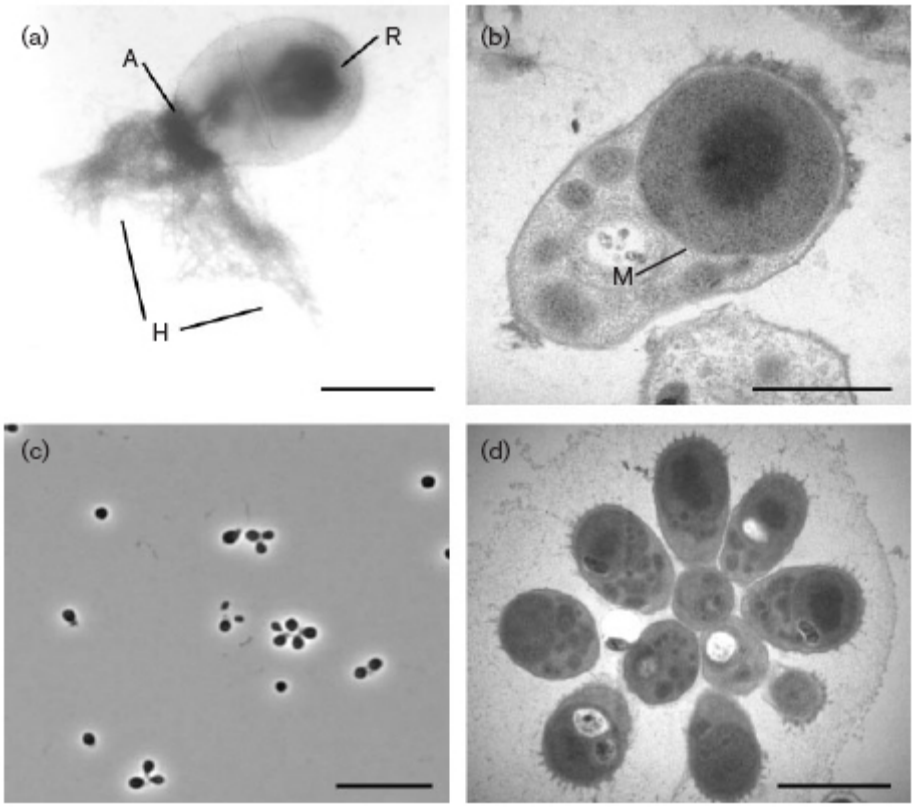


Figure I-5 : Images microscopiques de *R. baltica.*
Photos (a-c) : souches SH 1^T et *(d)* SH 796. a) Image de microscopie électronique d'une cellule présentant l'organisation polarisée des cellules de *R. baltica*. R : Pôle de reproduction; A : pôle d'attachement ; H : substance d'ancrage. Barre : 0,5 µm. b) Image de microscopie électronique présentant la compartimentation cellulaire. M : Membrane entourant la structure pirellulosomale ; Barre : 0,5 µm. c) Image de microscopie photonique d'agrégats typiques en rosette. Barre : 10 µm. d) Image de microscopie électronique d'une rosette montrant l'ancrage des cellules via le pôle d'attachement. Barre : 1 µm. Figure extraite de (Lindsay *et al.*, 2001).

Il apparaît donc que le phylum *Planctomycetes* constitue un groupe de bactéries extrêmement originales. Il est ainsi possible de citer des caractéristiques présentant de grandes ressemblances avec les cellules eucaryotes (reproduction par bourgeonnement, compartimentation cellulaire) ainsi qu'avec les cellules archéennes (enveloppe cellulaire constituée de glycoprotéines sans muréine).

III.A.2 - Perspectives environnementales

Beaucoup de programmes en écologie microbienne ont cherché à comprendre la structure et le rôle des macroagrégats détritiques marins dans les phénomènes de recyclage du carbone. Ces agrégats, qui sédimentent le long des colonnes d'eau dans les océans, sont également connus sous le nom de *neige marine*, en raison de l'aspect floconneux et de la

couleur blanchâtre de ces macroparticules (> 0.5 mm) (Fowler and Knauer, 1986). La neige marine est formée principalement à partir de l'agrégation de plus petites particules dans la colonne d'eau incluant phytoplancton, bactéries, fragments fécaux, réseaux trophiques de zooplancton, ainsi que d'autres débris organiques. De grandes quantités de carbone sont recyclées sous cette forme et comprendre comment les populations marines les transforment en particules dissoutes peut constituer une des clés de la compréhension du recyclage du carbone à l'échelle mondiale (Alldredge, 2000).

Les questions environnementales posées par l'existence des neiges marines et des organismes marins qu'elles hébergent sont donc de plusieurs ordres :

- Quels sont les phyla microbiens important dans leur minéralisation ?
- Quelles sont les bases moléculaires du rôle écologique de ces organismes ?

C'est pour tenter de répondre à ces questions que le Laboratoire de Biologie Marine du MPI a lancé un programme de métagénomique pour identifier les acteurs majeurs de ces communautés, ainsi que séquencer et annoter les génomes des acteurs identifiés. Les échantillonnages ont été réalisés dans la baie de Kiel en Allemagne, sur les rives occidentales de la mer Baltique. Parmi les organismes identifiés lors de ce programme de criblage, trois organismes ont été sélectionnés pour leur importance environnementale : *Rhodopirellula baltica*, et les *Deltaproteobacteria Desulfotalea psychrophila* et *Desulfobacterium autotrophicum* (ces deux organismes étant des bactéries sulforéductrices).

R. baltica a été choisie pour son appartenance au phylum des *Planctomycetes*, dans lequel aucun génome n'était connu au début des années 2000, mais dont l'importance environnementale et la répartition globale rendait l'analyse particulièrement intéressante (Fuerst *et al.*, 1997; Neef *et al.*, 1998).

III.A.3 - La mer Baltique

La mer Baltique est une petite mer d'environ 400 000 km^2, située au nord de l'Europe. Elle borde neuf pays (l'Allemagne, le Danemark, l'Estonie, la Finlande, la Lettonie, la Lituanie, la Pologne, la Russie, et la Suède), qui lui apportent les eaux de leurs 80 fleuves (Figure I-6). Il s'agit de la mer la plus jeune de la planète. Sa naissance est associée à la fonte de la calote polaire scandinave, il y a 15 000 à 8 000 ans. Elle présente une profondeur moyenne de 55 m, avec une valeur maximale de 459 m, au large de l'île de Gotland et proche des côtes lettones (source : Wikipedia)

Elle est relativement protégée des influences océaniques, ne communiquant avec les eaux mondiales, via la Mer du Nord, que par deux détroits situés à l'extrême sud de la Suède. Pour cette raison, le renouvellement complet de ses eaux est très lent, autour de 30 ans. Cette position isolée lui impose de fortes variations thermiques saisonnières : en hiver, la banquise recouvre une grande partie de sa surface (en particulier le golfe de Botnie, entre la Suède et la Finlande), tandis qu'en été, la température de l'eau avoisine les 15°C.

Les conditions environnementales de cette mer sont également particulières. En effet, la très forte influence fluviale qu'elle subit, alliée à une faible interaction avec les eaux océaniques lui imprime une faible salinité, entre 5‰ et 10‰ d'Est en Ouest (contre 35‰ en moyenne dans les eaux océaniques mondiales). Cette mer est également soumise à une importante anthropisation. Elle constitue en effet la seule voie maritime pour accéder à l'océan depuis les nations baltiques. De grand ports de commerce (tel que Saint-Pétersbourg) sont ainsi situés sur ses rives, imposant un trafic maritime soutenu.

Figure I-6 : La mer Baltique, vue depuis Google Earth.
L'étoile rouge indique la zone d'échantillonnage où a été isolée *R. baltica*.

D'une manière générale, la très forte présence humaine alliée à ses faibles capacités à se renouveler en ont fait une mer globalement très polluée. D'après le rapport "Clean Baltic within REACH?" de l'organisation WWF, datant de janvier 2005, les poissons de la Mer Baltique seraient tellement contaminés par plusieurs produits toxiques qu'ils pourraient être

interdits à la vente en Europe. Cette mer est également soumise à une forte eutrophisation (ses eaux présentant un fort excès de matières organiques), la transformant peu à peu en désert marin. Des programmes de réhabilitation ont été cependant lancés par les états baltiques depuis quelques années.

Les différentes caractéristiques de la mer Baltique en font une mer à part sur Terre et un véritable modèle de « mer douce ». En effet, le faible brassage de ses eaux (l'amplitude des marées n'excède pas 30 cm) induit un meilleur maintien des populations microbiennes, permettant de fait d'analyser plus facilement leurs interactions. Sa faible salinité, trop forte pour les microorganismes d'eaux douces, et trop faible pour les microorganismes océaniques, a également permis le développement de communautés assez originales. La forte influence de l'homme sur son environnement peut cependant apporter certains biais dans les études mais la détermination de la composition des populations microbiennes la peuplant fait également partie des clés de l'estimation de sa capacité de récupération (Khoroshko *et al.*, 2007).

III.B - Un génome à fort potentiel

III.B.1 - Rhodopirellula baltica, ou la naissance d'un modèle

L'annotation du génome de *R. baltica* a été réalisée manuellement par les équipes du LBM-MPI. L'analyse des ces gènes a révélé que plus de la moitié d'entres-eux étaient de fonction inconnue, étant donnée l'absence d'homologie qu'ils présentaient avec les séquences contenues dans les banques de données à l'époque. Au final, seulement 35% des gènes auront pu bénéficier d'une inférence fonctionnelle. L'analyse des gènes annotés a indiqué que *R. baltica* était une bactérie assez traditionnelle pour ses voies métaboliques centrales, mais présentait également des singularités assez exceptionnelles.

R. baltica présente donc les voies métaboliques classiques des bactéries hétérotrophes telles que les voies de la glycolyse, le cycle de Krebs, la phosphorylation oxydante, ainsi que le cycle des pentose-phosphates. Elle semble également capable de synthétiser l'ensemble des aminoacides. Elle présente en outre l'ensemble des gènes nécessaires à la formation d'un flagelle fonctionnel (Glöckner *et al.*, 2003).

Propriété du chromosome	
Taille totale, bases	7 145 576
Composition G+C, %	55.4
Séquences codantes	7,325
Densité codante, %	95
Longueur moyenne de gène, bases	939
Gènes présentant une similarité dans les banques de données*	3,380 (*46%*)
Gènes avec inférence fonctionnelle	2,582 (*35%*)
ARNr	1 x (16S) et (23S–5S)
ARNt	70
Autres ARNs stables	1 (ribozyme)

*Seuil de BLASTP $E_{value} \leq 1.10\text{-}3$, inclus les scores des protéines hypothétiques.

Tableau I-4 : Données de séquençage du génome de *R. baltica.*
Données extraites de (Glöckner *et al.*, 2003).

Pour autant, son génome présente également quelques familles de gènes intrigantes. *R. baltica* semble par exemple posséder une voie métabolique de dégradation de composés en C1 (monocarbonés), rencontré uniquement dans des génomes archéens. Elle présente également une grande partie des principaux gènes impliqués dans la synthèse de la muréine, ce qui semble suggérer que ce phylum aurait divergé de bactéries possédant ce peptidoglycane en développant une paroi cellulaire protéique. Le système de compartimentation, unique à ce phylum, ammène également quelques observations. Elle a ainsi parmi les plus hauts taux de protéines présentant un peptide signal dans un génome, ses protéines devant franchir la membrane du nucléoïde pour être adressées dans son paryphoplasme. Parmi les autres découvertes inattendues de ce génome, deux ont en particulier suscité un grand intérêt dans notre laboratoire : *R. baltica* possède en effet plus de 100 gènes codant pour des formylglycine-sulfatases et autant codant pour des polysaccharidases.

Posséder 100 gènes de sulfatases est déjà énorme pour un seul organisme, mais ce nombre l'est également vis-à-vis de l'ensemble des sulfatases trouvées dans les banques de données. En effet, à titre de comparaison, l'analyse de 70 génomes procaryotes séquencés en 2002 identifiait un maximum de six sulfatases (dans le génome de *Pseudomonas aeruginosa*). Depuis 2003, les génomes d'autres bactéries marines hétérotrophes ont été séquencés, et si certaines d'entres elles présentent un nombre de sulfatases du même ordre de grandeur (G. Michel, communication personnelle), *R. baltica* reste parmi les grands pourvoyeurs de ce type d'activité. Le fait que *R. baltica* possède un nombre de sulfatases supérieur de deux ordres de grandeur à ceux de la quasi-totalité des génomes procaryotes connus a soulevé bien évidemment la question de leur rôle biologique. Les sulfatases

bactériennes semblent être essentiellement utilisées pour subvenir aux besoins en sulfate de ces organismes (Kertesz, 2000). Etant donné que le milieu marin a de hautes concentrations en sulfate, il est très probable que *R. baltica* se serve de sa panoplie de sulfatases plutôt pour dégrader des substrats sulfatés que de se fournir en sulfate. Cette hypothèse est d'ailleurs soutenue par le fait que le quart de ces sulfatases présente un peptide signal, suggérant qu'elles sont, au moins paryphoplasmique, voire sécrétées dans le milieu extérieur. Certaines sulfatases ont de plus été trouvées dans un contexte génomique très orienté vers la dégradation de polysaccharides algaux, qui sont connus depuis longtemps pour contenir justement de nombreuses substitutions sulfatées (ce point sera abordé dans la section suivante).

Le fait que *R. baltica* possède plus de 120 polysaccharidases est également très intéressant, et ce d'autant plus qu'une majorité d'entre elles présentent un peptide signal, suggérant leur probable exportation dans le périplasme (le paryphoplasme chez les *Planctomycetes*) et une action sur des polysaccharides exogènes.

Ces deux données sont à mettre en perspective pour avoir accès aux substrats auxquels *Rhodopirellula baltica* a accès. En effet, *elle* a été isolée à partir des neiges marines dont le contenu exact en macroparticules n'est pas connu. Les algues cependant, en tant que macroorganismes marins, doivent contribuer en proportion non négligeable à leur formation. Il semble donc que *R. baltica* soit très bien armée pour dégrader des polysaccharides très variés et pour désulfater des substrats exogènes. Faisant partie d'un phylum contenant des organismes très répandus dans les océans, un rôle de grand dégradeur des polysaccharides marins semble donc s'imposer.

R. baltica est depuis son séquençage étudiée en tant qu'organisme modèle pour la dégradation aérobie des polysaccharides en milieu marin. De récentes analyses de son protéome ont ainsi été réalisées pour tenter de mettre en évidence un schéma global de catalyse et pour décrire les principales voies métaboliques de la croissance de *R. baltica* (Rabus *et al.*, 2002; Gade *et al.*, 2005a; Hieu *et al.*, 2008). L'une de ces études a en particulier étudié sa prise en charge des saccharides simples, afin de mettre en évidence ses voies métaboliques centrales (Gade *et al.*, 2005b).

Cependant, malgré son statut d'organisme modèle pour la dégradation des polysaccharides environnementaux, il n'existe aucune caractérisation des voies métaboliques impliquant les polysaccharides complexes chez *R. baltica*, qui sont pourtant concrètement les substrats qu'elle dégrade au quotidien. Cette étape est ainsi essentielle

pour répondre à la question fondamentale des bases moléculaires de la dégradation des polysaccharides par *R. baltica*.

III.B.2 - Les polysaccharides en question(s)

Une grande raison de l'absence d'étude du métabolisme des polysaccharides complexes chez *R. baltica* est, justement, que les polysaccharides auxquels elle a accès sont d'une grande complexité. Il convient avant d'aller plus loin d'introduire ce que j'entends par *polysaccharides complexes*. Considérons la paroi des végétaux supérieurs, et celle des algues, afin de les mettre dans une perspective commune.

La **paroi des végétaux supérieurs** est un enchevêtrement de polysaccharides réticulés entre eux et avec des complexes protéiques. Globalement, la partie squelettique de la paroi est composée de deux types de fibres polysaccharidiques : les microfibrilles de cellulose, responsables de la rigidité de la paroi, et les chaînes d'autres polysaccharides appelés collectivement hémicellulose. Ces différentes chaînes de polysaccharides peuvent s'assembler suivant deux grands schémas architecturaux, constituant les deux grands types de paroi végétale.

Le type I est rencontré chez la plupart des plantes dicotylédones et certaines monocotylédones. La paroi végétale de type I est caractérisée par des quantités à peu près égales de xyloglucanes (XyG) et de cellulose. Les XyG se fixent aux chaînes de glucanes de la cellulose, permettant d'orienter la microfibrille, en jouant sur la distance séparant deux microfibrilles adjacentes ou se lient entre eux. Le réseau XyG-cellulose est de plus encapsulé dans un maillage de pectine (homogalacturonane) présentant une forte méthyl-estérification. La pectine forme la partie amorphe de la paroi et comble les espaces intracellulaires.

Le type II est rencontré chez certaines plantes monocotylédones. Ce type présente une architecture semblable au type I, à cela près que les XyG sont remplacés par des glucuronoarabinoxylanes (GAX). Ces structures sont en général pauvres en pectines mais des contributions de charges sont apportées par les résidus glucuroniques de la chaîne de GAX. De plus, ces parois contiennent généralement peu de protéines structurales en comparaison avec le type I. Les parois de type II peuvent cependant accumuler une forte réticulation en développant avec l'âge des réseaux de phénylpropanoïdes.

La Figure I-7 présente une modélisation de la paroi des végétaux supérieurs.

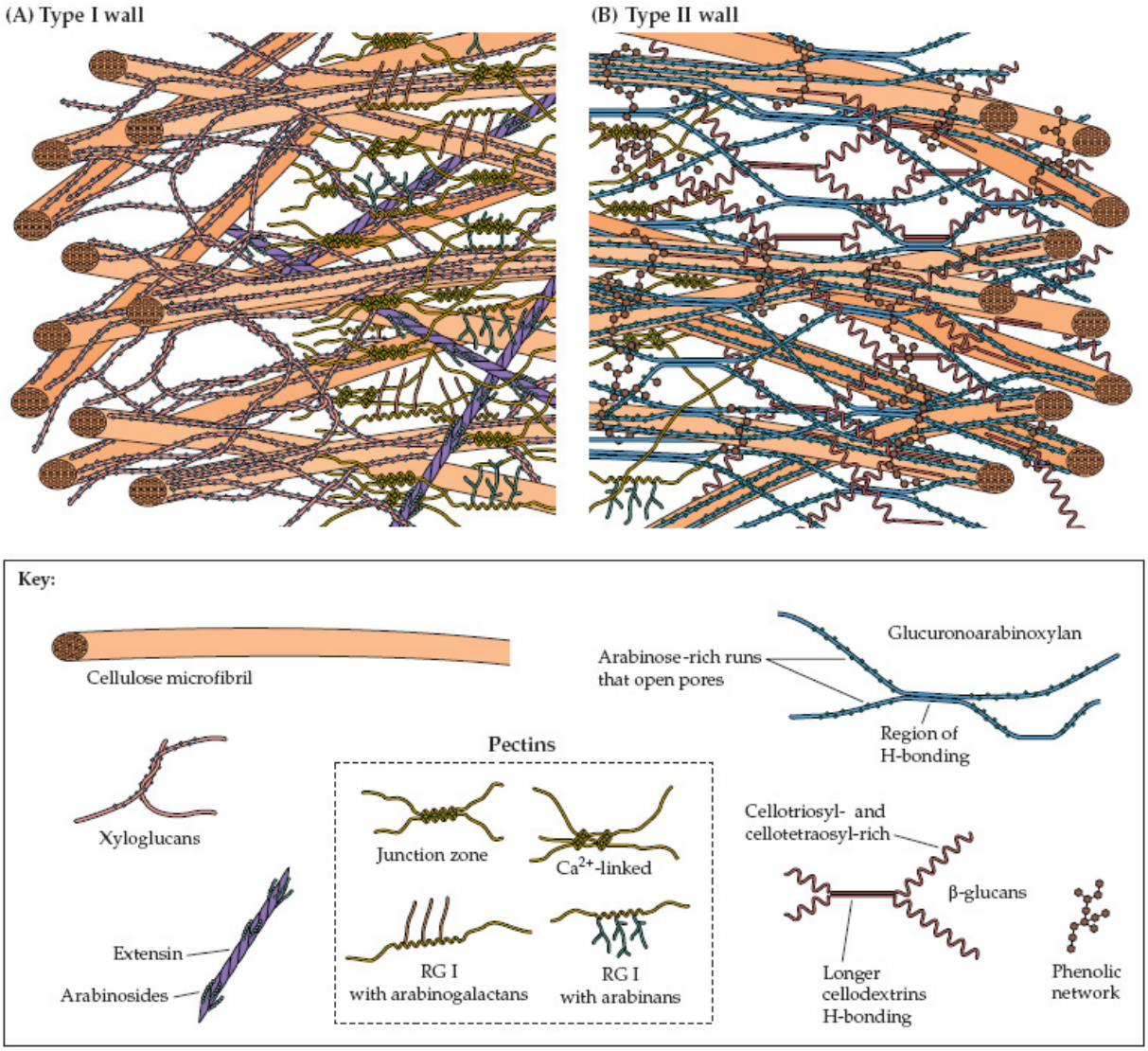


Figure I-7 : Modélisation de la paroi des végétaux supérieurs.
(A) Modèle de paroi de type I ; (B) Modèle de paroi de type II, en particulier présent dans l'herbe. Figure extraite de (Carpita and McCann, 2000).

La **paroi des algues** présente quelques ressemblances avec celle des plantes terrestres, avec en particulier une phase squelettique composée de structures microfibrillaires emmaillée dans une phase amorphe. La paroi des algues est surtout de nature très variée et dépend drastiquement du genre algal considéré. En effet, les algues sont séparées en trois grands types monophylogéniques qui présentent des divergences majeures, au nombre desquelles on compte la structure de leur paroi.

Chez les algues vertes (qui sont phylogénétiquement proches des plantes), les parois présentent de grandes variations selon les genres algaux. La cellulose est souvent rencontrée dans la partie microfibrillaire, mais il est aussi possible de trouver des β-mannanes, des xylanes, ou encore des glycoprotéines. Les polysaccharides de la partie

mucilagineuse ne sont pas très bien identifiés, mais la famille des *Ulvales* est connue pour produire un rhamnoglucuronane particulier, l'ulvane (Lahaye and Robic, 2007).

Chez les algues rouges, la partie microfibrillaire rigide est souvent composée de β-glycanes (cellulose, mannanes ou xylanes). La partie mucilagineuse est elle le plus souvent composée de galactanes plus ou moins sulfatés, de type agars ou carraghénanes. Il a cependant été rapporté des algues rouges présentant d'autres mucilages, tels que des mannanes sulfatés, ou encore des polysaccharides riches en mannose et faiblement ioniques. On notera que les algues rouges d'eau douce présentent souvent des mucilages peu ioniques, au contraire des algues marines, qui présentent des mucilages souvent sulfatés.

Enfin, chez les algues brunes, qui constituent un phylum à part des deux autres, la partie microfibrillaire est essentiellement cellulosique, tandis que la partie mucilagineuse est composée d'un maillage d'alginates et de polymères complexes à base de fucose.

Le Tableau I-5 présente une comparaison de la composition en polysaccharides ioniques caractéristiques de plusieurs phyla.

	Métazoaires	Plantes terrestres	Algues
Polysaccharides sulfatés	Héparines / Héparanes	Absents	Carraghénanes (algues rouges)
	Chondroïtines		Agars (algues rouges)
	Dermatanes		Fucanes (algues brunes)
Polysaccharides carboxylés	Acides hyaluroniques	Acides pectiques	Alginates (algues brunes)
			Ulvanes (algues vertes)

Tableau I-5 : Répartition des polysaccharides anioniques.
Présentation comparée de la composition en polysaccharides ioniques de la matrice extracellulaire des animaux, de la paroi des plantes terrestres ainsi que de la paroi des algues.

Il apparaît donc que le métabolisme des polysaccharides complexes est très loin d'être facile à étudier. Il implique de s'intéresser à des sources de carbone très différentes, présentant des structures bien définies au sein des végétaux, mais qu'il est difficile de reproduire en laboratoire. Par exemple les microfibrilles permettant la structuration des parois ne présentent pas du tout la même accessibilité que les polysaccharides mucilagineux

qui présentent souvent une forte solubilité. La notion d'accessibilité au substrat est pourtant une notion fondamentale dans la biochimie des enzymes : certaines tolèrent des substrats en milieu hétérogène, tandis que d'autres ne pourront prendre en charge le même substrat que dissout en solution.

III.B.3 - Une annotation à parfaire

Au vu de l'ensemble de ces données, il semble clair que les gènes codant les polysaccharidases chez *R. baltica*, qui sont les fondements même des bases moléculaires de son rôle écologique, rassemblent nombre des phénomènes pouvant générer traditionnellement des erreurs dans les annotations de génomes. D'une manière générale, les polysaccharidases appartiennent à des familles d'enzymes qu'il n'est pas facile d'annoter.

Un grand nombre d'entre elles présente ainsi une structure multimodulaire qui n'est régulièrement pas prise en compte au cours des annotations manuelles des génomes, voire pas du tout prise en compte par les méthodes d'annotation automatiques. En effet, la multimodularité des enzymes est un phénomène difficile à analyser. Certaines présentent des modules bien différenciés, d'autres des modules s'interpénétrant, d'autres présentent même des modules imbriqués dans d'autres modules (voir par exemple (Crennell *et al.*, 1994)). Certaines enzymes présentent plusieurs modules catalytiques et d'autres un unique accompagné d'une multitude de modules annexes. Tous ces cas de figure doivent être traités exhaustivement et surtout, au cas par cas. A notre époque d'accélération de la recherche et d'augmentation des échelles de mesure, cette approche n'est malheureusement pas beaucoup utilisée.

Ces enzymes ont de plus pour cibles des objets macromoléculaires complexes étant d'une part de natures très variées (cellulose, xylanes, carraghénanes, galacturonanes, glucurono-arabinoxylanes, …) et d'autre part de structures également variées (microfibrillaires, chargés, estérifiés, …). Ces activités sont décrites dans des familles référencées au sein de plusieurs banques de données (CAZy, PFAM, …) qui sont une aide très précieuse pour l'annotateur, étant donné qu'il peut y trouver une vision globale des caractéristiques biochimiques et structurales de ces familles. Cependant, les polysaccharidases voient souvent leur activité évoluer suite à de très fines variations des séquences protéiques, variations qui peuvent passer totalement à travers les filtres mis en place si la similitude des enzymes est à la base peu importante. C'est dans cette situation que l'esprit critique de l'annotateur fera toute la différence.

C’est ici qu’intervient le dernier point. En effet, *R. baltica* est justement une bactérie présentant une grande divergence par rapport aux autres phyla bactériens. Une divergence telle que plus de la moitié de son génome n’a pas pu être annotée, faute d’homologues dans les banques de données. Pour autant, plus d’une centaine de polysaccharidases ont pu être identifiées parmi ses gènes annotables, laissant présager que bien des activités n’ont pas encore été révélées. Cette grande divergence peut conduire à une qualité d’annotation rabaissée du simple fait que les enzymes trouvent difficilement leurs homologues. Encore une fois, seule une annotation experte et critique peut ici faire la différence. Ceci est d’autant plus important que *R. baltica* peut apporter, en tant qu’organisme modèle, des éclaircissements majeurs sur les voies de recyclage des polysaccharides complexes dans les écosystèmes marins.

Problématique de la thèse

Ces vingt dernières années ont été marquées en biologie par l'avènement de la **génomique**, la science de l'étude des génomes. L'ère de la post-génomique est née de l'interaction entre les méthodologies de la génomique, en particulier les analyses à grande échelle des systèmes, avec les autres champs disciplinaires de la biologie. C'est au cours d'études métagénomiques dans la mer Baltique que la planctomycète marine ***Rhodopirellula baltica*** a été identifiée. Son génome a été le tout premier génome séquencé d'une **bactérie marine hétérotrophe** (Glöckner *et al.*, 2003). Il a révélé un rôle potentiellement majeur de *R. baltica* dans le recyclage des polysaccharides. En particulier, plus de cent sulfatases et de **polysaccharidases** ont pu être identifiées, avec des fonctions à la fois larges et très **originales**. Le but de mes travaux au cours de ces années de recherche aura été de permettre la **validation** du potentiel de dégradation de cet acteur écologique que constitue *Rhodopirellula baltica*. Cette validation a pu s'inscrire dans une perspective résolument moderne combinant génomique structurale et fonctionnelle, avec le recours à une **approche à moyen débit** d'une sélection parmi ces enzymes. J'ai cherché à donner les outils moléculaires nécessaires pour commencer l'effort de validation systématique, ainsi qu'une révision préliminaire manuelle des annotations, par des méthodes éprouvées.

Ainsi, dans le chapitre II de ce manuscrit, je présenterai une étude à moyen débit réalisée sur les polysaccharidases de *R. baltica*. Ce travail a eu pour but d'exprimer en système hétérologue un ensemble de gènes centraux du métabolisme des polysaccharides complexes de *R. baltica*, afin de permettre des études biochimiques et structurales sur les enzymes qu'ils codent.

Dans le chapitre III, je présenterai les résultats de surexpression, purification, caractérisation biochimique et cristallographie de quatre enzymes issues des meilleures expressions solubles les plus intéressantes de l'étude à moyen débit.

Dans le chapitre IV, je présenterai les révisions manuelles des annotations de l'ensemble des gènes codant pour des polysaccharidases de *Rhodopirellula baltica*. Une description plus détaillée de certains métabolismes centraux de polysaccharides complexes sera également présentée.

Enfin, un résumé de l'ensemble du travail de thèse sera proposé dans le chapitre V. Fort des résultats de ces analyses, des conclusions sur le modèle et les méthodes de validation seront proposées, ainsi que des perspectives pour les travaux restant à effectuer.

Chapitre II

-

Etude à moyen débit des polysaccharidases de *Rhodopirellula baltica*

I - Etude à moyen débit des polysaccharidases de Rhodopirellula baltica

I.A - Recensement et sélection des protéines

I.A.1 - Recensement

Le recensement de l'ensemble des gènes de *R. baltica* susceptibles de coder des protéines actives sur les polysaccharides a été opéré à partir du génome publié (Glöckner *et al.*, 2003), et d'une recherche des familles enzymatiques connues dans la base de données CAZy. 128 séquences de protéines ont ainsi été recueillies, réparties dans les 4 grandes classes d'activités répertoriées dans les polysaccharidases (GT : glycosyl transferases, GH : glycoside hydrolases, CE : carbohydrate esterases et PL : polysaccharides lyase). Le Tableau II-6 présente la répartition de ces enzymes dans les différentes classes de polysaccharidases et le nombre de familles de la base CAZy identifiées.

Classe	GH	GT	PL	CE	Total
Nombre d'enzymes	39	61	5	23	128
Nombre de familles	17	14	3	7	41

Tableau II-6 : Polysaccharidases de *R. baltica*
Présentation des polysaccharidases de *R. baltica* en fonction de leur classe d'activité et de leur famille CAZy.

Le but du sujet était d'appréhender la complexité de reconnaissance des sources de carbone par *R. baltica*. *In fine*, les protéines sont destinées à une étude biochimique pour identifier leur(s) substrat(s) et à une étude cristallographique pour comprendre la nature de leur mode d'action. Ces deux objectifs ne sont pas compatibles avec la multimodularité fréquente des polysaccharidases et implique qu'il faille séparer physiquement les différents modules. En effet, si une analyse de l'activité globale d'une enzyme multimodulaire peut permettre de comprendre son rôle biologique, seule une étude indépendante de chacun de ses modules permet la compréhension des mécanismes de dégradation, de modification, et/ou de synthèse du substrat. Pour ce qui est de la cristallographie, elle aussi peut être très limitée par la multimodularité des enzymes. En effet, les modules sont très souvent reliés

entre eux par une séquence flexible, peu voire pas du tout repliée et donc peu susceptible de participer à la périodicité du cristal. Séparer les modules est non seulement indispensable à leur analyse par cristallographie (pour la raison citée plus haut) mais également possible étant donné que, de par la définition même de « module », ceux-ci sont structuralement indépendants (Henrissat and Davies, 2000).

Cette troncature des protéines implique que les modules soient délimités de façon réfléchie. Couper trop court dans la séquence peut engendrer des instabilités dans des éléments de structure secondaire et empêcher un repliement stable de la protéine, ce qui se traduit au mieux par un frein à la cristallogénèse (l'anisotropie et la non périodicité étant les poisons de la genèse cristalline), et au pire par la non solubilité de la protéine. Couper trop long dans la séquence peut ajouter des régions flanquantes à la protéine, qui ne se replieront potentiellement pas ou peu (étant elles-mêmes des fragments des modules voisins), ou bien qui, quoique repliées, auront un positionnement plus ou moins aléatoire. Ce type d'élément ajoutant une apériodicité peut complètement inhiber la croissance cristalline. Il était donc critique de cerner quelles protéines présentaient une structure modulaire d'une part, et de déterminer la limite la plus probable de ces modules d'autre part. Les points suivants détaillent l'investigation.

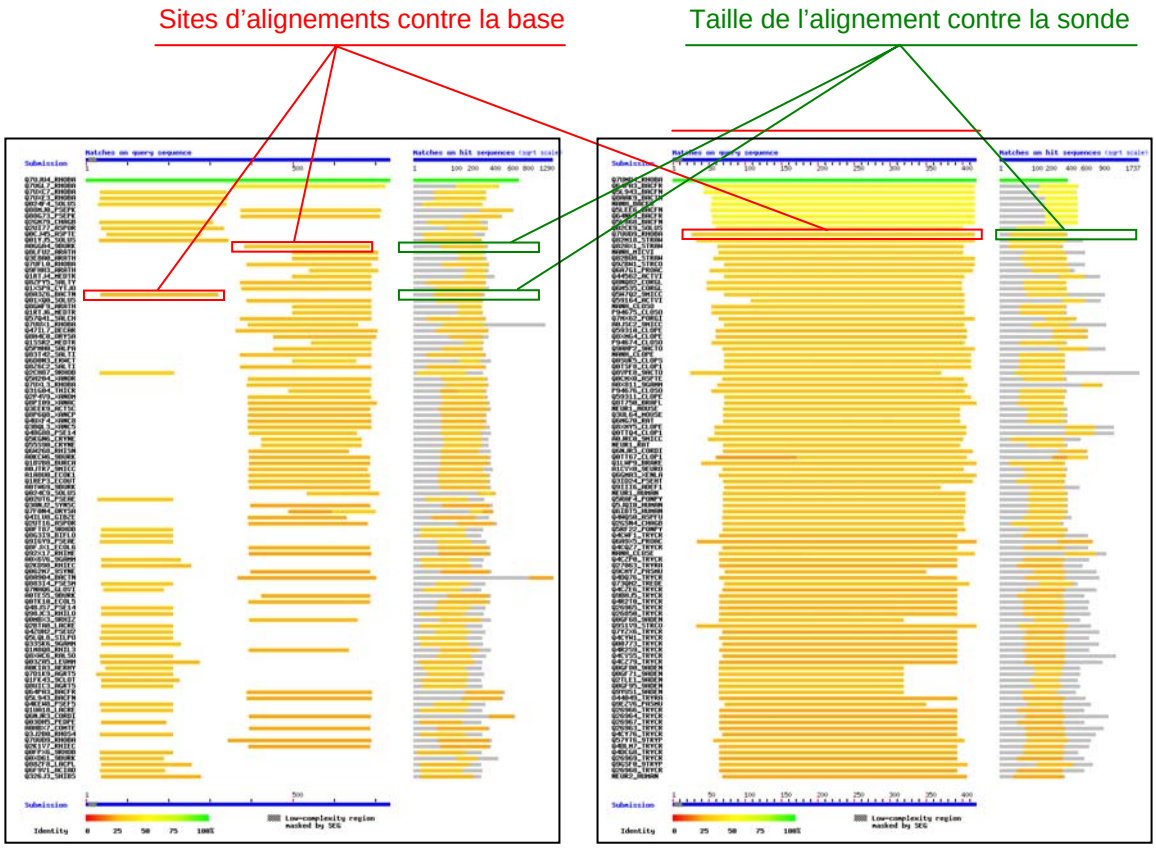
I.A.2 - Peptides signaux et hélices transmembranaires

Les éventuels peptides signaux ont tout d'abord été supprimés des séquences protéiques. En effet, conformément aux attentes que l'on pourrait avoir d'enzymes censées dégrader des polysaccharides exogènes, un nombre non négligeable des enzymes de *R. baltica* recensées présentaient ce type d'adressage : 80% des PL, 60% des CE, 45% des GH, et très peu de GT avec 3% d'entres elles.

Une recherche d'éventuelles hélices transmembranaires a également été effectuée sur les protéines. Très peu ont montré ce type de structure (7 au total) : 4 GT, 1 CE, 1 GH et 1 PL. Néanmoins, les différentes classes ont présenté de grandes différences dans le nombre et la disposition des hélices. Ainsi, les 4 GT possèdent toutes plusieurs hélices (de deux à une dizaine), situées au cœur de la séquence, faisant d'elles des protéines véritablement transmembranaires. La GH, la CE et la PL n'en présentent chacune qu'une seule qui est de plus très proche de l'extrémité N-terminale de la protéine. Ce dernier type d'hélices transmembranaires prédites ressemble beaucoup aux peptides signaux : elles peuvent éventuellement être supprimées par biologie moléculaire afin de générer une protéine recombinante plus courte mais non ancrée à la membrane.

I.A.3 - Modules et limites

La structure modulaire des enzymes a été analysée par une recherche de type BLAST contre la base de données UniProtKB/trEMBL-nr et UniProtKB/Swiss-Prot afin de déterminer si ces enzymes présentaient plusieurs sites d'alignement. Il s'est assez vite avéré qu'un grand nombre d'entre elles possédaient plusieurs modules (jusqu'à huit), de différentes natures (catalytiques ou non, agissant sur les sucres ou non). Ceci est illustré par la Figure II-8 (images extraites du BLAST hébergé sur www.expasy.org). Dans cet exemple, la protéine RB8895 présente un long domaine aligné sur la quasi-totalité de sa séquence, signe qu'elle est très probablement non modulaire. A l'inverse, la protéine RB11055 présente quant à elle clairement deux zones alignables. Typiquement, dans cette situation, un second jeu de recherche est lancé avec BLAST contre UniProtKB/trEMBL-nr et UniProtKB/Swiss-Prot pour chacune des différentes zones alignables détectées.



Résultat de BLAST pour la séquence de la protéine **RB11055**. Deux hits d'alignement sont ici visibles, signant la présence probable de deux modules.

Résultat de BLAST pour la pour la séquence de la protéine protéine **RB8895**. Cette protéine ne semble pas modulaire.

Figure II-8 : Exemples de résultats de BLAST.
Présentation de deux résultats de BLAST illustrant le caractère modulaire ou non de certaines polysaccharidases

La mise en évidence de l'existence de modules a constitué une première étape. La délimitation fine des bornes des modules a été réalisée en procédant par analyse HCA (Hydrophobic Cluster Analysis) (Gaboriaud *et al.*, 1987; Callebaut *et al.*, 1997) dont un exemple est présenté en Figure II-9**.**

Le but de ce type d'analyse est de prédire la présence des éléments de structure secondaire à partir de la séquence primaire de la protéine. Dans son principe, la séquence primaire de la protéine étudiée est enroulée comme s'il s'agissait d'une hélice α continue le long de la séquence (avec un tour tous les 3,6 acides aminés).

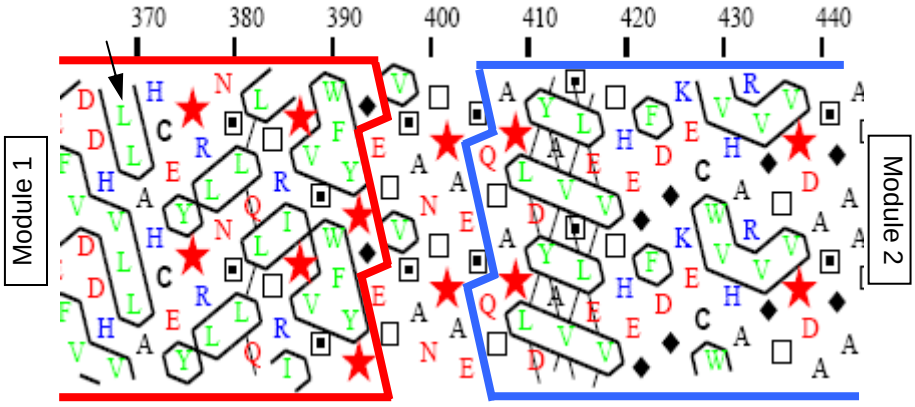


Figure II-9 : Présentation en diagramme HCA.
Présentation d'une limite typique entre deux modules. La flèche indique le sens de lecture du graphe.

Cette hélice α est ensuite coupée longitudinalement, étalée et dupliquée pour une meilleure visibilité. Le diagramme se lit donc de gauche à droite, « en diagonale » et en suivant l'ordre de la séquence primaire. Dans le formalisme HCA, les acides aminés sont représentés par leur code à une lettre, sauf certains d'entres eux qui sont représentés par des symboles (dans le but de les repérer facilement dans la séquence). Les glycines et les prolines, représentants les changements de direction dans la continuité de la séquence, sont ainsi respectivement symbolisées par un losange noir (♦) et une étoile (★). Les thréonines et les sérines, qui peuvent présenter une certaine amphipathie, sont elles respectivement représentées par un carré blanc (□) et un carré blanc avec un point noir (⊡). Les acides aminés acides ou polaires (DENQ) sont de plus représentés en rouge, les basiques (KRH) en bleu et les acides aminés hydrophobes (WFYLIMV) sont représentés en vert et entourés d'un liseré noir. Lorsque deux acides aminés hydrophobes (ou plus) se retrouvent côte à côte, leurs liserés fusionnent et forment des blocs appelés amas hydrophobes. Deux acides aminés hydrophobes appartiennent à deux amas différents lorsqu'ils sont séparés par au moins quatre résidus polaires ou par une proline.

Le diagramme HCA ainsi présenté met en évidence la périodicité de ces amas dont la disposition permet une prédiction souvent fiable d'un élément de structure secondaire. Les amas verticaux ont de fortes chances de représenter des brins β, tandis que les amas horizontaux ont de fortes chances de représenter des hélices α. Sur la Figure II-9, la séquence qui relie les deux modules est caractéristique d'une zone non repliée qui ne sert que de liaison entre les modules : il n'y a pas d'élément de structure secondaire visible (i.e. pas d'amas) et elle est essentiellement composée de petits acides aminés (S, T, G, P, A) et d'acides aminés chargés.

A l'issue de cette analyse de chaque séquence, 165 modules différents ont été identifiés. Dans leur majorité, ces modules, bien qu'étant au sein d'une protéine ayant sans ambiguïté une action liée au métabolisme des polysaccharides, se sont révélés soit conservés mais de fonction inconnue, soit orphelins. La grande divergence de séquence des protéines de *R. baltica* (Glöckner *et al.*, 2003), par rapport aux orthologues présents dans l'ensemble des génomes bactériens séquencés au début de ma thèse, a été un handicap à cette étape. Pour la même raison, la réannotation de ces différentes enzymes (voir chapitre IV) s'est avérée difficile, du fait de l'absence de comparaison possible avec d'éventuels homologues. Ceci a été en partie remédié par la publication de deux génomes de planctomycètes au cours de ma thèse : ceux de *Blastopirellula marina* et de *Planctomyces maris* (Woebken *et al.*, 2007) qui m'ont beaucoup apporté par leur potentiel de comparaison à *R. baltica*. Ainsi, des modules initialement orphelins se sont retrouvés conservés chez ces deux planctomycètes, suggérant des fonctions uniques à ce phylum bactérien. La délimitation par bioinformatique de ces modules a également pu être affinée, bien que la méthode HCA aie souvent permis des prédictions correctes.

A titre d'exemple, la protéine RB3006 (annotée en tant que « sialidase ») est trimodulaire. Elle présente un module catalytique de la famille GH33 (responsable de son annotation) et deux modules de taille comparable mais de fonctions inconnues, dont l'un d'entre eux était orphelin lors de mes analyses. Ce dernier module a néanmoins été sélectionné et a montré une très bonne expression sous forme soluble en système hétérologue, confirmant la prédiction de ses bornes. Une comparaison avec les génomes de *B. marina* et *P. maris* a permis de constater que non seulement des homologues existaient dans ces deux génomes, confirmant le caractère indépendant de ce module, mais qu'en plus, l'un de ces homologues était une protéine à part entière chez *B. marina,* ajoutant encore à l'intérêt que nous lui portions. L'ensemble des analyses précédentes (BLAST et

HCA) m'a ainsi permis de donner une prédiction fiable de l'existence de ce module ainsi que de ses bornes confirmant ainsi la validité de cette approche.

I.A.4 - Sélection et commentaires

Les gènes voués à entrer dans la sélection finale ont été choisis parmi ces 165 modules, en fonction de leur appartenance aux différentes stratégies de clonage décrites dans le paragraphe Chap II – II.B.3. Une trentaine de gènes a ainsi été éliminé en raison de la présence de sites de restriction inopportuns dans la séquence ; ces gènes pourront faire l'objet d'une étude approfondie, mais ils nécessiteront d'être traités au cas par cas. En ce qui concerne les gènes restants, la sélection s'est opérée sur l'intérêt scientifique particulier de chaque protéine. En l'occurrence, les critères retenus ont été la présence d'une activité originale pour ce type de bactérie, ou encore une séquence fortement divergente au sein d'une famille bien étudiée, ou au contraire, une séquence appartenant à une famille peu étudiée

Je ne vais pas détailler les raisons de l'intégration de chaque protéine à la liste finale mais je pense néanmoins qu'il serait très intéressant de souligner l'origine des choix pour quelques unes de ces protéines.

Comme il a été montré précédemment, le xylane est un polysaccharide majeur de la paroi des plantes supérieures. Il représente également une source de carbone potentielle pour *R. baltica*, puisque beaucoup d'enzymes ont une activité reliée à sa dégradation. Un grand nombre de ces enzymes ont donc été sélectionnées, en pariant sur la potentielle originalité de leur action. Par exemple, l'une d'entre elles, *RB10416*, possède un module de la famille GH10 et est annotée « endo-1,4-beta-xylanase ». Cette enzyme présente la particularité d'être extrêmement modulaire, avec huit modules identifiés. Les huit modules de cette protéine ont ainsi été sélectionnés afin de pourvoir cerner expérimentalement leur rôle.

Beaucoup de glycosyltransférases de la famille GT4 ont également été sélectionnées (14 gènes, soit environ 15% des peptides totaux). En effet, il s'agit de la famille la plus représentée au sein du génome de *R. baltica* (avec 28 paralogues recensés). Cette famille fait partie des familles de GT les plus grandes de la base de données CAZy (avec plus de 8000 séquences) mais également des moins étudiées structuralement puisqu'aucune structure n'avait encore été résolue lors du choix des protéines. Cette précision a son importance car, depuis le début de ma thèse (janvier 2005), pas moins de 7 structures ont été résolues : **2F9F** (non publié Zhou *et al.*, 2006), **2JJM** (Ruane *et al.*, 2008), **3C48** (Vetting *et al.*, 2008), **2IUY** et **2IV7** (Martinez-Fleites *et al.*, 2006), **2R60** (Chua *et al.*, 2008) et

2GEJ (Guerin *et al.*, 2007). Cela confirme bien l'intérêt porté par la communauté scientifique à cette famille, du fait notamment que ces enzymes sont entre autre impliquées dans la synthèse des polymères composant la paroi bactérienne.

La famille des glycosyltransférases GT2 est l'autre grande famille d'enzymes sélectionnées avec 12 peptides, pour sensiblement les mêmes raisons que la famille GT4.

Le Tableau II-7 présente les 96 peptides choisis, répartis tels que disposés au sein de la plaque de clonage multipuits utilisée au cours de ma thèse. Sont représentés en outre leur répartition dans les différentes stratégies de clonage et, par un code couleur, la taille des gènes clonés (en pb). L'ensemble des cibles sélectionnées représente différentes classes d'enzymes, couvrant un spectre d'activités le plus large possible, sous forme de module isolés ou bien de protéines entières:

- Glycosyltransférases : familles GT2, GT4, GT11, GT12, GT25, GT26, GT30, GT32, GT35 et GT non classées (GTNC);
- Glycoside hydrolases : familles GH5, GH10, GH16, GH20, GH33, GH43, GH57, GH78 et GH86 ;
- Carbohydrate esterases : familles CE1, CE4, CE9, CE6, CE11 et CE12 ;
- Polysaccharide lyase : familles PL1 et PL7 ;
- d'autres modules de fonctions variées : Laminine, Sulfatase, Phosphatase, CBMs, Sulfotransférase, … ;
- beaucoup de modules de fonction inconnue appelés UNK pour « UNKown » associés dans les protéines sauvages à des modules catalytiques et des domaines conservés de fonction inconnue (*Domain of Unkown Function* - DUF) identifiés par la base PFAM ;
- quelques protéines modulaires ont aussi été exprimées avec l'ensemble de leurs modules, dans une perspective d'analyse biochimique complémentaire. Des études structurales à basse résolution sont aussi envisageables par des techniques en solution comme le SAXS (Small Angle X-ray Scattering). Ces constructions sont appelées « ALL » dans la plaque.

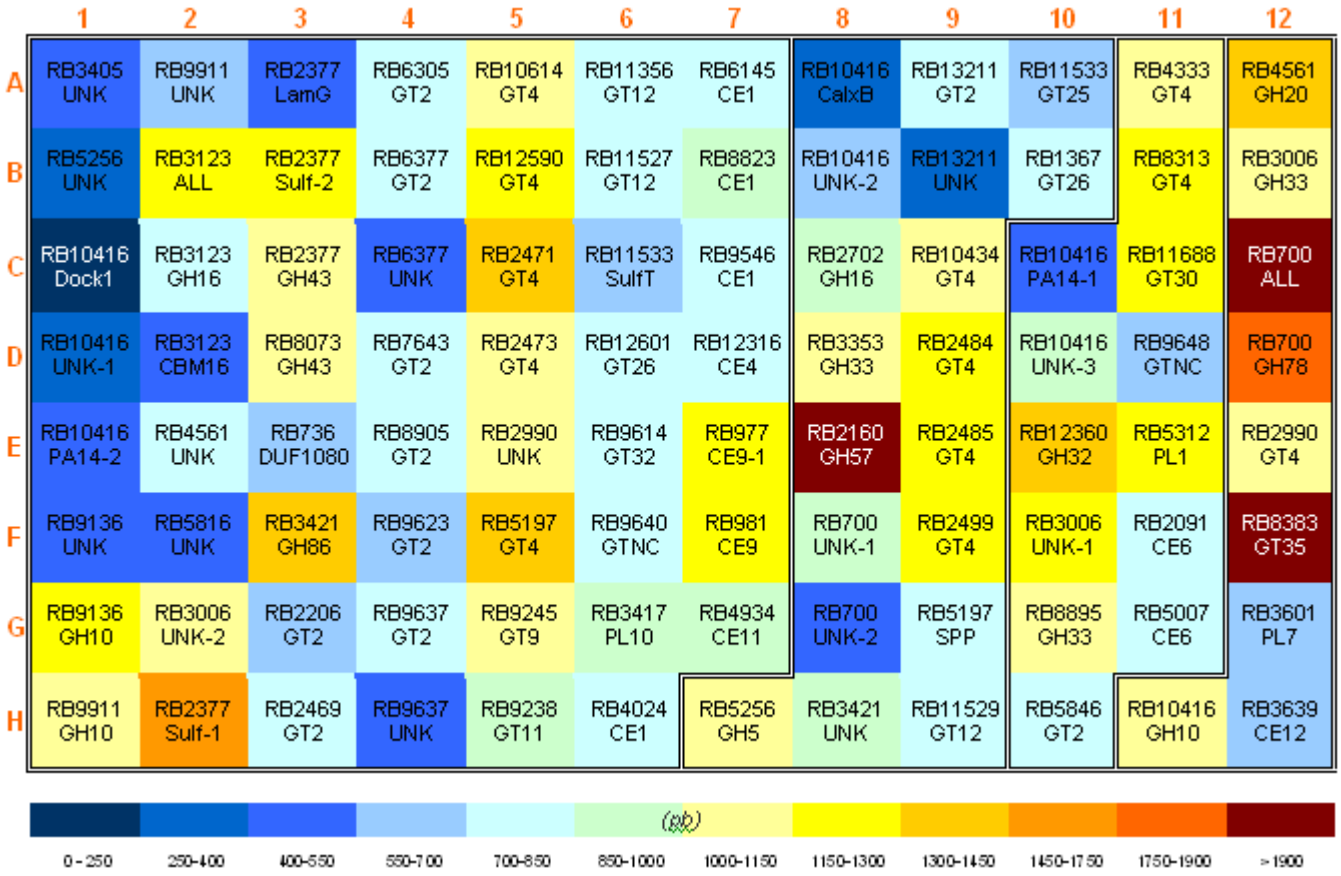


Tableau II-7 : Organisation de la plaque.

Présentation de la plaque avec les peptides dans leur disposition finale. Les couleurs indiquent la taille des fragments géniques clonés (en pb – cf échelle de couleurs ci-dessus) et les zones délimitées par « ‖ » correspondent aux stratégies de clonage avec de gauche à droite : *Bam*HI-*Eco*RI, *Bam*HI-*Mfe*I, *Bgl*II-*Eco*RI et *Bgl*II-*Mfe*I.

Les codes utilisés pour définir chaque module sont les suivants (avec éventuellement leur numéro PFAM):

ALL	Protéine entière (sans hélice transmembranaire ni peptide signal)		
GH	Glycoside hydrolase	CalxB	Calx-beta domain (PF03160)
GT	Glycosyl transférase	PA14	Adhésion (PF07691)
GTNC	GT Non-Classée	LamG	Laminine G (PF00054)
CE	Carbohydrate esterase	SPP	Sucrose-6P phosphatase
PL	Polysaccharide lyase	Sulf	Sulfatase (PF00884)
CBM	Carbohydrate Binding Module	SulfT	Sulfotransferase(PF00685)
UNK	Fonction inconnue	Dock1	Dockerine type I (PF00404)

Les chiffres suivant un domaine (exemple : *UNK-1*) servent à différencier plusieurs domaines du même type présents dans la protéine.

I.B - Clonage et expression à moyen débit

I.B.1 - Principes et mises au point

I.B.1.1 Philosophie d'une approche à grande échelle

Il convient de présenter la philosophie du projet d'étude à moyen débit. Le terme a été choisi par référence aux programmes de « hauts débits » des grands centres de génomique structurale (Bochkarev and Tempel, 2008; Busso *et al.*, 2008; Haquin *et al.*, 2008; Marsden and Orengo, 2008). Le terme haut débit sert lui-même de référence par opposition au traitement au cas par cas, qui est souvent utilisé quotidiennement dans les laboratoires. Mon approche s'est ainsi inspirée de l'approche des grands programmes de génomique, sans atteindre le gigantisme du nombre de protéines (qui atteint jusqu'à plusieurs milliers) que ces centres sont capables d'étudier en parallèle.

Le passage d'une stratégie au cas par cas à une stratégie à plus haut débit a cependant beaucoup d'implications en terme de traitement des échantillons d'une part, mais également en terme d'acceptation de l'échec d'autre part. En effet, il n'a pas été question d'exprimer sous forme soluble l'*ensemble* des cibles mais bel et bien le *maximum*. Certaines cibles, ayant besoin d'un traitement particulier (à quelque niveau que ce soit : amplification, clonage ou expression), ont tout simplement été mises de côté. Il est possible de résumer les contraintes liées à une étude à haut débit à trois mots-clefs : la ***normalisation*** des techniques, la ***pureté*** des échantillons et la ***qualité*** des réactifs utilisés.

Par normalisation, j'entends un traitement unique pour l'ensemble des peptides et ce, à chaque étape. Ceci a nécessité par exemple l'utilisation de la plaque regroupant l'ensemble des peptides ou encore, comme décrit précédemment, que les peptides ne pouvant suivre le traitement global soient abandonnés. La normalisation a notamment pour conséquence directe la suppression de toute source de perturbation aléatoire et donc que les échantillons soient les plus purs possibles. Par exemple, il est aisé d'extraire l'ADN et de le purifier (des kits commerciaux permettent une purification très rapide), mais quand il s'agit d'une plaque de 96 échantillons, l'augmentation d'échelle change radicalement la manière de procéder (l'utilisation d'une centrifugeuse est beaucoup moins triviale par exemple). Enfin, il est essentiel d'être confiant dans le matériel utilisé. Si un réactif se trouve défectueux (enzymes de restriction non fonctionnelles par exemple), ce n'est pas un peptide mais l'ensemble qui est au mieux retardé, au pire jeté. En d'autre termes, le coût de l'échec est

démultiplié (coût en temps, en produits, en consommables, …). Ces trois contraintes sont détaillées dans les sections suivantes.

I.B.1.2 La normalisation

La normalisation des expériences a impliqué en tout premier lieu l'unicité de la température de semi-hybridation (Tm) des amorces oligonucléotidiques. Celle-ci a été choisie à 70°C ± 2°C (ce qui correspond à une température moyenne assez standard). Les amorces ont été générées en conséquence. L'utilisation d'une stratégie de clonage unique (basée sur les sites de restrictions des endonucléases *Bam*HI et *Eco*RI) s'est inscrit dans cette continuité. L'utilisation en remplacement de leur isocaudamères respectifs m'a permis d'élargir le spectre d'inclusion de cibles, avec une contrainte cependant : les sites de restriction hybrides formés par les ligatures des sites *Bam*HI/*Bgl*II et *Eco*RI/*Mfe*I ne sont plus réouvrables par aucune de ces enzymes (ceci est illustré sur les Figure II-10 et Figure II-11). Il a cependant été considéré que ceci aurait un impact mineur sur la poursuite de l'étude. Au final, 192 amorces ont été générées sur un protocole standardisé que je rappelle ici :

- amorce sens : $^{5'}$-gggggg-[site de restriction]-[séquence d'hybridation]-$^{3'}$

- amorce anti-sens : $^{5'}$-CCCCCC-[site de restriction]-stop-[séquence d'hybridation]-$^{3'}$

Ces amorces vont permettre non seulement l'amplification sélective d'un gène, mais aussi de façonner le produit PCR qui en sera généré. Dans le cas des amorces sens, la méthionine initiale est apportée par le plasmide, la séquence correspondant au peptide d'intérêt devra donc démarrer en aval de celle-ci (Figure II-10); dans le cas des amorces anti-sens, le codon stop devra être porté par l'insert, impliquant de l'inclure dans l'amorce, en amont du site de restriction (Figure II-11).

La normalisation a également impliquée l'unicité de la concentration des produits, en tout premier lieu des amorces. Ainsi, la commande et la livraison en plaque des amorces à la concentration standard de 100 µM a été dictée par cette contrainte. L'utilisation d'un unique ADN génomique (celui de *R. baltica*) suit également la même logique, et, étant donné que les amorces et la matrice d'ADN se sont retrouvées dans des concentrations standardisées pour l'ensemble des produits PCR dessinés, un ratio amorces/ADN unique a été appliqué à la plaque.

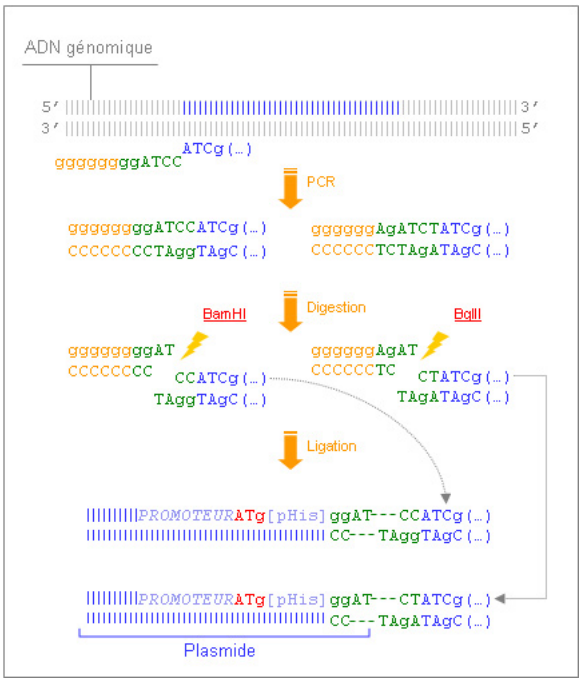


Figure II-10 : Amorces sens.
Schéma de principe de l'utilisation des amorces sens.

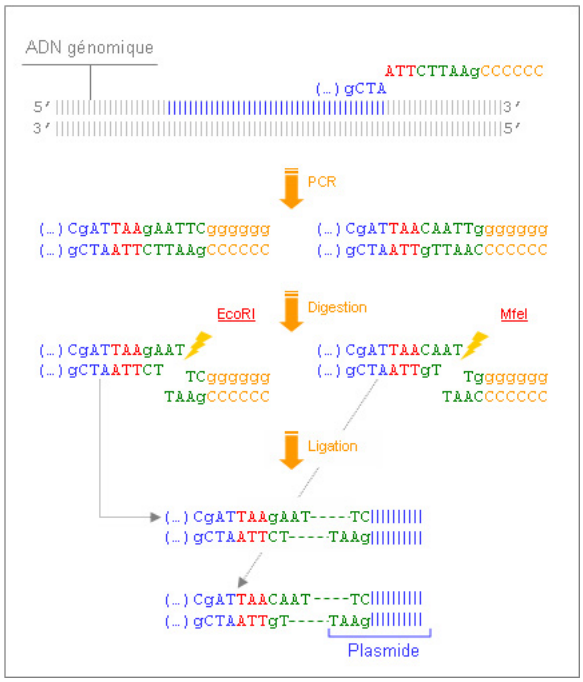


Figure II-11 : Amorces anti-sens.
Schéma de principe d'utilisation des amorces anti-sens.

I.B.1.3 Pureté

Je viens de commenter le fait qu'un unique ADN génomique a été utilisé pour l'ensemble des produits de PCR dessinés. Cela implique que cet ADN doit être reproductiblement amplifiable par PCR. Il s'est avéré que cela n'était pas aussi évident (ceci est détaillé dans la section suivante). Afin d'assurer aux ADN polymérases une accessibilité uniforme à l'ADN, une de mes premières actions aura été de le purifier sur gradient de chlorure de césium. Dans le même souci de supprimer au maximum les sources d'erreur aléatoire, un effort accru a été fourni sur la purification des produits PCR au cours du clonage, notamment avant et après coupure de restriction de leurs extrémités, étant donné que les oligonucléotides utilisés comme amorces peuvent avoir une action inhibitrice sur les nucléases et/ou les ligases.

I.B.1.4 Contrôles qualité

Pour chaque stratégie de clonage, un gène a été sélectionné comme contrôle qualité. L'ensemble des étapes a été ainsi mis en oeuvre sur quatre gènes. Ce contrôle a très vite montré son caractère essentiel. En effet, il est apparu, dès l'amplification sur l'ADN génomique, qu'un biais générait les amplifications certes à la bonne taille mais surtout

aléatoirement. Après une période de tests intensifs, la non reproductibilité des résultats m'a conduit à penser qu'une purification pourrait être nécessaire. Effectivement, un gradient isopycnique au chlorure de césium a permis d'obtenir des résultats reproductibles. C'est également au cours de ce test que le ratio amorces/ADN a été établi. La suite des étapes s'est avérée tout aussi essentielle et a permis de dépister différents écueils (compétence des cellules, fonctionnalité des nucléases, de la ligase et des polymérases, dNTP pour la PCR en quantité suffisante, …).

I.B.1.5 Optimisation pratique

Finalement, l'augmentation d'échelle par rapport à un traitement au cas par cas s'est également traduite pas une adaptation des outils usuels : utilisation de systèmes de purification par pompe à vide, utilisation de gels d'agarose beaucoup plus grands (104 puits) que d'ordinaire (dizaine de puits), limitation du nombre d'étapes, réalisation du choc thermique en plaque ou bien encore utilisation de miniplaques de boites de pétri pour les étalements de cellules. Je présente ci-après quelques-unes de ces optimisations.

La réalisation du choc thermique, qui est aisément réalisable sur quelques flacons de 2 mL, doit être adaptée à l'échelle de la plaque. L'homogénéité de la température est ici primordiale : en quelques secondes l'intégralité des puits doit passer de 0°C à 42°C, puis retourner de 42°C à 0°C. Les premiers essais ont confirmé que le refroidissement sur glace, classiquement utilisé pour des tubes séparés, n'était pas efficace car la température n'était pas homogène dans chaque puit de la plaque. Le refroidissement en milieu liquide s'est révélé très efficace pour optimiser les échanges thermiques : un mélange eau / glace pillée / NaCl a été utilisé pour les étapes à 0°C.

L'étalement sur boite de pétri a été une autre difficulté. En effet, afin de ne pas avoir à traiter 96 boites de pétri pour chaque expérience, j'ai opté pour des plaques de 6 boites de pétri. Elles présentent l'indéniable avantage d'être de plus petite contenance que les boites classiques, et donc nécessitent moins de matériel (tant pour la gélose que pour la quantité de cellules à déposer). De plus, elles permettent une manipulation plus rapide et sont empilables, ce qui a grandement facilité leur utilisation. La Figure II-12 présente une de ces plaques, ornée de quelques colonies.

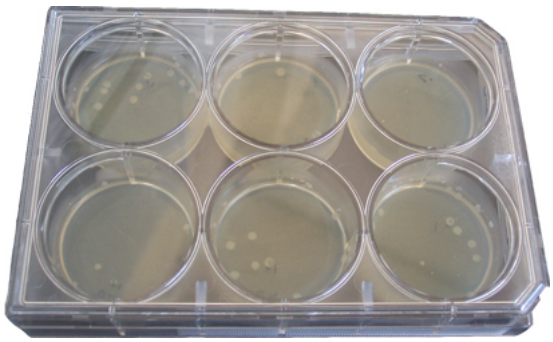


Figure II-12 : Colonies.
Présentation d'une plaque de 6 boites de pétri type avec quelques colonies.

Pour autant, l'inoculation et l'observation n'ont pas été des plus aisées de par le nombre néanmoins important de ces plaques : pour une expérience de transformation, ce sont 16 plaques qui sont générées. Sachant que les clonages ont été réalisés en parallèle dans le plasmide pFO4 et dans le plasmide pGEX-4T-1 plus d'une quarantaine de plaques devait être manipulée.

Enfin, le criblage par PCR sur colonies bactériennes aura été d'un grand secours, permettant une analyse rapide, fiable et en parallèle d'une multitude de clones. Il a également montré ses limites, car étant très facilement inhibé par des détails techniques (trop de cellules prélevées sur la colonie, ou pire, prélèvement d'un peu de gélose avec les cellules). Il est d'ailleurs intéressant ici de constater que le nombre de colonies par boite n'a pas été un facteur déterminant pour décider *a priori* de la réussite de la transformation : parfois peu de colonies (moins de cinq) ont poussé mais toutes présentaient des bandes amplifiées à la taille de l'insert alors que parfois, un nombre conséquent de colonies, était obtenu mais celles-ci s'avéraient négatives.

I.B.2 - Résultats

I.B.2.1 Résultats de PCR

L'**annexe 1** présente l'ensemble des amorces oligonucléotidiques qui ont été générées pour réaliser les PCR.

Deux séries de réactions à deux températures d'hybridation différentes ont été réalisées : la première série à 50°C a donné 72 produits de PCR et la seconde, à 60°C, en a donné 90. Au total, le cumul des deux plaques a permis d'obtenir 92 fragments amplifiés

sans bande parasite. La Figure II-13 présente la série de gels issus de la réaction de PCR à 50°C (« *plaque A* »).

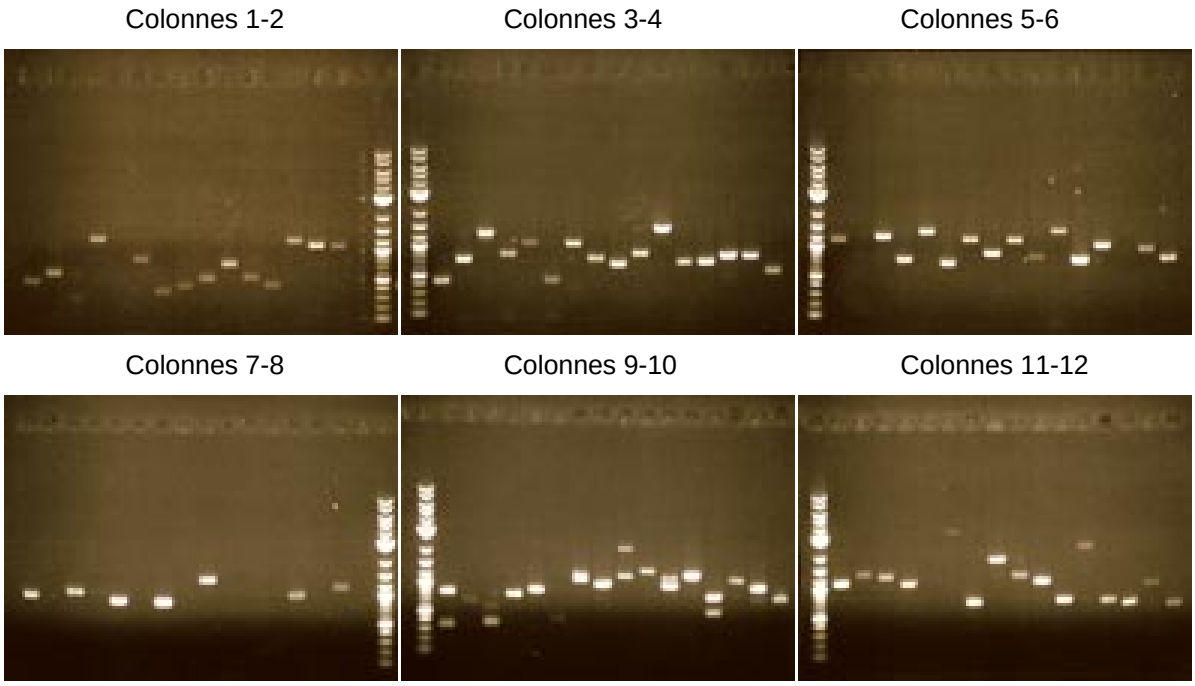


Figure II-13 : Gels d'agarose résumant les résultats de PCR.
1ers gels d'agarose 0.8% de l'amplification des cibles

Les fragments qui n'ont pas pu être amplifiés sont contenus dans les puits H02 :rb2377 module SufCa, F07 : rb981 module CE09, C09 : rb10434 module GT4 et C11 : rb11688 module GT30 (Figure II-14). En réalisant des réactions séparées pour chacune d'entre elles, aucun résultat probant n'a été obtenu. Elles ont finalement été laissées de côté et ont constitué des témoins négatifs pour les étapes suivantes en plaque (leur puit a été laissé en eau).

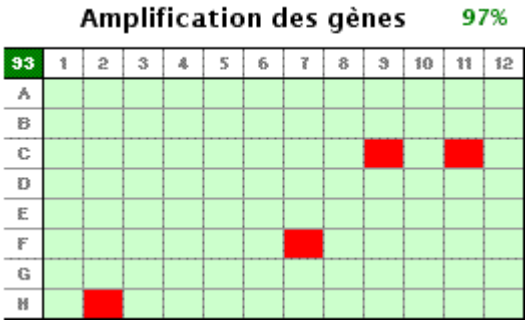


Figure II-14 : Résultats d'amplification des gènes.
Résultat d'amplification des cibles sur l'ADN génomique. vert : gènes correctement amplifiés ; rouge : gène non amplifié

I.B.2.2 Résultats de digestion

Comme expliqué précédemment, un soin particulier a été apporté à la purification des produits PCR, avant et après digestion.

Toujours dans la philosophie du moyen débit et pour minimiser les pertes de matériel, la digestion a été réalisée en présence simultanée des deux enzymes propres à chaque stratégie de clonage. Le problème qui se pose ici est alors la compatibilité des enzymes vis-à-vis du tampon réactionnel. Chaque enzyme fonctionne à son optimum pour une concentration saline et à un pH précis. Ces optima de tampon ne sont en fait pas totalement compatibles entre eux parmi les enzymes choisies. Le Tableau II-8 résume ces données (fournies par le revendeur) par rapport aux quatre tampons commerciaux disponibles (NEB 1 à 4 – New England Bioloabs).

	NEB1	NEB2	NEB3	NEB4
*Bam*HI	75	100	50	75
*Bgl*I	10	75	100	10
*Eco*RI	100	100	100	100
*Mfe*I	75	50	10	100

Tableau II-8 : Tampons de PCR. Présentation des activités relatives des endonucléases de restriction en fonction du tampon choisi (exprimées en pourcentage de l'activité maximum).

Dans le cadre d'une digestion en série, il aurait fallu procéder à un changement de tampon (i.e. procéder à la purification des cibles) entre chaque coupure impliquant des enzymes non compatibles en tampon. L'utilisation de tampons uniques, moyennant quelques compromis en rendement, a été préférée pour réaliser les coupures. Le tampon NEB2 a été choisi comme tampon réactionnel pour les couples *Bam*HI/*Eco*RI (100% / 100%) et *Bgl*II/*Mfe*I (75% / 50%), le tampon NEB3 pour le couple *Bgl*II/*Eco*RI (100% / 100%) et le tampon NEB4 pour le couple *Bam*HI/*Mfe*I (75% / 100%).

I.B.2.3 Résultats des transformations

Les transformations et la culture des cellules recombinantes ont été des étapes également critiques. Elles ont notamment nécessité plusieurs essais avant de se révéler probantes. Ceci a tenu en grande partie au fait qu'elles ont impliqué une augmentation drastique d'échelle, tant temporelle que matérielle.

Il aura fallu au final deux séries complètes de transformation pour obtenir les clones, plus quelques traitements au cas par cas. Les tableaux suivants présentent les résultats de transformation de la souche *E. coli* DH5 avec les plasmides pFO4 (Tableau II-9) et pGEX-4T-1 (Tableau II-10).

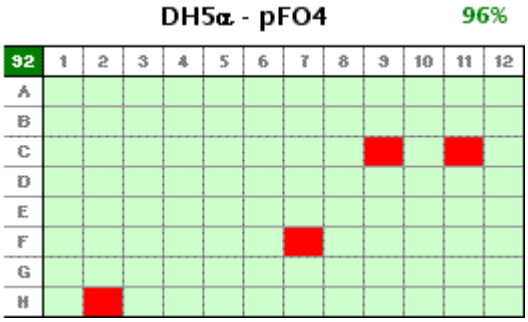


Tableau II-9 : Transformation en DH5α du plasmide pFO4.
Résultats de transformation de la souche *E. coli* DH5α par le plasmide pFO4.

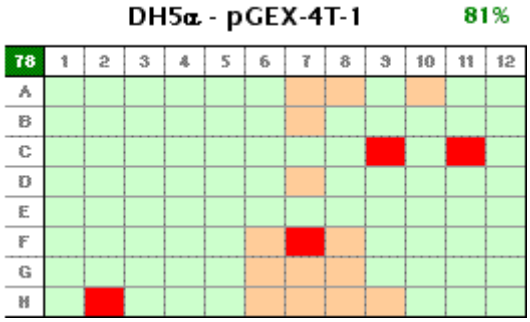


Tableau II-10 : Transformation en DH5α du plasmide pGEX-4T-1.
Résultats de transformation de la souche *E. coli* DH5α par le plasmide pGEX-4T-1.

Tableau II-9 et Tableau II-10 : Résultats de transformation de la souche *E. coli* DH5α par les plasmides. Vert: cibles transformées et validées par criblage PCR ; orange : cibles non transformées après deux tentatives ; rouge vif : témoins négatifs

Les transformations avec le plasmide pFO4 ont montré un rapide succès sur l'intégralité de la plaque (100% des gènes amplifiés). Les résultats avec le plasmide pGEX-4T-1 ont été un plus mitigés (85% des gènes amplifiés après la seconde série de transformation). Ceci m'a amené à considérer exclusivement les cibles en fusion pHis pour la suite du projet. Les gènes clonés dans le vecteur pGEX-4T-1 ont été réservés à éventuel « repêchage » ultérieur de protéines intéressantes, présentant une faible solubilité en étiquette histidine. De plus, ce choix m'a permis d'avoir un traitement uniforme des purifications des protéines, toutes se réalisant par une première étape de chromatographie d'affinité au nickel.

La dernière étape de transformation de la souche d'expression *E. coli* BL21(DE3) s'est réalisée, forte de l'expérience accumulée avec la souche DH5α. De bons rendements ont été obtenus (100% des cibles transformées en DH5α) puisque l'intégralité des plasmides transformés ont été retrouvés dans la souche d'expression *E. coli* BL21(DE3) (Tableau II-11).

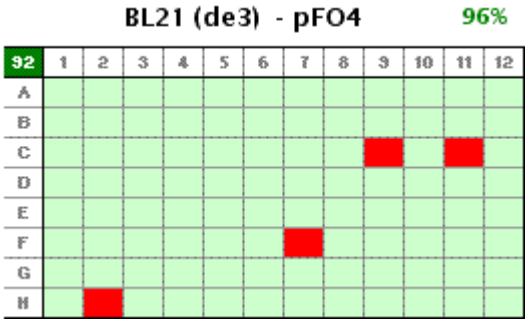


Tableau II-11 : Transformation BL21(DE3).
Résultat de transformation de la souche *E. coli* BL21 (DE3) par le plasmide pFO4.

I.B.2.4 Culture, expression et test de solubilité

Une des dernières contraintes imposée par la stratégie « moyen débit » a concerné tout particulièrement l'expression, étape critique s'il en est puisque toute la suite du projet dépendait de ses résultats. Une culture classique en milieu LB avec induction à l'IPTG n'aurait pas ici été satisfaisante, et ce pour plusieurs raisons. Tout d'abord, cela impliquait de manipuler la plaque au cours de l'expérience afin d'ajouter l'IPTG pour induire l'expression, multipliant les sources d'erreurs éventuelles ou de contamination. Deuxièmement, ce milieu ne garantit pas l'absence d'expression de fuite qui aurait pu biaiser la lecture des résultats. Enfin, les rendements d'expression en culture liquide étant limité à la capacité des bactéries à croître dans leur milieu, un milieu permettant une meilleure croissance bactérienne nous a semblé tout a fait indiqué pour permettre une meilleure lisibilité de l'expression et l'utilisation de bien plus petits volumes.

Le milieu ZYP-5052 (Studier, 2005) nous a semblé la solution adaptée. Ce milieu très riche permet d'atteindre des densités optiques de l'ordre de 25 contre 3 à 5 en milieu LB. Autrement dit, à volume égal, de quasiment décupler le rendement d'expression. Ce milieu a en outre la propriété très intéressante d'être « autoinductible ». De plus, il a été conçu pour inhiber l'expression de fuite, observés dans les systèmes dont l'induction est basée sur la présence d'IPTG. Cela a été réalisé par la présence de glycérol (0,5%), de glucose (0,05%) et de lactose (0,2%). Le glucose est le saccharide consommé en premier. Il permet tout à la fois un apport énergétique efficace stimulant la croissance initiale de la culture et l'inhibition de l'expression de fuite par répression catabolique. Lorsqu'il est entièrement consommé, les bactéries métabolisent le glycérol, qui assure une montée en phase exponentielle plus lente et plus contrôlée. Lorsque le glycérol est à son tour épuisé, le lactose prend le relais et permet l'induction des protéines recombinantes.

Les cultures ont en outre été réalisées à basse température (20°C), afin de faciliter le repliement des protéines exprimées.

Après trois jours de culture, les cellules sont lysées et les protéines sont déposées sur gels SDS-PAGE sous la forme de deux échantillons (un correspondant à l'extrait cellulaire total et l'autre à la fraction soluble de cet extrait), soit un nombre total de 192 échantillons. Les gels acceptant jusqu'à 10 dépôts, 8 extraits protéiques (et donc 4 échantillons) ont été analysés par gel, pour un total de 24 gels. Par soucis de clarté, l'ensemble des différents gels ne sera pas présenté. La Figure II-15 présente un gel qui donne une bonne idée des résultats et de l'effort à réaliser pour leur interprétation.

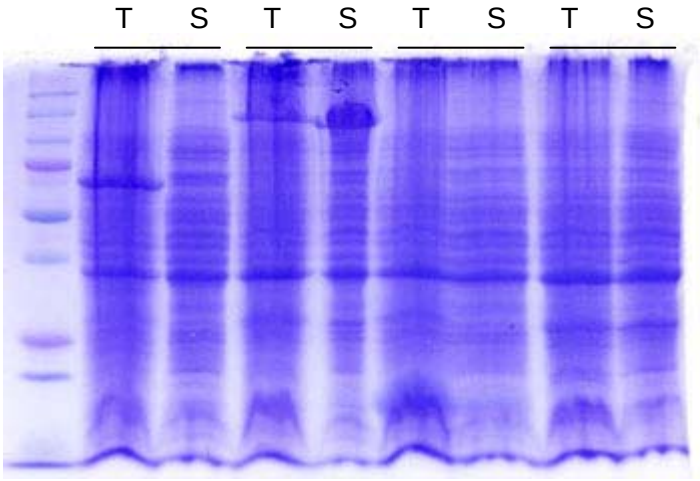


Figure II-15 : Exemple de gel SDS-PAGE issu de l'analyse à moyen débit.
Présentation du gel de la première moitié de la colonne 12 (de gauche à droite : D12, C12, B12, A12). Ce gel présente trois cas de figures d'expression. D12 : La protéine est présente dans la fraction T, mais pas dans la fraction S ; l'expression est bonne, mais elle est insoluble. C12 : l'expression est bonne et est présente dans la fraction S ; il s'agit donc d'une cible exprimée sous forme soluble. B12 et A12 : pas d'expression, ou trop peu pour être décelée.

Parallèlement aux analyses en gels de polyacrylamide, une expérience de dot-blot (Vincentelli *et al.*, 2005) a été réalisée afin de tester, par une autre méthode, le niveau d'expression. La Figure II-16 présente les deux membranes réalisées dans le cadre de cette expérience.

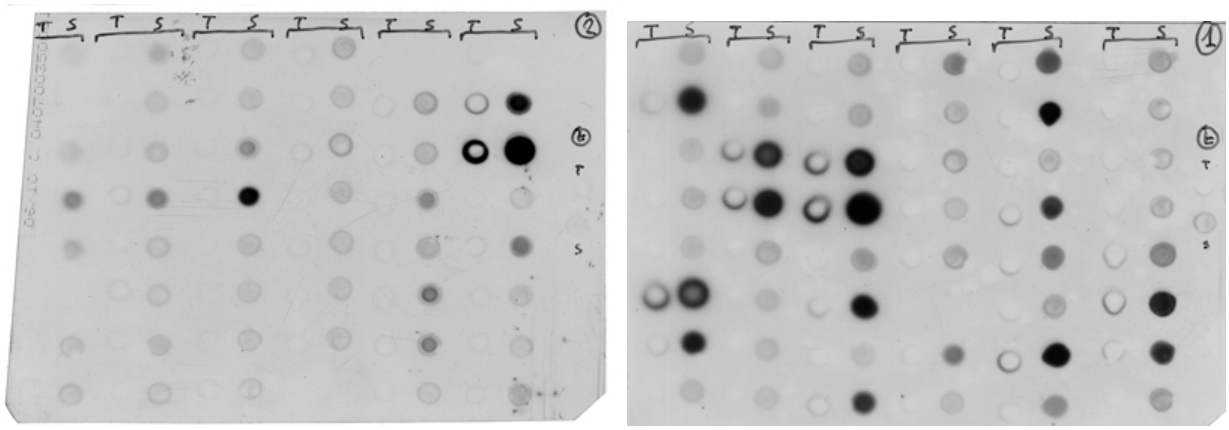
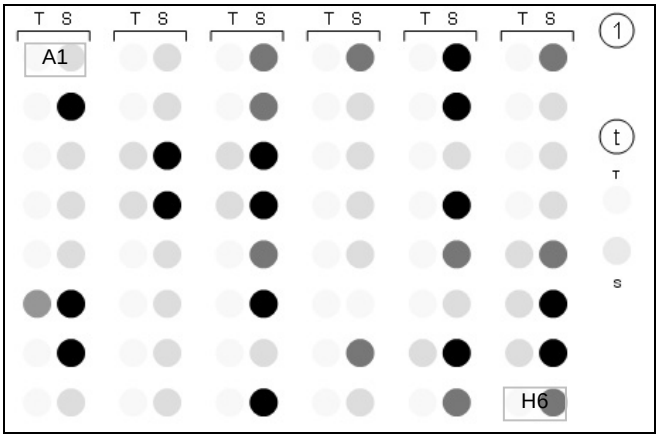


Figure II-16 : Membranes de Dot-Blot.
Présentation des membranes de Dot-Blot.



Représentation schématique de la membrane 1 de dot-blot.

Les cercles correspondent à un dépôt d'échantillon révélé sur le film photographique (un cercle noir correspond à une forte impression du film photographique, due à une forte concentration d'échantillon). Les colonnes T et S correspondent aux dépôts des fractions totale et soluble. La membrane 1 contient les échantillons A1 → H6. la membrane 2 contient les échantillons A7 → H12. Le « t » minuscule indique les dépôts d'échantillons témoins.

Ces deux méthodes ont été très complémentaires, de part leur sensibilité différentes et le type d'information qu'elles ont fourni.

Ainsi, les gels SDS-PAGE donnent une idée très claire du niveau d'expression. En revanche, leur lecture s'est avérée assez aléatoire, puisque globalement, chaque puit des gels était très chargé. Une bande peu intense à la taille attendue n'a ainsi pas nécessairement signifié une expression (même faible), tandis qu'une absence de bande intense à cet endroit a indiqué clairement une absence d'expression (ou tout du moins une expression très faible). Le dot-blot s'est, au contraire, révélé extrêmement sensible dans sa détection et les protéines ayant bénéficié d'une expression sous forme soluble, même moyenne, ont été plutôt bien visibles. Le biais technique du dot-blot, a été certaines surdétections, dues à des marquages non spécifiques des anticorps. Cela s'est particulièrement vu dans certains cas où les deux techniques donnaient des informations opposées, comme majoritairement une expression bonne en dot-blot et pas de bande visible

sur gel. Je pense que ce qui a péché à cette étape est l'absence de purification au niveau de la plaque. Cela m'aurait certainement permis d'être beaucoup plus catégorique dans l'estimation du niveau d'expression.

Finalement, les cibles présentant une convergence entre les deux techniques (le dot-blot a été particulièrement déterminant et a dirigé l'analyse) pour une expression allant de moyenne à très bonne, ont été au nombre de 32 (soit le tiers de la plaque). Ce nombre est raisonnablement bon étant donné les conditions expérimentales et le fait que la majorité des protéines sont des modules isolés du reste de leur protéine d'origine (et donc peu exprimables). Ce nombre correspond d'ailleurs aux valeurs des grands centres de génomique structurale (O'Toole *et al.*, 2004). Il m'a donc conforté dans les choix réalisés. Le Tableau II-12 reprend les meilleures expressions en protéines solubles (d'après les analyses des gels et des dot-blots).

L'**annexe 2** présente l'ensemble des cibles et de leurs rendements respectifs dans chaque test d'expression : SDS-PAGE total et soluble, et Dot-Blot total et soluble.

Catégorie 1		
F01	RB9136	UNK
B08	RB10416	UNK-2
G08	RB700	UNK-2
C10	RB10416	PA14-1
E10	RB12360	GH32
F10	*RB3006*	*UNK-1*
E11	*RB5312*	*PL1*
F11	RB2091	CE6
C12	RB700	ALL
G12	RB3601	PL7

Catégorie 2		
G01	RB9136	GH10
E04	RB8905	GT2
B07	RB8823	CE1
C07	RB9546	CE1
E08	*RB2160*	*GH57*
H08	RB3421	UNK
A10	RB11533	GT25
B12	*RB3006*	*GH33*

Catégorie 3		
D01	RB10416	UNK-1
B02	RB3123	ALL
C02	*RB3123*	*GH16*
E02	RB4561	UNK
F04	RB9623	GT2
E05	RB2990	UNK
E06	RB9614	GT32
A07	RB6145	CE1
D07	RB12316	CE4
H07	RB5256	GH5
A08	RB10416	CalxB
D08	RB3353	GH33
B09	RB13211	UNK
H09	RB11529	GT12

Tableau II-12 : Meilleure expressions solubles réparties en 3 catégories.
Catégorie 1 : expression soluble excellente (10 protéines)
Catégorie 2 : expression soluble bonne (8 protéines)
Catégorie 3 : expression soluble faible mais raisonnable (14 protéines)

En jaune sont marquées les cibles finalement retenues pour l'analyse cas par cas (voir ci-après)

Le choix des cibles retenues pour la poursuite de l'analyse s'est effectué en cherchant le meilleur compromis entre l'intérêt que nous avions pour le module et sa solubilité. Ainsi,

beaucoup d'enzymes présentant une excellente expression n'ont pas été traitées, car leur étude n'avait de sens que dans le cas où l'ensemble des modules de la protéine présentait une bonne expression. Je pense par exemple au module UNK de la protéine RB9136 qui n'avait d'intérêt que si le module catalytique de la protéine s'exprimait également correctement, ou bien encore aux modules de la protéine RB10416 (dont 4 montrent une expression très correcte) et qui ont été mis de côté pour les mêmes raisons.

Le consensus sur le choix des cibles s'est arrêté sur les protéines suivantes : RB3006 module GH33 et UNK-1, RB2160 module unique GH57 et RB5312 module unique PL1. Ci-après est présentée une courte introduction de ces différentes enzymes (l'explication précise de l'intérêt que j'ai porté à ces cibles sera détaillée dans leur chapitre respectif) :

- **RB3006 GH33 et UNK-1** : annotée comme « sialidase - EC 3.2.1.18». *R. baltica* possède sept sialidases de la famille GH33, assez divergentes entre elles d'une part et très divergentes par rapport aux orthologues de la famille GH33. Le module UNK-1 n'est retrouvé que chez trois bactéries (deux planctomycètes et une *Verrucomicrobia*) sous forme de protéine à part entière ;

- **RB2160 GH57** : annotée comme « α-amylase – EC 3.2.1.1». La famille GH57 est une famille de glycoside hydrolase dans laquelle des séquences issues exclusivement de bactéries ou d'archées sont présentes. RB2160 de *R. baltica* présente en outre un domaine additionnel que seul un sous-groupe de cette famille possède ;

- **RB5312 PL1** : annotée comme « pectate lyase - EC 4.2.2.2». Cette activité m'a parue pour le moins curieuse pour une bactérie marine. De plus, quatre activités directement liées à la dégradation de la pectine ont été prédites chez *R. baltica* dont deux appartenant à la famille PL1 : il s'agirait donc d'une source de carbone non négligeable pour elle ; de plus, cette enzyme est très divergente non seulement au sein de sa famille mais également par rapport à son paralogue ;

- **RB3123 GH16** : annotée comme « κ-carraghénase - EC 3.2.1.83 ». *R. baltica* possède deux enzymes de la famille GH16. Cependant, cette protéine est très divergente par rapport aux orthologues présents dans la famille GH16 (contrairement à son paralogue) et cette annotation semblait douteuse. Cette enzyme a présenté une faible expression soluble, mais étant donné la forte expérience de notre équipe dans l'étude des relations structure/fonction de la famille GH16, l'intégrer à la liste a semblé tout à fait approprié.

II - Matériels & Méthodes

II.A - Recensement des enzymes du métabolisme des sucres de *R. baltica*

II.A.1 - Identification des protéines et analyse de leur architecture modulaire

Les enzymes de *R. baltica* qui créent ou clivent les liaisons glycosidiques (glycosyltransférases, glycoside hydrolases et polysaccharide lyases) ou qui catalysent la désestérification des polysaccharides méthylés ou acétylés (carbohydrates esterases) ont été sélectionnées à l'aide de la banque de données CAZy (http://www.cazy.org) (Coutinho and Henrissat, 1999).

Les autres classes d'enzymes, impliquées dans le métabolisme des sucres, ont été identifiées par similitude de séquence avec des protéines biochimiquement caractérisées, elles-mêmes sélectionnées dans la base de données UniProtKB/Swiss-Prot (http://www.uniprot.org) (Bairoch *et al.*, 2004). Pour chaque protéine identifiée, une recherche de modules conservés a été effectuée contre la base de données PFAM (Bateman *et al.*, 2004). La présence éventuelle de modules orphelins additionnels a été détectée par des recherches utilisant BLAST contre la base UniProt.

II.A.2 - Délimitation fine des modules protéiques

Les limites les plus probables des modules ont été affinées manuellement en effectuant une analyse par la méthode HCA (Hydrophobic Cluster Analysis - (Gaboriaud *et al.*, 1987; Callebaut *et al.*, 1997). Les éventuels peptide signaux et les hélices transmembranaires ont été respectivement prédit en utilisant les outils SignalP v2.0 (Nielsen *et al.*, 1999) et TMHMM v2.0 (Sonnhammer *et al.*, 1998; Krogh *et al.*, 2001).

II.B - Clonage des gènes par une approche à moyen débit

II.B.1 - Purification de l'ADN génomique de *R. baltica*

L'ADN génomique de *R. baltica* (fourni par le laboratoire de Biologie Marine du Max Planck Institute) a été purifié par un gradient au chlorure de césium (Williamson, 1969). 1,633 g de CsCl ont été pesés et ajoutés à 500 µL d'une solution de 10 mM tris-HCl pH 8,0 ; 1 mM ethylene diamine tetraacetic acid (EDTA) (TE). La solution a été ajustée à 750 µg/mL de bromure d'éthidium (BET) et 200 µL de la solution d'ADN génomique (concentration estimée à 20 µg/mL par spectrométrie-UV à 260 nm) a été ajoutée. La solution a été complétée à un volume final de 2 mL. La densité finale obtenue a été de 1,6. Le tube a été scellé puis centrifugé à 90000 rpm et 20 °C pendant 24 h dans un rotor TLA100 (Beckman). La bande d'ADN formée a été prélevée en chambre noire sous UV avec une seringue.

Le BET a été extrait plusieurs fois à l'aide de butanol saturé en eau. L'ADN a alors été dilué dans 2 volumes de H_2O et précipité par 2 volumes d'éthanol absolu durant une nuit à -20 °C. Après précipitation, la solution a été centrifugée à 10000 ***g*** à 4 °C pendant 30 min. Le culot obtenu a été séché à l'air libre et à température ambiante, puis dissout dans 10µL de TE et stocké en chambre froide à 4 °C.

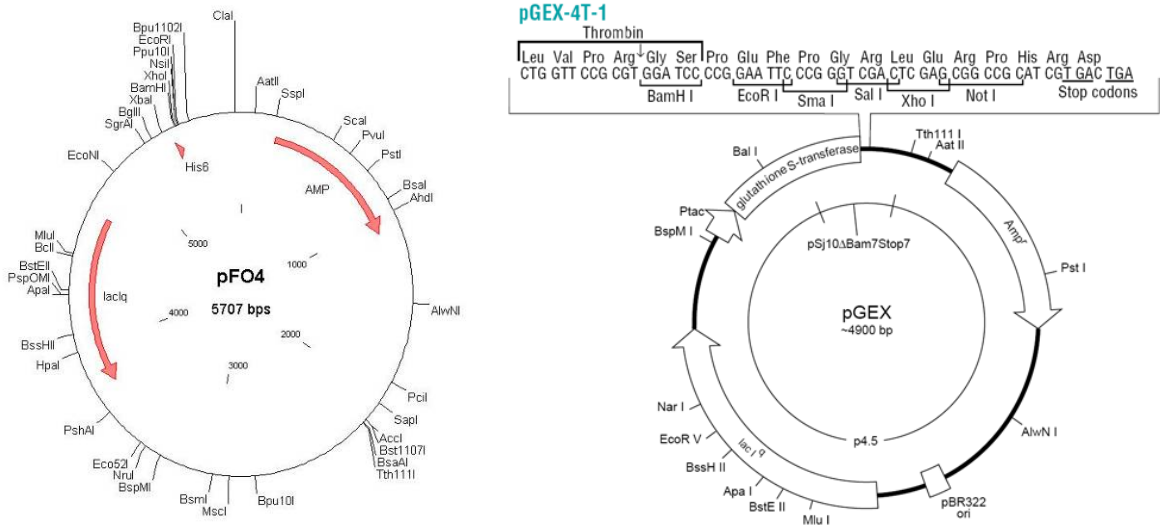
II.B.2 - Préparation des plasmides d'expression

Notre stratégie de clonage est basée sur la ligature d'un produit PCR unique pour chaque gène dans deux plasmides d'expression différents. Ainsi, notre choix s'est porté sur les plasmides pFO4 (don de Mr Robert Laroque, BRI, Montréal) et pGEX-4T-1 (GE HealthCare Life Science).

Le plasmide pFO4 (**Figure II-17**) est un dérivé du plasmide pET-15b (Novagen), le gène est ainsi sous contrôle du promoteur de l'ARN polymérase du phage T7. Ce vecteur nécessite l'utilisation d'une souche *Escherichia coli* lysogénisée par le phage λDE3 qui contient le gène de la T7 ARN polymérase sous le contrôle du promoteur *lac*, inductible à l'IPTG ou au lactose. Le plamside pET-15b a en outre été muté pour introduire les sites de restriction des enzymes *Bam*HI et *Eco*RI dans son site de clonage multiple. Le gène cloné dans le plasmide pFO4 résultant sera fusionné à une séquence codant pour un peptide hexahistidinyl (pHis) situé en 5' du gène.

Le plasmide pGEX-4T-1 (Figure II-17) contient le promoteur *tac*, également inductible à l'IPTG ou au lactose et peut-être utilisé avec une souche *E. coli* non lysogénisée. Il possède les sites des enzymes *Bam*HI et *Eco*RI dans son site de clonage multiple. Le gène cloné dans ce plamside sera fusionné en 5' à une séquence codant pour une Glutathion S-Transférase (GST).

Les plasmides ont été purifiés par un gradient au chlorure de césium. Après linéarisation par les enzymes de restriction *Bam*HI et *Eco*RI pendant 3 H à 37 °C, dans un tampon NEB2, les plasmides ont été déphophorylés à l'aide d'une phosphatase alcaline (Shrimp Alkaline Phosphatase, Promega) pendant 45 min à 37 °C. Les plasmides ont ensuite été précipités, repris dans l'eau et stockés à -20 °C.



carte du plasmide pFO4 carte du plasmide pGEX-4T-1

Figure II-17 : Présentation des cartes des plasmides pFO4 et pGEX-4T-1

II.B.3 - Dessin des amorces oligonucléotidiques

Pour chaque gène, les amorces oligonucléotidiques ont été dessinées selon le schéma suivant :

→ Amorces sens : 5'-gggggg-[site de restriction]-[séquence du gène]-3'

→ Amorces anti-sens : 5'-CCCCCC-[site de restriction]-stop-[séquence-complémentaire du gène]-3'

Les longueurs des séquences des gènes ont été déterminées afin d'obtenir une température de semi-hybridation $T_m|_1$ = 70 °C ou $T_m|_2$ = 72 °C. Par défaut, les sites de restriction des amorces sens et anti-sens ont été choisis pour être respectivement ceux de *Bam*HI et *Eco*RI. Cependant, en cas de présence de l'un de ces sites dans le gène, ils ont été remplacés par celui d'un de leur isocaudamère, respectivement *Bgl*II et *Mfe*I. En conséquence, les gènes sélectionnés ont été classé en quatre stratégies de clonage, selon leur compatibilité avec ces enzymes de restriction :

- Groupe 1 : compatible avec *Bam*HI/*Eco*RI ;
- Groupe 2 : compatible avec *Bam*HI/*Mfe*I ;
- Groupe 3 : compatible avec *Bgl*II/*Eco*RI ;
- Groupe 4 : compatible avec *Bgl*II/*Mfe*I.

Les gènes ne correspondant à aucun de ces groupes ont été écartés de l'analyse. Au final, 96 séquences ont été sélectionnées et les 192 amorces générées ont été synthétisées par Operon (Allemagne) et livrées sous forme de deux plaques de 96 puits à la concentration standard de 100 µM.

II.B.4 - Amplification des gènes par PCR

La réaction de PCR (Polymerase Chain Reaction), ainsi que **toutes les étapes suivantes ont été réalisées en plaque 96 puits.** Pour chaque réaction de PCR (*i.e.* pour chaque puit de la plaque) ont été mélangés :

- 20 ng d'ADN génomique purifié de *R. baltica* ;
- 10 mM de dNTP (dATP, dTTP, dCTP, dGTP) ;
- 1 unité d'ADN Polymérase Pfu (Proméga) ;
- 1 X du tampon commercial de la polymérase ;
- 10 µM de chaque amorce (sens et anti-sens) ;
- H_2O q.s.p. pour un volume final de 50 µL.

Les cycles de PCR ont été réalisés comme suit : 2 min à 94 °C, suivi de 30 cycles de 30 s à 94 °C (température de dissociation), 30 s à 50 °C (température d'hybridation) et 6 min à 72 °C (température d'élongation), pour finir par une dernière étape de 10 min à 72 °C. La plaque réalisée à partir de cette série de cycles a constitué ce que j'appellerais par la suite la « *plaque A* ». Une seconde PCR, référencée en tant que « *plaque B* » a également été réalisée avec une température d'hybridation de 60 °C. Les produits de PCR ont été analysés

par électrophorèse sur gel 0.8% agarose polymérisé en TAE (40 mM Tris-HCl pH 8,0 ; 20 mM acétate de sodium ; 1 mM EDTA) après coloration au BET.

Les réactions de PCR ayant échoué dans la plaque A ont été remplacées par leur équivalent réussi dans la plaque B. Les gènes pour lesquels les réactions de PCR des deux plaques ont échoué, ont été abandonnés dans la suite du projet. Toutes les étapes suivantes ont été réalisées à partir de la plaque A complétée si nécessaire.

II.B.5 - Préparation des produits PCR pour le clonage

Les produits PCR de la plaque A complétée ont été purifiés sur un système de traitement de plaques Qiaquick 96® (Qiagen) utilisant une pompe à vide et en suivant le protocole commercial.

Les fragments de PCR ont ensuite été digérés pendant 3 H à 37 °C par les couples d'endonucléases de restriction adaptés à chaque groupe dans un tampon commercial adapté aux enzymes : NEB2 pour *Bam*HI / *Eco*RI, NEB4 pour *Bam*HI / *Mfe*I, NEB3 pour *Bgl*II / *Eco*RI et NEB2 pour *Bgl*II / *Mfe*I. Après coupure, ils ont été repurifiés sur le système Qiaquick 96®. La plaque ainsi constituée de produits PCR digérés et purifiés a été stockée au froid (4 °C).

II.B.6 - Ligature dans les plasmides d'expression

Deux plaques de ligature de 96 puits ont été constituées : une plaque avec pFO4 et l'autre avec pGEX-4T-1. Dans chaque puit de ces plaques ont été ajoutés :

- 2 µL de produit PCR digéré ;
- 5 ng de vecteur linéarisé et déphophorylé ;
- 1 unité de T4 DNA ligase (New England Biolabs);
- 1 X de tampon de ligature commercial (New England Biolabs) ;
- H_2O q.s.p. pour un volume total de 10 µL.

Les plaques ont alors été laissées 12 H à température ambiante.

II.B.7 - Préparation des cellules compétentes

Les cellules compétentes ont été rendues compétentes par choc calcique à partir du protocole extrait de (Hanahan, 1983). Une préculture de 2,5 mL de milieu LB a été inoculée à partir de colonies sur boite de pétri et laissée 12h sous agitation 250 rpm à 37 °C. Un volume de 125 mL de LB supplémenté avec 20 mM $MgSO_4$ préalablement incubé à 37 °C a été alors inoculé avec 1,25 mL de la préculture. Lorsque l'absorbance à 600 nm de la culture bactérienne a atteint une valeur comprise entre 0,4 et 0,6, la solution a été centrifugée à 2000 ***g*** pendant 10 min à 4 °C. Le culot bactérien a été resuspendu dans 30 mL de tampon TFB1 (30 mM acétate de potassium ; 100 mM KCl ; 10 mM $CaCl_2$; 50 mM $MnCl_2$; 15% glycérol) préalablement stérilisé par filtration à 0,22 µm et refroidi dans la glace. Le mélange est laissé dans la glace pendant 10 min puis centrifugé à 2000 ***g*** pendant 10 min à 4 °C. Le culot bactérien est alors resuspendu dans 5 mL de tampon TFB2 (10 mM 3-(N-morpholino)-propanesulfonic acid (MOPS) pH 7,0 ; 75 mM $CaCl_2$; 10 mM KCl ; 15% glycérol) également stérilisé par filtration à 0,22 µm et préalablement refroidi dans la glace. Le mélange est laissé 30 min sur glace. Les cellules compétentes sont alors aliquotées en fraction de 100 µL et les tubes stockés au congélateur -80 °C.

II.B.8 - Transformation des plasmides dans la souche de stockage

Dans chaque puit des plaques de ligature ont été ajoutés 50 µL de cellules compétentes (cf 1-c) *E. coli* DH5α et la transformation des cellules à été réalisée par choc thermique : le mélange cellules - produit de ligature est incubé pendant 30 min à 0 °C ; il est ensuite exposé à une température de 42 °C pendant 45 s et réincubé pendant 10 min à 0 °C (Hanahan, 1983). Afin d'obtenir une bonne homogénéité de température dans les puits, les étapes à 0 °C ont été réalisées en plongeant les plaques dans un mélange eau/NaCl/glace ; l'étape à 42 °C a été réalisée en plongeant les plaques dans un bain marie. Les cellules ont ensuite été incubées dans du milieu SOC (2% bacto-tryptone ; 0,5% extrait de levure ; 0,05% NaCl ; 0,2% glucose) q.s.p. 2 mL sous agitation douce pendant 45 min à 37 °C.

II.B.9 - Criblage par PCR des transformants

Les cellules transformées (100 µL) ont ensuite été étalées sur des plaques Falcon (Becton Dickinson Labwares) de 6 boites de pétri contenant 6 mL d'un gel agar-LB-ampicilline (100 µg/mL) et les 32 plaques de boites de pétri ont été laissées en étuve à 37 °C pendant 12 H. Le succès du clonage a été testé par criblage PCR sur les colonies bactériennes obtenues. Le criblage PCR consiste à prélever une petite partie d'une colonie bactérienne à partir de laquelle une réaction de PCR est réalisée. Le mélange réactionnel du crible PCR est réalisé comme suit :

- quelques cellules bactériennes ;
- 1 unité de Taq DNA polymerase (Promega) ;
- 10 µM de la paire d'amorces spécifiques flanquant le site de clonage multiple de chaque vecteur (Tableau II-13) ;
- 10 mM de dNTP (dATP, dTTP, dCTP, dGTP) ;
- 1 X du tampon commercial de la polymérase ;
- H_2O Q.s.p. 20 µL

Vecteur	Amorce sens	Amorce anti-sens
pFO4	GCA GCA GCC ACCATC ACC ATC ACC	CCT TTC GGG CTT TGT TAG CAG CCG G
pGEX-4T-1	CCT CCA AAA TCG GAT CTG GTT CCG CG	CGA TGC GGC CGC TCG AGT CGA CCC G

Tableau II-13 : Amorces universelles des plasmides.
Amorces correspondants aux zones flanquant le site de clonage multiple des vecteurs pFO4 et pGEX-4T-1

Les cycles de réactions PCR sont alors : 2 min à 94 °C, suivi de 25 cycles de 30 s à 94 °C, 30 s à 50 °C et 2 min à 72 °C, et pour finir 10 min à 72 °C. Si le criblage par PCR s'est avéré infructueux, une extraction de plasmide est réalisée en sélectionnant cinq colonies bactériennes. En cas de nouvel échec, l'extraction des plasmides a été réalisée sur l'ensemble des colonies ayant poussées sur la boite de pétri.

Une extraction du plasmide a été réalisée sur chaque colonie en utilisant le kit d'extraction *MiniPrep SV purification kit* (Promega).

Des stocks de cellules en 20% glycérol et de plasmides purifiés ont été réalisés en cryoplaque et individuellement en cryotubes, et placés au congélateur à -80 °C.

II.B.10 - Transformation des plasmides dans les souches d'expression

Les souches *E. coli* BL21 et BL21(DE3) ont été utilisées pour être transformées respectivement par les plasmides recombinant pGEX-4T-1 et pFO4, en utilisant le protocole de transformation précédemment cité.

Des stocks de cellules en 20% glycérol ont été également réalisés en cryoplaque et individuellement en cryotubes, et placés au congélateur à -80 °C.

II.C - Tests d'expression protéique

II.C.1 - Caractérisation biophysique des niveaux d'expression protéique

II.C.1.1 Electrophorèse sur gel de polyacrylamide

Des gels de polyacrylamide (de 0,5 mm d'épaisseur) en conditions dénaturantes avec du dodécylsulfate de sodium (SDS) ont été polymérisés en utilisant un mélange acrylamide – bis-acrylamide (30:0.8) à la concentration standard de 12%. Les gels avaient 10 ou 15 puits de dépôt pour un volume maximum d'échantillon respectivement de 20 µL et 12 µL. Les échantillons ont été bouillis 5 min en présence de 30 mM tris-HCl pH 6,8 ; 1% SDS ; 5% β-mercaptoéthanol ; 10% glycérol et 0,05% bleu de bromophénol avant leur dépôt.

Les électrophorèses en conditions dénaturantes (SDS-PAGE) ont été réalisées à température ambiante et à ampérage constant de 25 mA par gel dans un tampon 20 mM tris-HCl ; 200 mM glycine ; 0.1% SDS. La séparation électrophorétique a été considérée comme réalisée lorsque le front de migration, matérialisé par la migration du bleu de bromophénol, a atteint la limite inférieure du gel (en environ 1 h).

Les échantillons ont été révélés par coloration du gel pendant 30 min dans une solution de bleu de Coomassie (20% éthanol ; 10% acide acétique ; 0,25% bleu de Coomassie) puis décoloration 1 h dans un mélange 20% éthanol, 10% acide acétique.

II.C.1.2 Transfert sur membrane de type « Dot-Blot »

Une membrane de nitrocellulose a été découpée aux dimensions de l'échantillon. Une grille à été dessinée au crayon sur la membrane pour s'assurer de la précision du dépôt. Il faut compter un point de dépôt tous les centimètres, ceux-ci ayant tendance à s'étaler. La membrane a été déposée dans une boite de pétri adaptée à sa taille et 2 µL de chaque échantillon a été déposé sur la membrane au centre de chaque point de la grille, en faisant attention à les déposer doucement (pour limiter leur étalement). La membrane é été laissée à l'air libre à température ambiante le temps de sécher (une quinzaine de minutes).

Les sites non spécifiques de fixation des anticorps ont été alors bloqués en incubant la membrane à température ambiante pendant 30 min, dans un mélange 5% BSA en tampon TBS-T (20 mM tris-HCl pH 7.5 ; 150 mM NaCl ; 0.05% Tween20). Noter qu'il est également possible d'utiliser de la caséine ou du lait en poudre dilués dans du TBS-T.

Le tampon de blocage a été enlevé, et un lavage rapide au TBS-T a été réalisé. La membrane a alors été incubée à température ambiante pendant 30 min, avec l'anticorps primaire dissous dans 0,1% BSA en TBS-T. La concentration typique utilisée dans le cas d'un anticorps purifié est de 0,1-10 µg/mL. Nous avons cependant utilisé des anticorps d'antisérum à la dilution 1:20000. La membrane a ensuite été lavée trois fois 5 min en TBS-T.

Elle a ensuite été incubée à température ambiante 30 min, avec l'anticorps secondaire (conjugué avec la peroxydase HRP), en tampon TBS-T, puis relavée trois fois dans un tampon TBS-T (15 min, 5 min, 5 min) et enfin une fois 5 min en TBS (20 mM tris-HCl pH 7.5 ; 150 mM NaCl).

Le réactif de luminescence électrochimique (ECL) nécessaire à la photoimpression du film radiosensible doit être préparé extemporanément (il s'agit d'un mélange 50:50 de deux solutions commerciales – ECL Western Blotting Analysis System, Amersham). Les solutions sont assez onéreuse et ne peuvent servir qu'une fois, il est donc conseillé de prévoir les quantités à utiliser. Un volume final de mélange de 1 mL/cm^2 de membrane est normalement amplement suffisant. Le mélange ECL a été déposé sur la membrane et incubé 1 min. L'excès de réactif doit être enlevé pour permettre une lecture homogène. La membrane a été ensuite recouverte avec un film plastique de type Saran et un film photosensible a été déposé sur la membrane en chambre noire. Le temps d'impression du film dépend beaucoup de l'échantillon. Il faut le laisser au moins 1 min et jusqu'à une quinzaine si le signal est faible. Une marque a été réalisée sur le film pour retrouver son sens de lecture après révélation.

La révélation a été réalisée en chambre noire. Le film impressionné a été trempé 10 min successivement dans trois bains : un bain de révélation, un bain de fixation et dans de l'eau. Un lavage à l'eau a été réalisé entre chaque trempage. Le film a ensuite été séché à l'air libre dans la chambre noire. Un soin particulier a été apporté à ne pas toucher le film les mains nues avant son séchage étant donné qu'il est très facile de lui laisser des marques indélébiles, qui peuvent compromettre son interprétation.

II.C.2 - Expression à moyen débit

Les souches d'expression recombinantes BL21 et BL21(DE3) ont été cultivées en plaque 96 puits (de contenance de 3 mL par puit) dans 2 mL de milieu autoinductible ZYP5052 (Studier, 2005), complété par 100 µg/mL d'ampicilline. L'analyse de la croissance des cultures bactériennes a été réalisée à 600 nm, sur quatre cibles réparties sur la plaque choisies au début de la culture, deux fois par jour afin de constater la sortie de la phase exponentielle de croissance pour stopper l'expression.

Les plaques ont alors été centrifugées à 700 *g* à 4 °C pendant 20 min et le surnageant enlevé. Dans chaque puit ont été ajoutés 1 mL de tampon de lyse BugBuster® Protein Extraction Reagent (Novagen) ; ceci a constitué « *l'extrait total* ». Cet extrait cellulaire total ainsi obtenu a été centrifugé à 14000 *g* à 4 °C pendant 30 min. Le surnageant de cette étape de centrifugation sera par la suite appelé « *l'extrait soluble* ».

Les résultats des différentes extractions (*total* et *soluble*) ont été systématiquement analysés par électrophorèse en gels de polyacrylamide 12%, en conditions réductrices de type SDS-PAGE, colorés au bleu de Coomassie. Les protéines issues des cultures en BL21(DE3), produites avec le plasmide pFO4, et présentant donc une étiquette poly-Histidine ont été analysées par la technique de Dot-Blot sur membrane de nitrocellulose, avec des anticorps de souris anti-pHis (GE Healthcare). Les protéines issues des cultures en BL21, produites avec le plasmide pGEX-4T-1, et présentant donc une fusion GST ont été quant à elles partiellement purifiées sur mini-colonnes de glutathion-sepharose, selon les recommandations du fabricant, afin de faciliter leur lecture en SDS-PAGE.

La Figure II-18 présente le schéma global de la stratégie d'analyse à partir des cultures bactériennes.

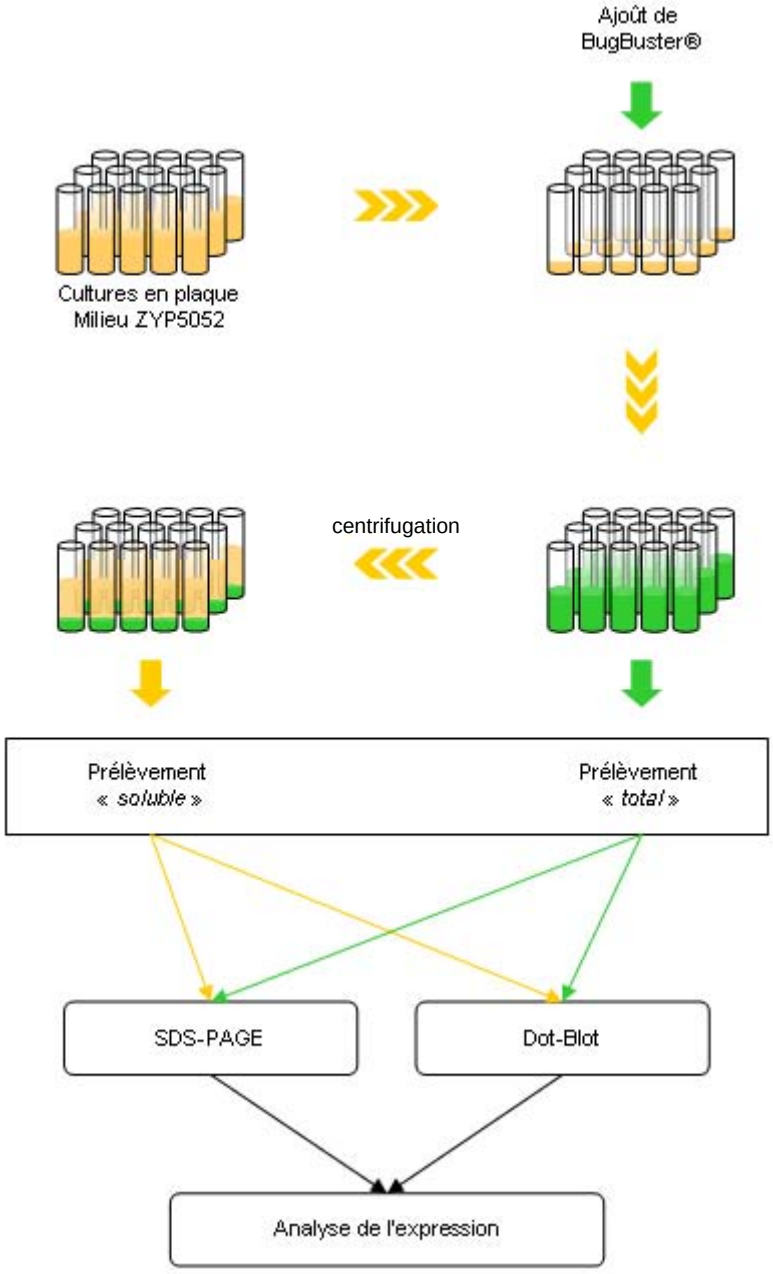


Figure II-18 : Schéma de principe de l'étude à moyen débit
Présentation du traitement à moyen débit des cibles pour évaluer leur degré d'expression soluble

Chapitre III

-

Caractérisation fonctionnelle et structurale de polysaccharidases originales de *Rhodopirellula baltica*

I - RB3123 : Une nouvelle glycoside hydrolase de la famille GH16 ?

I.A - La famille GH16

Les hydrolases forment l'immense majorité des enzymes de dégradation de saccharides. Leurs mécanismes catalytiques impliquent un ou deux déplacements de groupes nucléophiles selon leur mode d'action, via des intermédiaires réactionnels cationiques (voir Figure III-19). Elles peuvent catalyser l'hydrolyse de toute liaison glysosidique (Koshland, 1953; Sinnott, 1990; Davies and Henrissat, 1995).

Leur mode d'action peut être divisé en deux catégories : un mode qui libère après hydrolyse des saccharides présentant un carbone réducteur dans la même configuration anomérique que la liaison hydrolysée ; et un mode qui inverse la configuration anomérique. Ces deux modes d'actions sont respectivement appelés *mécanisme de rétention* et *mécanisme d'inversion* de la configuration anomérique.

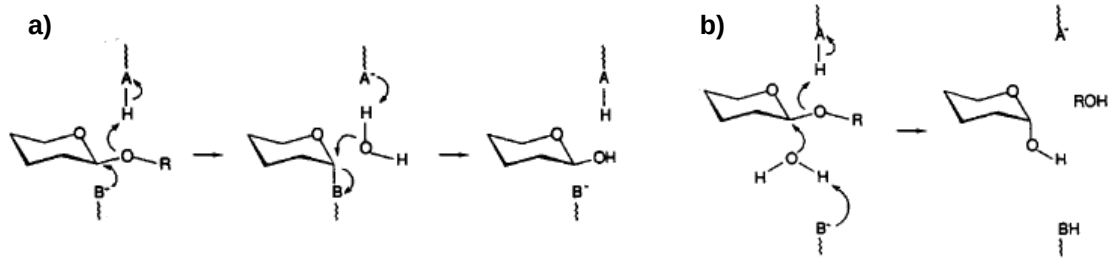


Figure III-19 : Modes d'action des glycoside hydrolases.

Présentation des modes d'action des hydrolases tel que décrit par Koshland en 1953. Figure extraite de (Davies and Henrissat, 1995). a) Mécanisme de rétention de la configuration anomérique: l'oxygène glycosidique est protoné par un catalyseur acide et le départ de l'aglycone est assisté par une catalyseur nucléophile. La liaison du glycosylenzyme formé est ensuite hydrolysée par une molécule d'eau et cette seconde substitution nucléophile génère un produit avec la même stéréochimie que le substrat initial. b) Mécanisme d'inversion de la configuration : la protonation de l'oxygène glycosidique est ici concomitante à l'attaque nucléophile d'une molécule d'eau stabilisée par la base. La stéréochimie du produit est ici inversée.

La famille 16 des glycoside hydrolases est une famille multispécifique comprenant actuellement environ 900 séquences. Huit activités enzymatiques y sont actuellement

recensées (Tableau III-14) dont sept ont en commun de dégrader des β-D-glucanes ou des β-D-galactanes. La dernière classe d'enzymes catalyse la transglycolyse des xyloglucanes (activité xyloglucan:xyloglucosyltransferase, XET).

Activité	Numéro EC	Nombre de structures (et codes PDB publiés)
Endo-1,3(4)-β-D-glucanase	3.2.1.6	1 (1)
Laminarinase (endo-1,3-β-D-glucanase)	3.2.1.39	1 (1)
Lichenase (endo-1,3-1,4-β-D-glucanase)	3.2.1.73	3 (13)
β-agarase	3.2.1.81	2 (3)
κ-carraghénase	3.2.1.83	1 (1)
Keratan-sulfate endo-1,4-β-galactosidase	3.2.1.103	-
Xyloglucanase	3.2.1.151	1 (3)
Xyloglucan:xyloglucosyltransferase	2.4.1.207	2 (5)

Tableau III-14 : Activités de la famille GH16.
Récapitulatif des activités et structures retrouvées dans la famille GH16.

D'un point de vue distribution, ces enzymes sont présentes dans tous les phyla (bactéries, archées, eucaryotes et virus) et, même si le milieu terrestre est très représenté notamment à travers l'activité XET, certains des polysaccharides-substrats de cette famille sont retrouvés en abondance dans les océans. Une courte présentation de l'ensemble des substrats rencontrés dans cette famille est proposée ci après.

Deux types de glucanes servent de substrats aux enzymes de la famille GH16 : les β-(1,3)-D-glucanes et les β-(1,3)-(1,4)-D-glucanes (également appelés β-D-glucanes à liaisons variées ou « mixed-linked » β-D-glucanes - MLG). Ces glucanes constituent une famille assez hétérogène de polysaccharides retrouvés dans de nombreux organismes avec des fonctions biologiques très différentes. Ainsi, nous pouvons citer parmi eux la **laminarine** (Figure III-20), qui est un β-(1,3)-D-glucane à chaînes courtes faiblement branchées en β(1,6). Elle présente la particularité d'être le principal polysaccharide de réserve des macroalgues brunes, à l'encontre de la majorité des autres organismes photosynthétiques qui utilisent principalement l'amidon. D'autres types de β-(1,3)-D-glucanes présentant des structures légèrement différentes existent : le pachyman (polysaccharide de la paroi des champignons), le paramylon (polysaccharide de réserve des euglènes) ou encore le curdlane qui est un β-(1,3)-D-glucane bactérien non ramifié. Les β-(1,3)-(1,4)-D-glucanes

sont des polysaccharides linéaires présentant une alternance entre des liaisons β-1,3 et β-1,4. Le **lichenane** (Figure III-21) est un MLG présent chez certains champignons lichénisés dont il est un des constituants pariétaux. Chez les végétaux supérieurs de l'ordre des *Poales*, qui contient notamment les céréales, des β-(1,3)-(1,4)-D-glucanes forment des hémicelluloses de la paroi cellulaire. Ces polysaccharides sont alors non branchés et sont formés de 90 % d'unités cellotriose et cellotétraose connectées par des liaisons β(1,3) avec un ratio 2:1.

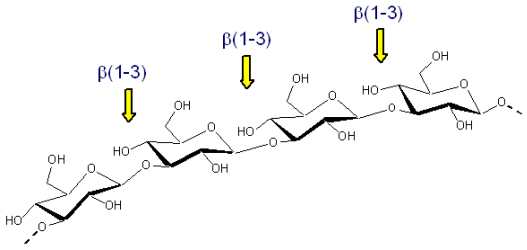


Figure III-20 : Présentation de la laminarine
Présentation de la laminarine β-1,3-D-glucane à chaînes courtes, faiblement branché en β-1,6.

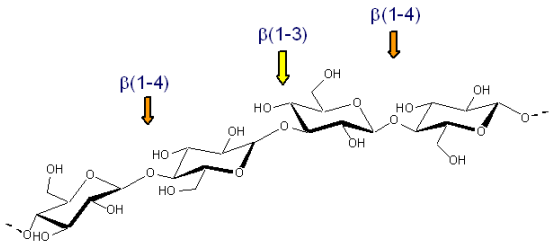


Figure III-21 : Présentation du lichenane.
Présentation du lichenane oligoside Glc-β-1,4-Glc-β-1,3-Glc.

La multitude d'activités enzymatiques, répertoriées dans l'hydrolyse de ces polysaccharides, reflète bien cette diversité. En effet, pour le cas de la famille GH16, les activités 3.2.1.6 (clivage des liaisons β(1,3) dans les β-(1,3)-D-glucanes et des liaisons β(1,4) dans les β-(1,3)-(1,4)-D-glucanes si l'ose non réducteur est substitué en C_3), 3.2.1.39 (clivage des liaisons β(1,3) dans les β(1,3)D-glucanes) et 3.2.1.73 (clivage des liaisons β(1,4) dans les β-(1,3)-(1,4)-D-glucanes) sont toutes liées à l'activité β-D-glucanase. Cependant chacune présente une spécificité qui rend ces enzymes souvent inactives sur les autres substrats de ce type. Elles sont retrouvées assez diversement dans la nature, tant chez les procaryotes et certains eucaryotes (champignons, nématodes, plantes, algues brunes, …) que chez les virus.

Les **agars** sont des galactanes sulfatés présents dans la paroi d'algues rouges, appelées agarophytes, où ils composent la phase amorphe de leur paroi, la phase cristalline étant composée de β-glucanes de type cellulose (Kloareg and Quatrano, 1988). Les agars sont des polysaccharides linéaires constitués de D-galactose et de 3,6-anhydro-L-galactose alternativement liés par des liaisons β(1,4) et α(1,3). Ces unités disaccharidiques peuvent être substituées par des groupements ester-sulfates, méthyles ou encore pyruvates. L'agarose correspond à des agars neutres (Figure III-22). Il est en particulier connu pour ses propriétés hautement gélifiantes (Armisen, 1991; de Reviers, 2002). La famille GH16 ne

contient que des agarases spécifiques de la liaison β(1,4) des agars (d'où leur nom de β-agarases). Elles n'ont été retrouvées jusqu'à présent que dans des espèces bactériennes.

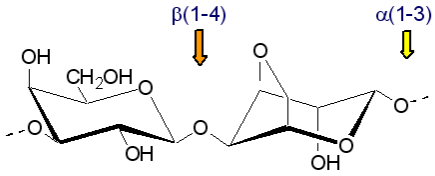


Figure III-22 : Présentation de l'agarose
Présentation de l'agarose oligodioside idéal α-1,3-D-Gal-β-1,4-L-(3,6-anhydro)Gal.

Les **carraghénanes** forment une large famille de galactanes plus ou moins sulfatés (jusqu'à 30 % de sulfatation), présents das la paroi de certaines algues rouges appelées les carraghénophytes, comme les algues appartenant aux genres *Chondrus, Gigartina, Eucheuma, Hypnea, kappaphycus* où ils composent, tout comme l'agar chez les agarophytes, la phase amorphe de la paroi de l'algue. La différence entre les carraghénanes et les agars, réside dans le fait que le 3,6-anhydro-galactose est en conformation D dans les carraghénanes. L'unité disaccharidique idéale du κ-carraghénanes est le α-(1,3)-D-galactose-4-sulfate-β-(1,4)-3,6-anhydro-D-galactose (Figure III-23). les carraghénanes sont également connus pour leurs propriétés gélifiantes (Kloareg and Quatrano, 1988). L'activité κ-carraghénase est peu représentée au sein de la famille GH16 avec seulement 3 enzymes, toutes de bactéries marines.

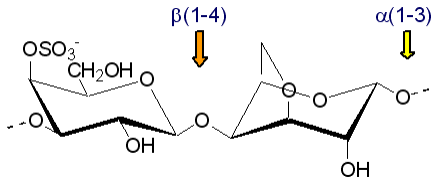


Figure III-23 : Présentation du κ-carraghénane
Présentation du κ-carraghénane oligodioside idéal α-1,3-D-(4-Sulfo)Gal-β-1,4-D-(3,6-anhydro)Gal.

Le **kératane sulfate** (Figure III-24) est un glycoaminoglycane complexe de la matrice extracellulaire des animaux. C'est un polymère linéaire de lactosamine sulfaté dont la structure idéale est le disaccharide β-(1,3)-galactose-β-(1,4)-N-acétyl-glucosamine-6-sulfate. Plus de quinze protéines différentes ont été identifiées dans la fraction peptidique qui lui est liée (Funderburgh, 2000). Sa très grande représentation dans l'ensemble des tissus des animaux a justifié les nombreuses recherches, notamment en médecine, depuis les années 1970 (Funderburgh, 2000).

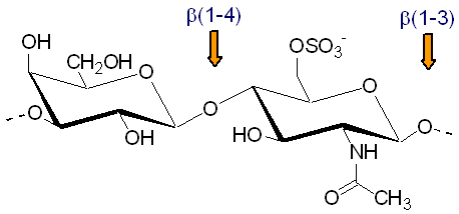


Figure III-24 : Présentation du kératane sulfate.
Présentation du kératane sulfate oligodioside idéal β-1,3-Gal-β-1,4-(6-sulfo)GlcNac.

Enfin, les **xyloglucanes** (Figure III-25) sont parmi les constituants majeurs de la paroi primaire des plantes supérieures (Carpita and McCann, 2000). Ils sont constitués d'un squelette β-1,4-D-glucane régulièrement substitué avec des résidus α-1,6-D-xylopyranosyl (Baumann *et al.*, 2007).

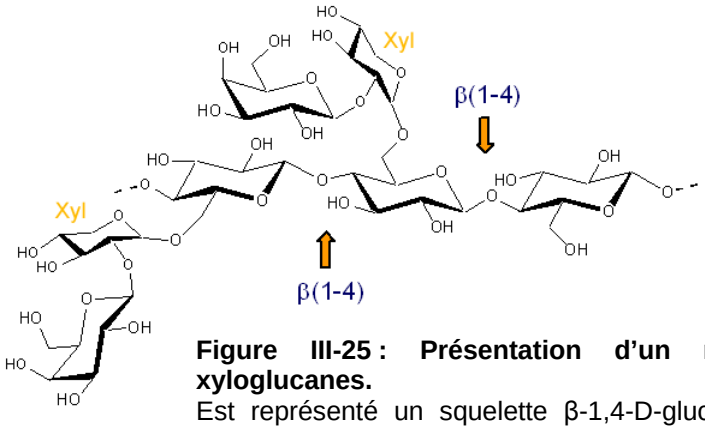


Figure III-25 : Présentation d'un motif trouvé dans les xyloglucanes.
Est représenté un squelette β-1,4-D-glucane substitué avec des résidus α-1,6-D-xylose (Xyl).

La variété et l'aspect ubiquitaire des substrats des enzymes de la famille GH16 ont fortement contribué à l'essor des études sur cette famille. La famille GH16 est ainsi structuralement étudiée depuis de nombreuses années. La première structure publiée date de 1993, avec la β-glucanase de *Paenibacillus macerans* (Keitel *et al.*, 1993). Les structures de dix autres enzymes de la famille GH16 ont été caractérisées depuis, dont quatre par notre équipe : la κ-carraghénase *pc*Kcar de *Pseudoalteromonas carrageenovora* (Michel *et al.*, 2001), les β-agarases AgaA et AgaB de *Zobellia galactanivorans* (Allouch *et al.*, 2003 ; Allouch *et al.*, 2004) et la xyloglucanase de *Tropaleolum majus* (capucine) (Baumann *et al.*, 2007). Le repliement est constitué d'une architecture de type sandwich β, et plus précisément du sous-type β-jelly roll. Les Figure III-26 et Figure III-27 présentent les structures de la κ-carraghénase de *P. carrageenovora* et de la xyloglucanase *tm*XGH de *T. majus*.

a) b)

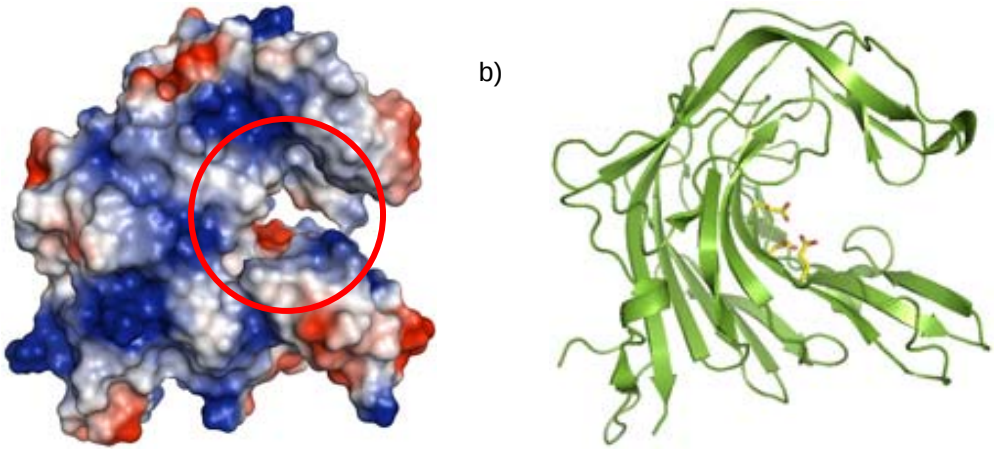


Figure III-26 : Structure de la κ-carraghénase *P. carrageenovora*.
Structure de la κ-carraghénase *pc*Kcar de *P. carrageenovora* (code PDB : 1DYP).
a) Présentation de la distribution des charges de surface. Le site actif (cercle rouge) est encapsulé dans un tunnel. b) Présentation des éléments de structure secondaire avec mise en évidence des résidus catalytiques.

a) b)

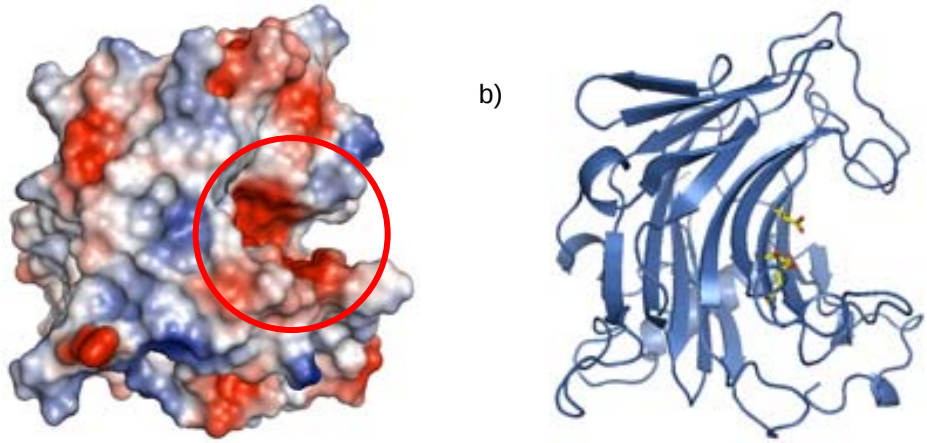
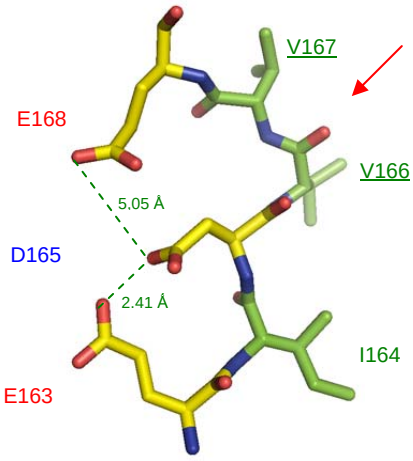


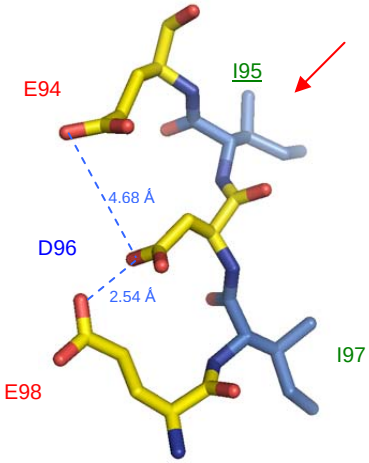
Figure III-27 : Structure de la xyloglucanase de *T. majus*.
Structure de la xyloglucanase *tm*XGH de *T. majus* (code PDB : 2UWA).
a) Présentation de la distribution des charges de surface. Le site actif (cercle rouge) est exposé vers l'extérieur via une gorge. b) Présentation des éléments de structure secondaire avec mise en évidence des résidus catalytiques.

Des études de mutagenèse dirigée sur les lichenases ont permis de déterminer la nature et la position des acides aminés catalytiques (Malet *et al.*, 1993), qui sont strictement conservés dans la famille GH16. Dans le cas de *pc*Kcar, l'attaque nucléophile est ainsi réalisée par un acide glutamique en position 163 (**E163**), tandis que le donneur de proton a été identifié comme étant un acide glutamique en position 168 (**E168**). Un acide aspartique en position 165 (**D165**) contribue en outre à l'activation de E163 par une liaison hydrogène. Le site actif de l'ensemble des enzymes de cette famille est ainsi constitué d'une des séquences conservées **E**X**D**X**E** ou **E**X**D**XX**E**, où X est un aminoacide hydrophobe (de type valine, isoleucine, méthionine, …). Il apparaît que la différence entre les deux motifs catalytiques est la présence dans le second d'un acide aminé supplémentaire pointant vers le cœur hydrophobe de la protéine entre l'acide aspartique D165 et l'acide glutamique E168,

entraînant l'apparition d'un renflement β (β-bulge en anglais) (Figure III-28). Sur la base d'arguments phylogénétiques et structuraux, il a été proposé que les enzymes de la famille GH16 présentant les résidus catalytiques dans un brin β régulier auraient divergé à partir d'un ancêtre commun présentant un site catalytique avec un renflement β (Michel *et al.*, 2001). Enfin, il a également été démontré que le mécanisme catalytique de la réaction libérait des produits en rétention de la configuration anomérique (Keitel *et al.*, 1993; Juncosa *et al.*, 1994).



Zoom sur le site actif de *pc*Kcar, constitué de la séquence **E**I**D**V**V**E contenant le renflement β occasionné par la présence des deux valines V166 et V167.



Zoom sur le site actif de *tm*XGH, constitué de la séquence **E**I**D**I**E** qui ne présente pas de renflement β.

Figure III-28 : Présentation des résidus catalytiques la famille GH16.
Noter la conservation des distances entre les trois résidus catalytiques **E**<--->**D**<--->**E**

Un dernier point concerne le contenu en résidus des différents sous-sites du site actif de ces enzymes. En effet, si l'on se base sur les différences entre les deux substrats finalement assez similaires que sont l'agarose et le κ-carraghénane, il apparaît que leur liaison glycosidique β-(1,4) (qui est celle clivée au cours de l'hydrolyse) présente un environnement chimique radicalement différent (Figure III-29).

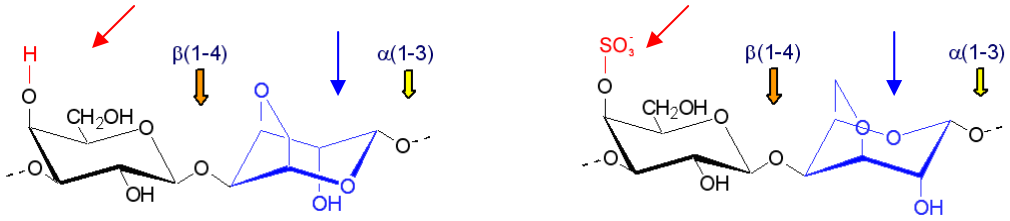


Figure III-29 : Comparaison de l'agarose et du κ-carraghénane
Mise en évidence des différences de substitution (en rouge) et de configuration (en bleu) entre l'agarose et le κ-carraghénane.

Ces différences sont à l'origine de remaniements majeurs dans les sites de fixation du substrat. En effet, la position des résidus catalytiques étant strictement conservée, la liaison hydrolysée des substrats doit être stabilisée dans des positions identiques. La grande variété de substrat de la famille GH16 implique que les résidus des sous-sites -1 et +1 soient particulièrement critiques pour le maintien de l'activité de ces enzymes. A titre d'exemple, le sulfate du résidu galactose-4-sulfate du κ-carraghénane impose aux κ-carraghénases non seulement un site -1 plus ouvert que celui des agarases, mais également capable de maintenir une charge négative. Ainsi Michel et al. (Michel *et al.*, 2001) ont proposé l'importance du résidu arginine R260 de *pc*Kcar dans l'interaction avec le groupement sulfate. Cette hypothèse a par ailleurs été récemment confirmée par l'obtention de la structure de *pc*Kcar en complexe avec un tétrasaccharide de κ-carraghénane (Czjzek M., communication personnelle). Concernant le sous-site +1, l'inversion de configuration du second résidu de l'oligosaccharide idéal entre l'agarose et le κ-carraghénane impose, quant à elle, que ce sous-site soit radicalement différent, étant donné que les zones hydrophobes et polaires du résidu 3,6-anhydro-galactose sont inversées.

La famille GH16 offre ainsi un large spectre d'activités enzymatiques sur des substrats dont beaucoup sont retrouvés dans le milieu marin. Elle constituait donc une cible très intéressante dans le cadre de l'étude du métabolisme global des carbohydrates d'une bactérie marine présentant un fort potentiel de dégradation de ce type de substrat.

I.B - Résultats des analyses

I.B.1 - Analyse bioinformatique

R. baltica présente deux séquences appartenant à la famille GH16, représentées dans le Tableau III-15.

Protéine	Modularité	Taille résidus	Taille kDa	Annotation initiale
RB3123		432	49	Probable glycosyl hydrolases-putative kappa-carrageenase (EC 3.2.1.-)
RB2702		307	35	Kappa-carrageenase [precursor] (EC 3.2.1.83)

Tableau III-15 : Protéines membres de la famille GH16 chez *R. baltica*.
Les peptides signaux sont représentés par un module grisé.

Les séquences de ces deux protéines ont été alignées avec les séquences des modules catalytiques GH16 de Q9RGX9_9FLAO (β-agarase A de *Z. galactanivorans*), Q874E3_PHACH (laminarinase de *Phanerochaete chrysosporium*), O84907_9FLAO (κ-carraghénase de *Z. galactanivorans*) et O73951_PYRFU (laminarinase de *Pyrococcus furiosus*), soit deux séquences bactériennes, une séquence eucaryote et une séquence archéenne. Cet alignement fait clairement apparaître, d'une part que les deux séquences de *R. baltica* possèdent les acides aminés catalytiques, et d'autre part qu'elles présentent un renflement β au sein de leur site actif (Figure III-30).

Les domaines catalytiques des deux paralogues de *R. baltica* présentent une faible similitude entre eux avec 25 % d'identité pour 40 % de similitude. RB2702 présente cependant une forte similitude avec les κ-carraghénases de *P. carrageenovora* (avec 46 % d'identité) et *Z. galactanivorans* (avec 43 % d'identité). Elle possède de plus l'arginine conservée du sous-site -1 R260 de *pc*Kcar (R269 chez RB2702) proposée pour interagir avec le groupement sulfate du galactose-4-sulfate du κ-carraghénane. RB3123 n'a elle qu'une faible similitude avec les autres membres de la famille GH16. Elle présente une identité à hauteur de 25 % avec la plupart de ses homologues, les plus proches (autour de 35 %) n'étant pas caractérisés à l'heure actuelle.

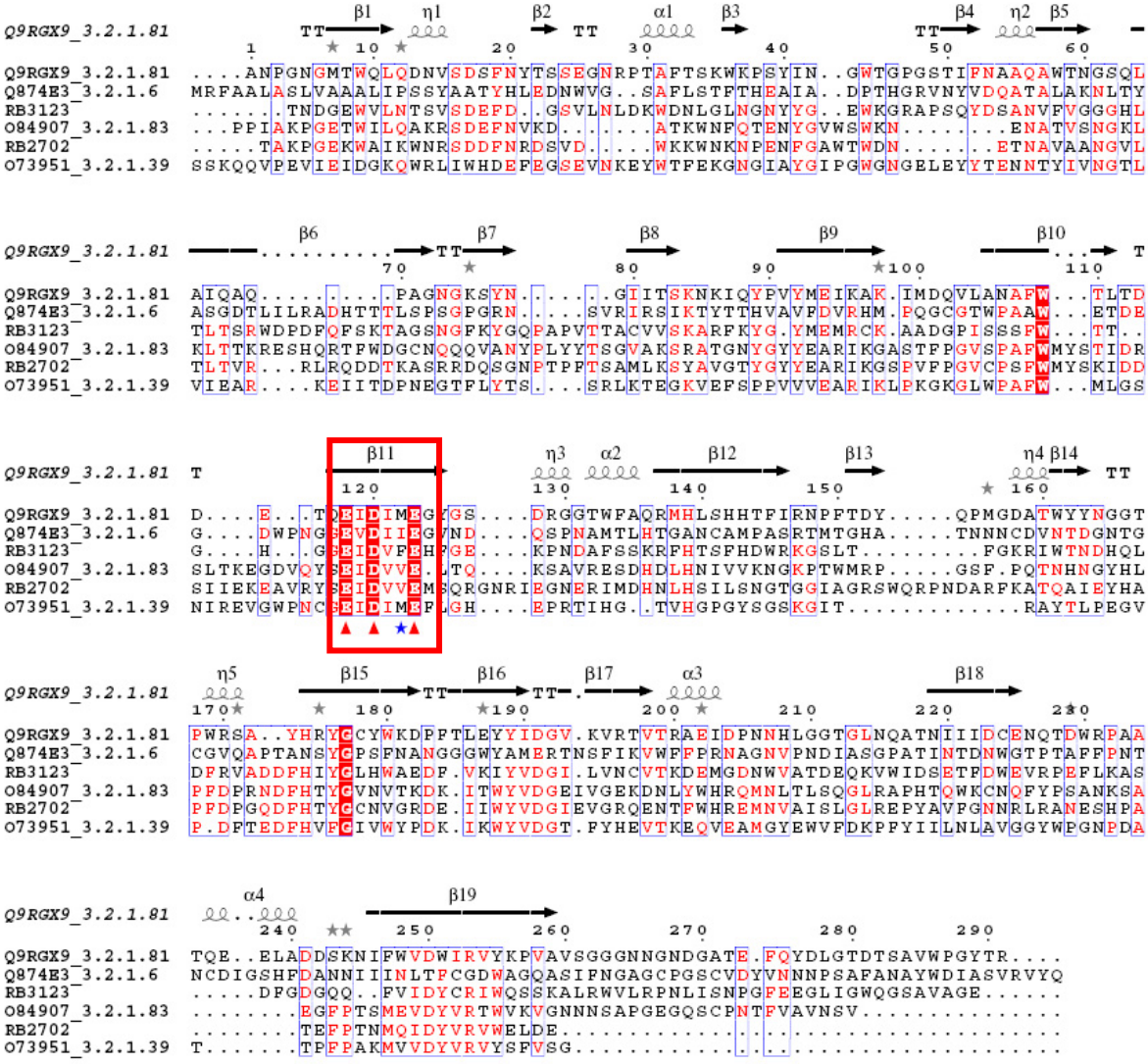


Figure III-30 : Alignement de séquences dans la famille GH16.
Alignement de séquences représentatives des domaines catalytiques de la famille GH16 présentant un renflement β (étoile bleue ★) avec les deux représentant de *R. baltica*. Noter que ces deux séquences présentent également un renflement β. Les acides aminés catalytiques au sein du site actif (carré rouge) sont marqués par des flèches rouges ▲. Figure extraite de ESPript après alignement avec Multalign.

Une analyse phylogénique a été menée sur l'ensemble de la famille 16. L'arbre phylogénique montre que ces deux séquences, au sein de la famille GH16, appartiennent à des sous-groupes différents (Figure III-31). RB2702 se trouve au centre d'un sous-groupe robuste (valeur de nœud de 84) contenant exclusivement les κ-carraghénases. RB3123 ne se retrouve dans aucune branche particulière et semble bien s'exclure des autres séquences de la famille.

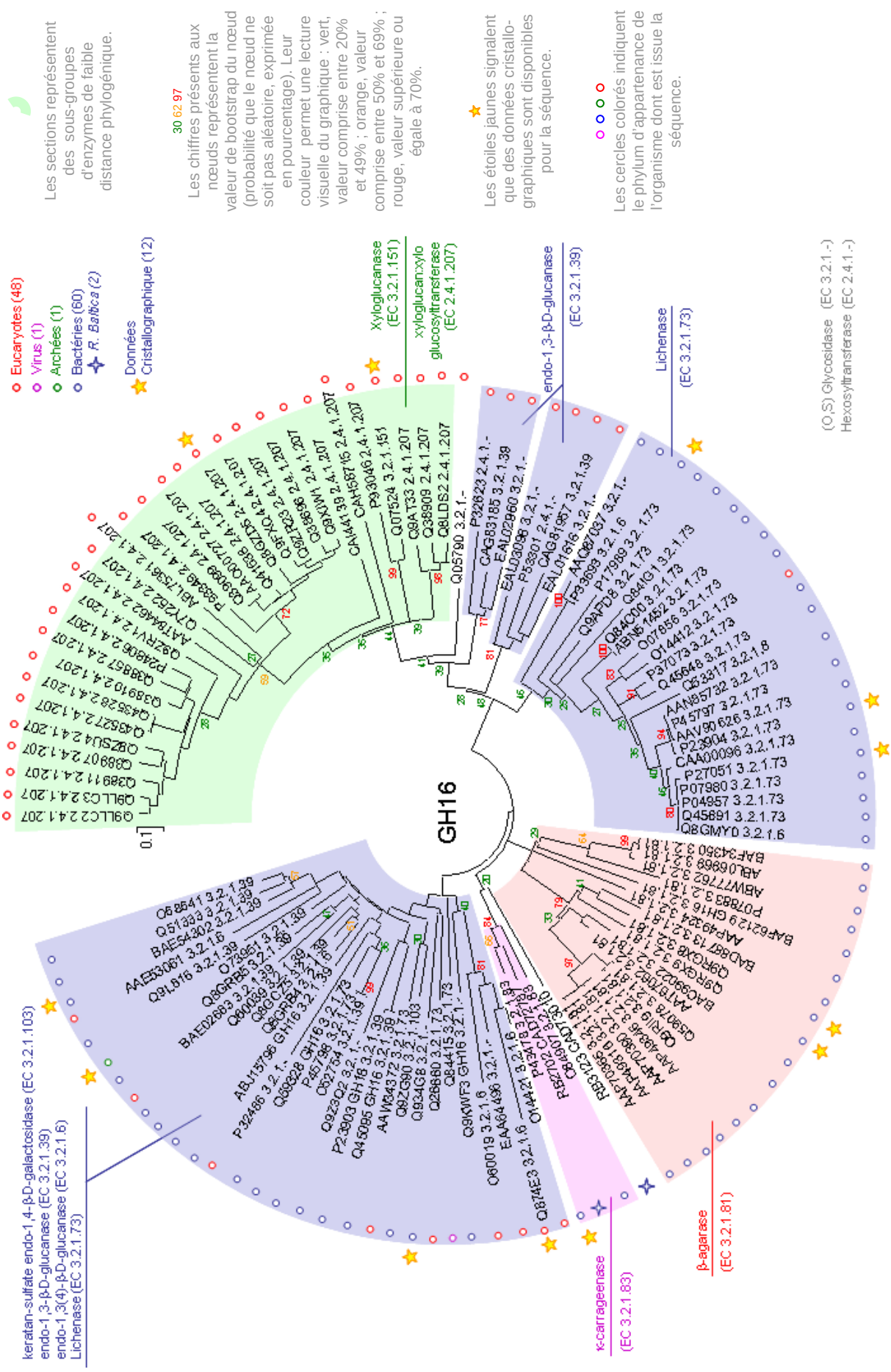


Figure III-31 : Arbre phylogénique de la famille GH16

Arbre phylogénique généré par la méthode d'évolution minimum, avec délétion des indels., à partir d'un alignement de type FFT-NS-i par MAFFT.

De plus, si les homologues du second module identifié de RB3123, un CBM de la famille 16, sont connus pour lier la cellulose et les glucomannanes (Bae *et al.*, 2008), la faible similitude entre eux et le module de *R. baltica* (30 % d'identité en moyenne vis à vis de la quasi-totalité des séquences de la famille CBM16) ne permet pas de conclure sur son motif de reconnaissance.

C'est ainsi que RB3123 a retenu notre attention dès la sélection des protéines. L'équipe travaille sur les enzymes de cette famille depuis longtemps et le laboratoire en possède quasiment tous les substrats connus. Même s'il s'est avéré, d'après les tests d'expression à moyen débit, qu'elle était faiblement produite sous forme soluble, j'ai décidé de l'inclure dans les protéines destinées à être caractérisées. D'une part, pour démontrer l'erreur d'annotation supposée et d'autre part, pour permettre la détermination de son substrat réel.

RB3123 présente un peptide signal qui semble confirmer son rôle dans la dégradation de polysaccharides exogènes. Elle est composée de deux modules :

- un module catalytique de la famille GH16 de 268 résidus, en position N-terminale, connecté au module suivant par une courte région charnière.
- un module appartenant à la famille des CBM16 de 135 résidus lui faisant suite en position C-terminale,

Elle a été clonée sous trois formes : la protéine entière, privée du peptide signal (emplacement B02), le module catalytique GH16 seul (emplacement C02) et le module CBM16 seul (emplacement D02). La séquence de RB3123 est présentée en Figure III-32. Un résumé de sa composition en aminoacides ainsi que de ses propriétés physico-chimiques théoriques est présenté dans le Tableau III-16.

Séquence protéique de RB3123 :

```
  1 MIKHLCTVLC FVLAQGNPLA AQEHPFSDPT NDGEWVLNTS VSDEFDGSVL
 51 NLDKWDNLGL NGNYYGEWKG RAPSQYDSAN VFVGGGHLTL TSRWDPDFQF
101 SKTAGSNGFK YGQPAPVTTA CVVSKARFKY GYMEMRCKAA DGPISSSFWT
151 TGHGGEIDVF EHFGEKPNDA FSSKRFHTSF HDWRKGSLTF GKRIWTNDHQ
201 LDFRVADDFH IYGLHWAEDF VKIYVDGILV NCVTKDEMGD NWVATDEQKV
251 WIDSETFDWE VRPEFLKASD FGDGQQFVID YCRIWQSSKA LRWVLRPNLI
301 SNPGFEEGLI GWQGSAVAGE DVRSGRGSAV MESSGTIHQT VPVKPNTTYI
351 LSGWVSSPKT NGKDLWYNAY LGVRSYGGEE TKARFFFPYF HQKSLQFRTG
401 PEASKAIIFF TNNPQDQKAF IDDISLVEAE QP
```

Figure III-32 : Séquence protéique de RB3123.
Le module GH16 est représenté en bleu, le module CBM16 est représenté en vert, les zones grises représentent respectivement le peptide signal et la zone charnière entre les deux modules.

	Protéine sauvage	Protéine entière clonée	Module GH16 cloné	Module CBM16 cloné
Poids moléculaire	48 700 Da	49 800 Da	32 000 Da	16 100 Da
pI théorique	5,34	5,75	5,35	6,70
Nombre d'acides aminés	432	440	277	144
Asp + Glu	55	55	41	14
Arg + Lys	41	42	26	13
Cystéines	6	6	4	2
$\varepsilon_{280\,nm}$ ($M^{-1}.cm^{-1}$)	109 230	109 230	79 660	23 950

Tableau III-16 : Récaputulatif des données biochimiques théoriques de RB3123.
Données extraites à partir du logiciel ProtParam disponible en ligne sur le site d'ExPASy.

I.B.2 - Résultats d'expression, de purification et de caractérisation biophysique

La production à plus grande échelle de cette protéine a confirmé ce que les tests d'expression en plaque indiquaient : cette protéine s'est révélée assez difficile à produire d'une part, et à purifier d'autre part.

Pour ce qui est de l'expression, la production du domaine catalytique de RB3123 a montré de faibles niveaux d'expression avec le milieu ZYP-5052 (difficilement discernable sur gel SDS-PAGE, après une culture de 200 mL ayant atteint une densité optique à 600 nm de 15). Plusieurs tentatives ont été réalisées en variant les conditions de culture. Une culture

de 1 L en fermenteur a notamment été réalisée. Les conditions physicochimiques de ce type de culture étaient à chaque instant suivies par pilotage informatique et, si nécessaire, réajustées (pH, température, oxygénation). L'oxygénation était assurée par un débit d'air stérile dans le milieu de culture, alliée à un dispositif de perturbation des écoulements. Dans ces conditions nous avions bon espoir d'améliorer les conditions de croissance et d'expression. Le résultat a effectivement montré une amélioration sensible, mais le niveau d'expression est resté très bas, bien en deçà de ce dont on peut espérer obtenir à partir d'une surexpression en système hétérologue, c'est-à-dire une quantité de protéine estimée, en fin de purification, à 0,5 mg par litre de culture. La production de la protéine entière n'a pas modifié les niveaux d'expression.

La chromatographie d'affinité au nickel a donné de bons résultats. La protéine a été éluée aux alentours de 250 mM imidazole, sur un gradient d'imidazole allant de 50 mM à 350 mM.

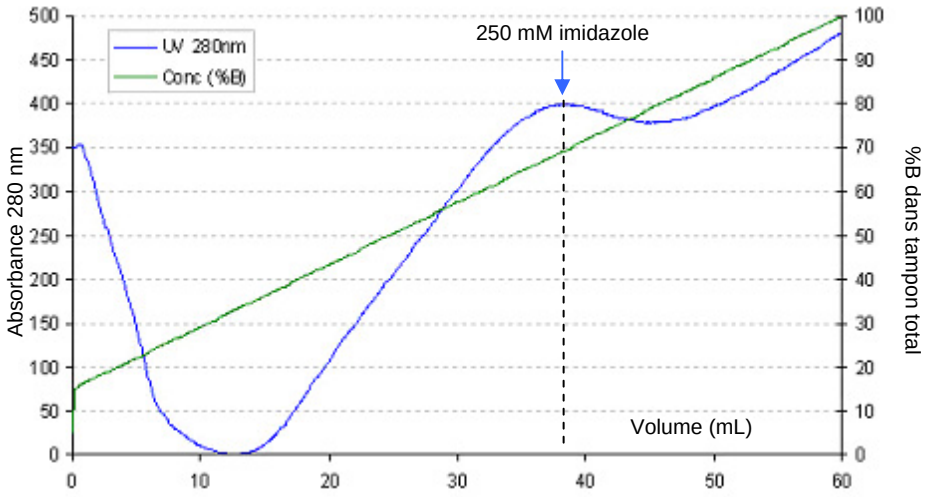


Figure III-33 : Purification de RB3123 (GH16) (Colonne d'affinité).
Chromatogramme de purification sur colonne d'affinité au nickel du module GH33 de RB3006 ;

Les étapes suivantes se sont révélées plus problématiques. La séparation par chromatographie d'exclusion de taille a été réalisée sur deux types de colonnes : une superdex 75 et une superdex 200 (dont, je le rappelle, les pouvoirs de séparation sont différents : 10 kDa – 80 kDa pour la Superdex 75 ; 20 kDa – 200 kDa pour la Superdex 200). Les chromatogrammes de sortie de ces colonnes a en effet montré que le volume d'élution de la protéine correspondait au volume mort de la colonne, et ce, pour les deux colonnes utilisées. Ce comportement suggère que le poids moléculaire de RB3123 dépasse de loin la taille attendue de 32 kDa.

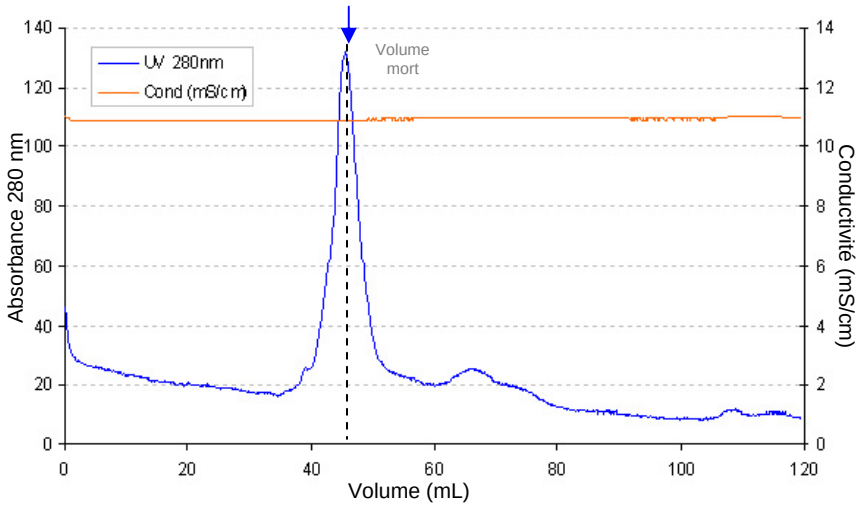


Figure III-34 : Purification de RB3123 (Colonne d'exclusion de taille). Chromatogramme de purification du module GH16 de RB3123, par exclusion de taille sur une colonne Superdex 75. La protéine sort dans le volume mort.

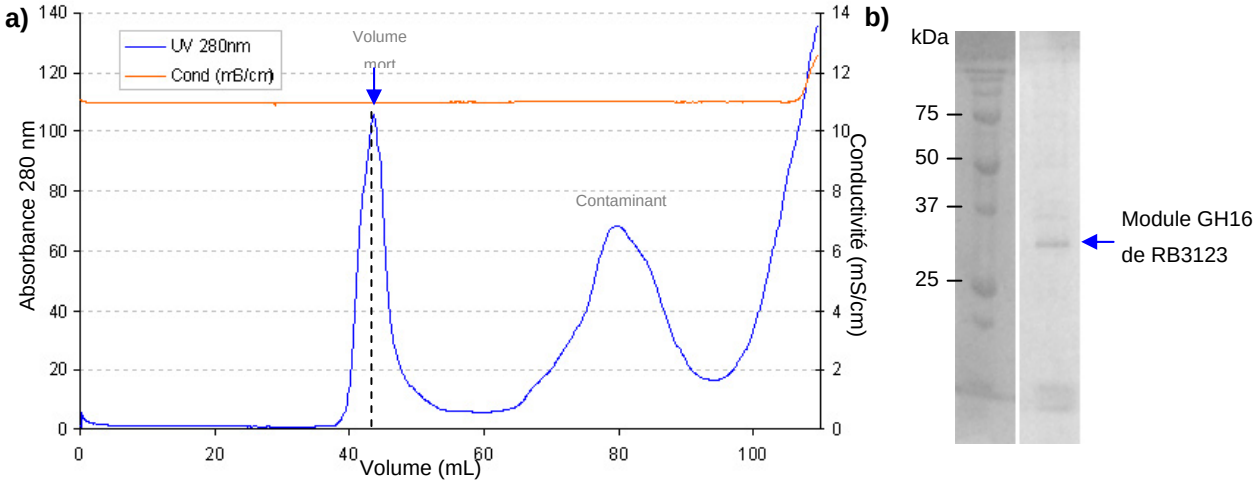


Figure III-35 : Purification de RB3123 (Colonne d'exclusion de taille). a) Chromatogramme de purification du module GH16 de RB3123, par exclusion de taille sur une colonne Superdex 200. La protéine sort également dans le volume mort; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

Le fait que RB3123 soit éluée dans le volume mort de ces colonnes semble être dû à une agrégation de la protéine. Ces agrégats supposés sont néanmoins probablement des agrégats solubles (dans lesquels la protéine peut rester active) puisqu'ils ne sont mis en évidence que lors de la manipulation d'un extrait soluble de la protéine. RB3123 présentant de plus une pureté acceptable, les tests d'activité ont donc été conduits sur la fraction du volume mort en sortie de Superdex 200.

I.B.3 - Tests enzymologiques

Deux techniques ont été utilisées pour déterminer l'activité du domaine catalytique de RB3123 : le dosage des sucres réducteurs et l'utilisation d'oses marqués par un fluorophore. Le dosage des sucres réducteurs a permis de cribler onze substrats naturels potentiellement accessibles dans l'environnement de *R. baltica*. Les oses marqués au p-nitrophénol ont permis de tester une activité de type exo (polysaccharides hydrolysés depuis leur extrémité).

Les substrats soumis à l'enzyme par le dosage des sucres réducteurs ont été choisis parmi les substrats connus et plausibles de la famille GH16 : l'agarose, le κ-carraghénane et la laminarine. Les activités XET et xyloglucanase n'ont pas été testées, étant donné leur très faible probabilité. D'autres substrats plus originaux et non concernés par la famille GH16 ont également été ajoutés à la liste : le ι-carraghénane et le λ-carraghénane en raison de leur ressemblance avec le κ-carraghénane, un β-D-galactane non chargé issu de la paroi de végétaux (Aldrich), ainsi que le xylane, la cellulose soluble (non fibrilaire), la chitine et le chitosane afin d'investiguer de nouvelles activités. Les résultats du dosage des sucres réducteurs sont résumés dans la

Figure III-36.

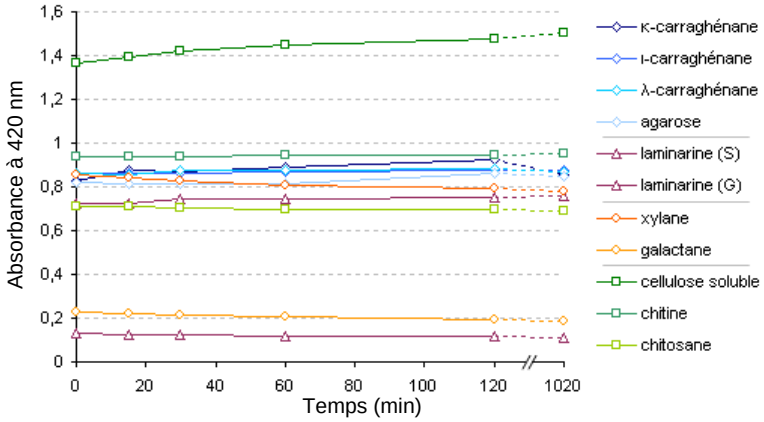


Figure III-36 : Résultats du dosage des sucres réducteurs
Vue graphique des résultats du dosage des sucres réducteurs après action du domaine catalytique de RB3123 sur divers polysaccharides. Les laminarines S et G correspondent respectivement à la laminarine des sociétés Sigma et Goëmar.

Quel que soit le substrat utilisé, le dosage des sucres réducteurs n'a montré aucune décroissance de l'absorbance à 420nm des solutions, et ce même après 17 heures d'incubation. Le module catalytique de RB3123 ne présente donc aucune activité sur les substrats usuels de la famille GH16.

Afin d'observer une éventuelle activité exo-α/β-galactanase ou exo-α/β-glucanase, des oses marqués au p-nitrophénol ont également été testés : p-nitrophenyl-α-D-galactopyranoside, p-nitrophenyl-β-D-galactopyranoside, p-nitrophenyl-α-D-glucopyranoside et p-nitrophenyl-β-D-glucopyranoside. Aucune activité n'a été détectée avec ces quatre substrats. Le domaine catalytique de RB3123 n'a donc pas d'activités exo-glucanase ou exo-galactanase détectables.

Les tests enzymologiques effectués ont donc confirmé la place toute particulière de RB3123 dans la famille GH16. Cependant, l'activité de la protéine est potentiellement affectée par sa tendance supposée à former des agrégats en solution. Ceci sera discuté à la fin de cette partie.

I.B.4 - Modélisation de la structure de RB3123 à partir de la structure de la β-agarase AgaA de Z. galactanivorans

Afin d'étudier le potentiel de dégradation de RB3123 et de pouvoir formuler des hypothèses de travail guidant l'identification d'une nouvelle spécificité, une approche théorique basée sur la génération d'un modèle structural a été réalisée.

Le recours à un modèle par homologie n'est pas l'équivalent d'une structure et son interprétation doit être menée avec circonspection. Néanmoins, dans le cas d'un modèle de bonne qualité, il est possible d'être guidé vers la découverte de certaines affinités potentielles entre la protéine et une classe de substrat. Une étude de ce type à été réalisée avec succès par Song en 2007 (Song *et al.*, 2007) sur des membres divergents de la superfamille des énolases. Il leur a été possible de procéder à des modélisations d'ancrage de ligands dans des modèles de protéines générés par homologie et d'affiner ainsi les cribles expérimentaux réalisés en parallèle.

Le principe de ce type de modélisation est basé sur la possession de données de structure d'un homologue de la protéine étudiée et d'un alignement de séquences primaines reposant sur la structure de cet homologue. Dans le cas de RB3123, afin de générer un modèle de confiance, le choix de la protéine de référence a été effectué en criblant les onze structures publiées dans la famille GH16. En effet, la clef de la finesse de l'analyse d'un modèle structural réside dans la confiance que l'on peut avoir dans le modèle créé. Cette fiabilité découle de plusieurs choses : tout d'abord, l'alignement initial avec la séquence de

référence doit être de bonne qualité. C'est-à-dire que le meilleur compromis possible entre le plus haut degré de similitude et le plus faible taux insertions/délétions dans les séquences, doit être atteint. La similitude garantit que des éléments de structures secondaires occuperont des places similaires dans les deux séquences comparées. Les insertions/délétions sont, quant à elles, les principales causes de perturbation dans ces éléments de structures secondaires (apparition de nouvelles contraintes stériques, glissement de fragments de séquences le long de ces éléments, déplacement d'éléments, …). De plus, ces insertions et délétions, par rapport à la séquence de référence sortent des contraintes imposées par l'alignement structural et doivent être recalculées par dynamique moléculaire et recuit simulé. La qualité du modèle dans ces zones ne correspond donc pas à une équivalence expérimentale, contrairement au reste de la structure générée. Cela peut introduire divers biais de lecture, selon la taille de la zone à modéliser ou l'algorithme utilisé. D'une manière générale, plus les séquences sont proches, plus le modèle sera bon. Une fois le modèle généré, il doit ensuite être cohérent vis-à-vis de lui-même, son repliement étant dicté par une autre séquence que la sienne. Cela implique de s'assurer que des acides aminés hydrophobes ne sont pas exposés vers le solvant, ou bien que des acides aminés chargés n'entrent pas en conflit électrostatique avec d'autres.

Après criblage des séquences candidates pour être utilisées comme références, le choix s'est porté sur la β-agarase A de *Z. galactanivorans* (AgaA). En effet, malgré une identité de séquence peu importante (autour de 24 %) et à même hauteur que d'autres représentants de la famille GH16, suggérant que RB3123 n'est probablement pas une agarase, l'alignement généré entre ces deux enzymes s'est révélé être de bonne qualité, et en particulier ne pas contenir beaucoup d'insertions (Figure III-37).

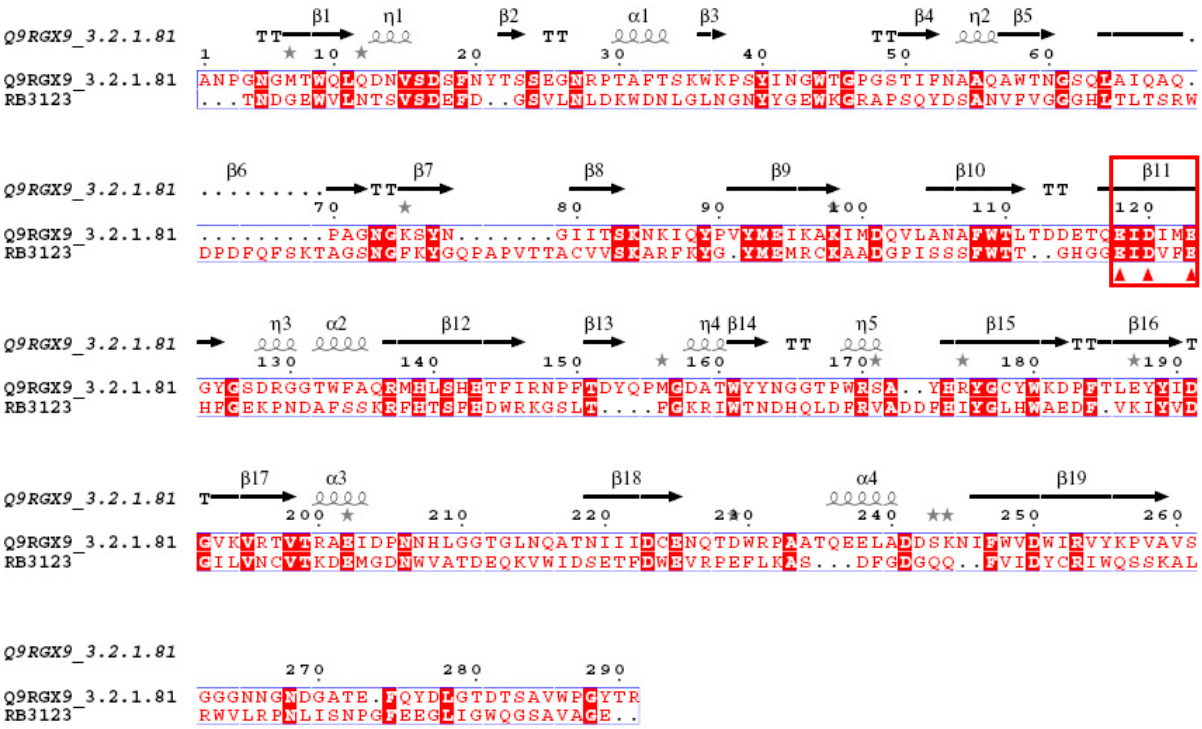


Figure III-37 : Alignement entre RB3123 et la β-agarase A de *Z. galactanivorans*. Les acides aminés catalytiques sont marqués par des triangles rouges(▲) et le site actif est encadré en rouge. Figure extraite de ESPript après alignement avec Multalign

Le modèle a été généré en utilisant Modeller v9.2 sur la base, d'une part, d'un alignement multiple de séquences des modules catalytiques de RB3123 et AgaA, et, d'autre part, le fichier de la structure du mutant E147S de l'AgaA (clef pdb : 1O4Y) en complexe avec son substrat (Allouch *et al.*, 2004) et épuré des aminoacides ne rentrant pas dans l'alignement. Le modèle généré (Figure III-38) ressemble bien évidemment très fortement à l'AgaA, avec une déviation moyenne (*Root Mean Square Deviation*, RMSD) calculée sur les C_α de 1,33 Å. Plusieurs éléments montrent clairement que ce modèle est plutôt de bonne qualité.

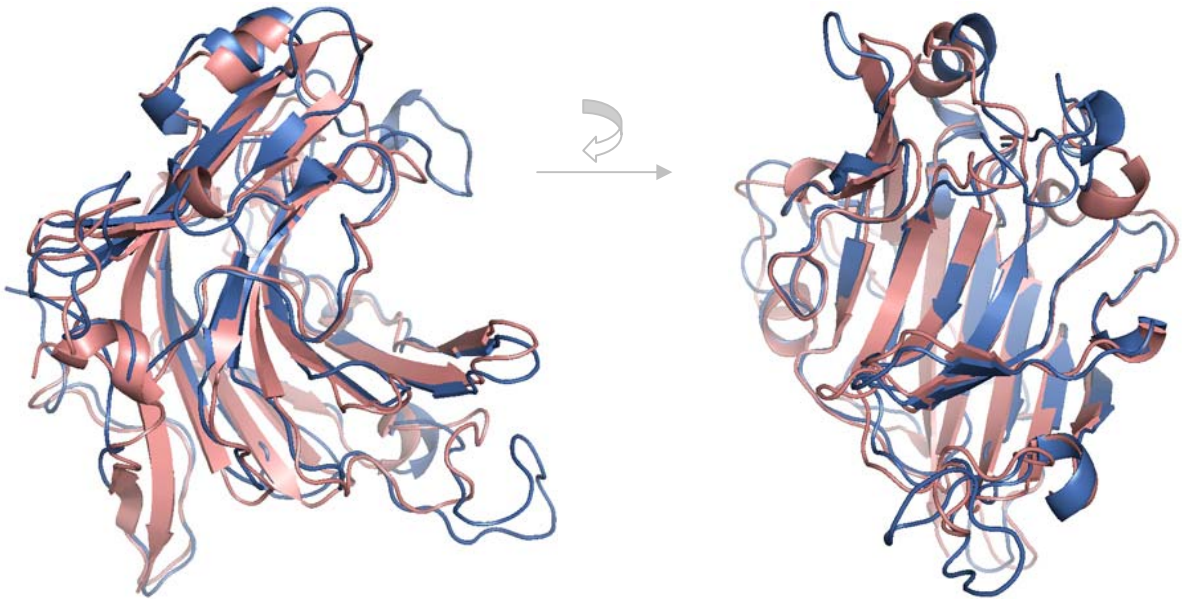


Figure III-38 : Modèle du domaine catalytique de RB3123.
Superposition du modèle du domaine catalytique de RB3123 (en bleu) avec la structure de AgaA (en rouge). Les deux images présentent la même superposition tournée horizontalement de 90°.

La qualité du modèle a pu être estimée à partir de l'analyse de certains éléments. Tout d'abord, la répartition des acides aminés hydrophobes, et en particulier les cycles aromatiques, se révèle cohérente avec celle de AgaA, avec de nombreux aminoacides aromatiques se superposant (**Erreur ! Source du renvoi introuvable.**). Il est également intéressant de noter que certains acides aminés aromatiques non conservés avec l'AgaA se trouvent dans des positions très favorables, avec notamment un probable empilement aromatique dans une poche hydrophobe derrière le site actif.

Figure III-39 : Résidus aromatiques dans le modèle de RB3123.
Présentation d'une vue globale du modèle montrant les acides aminés aromatiques sous forme de lignes. Les résidus superposés avec la référence sont montrés en bâtons avec celui de leur homologue (en rouge).

Ci-dessous sont présentées deux vues plus détaillées en stéréo. La vue présentée en Figure III-40 montre une poche hydrophobe derrière le site actif où trois résidus aromatiques (deux phénylalanines et une tyrosine) pourraient s'empiler pour stabiliser les feuillets adjacents. La phénylalanine colorée en vert foncé fait partie du renflement β du site actif de RB3123.

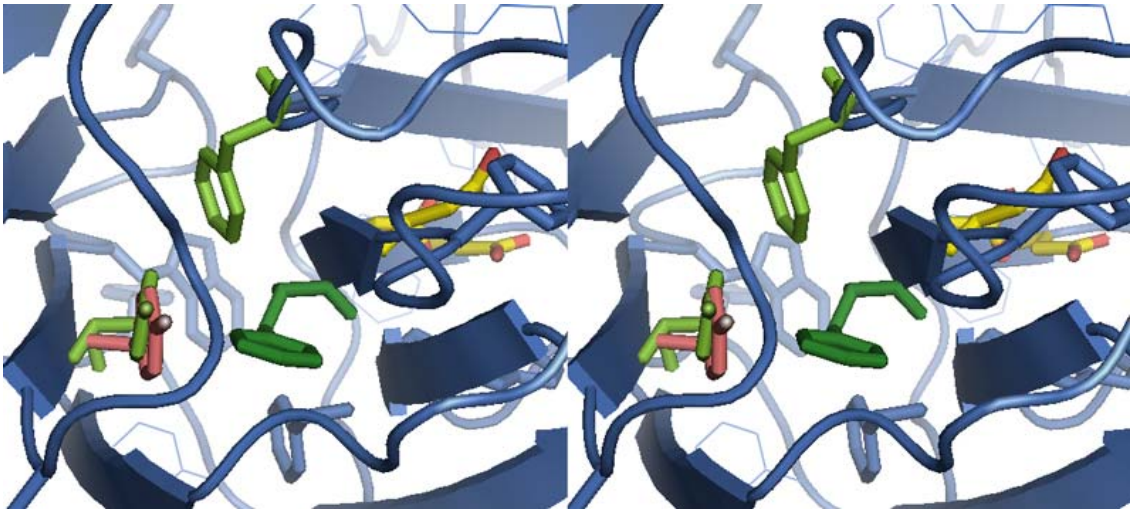


Figure III-40 : Détail du modèle de RB3123 (1).
Présentation en vue stéréoscopique de deux phénylalanines et une tyrosine du cœur hydrophobe de RB3123 dans des positions cohérentes, malgré leur non conservation. Le site actif est visible en jaune à l'arrière plan.

La vue suivante présentée en Figure III-41 détaille le site actif du modèle de RB3123. Sont surlignés en vert des acides aminés aromatiques proches du site de reconnaissance du substrat.

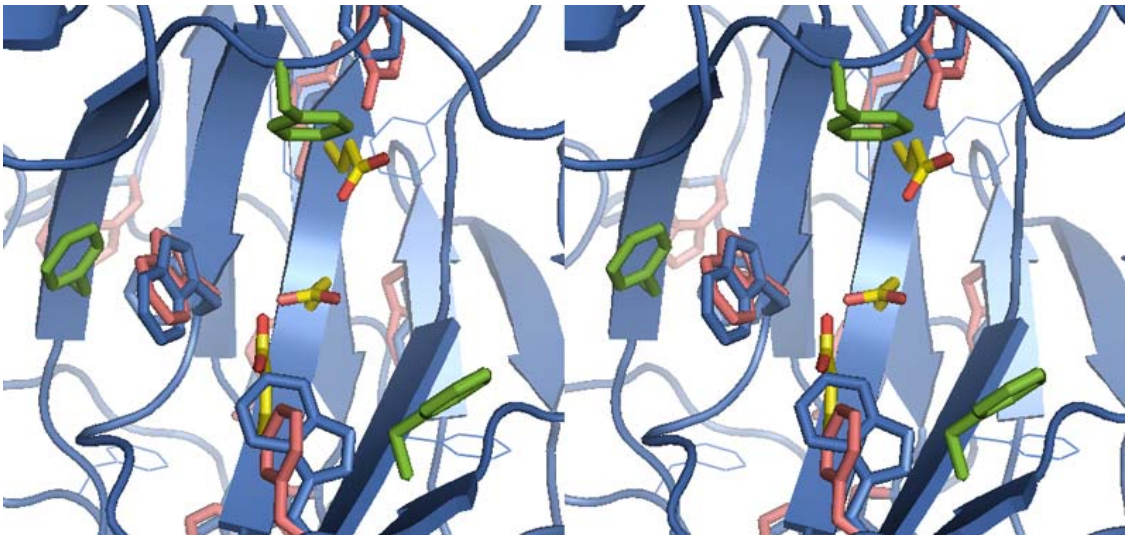


Figure III-41 : Détail du modèle de RB3123 (2).
Présentation en vue stéréoscopique su site actif de RB3123. Le site actif est visible en jaune à l'arrière plan. Les résidus non conservés avec la référence sont en vert.

D'autres éléments penchent en faveur d'un modèle de bonne qualité. Cette fois-ci, ce sont les acides aminés chargés (basiques : R, H, K et acides : D, E) qui présentent des configurations cohérentes, parfois conservées et parfois originales par rapport à la référence. Les figures suivantes présentent plusieurs vues du modèle de RB3123 mettant en évidence la disposition de ces acides aminés chargés.

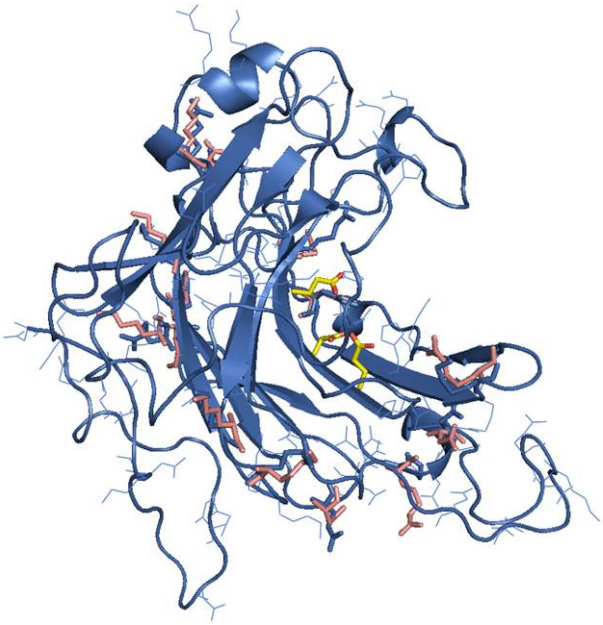


Figure III-42 : Résidus chargés dans le modèle.

Présentation d'une vue globale du modèle montrant les acides aminés chargés sous forme de lignes. Les résidus superposés avec la référence sont montrés en bâtons, avec celui de leur homologue (en rouge).

La vue présentée en Figure III-43 montre une vue latérale en stéréo du site actif mettant en évidence l'absence de résidus chargés négativement dans la gorge catalytique (la structure de AgaA en fait, elle, apparaître quelques-uns).

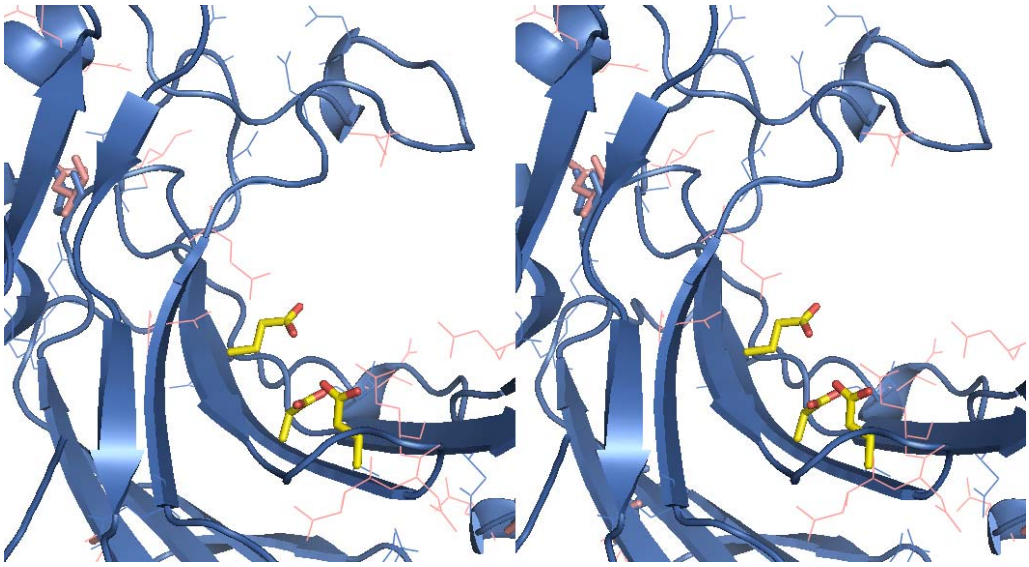


Figure III-43 : Détail du modèle de RB3123 (3).
Mise en évidence de la disposition (et en l'occurrence de l'absence) des résidus acides proches du site actif.

La vue présentée en Figure III-44 est une vue globale du site actif (encore une fois en stéréo) mettant, cette fois, en évidence les résidus basiques. Il est possible de souligner ici que, contrairement aux résidus acides, ces résidus semblent peupler très largement la gorge catalytique, en particulier ils semblent tapisser le sous-site +1.

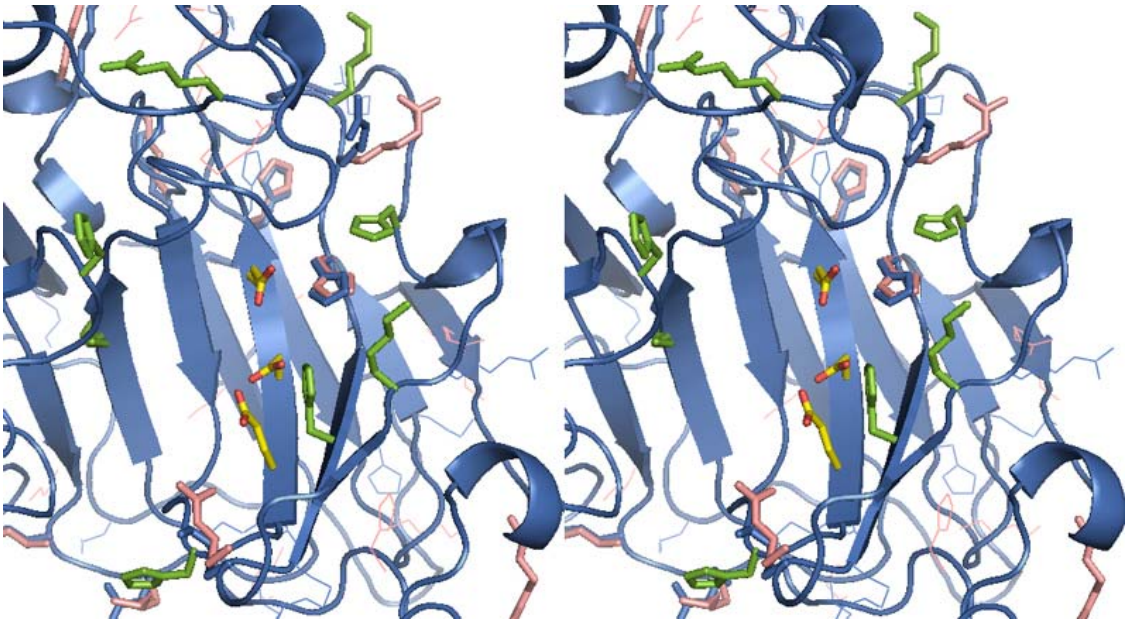


Figure III-44 : Détail du modèle de RB3123 (4).
Vue globale du site actif mettant en évidence la disposition des résidus basiques (les résidus non conservés avec l'AgaA sont représentés en vert).

I.C - Discussion

Les différentes résultats des tentatives de caractérisation biochimique de RB3123 montrent très clairement que cette enzyme n'est pas active sur les substrats qui lui ont été soumis, que ce soit les activités exo-glucanase et exo-galactanase, ou les activités sur les substrats concernant ou non (mais potentiellement rencontrés par *R. baltica*) les enzymes de la famille GH16. Il semble cependant que cette enzyme ait tendance à s'agréger en solution, ceci laissant ouverte la question de son bon repliement tertiaire et de l'existence d'une activité mesurable.

L'étude du modèle structural généré parallèlement montre une grande quantité de résidus basiques présents au niveau des sites de reconnaissances de la gorge du site actif. Cela semble indiquer la prise en charge d'un substrat fortement chargé négativement. Il

pourrait s'agir d'un polysaccharide de type carraghénane ou tout du moins contenant des ester-sulfates ou des acides uroniques. Avec toute la réserve nécessaire à l'étude d'un modèle non expérimental, il serait cependant intéressant de tenter des modélisations de substrats dans la gorge catalytique afin d'être guidé dans les tests expérimentaux.

S'il semble acquis que le substrat de RB3123 est original, comparé à ceux qui sont hydrolysés par les enzymes de la famille GH16, la caractérisation de l'activité de cette enzyme nécessitera néanmoins un travail approfondi pour déterminer des conditions de culture satisfaisantes permettant une expression, sous forme soluble, en plus grande quantité. Des tests d'expression en fusion avec la GST pourraient s'avérer un bon début. En cas de faible expression, il faudra peut-être se tourner vers d'autres systèmes d'expression hétérologue comme l'emploi de la levure. D'autres constructions seront également à envisager, notamment la forme entière de la protéine avec son CBM, qui est peut-être une des clefs de la solubilité de l'ensemble.

Lorsqu'une expression en grande quantité, sous forme soluble, sera obtenue, le criblage de banques de substrats pourrait s'avérer une solution efficace, étant donnée l'immense variété des substrats accessibles dans le monde marin, qui pour la plupart sont chargés négativement (carraghénanes, agars, fucanes, alginates, …). Les recherches autour de la création de ce type de banque en sont encore à leur début, même si certaines banque de substrats terrestres existent déjà (Willats et al.,). Une banque de polysaccharides d'origine marine est en cours d'élaboration au laboratoire dans le cadre d'un projet ANR CRAZY (coordonné par le Dr W. Helbert). Il est ainsi tout à fait envisageable de pouvoir utiliser cette ressource d'ici quelques années pour élargir encore le champ d'investigation.

Enfin, il apparaît que la famille 16 des glycoside hydrolases n'a pas encore livré tous ses secrets structuraux. La forte similitude entre la structure tridimensionnelle de RB3123 et celles de certains de ses homologues n'exclue pas que ces enzymes puissent avoir des spécificités de substrat radicalement différentes. Une analyse cristallographique de cette enzyme sera menée lorsque les niveaux d'expression seront suffisants, afin de comprendre ces fines variations de séquence et leurs effets sur cette spécificité.

II - RB2160 : une nouvelle glycoside hydrolase de la famille GH57 ?

II.A - La famille GH57

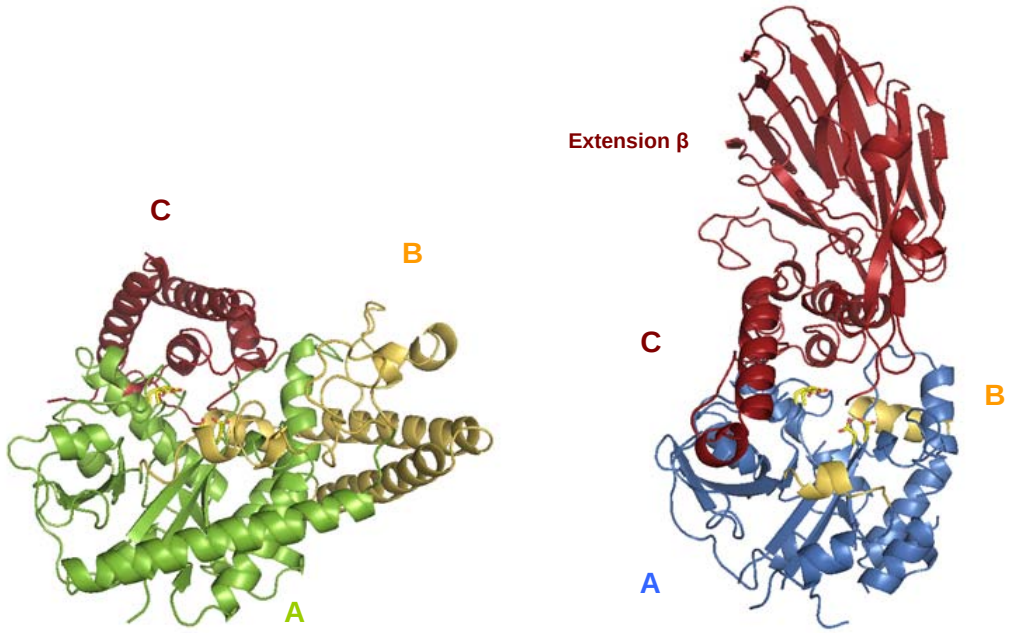
La famille GH57 est une famille multispécifique comprenant un peu moins de 300 séquences avec cinq activités enzymatiques recensées (Figure III-45) dont quatre sont liées au métabolisme de l'amidon (α-amylases, enzymes de branchement, 4-α-glucanotransférases, amylopullulanases) et une à la dégradation d'α-galactanes (voir http://www.cazy.org/fam/GH57.html).

Activité	Numéro EC	Nombre de structures (et codes PDB publiés)
(non déterminée)	-	1 (1)
α-galactosidase	3.2.1.22	
α-amylase	3.2.1.1	1 (1)
amylopullulanase	3.2.1.41	
4-α-glucanotransferase	2.4.1.25	1 (3)
branching enzyme	2.4.1.18	

Figure III-45 : Activités de la famille GH57.
Activités de la famille GH57 et nombre des structures disponibles par activité

La première séquence caractérisée date de 1988 avec l'α-amylase de *Dictyoglomus thermophilum* (Fukusumi *et al.*, 1988) mais il faudra attendre plus de dix ans pour qu'une première structure tertiaire, celle de la 4-α-glucanotransférase TLGT de l'archée hyperthermophile *Thermococcus litoralis* (code PDB 1K1W, Imamura *et al.*, 2003), soit publiée. Depuis, deux autres ont été publiées : l'enzyme non caractérisée TT1467 de *Thermus thermophilus* (code PDB 1UFA, Idaka, 2003) et l'α-amylase de *Thermotoga maritima* (code PDB 2B5D, Dickmanns *et al.*, 2006). Le repliement de la famille GH57 a ainsi pu être caractérisé comme présentant une architecture en tonneau de type $(\beta/\alpha)_7$. Ce repliement est assez rare au sein des glycoside hydrolases, mais il est néanmoins rencontré dans quelques familles, notamment la famille GH38 (composée de β-mannosidases), qui semble d'ailleurs apparentée à la famille GH57 (Imamura *et al.*, 2003). La Figure III-46

présente les structures tertiaires de l'α-amylase de la bactérie hyperthermophile *Thermotoga maritima*, (Dickmanns *et al.*, 2006) et celle de la 4-α-glucanotransferase TLGT.



Structure de l'α-amylase de *T. maritima* 2B5D

Structure de la 4-α-glucanotransferase de *T. littoralis* 1K1W

Figure III-46 : Structures caractéristiques de la famille GH57.

Sur les structures de la **Figure III-46**ci-dessus, les domaines catalytiques en TIM barrel $(\tilde{\beta}/\alpha)_7$ ont été représentés en vert (2B5D) et en bleu (1K1W). Topologiquement, ces domaines centraux constituent le domaine A. Les deux structures présentent en outre deux domaines additionnels, peu similaires, mais superposables : le domaine B, qui est un domaine fonctionnel constitué d'hélices α insérées entre des éléments de structure secondaire du domaine A (en jaune), et le domaine C, qui est de composition variable et qui est situé en position C-terminale (en rouge). La 4-α-glucanotransferase TLGT présente en outre un module additionnel dont je parlerai plus tard, situé dans le prolongement du domaine C et constitué d'un sandwich β (extension rouge au-dessus de la structure).

Les acides aminés catalytiques de la famille GH57 ont été déterminés. Dans la structure de la 4-α-glucanotransferase TLGT, il apparaît que l'attaque nucléophile est réalisée par l'acide glutamique E123. L'acide aminé donneur de proton n'a lui toujours pas été identifié expérimentalement, même si l'acide aspartique D214, conservé dans le voisinage stérique de E123, est un bon candidat du fait de son orientation, très favorable à cette fonction, et de sa conservation dans tous les représentants de la famille. De plus, sa position à environ 7 Å du résidu nucléophile est cohérente avec un mécanisme catalytique libérant des produits avec conservation de la configuration anomérique (Imamura *et al.*, 2003; Zona *et al.*, 2004).

Le repliement $(\beta/\alpha)_7$ peut être considéré comme une version incomplète du repliement TIM barrel $(\beta/\alpha)_7$ (classification CATH code : 3.20.20.80), qui est très fréquemment rencontré dans les glycoside hydrolases, et en particulier dans le clan structural GH-H. Ce clan est composé des familles GH13, GH70 et GH77, qui sont connues sous le nom global de « famille des α-amylases » car elles regroupent la majorité des enzymes amylolytiques connues (α-amylases ou autres). Les ressemblances entre la topologie des enzymes du clan GH-H et celle des enzymes de la famille GH57 sont assez frappantes : un domaine A constitué du centre catalytique de type TIM barrel $(\beta/\alpha)_n$, un domaine B essentiellement constitué d'hélices α imbriquées dans le domaine A et permettant la fixation de cations divalents, et un domaine C de taille et composition variables. La Figure III-47 présente un diagramme topologique de l'α-amylase de Thermotoga maritima illustrant l'imbrication des domaines A et B.

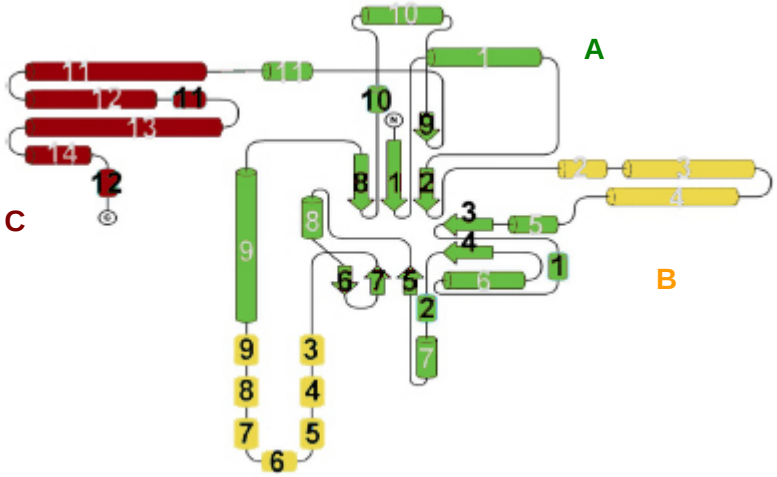


Figure III-47 : Topologie de 2B5D.
Diagramme extrait de (Dickmanns *et al.*, 2006)

Cependant, la famille GH57, malgré une prédominance d'activités amylolytiques, ne présente pas de similitude de séquence avec les familles du clan GH-H, et en particulier, ne possède pas les régions conservées caractéristiques des enzymes de la famille GH13 (Janecek, 2002).

Il apparaît également que si les différents domaines définis dans la Figure III-47 sont situés aux mêmes places topologiques, les tailles de ces différents domaines peuvent être assez variables. Ainsi, le domaine B de l'α-amylase de *Thermotoga maritima* est très étendu (120 résidus répartis sur deux zones constituées d'une série d'hélices α ; Figure III-47) en comparaison de celui de la 4-α-glucanotransferase de *Thermococcus litoralis* (seulement 20 résidus, également sur deux zones, constituées chacune d'une petite hélice α). Leur domaine C est de taille comparable (une centaine de résidus) et est constitué d'une série

d'hélices α. Il constitue un domaine de fonction inconnue avec un repliement de type immunoglobine, référencé comme DUF1925 dans la base PFAM. Cet alignement met de plus en évidence que la structure de la 4-α-glucanotransferase de *Thermococcus litoralis* présente un module supplémentaire de 280 résidus constitué d'une dizaine de brins β repliés en sandwich β. Cette extension β du domaine C, référencée en tant que domaine DUF1926 dans la base PFAM, est, non seulement exclusivement rencontrée dans la famille GH57, mais est également assez conservée. La **Figure III-48** présente un alignement des domaines $(\beta/\alpha)_7$ de ces deux structures et permet de mettre en évidence les topologies semblables de ces enzymes.

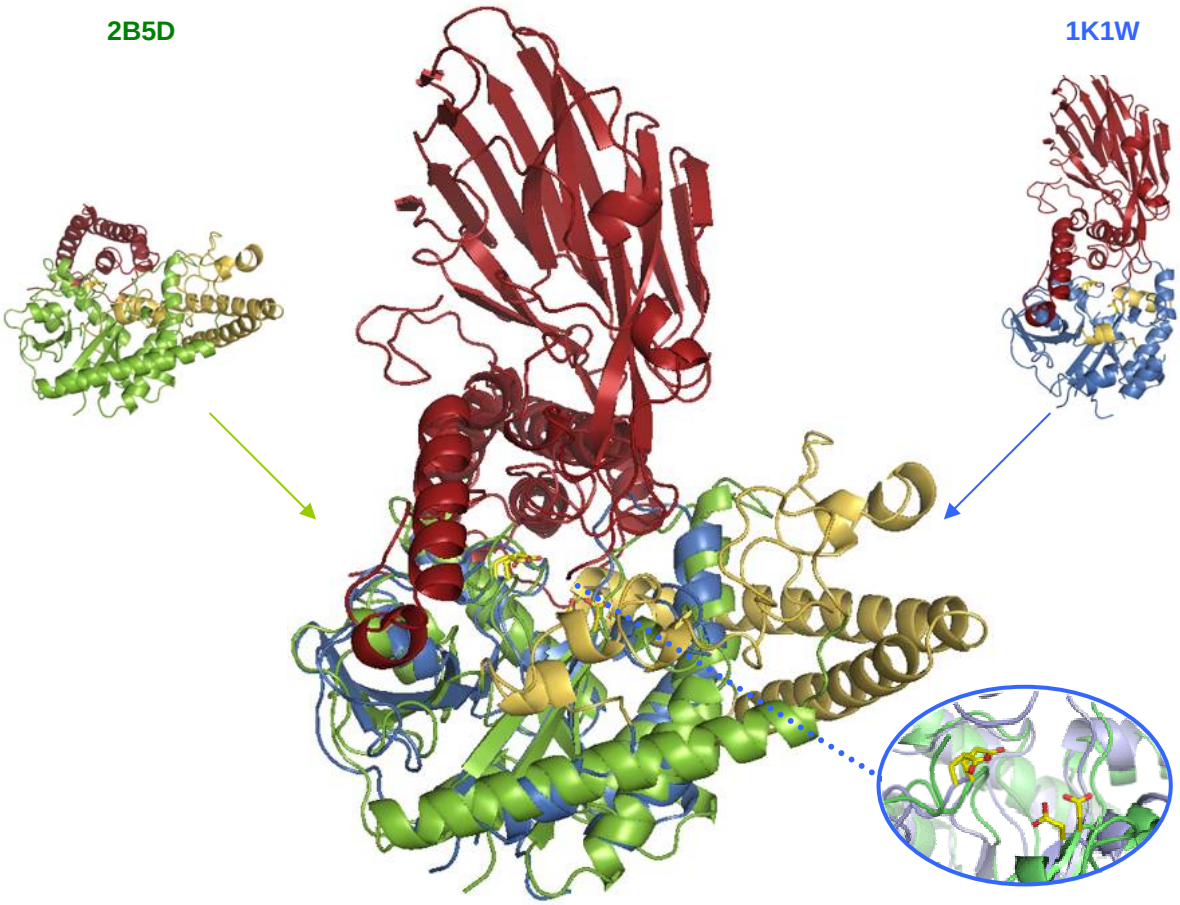


Figure III-48 : Superposition des structures 1K1W et 2B5D.
Superposition des structures 1K1W (bleu) et 2B5D (vert) avec alignement structural de leur domaine A. Un zoom sur le site actif est également présenté avec les acides aminés catalytiques superposés sur leur structure respective.

Au-delà de son repliement, cette famille présente plusieurs originalités. Elle n'est en effet présente que chez les procaryotes, dont un grand nombre est issu de milieux extrêmes, en particulier de milieux à hautes températures. De plus, même si les 3 structures publiées depuis 2001 ont permis de lui attribuer le repliement $(\beta/\alpha)_7$ ainsi que la détermination de ses acides aminés catalytiques, des interrogations fonctionnelles subsistent. Le domaine C est

en particulier source de nombreuses questions. En effet, moins d'une vingtaine de séquences de la famille GH57 présentent cette extension modulaire d'architecture en sandwich β. De plus, il semble qu'il s'agisse d'un module structuralement indépendant, présentant une large surface d'interaction avec le tonneau $(\beta/\alpha)_7$ du domaine A, ce qui suggère qu'il pourrait ne pas être indépendant du reste de la structure. Sa fonction n'est pas connue, mais il est actuellement considéré comme pouvant participer à la transglycolylation (Imamura *et al.*, 2003). Cela est néanmoins à relativiser avec le fait que l'α-amylase AMY1 de *Dictyoglomus thermophilum* (AmyA ; P09961) et la malto-oligasaccharidase de *Pyrococcus furiosus* (Q8U2G5) possèdent également ce domaine, ce qui pourrait suggérer une activité biologique sensiblement différente.

II.B - Résultats et discussion

II.B.1 - Analyse bioinformatique

R. baltica ne présente qu'une seule séquence appartenant à la famille GH57, RB2160, une protéine bimodulaire probablement non sécrétée (aucun peptide signal détecté) et présentant le module sandwich β précédemment décrit en II.A. La Figure III-49 résume les caractéristiques de RB2160.

Protéine	Modularité	Taille		Annotation initiale
		résidus	kDa	
RB2160	GH57 – DUF1926	719	81	alpha-amylase (EC 3.2.1.1)

Figure III-49 : Famille GH57 chez *R. baltica*.

Un alignement multiple, a été réalisé avec MAFFT, comprenant la séquence RB2160 de *R. baltica* et les enzymes 4-α-glucanotransferase TLGT de *Thermococcus litoralis* (O32462 ; n° EC 2.4.1.25), l'α-amylase AMY1 (AmyA) de *Dictyoglomus thermophilum* (P09961 ; n° EC 3.2.1.1) et la Cyclomaltodextrin glucanotransferase de *Archaeoglobus fulgidus* (A9QMB3 ; n° EC2.4.1.19) (Figure III-50). Ces trois enzymes présentent l'avantage de couvrir des activités différentes dans la famille GH57, et de posséder, elles aussi, l'extension β suivant le domaine C. Elles font de plus partie des meilleures similitudes de

séquence de la famille GH57 avec RB2160 (respectivement 42%, 41% et 42% sur le domaine catalytique, sans l'extension β).

Il apparaît que l'alignement de ces enzymes est peu fragmenté et quasiment sans insertion (excepté sur une courte région C-terminale). Les acides aminés catalytiques E123 et D214, dans la 4-α-glucanotransferase, ainsi que ceux impliqués dans la reconnaissance de l'acarbose, présentent eux aussi une grande conservation dans quasiment toutes les séquences. Avec leur analyse de la famille GH57, (Zona *et al.*, 2004) ont montré que cinq motifs très conservés, propres à cette famille, sont également observables. Ces régions contiennent les résidus catalytiques, ainsi que d'autres résidus très conservés (à plus de 90%) qui sont proposés pour jouer un rôle dans l'activité (Figure III-50).

L'alignement présente également les résidus impliqués dans la fixation de l'acarbose dans la structure de la 4-α-glucanotransferase TLGT (Imamura et al., 2003). Il est intéressant de constater que les prédictions de Zona et al. se sont révélées assez justes, puisqu'une grande partie des acides aminés, prédits pour interagir avec le substrat, fixent l'acarbose dans la 4-α-glucanotransferase TLGT (E216, W221 et D354).

Cet alignement a permis de mettre évidence que, d'une part RB2160 présentait bien les acides aminés catalytiques, au sein de motifs strictement conservés pour ces quatre séquences, et, d'autre part, que l'alignement, à lui tout seul, ne pouvait pas préciser sa fonction. Ici, apparaît une originalité de la protéine RB2160 : elle fait partie des rares séquences de la famille GH57 à ne pas présenter l'intégralité des acides aminés de reconnaissance du substrat, ni même l'intégralité des résidus conservés, présentés précédemment. Ainsi, parmi les résidus fixant l'acarbose, l'arginine R371 et la tyrosine Y601 de la 4-α-glucanotransferase TLGT (non représentée sur l'alignement) ne sont pas conservées dans RB2160 de *R. baltica*.

On remarque de plus que l'acide aspartique D354, pourtant au sein d'un motif très conservé dans les quatre séquences (Figure III-50 ;Figure III-51), est ici remplacé par une cystéine dans RB2160. En tout, trois séquences, parmi l'ensemble des séquences de la famille GH57, n'ont pas d'acide aspartique ou d'acide glutamique à cette position. RB2160 est la seule à y avoir une cystéine (les autres présentant une asparagine ou une phénylalanine).

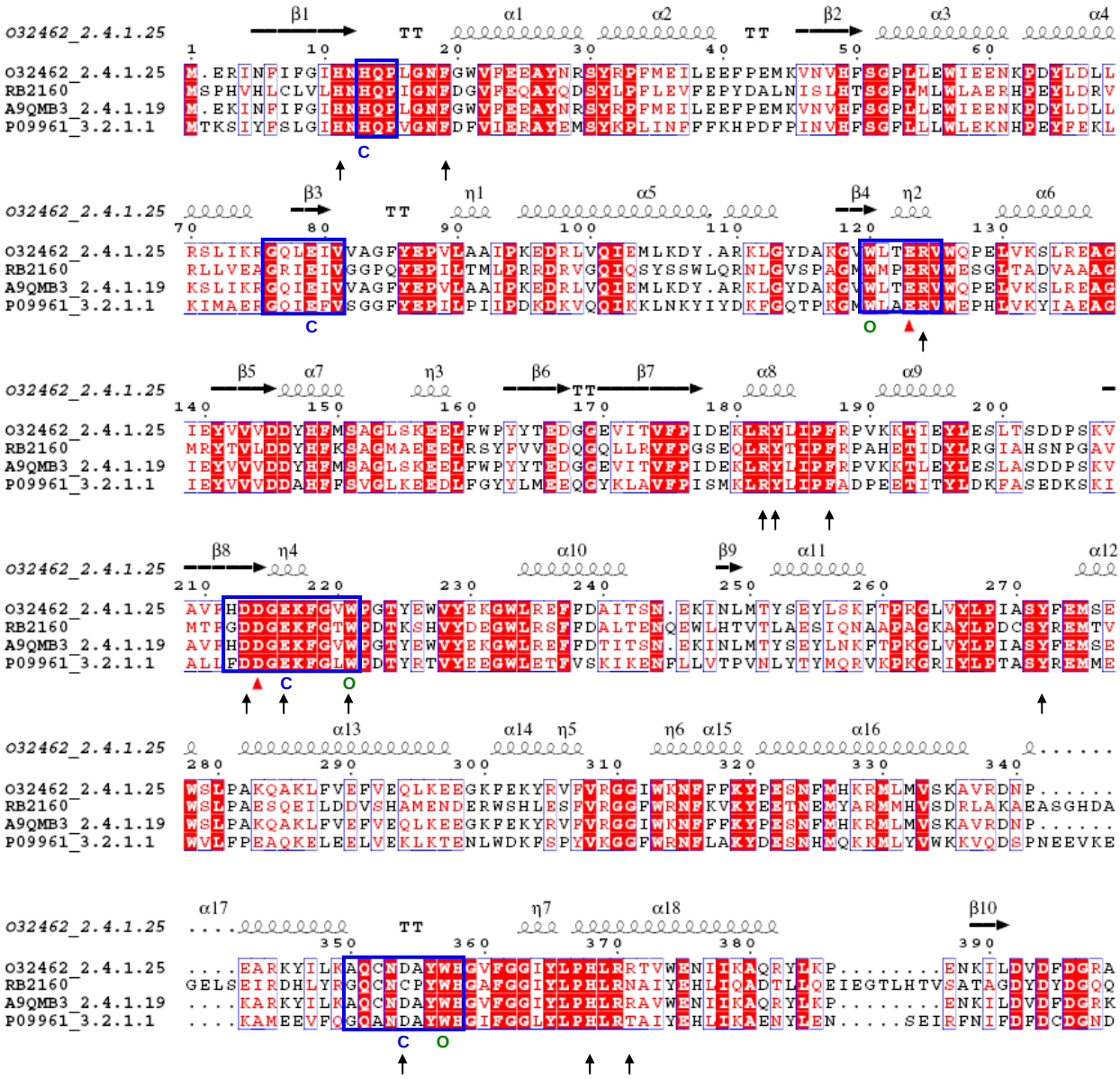


Figure III-50 : Alignement multiple dans la famille GH57.

Alignement multiple des séquences des modules catalytiques de la famille GH57. Les flèches rouges (▲) indiquent les acides aminés catalytiques. Les régions conservées dans la famille GH57 sont entourées de bleu. Les résidus hautement conservés, jouant probablement un rôle dans l'activité, sont matérialisés par des C bleus (résidus chargés) et des O verts (résidus aromatiques) (Zona et al., 2004). Les flèches sur la seconde ligne représentent les acides aminés liés à l'acarbose dans la structure de la 4-α-glucanotransferase TLGT (Imamura *et al.*, 2003).

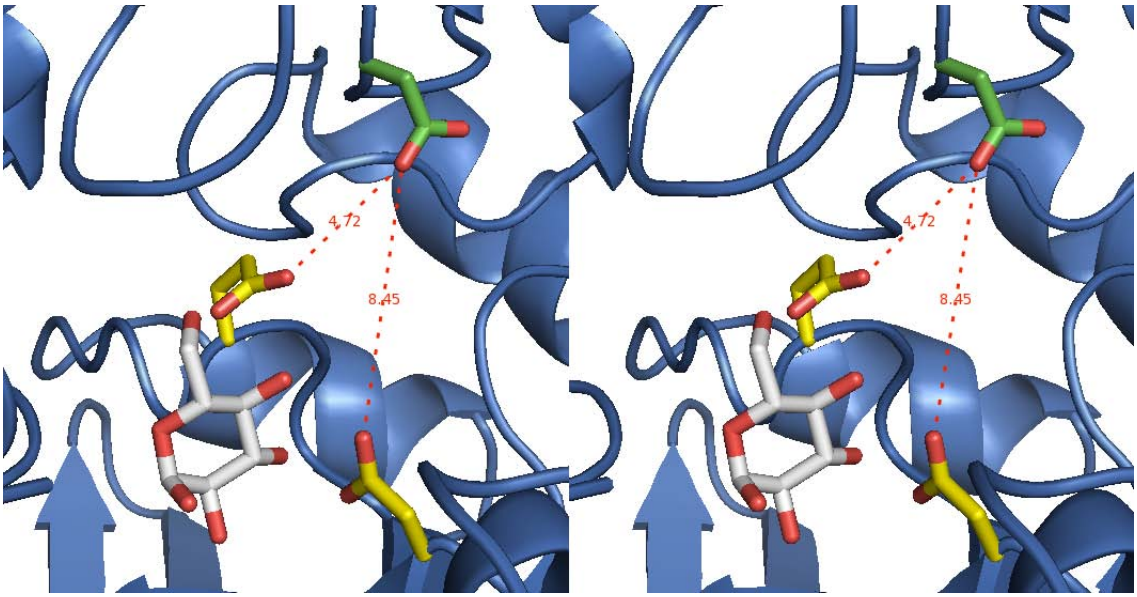


Figure III-51 : Vue stéréoscopique du site actif de TLGT.
L'acide aspartique D354 est représenté en vert. Les acides aminés catalytiques E123 et D214 sont représentés en jaune. Un des glucoses de l'acarbose est représenté en blanc. Les distances D354-E123 et D354-D214 sont matérialisées par une ligne discontinue rouge. Il apparaît que D354 est très proche des résidus catalytiques et il n'est pas impensable qu'il puisse interagir étroitement avec le substrat.

Une analyse phylogénétique complémentaire a été menée et un arbre phylogénique issu de l'alignement multiple des domaines catalytiques de 240 séquences de la famille GH57 a été généré (Figure III-52). Son analyse a permis de compléter les résultats de Zona et al. (Zona *et al.*, 2004) avec les séquences incluses dans la famille depuis 2004. Comme il était possible de s'y attendre, RB2160 branche dans le sous-groupe des enzymes possédant le domaine β et, au sein de ce domaine, avec les séquences issues de bactéries. Il apparaît également qu'il n'y a pas de consensus fort au sein de l'arbre et que les groupes identifiables sont peu informatifs sur les spécificités des séquences qui les composent.

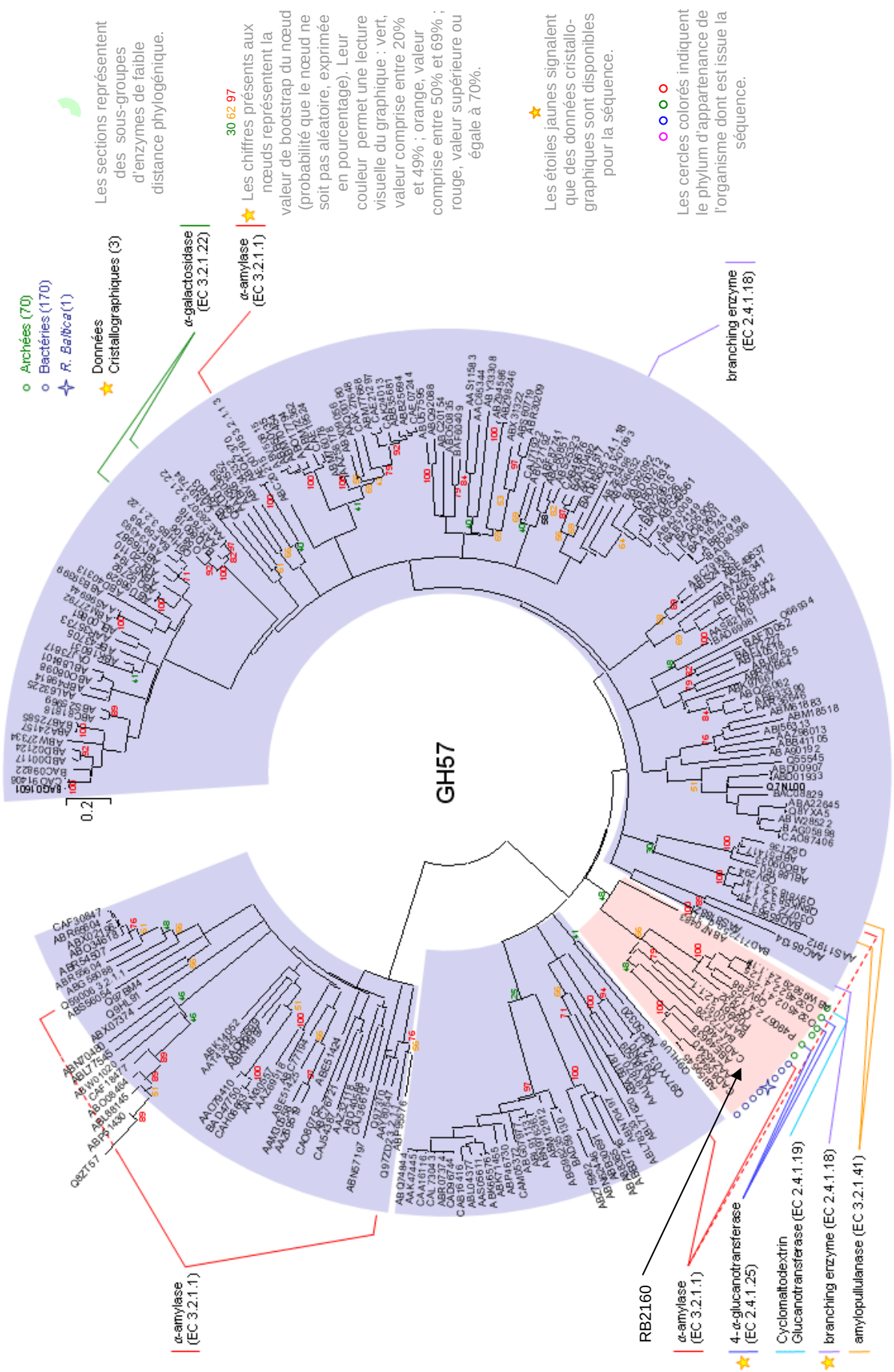


Figure III-52 : Arbre phylogénique de la famille GH57

Arbre phylogénique généré par la méthode d'évolution minimum, avec délétion des indels,, à partir d'un alignement de type FFT-NS-i par MAFFT.

Il ressort de ces différentes analyses que la protéine RB2160 est une enzyme relativement conservée dans sa famille, qu'elle présente les résidus catalytiques dans des régions assez conservées, mais qu'elle présente aussi des originalités qui la placent hors de tout cadre prédictif fiable. En effet, elle possède un module additionnel de fonction inconnue, spécifique des GH57 mais non systématiquement rencontré dans cette famille, qui vient recouvrir le site actif, sans que cela semble conférer, aux quelques enzymes qui le possèdent, une activité particulière. De plus, elle présente la caractéristique d'avoir une modification sur un résidu de la gorge catalytique, très conservé dans la famille GH57.

Faisant partie des meilleures expressions sous forme soluble, dans la plaque des tests d'expression, cette enzyme a donc été sélectionnée pour rejoindre la liste des protéines à caractériser en raison du rôle difficilement cernable qu'elle joue dans l'arsenal enzymatique de *R. baltica*.

RB2160 a été clonée sous une seule forme : la protéine entière, comprenant donc le domaine β additionnel (emplacement E08). Ci-après sont présentés la séquence entière de RB2160 et un tableau récapitulatif de ses propriétés biochimiques prédites par ProtParam sur le site ExPASy. LaFigure III-53 et le Figure III-54 présentent respectivement la séquence et les propriétés biochimiques théoriques de RB3006.

Séquence protéique de RB2160 :

```
  1 MSPHVHLCLV LHNHQPIGNF DGVFEQAYQD SYLPFLEVFE PYDALNISLH
 51 TSGPLMLWLA ERHPEYLDRV RLLVEAGRIE IVGGPQYEPI LTMLPRRDRV
101 GQIQSYSSWL QRNLGVSPAG MWMPERVWES GLTADVAAAG MRYTVLDDYH
151 FKSAGMAEEE LRSYFVVEDQ GQLLRVFPGS EQLRYTIPFR PAHETIDYLR
201 GIAHSNPGAV MTFGDDGEKF GTWPDTKSHV YDEGWLRSFF DALTENQEWL
251 HTVTLAESIQ NAAPAGKAYL PDCSYREMTV WSLPAESQEI LDDVSHAMEN
301 DERWSHLESF VRGGFWRNFK VKYEETNEMY ARMMHVSDRL AKAEASGHDA
351 GELSEIRDHL YRGQCNCPYW HGAFGGIYLP HLRNAIYEHL IQADTLLQEI
401 EGTLHTVSAT AGDYDYDGQQ EIRLSNESMV AWIDPAQGGR MYEWDLRGIN
451 HNLLATLQRR PEAYHRKVLA GPSSAGGDVA SIHDRVVFKQ EGLDQMIQYD
501 RYARKSLMDH FFDNEATLES VSRGESPERG DFVELPFEAK LRRGSDRVQA
551 QLRRDGNAWG IPITLTKAVT LQEGSGNLSV TYLLENLPPA SPLHFAVEWN
601 FAGLPSGADD RYFSDVDGNQ LGQLGERLDL TDVRGLSLSD RWLGVDIDLR
651 TNRDSGVWAF PVETVSQSEA GFELVHQSVC VMPHWIITAD AEGRWAVTID
701 IATRCENSVE LQSHDHVNA
```

Figure III-53 : Séquence protéique de RB2160.

	Protéine sauvage	Protéine clonée
Poids moléculaire	81 500 Da	82 500 Da
pI théorique	4,95	5,1
Nombre d'acides aminés	719	727
Asp + Glu	107	107
Arg + Lys	61	62
Cystéines	6	6
$\varepsilon_{280\,nm}$ ($M^{-1}.cm^{-1}$)	146 600	146 600

Figure III-54 : Récapitulatif des données biochimiques théoriques de RB2160.
Données extraites à partir du logiciel ProtParam sur le site www.expasy.org.

II.B.2 - Résultat d'expression, de purification et de caractérisation biophysique

Le fort niveau d'expression sous forme soluble de RB2160 en plaque s'est confirmé avec la surexpression à plus grande échelle. Cette protéine s'est révélée parmi les meilleures expressions sous forme soluble de l'ensemble des cibles avec des productions moyennes autour de 12 mg de protéine en fin de purification (soit autour de 60 mg/L de culture).

Il a néanmoins fallu ajuster les protocoles classiques. Cette enzyme semble notamment instable en faible molarité de sels (en dessous de 100 mM NaCl) et précipite en quelques jours. Cela a été particulièrement visible lors de la phase de concentration pour la cristallogénèse. Le tampon final de conservation a donc été pour cette enzyme 50 mM tris-HCl pH 7,5 ; 200 mM NaCl ; 2% glycérol.

La protéine a été purifiée à l'homogénéité en deux étapes : par chromatographie d'affinité avec une colonne de nickel (Figure III-55) d'où elle a été éluée à la concentration en imidazole de 155 mM environ (gradient d'imidazole de 50 mM à 500mM) et par chromatographie d'exclusion de taille sur une colonne Superdex 200 (Figure III-56). La purification a été contrôlée par électrophorèse sur gel de SDS-PAGE 12%.

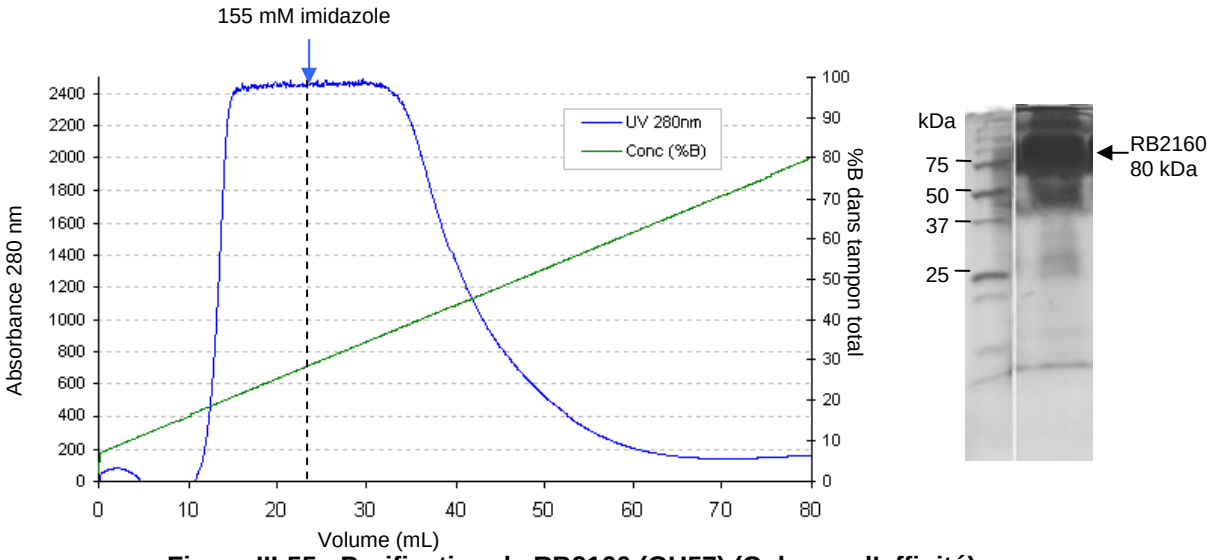


Figure III-55 : Purification de RB2160 (GH57) (Colonne d'affinité)
a) Chromatogramme de purification sur colonne d'affinité au nickel de RB2160 ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

A partir de la colonne d'exclusion de taille calibée, RB2160 élue à 67 mL, ce qui correspond à une protéine de 80 kDa. Par conséquent, RB2160 forme un monomère en solution. Ceci a été confirmé par une mesure de diffusion de la lumière (DLS).

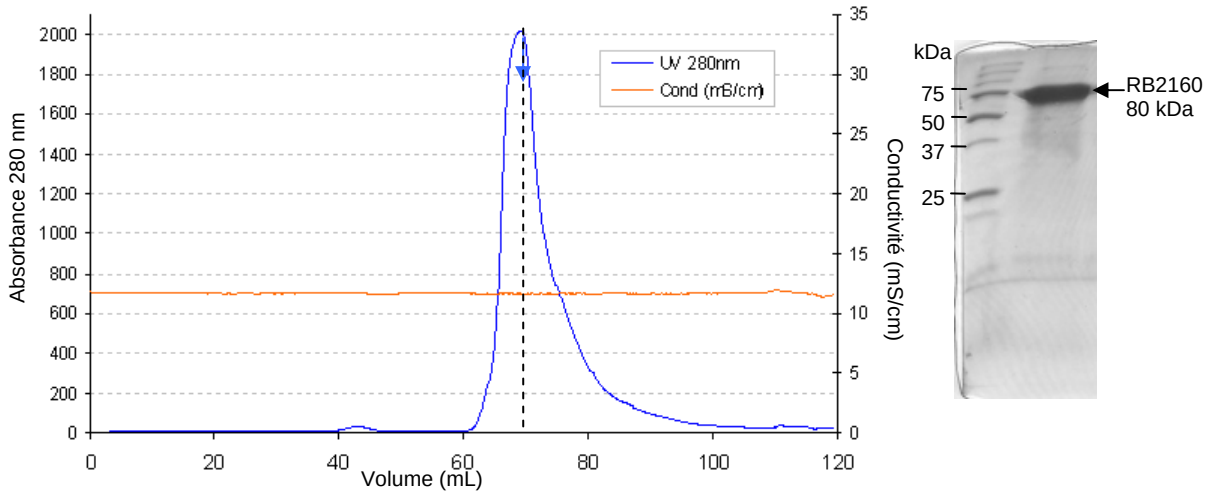


Figure III-56 : Purification de RB2160 (GH57) (Colonne d'exclusion de taille)
a) Chromatogramme de purification sur colonne d'exclusion de taille de RB2160 ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

II.B.3 - Tests enzymologiques

L'activité de RB2160 a été testée sur plusieurs substrats en utilisant plusieurs techniques. Des dosages des sucres réducteurs ont tout d'abord été réalisés sur de l'amylose (amidon soluble), du pullulane et du glycogène à des concentrations allant de 0,05% à 0,3%. Aucun test n'a cependant été positif démontrant que RB2160 n'est pas une α-amylase, contrairement à la fonction prédite par l'annotation initiale de la protéine.

Pour tester d'autres activités, la technique du zymogramme a été employée. Des essais ont été menés sur des gels de polyacrylamide à 7% copolymérisés avec 0,2% d'amidon ou 0,2% de glycogène, ainsi que sur des gels natifs non supplémentés en polysaccharide. Cependant, Il est apparu que, au cours des électrophorèses réalisées, la protéine RB2160 ne migrait pas dans le gel.

Ce résultat était d'autant plus inattendu que le pI théorique de l'enzyme, calculé par ProtParam sur le site ExPASy, était estimé à 5 et que l'électrophorèse s'était déroulée à pH 8. Cependant, d'autres éléments peuvent expliquer ce comportement non prévisible en solution. En effet, lors de tentatives de purification sur colonne échangeuse d'anions (monoQ), j'ai observé que la protéine RB2160 ne se fixait pas à la colonne. Un comportement similaire a été observé en utilisant une colonne échangeuse de cations (monoS). Au cours de ces chromatographies d'échange d'ions, l'enzyme a toujours été éluée à des concentrations salines faibles et son passage a systématiquement provoqué de fortes perturbations des mesures de conductivité de la solution. Il semblerait donc que RB2160 pourrait avoir un point isoélectrique réel proche de la neutralité. Au cours des études futures sur cette protéine, cette hypothèse devra être vérifiée expérimentalement par isoélectrofocalisation (IEF).

Afin de contourner ce problème, des zymogrammes en conditions dénaturantes ont été réalisés, cependant, aucune activité n'a été décelée.

II.C - Cristallogénèse

Parallèlement à la caractérisation biochimique, plusieurs productions d'enzyme ont été utilisées pour les études de cristallisation. Les tests ont été réalisés avec une solution enzymatique concentrée à environ 6 mg/mL. En tout, 672 conditions de cristallisation ont été testées à l'aide de kits commerciaux.

Sur l'ensemble des conditions testées, quelques-unes ont données un précipité. Cependant, aucune condition de cristallisation plus efficace qu'une autre n'a été mise en évidence. Des purifications à des concentrations en sels plus faibles et/ou sans glycérol ont été également testées mais sans beaucoup plus de succès.

II.D - Discussion autour de l'activité de RB2160

L'enzyme RB2160 aura été parmi les plus faciles à manipuler de toutes celles que j'ai étudiées. Le haut niveau d'expression sous forme soluble laisse à penser que la protéine adopte sa structure tertiaire correctement. De plus, les mesures de dispersion de la lumière sur DLS montrent clairement, qu'en solution, l'enzyme RB2160 est sous forme monomérique.

RB2160 ne présente cependant pas d'activité α-amylolytique, que ce soit sur l'amidon ou le glycogène. Elle aura en outre créé la surprise de ne présenter aucune activité sur tous les autres substrats qui lui ont été soumis. Il est donc envisageable soit que sa spécificité soit nouvelle, ou tout du moins différente des activités amylolytiques déjà caractérisées, soit qu'elle n'ait pu être détectée par les moyens mis en œuvre au cours de ce travail (par exemple, les produits de la réaction pourraient nécessiter d'autres méthodes de détection). Il faut souligner ici que beaucoup d'enzymes de la famille GH57 ont été caractérisées après avoir testé la croissance de l'organisme d'origine sur différents substrats. Parmi ces substrats, l'amidon est très fréquemment testé. Il est donc possible que la famille GH57 souffre d'un biais vers la caractérisation d'enzymes liées au métabolisme de l'amidon, du simple fait que ce sont celles-là qui sont recherchées préférentiellement. Il est donc très possible que d'autres spécificités de substrat, en lien ou non avec la dégradation et/ou la modification d'α-glucanes, puissent être ajoutées à la liste des activités de la famille GH57. Plusieurs preuves peuvent étayer cette hypothèse.

Tout d'abord, il existe déjà des activités non-amylolytiques dans la famille GH57 comme les exo-α-galactosidases (GalA de *P. furiosus* est une séquence brevetée - US 5958751 - de *Thermococcus alcaliphilus*). De plus, très récemment (début 2008), une nouvelle activité dans la famille GH57 a été caractérisée chez *Pyrococcus furiosus*. La protéine PF0870 semble en effet présenter une activité *β-amylolytique* (Comfort *et al.*, 2008). Cette protéine est assez divergente au sein de la famille (elle ne figure pas dans l'arbre phylogénique de la Figure III-52), même par rapport aux autres séquences archéennes. La famille GH57 étant connue pour libérer ses produits par *rétention* de la configuration anomérique, la protéine PF0870 présente une activité très surprenante. En effet, Les β-amylases (EC 3.2.1.2) sont des enzymes qui hydrolysent les chaînes d'amylose par leur extrémité non-réductrice, selon un mode d'action exo, et en libérant des molécules de β-maltose (disaccharide 4-Glc-α-(1,4)-Glc-β-1). Ceci implique un mécanisme d'hydrolyse contraire s'effectuant par *inversion* de la configuration anomérique. Il est ainsi plus probable, étant donné les conditions expérimentales utilisées par Comfort et collaborateurs (dégradation de maltotriose et de p-nitrophényl-maltose et absence de dégradation du glycogène ou de l'amidon), et les connaissances portant sur le mécanisme d'hydrolyse de la famille GH57, que la protéine PF0870 soit une exo-maltooligosaccharidase. Il est également intéressant de noter que si PF0870 présente une grande partie des résidus conservés de la famille GH57, un certain nombre n'est pas conservé. En particulier, l'acide aspartique D354, remplacé par une cystéine chez *R. baltica*, est ici remplacé par une sérine. Cependant, ces deux enzymes (PF0870 et RB2160) ne montrent que très peu de similitude entre-elles.

Enfin, il a récemment été démontré, chez l'archée hyperthermophile *Archeoglobus fulgidus*, un mécanisme de dégradation de l'amidon exogène par internalisation de maltocyclodextrine (Labes and Schonheit, 2007) (Figure III-57). Ces chaînes cycliques sont synthétisées par une maltocyclodextrine glucanotransférase de la famille GH57 qui est très semblable à la 4-α-glucanotransferase TLGT (voir partie II.B.1), ouvrant encore le champ d'action des GH57.

Cette découverte montre que des voies métaboliques originales de la dégradation de l'amidon peuvent exister et que la famille GH57 peut contenir des activités non encore découvertes. J'aborderai le métabolisme de l'amidon chez *R. baltica* dans le chapitre IV de ce manuscrit, puisqu'un certain nombre d'enzymes de son génome ont été annotées comme α-amylases ou comme ayant une activité en lien avec l'amidon.

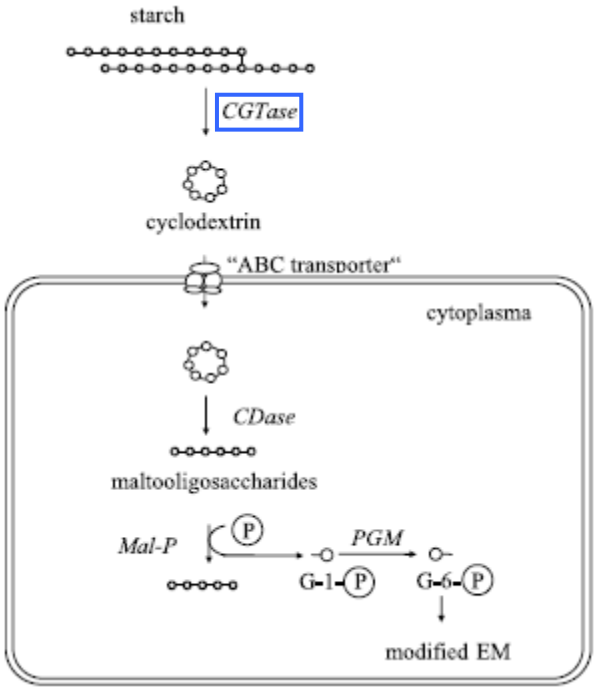


Figure III-57 : Dégradation de l'amidon par *Archaeglobus fulgidus*
D'après (Labes and Schonheit, 2007). La cyclomaltodextrine glucano-transférase (CGTase) appartenant à la famille GH57 est entourée en bleu.

Pour ce qui est de la cristallogénèse, il est surprenant qu'aucune condition n'ait donné de cristaux. L'absence de résultats en cristallogénèse est peut-être à mettre en corrélation avec l'étrange comportement de RB2160 en présence de sels (voir II.B.2). Il faudra probablement repenser le tampon utilisé lors des chromatographies d'exclusion de taille, sélectionner quelques conditions de cristallisation parmi celles présentant des précipités en modulant les concentrations en sels, pH et PEG manuellement, pour cerner la meilleure condition afin d'obtenir des cristaux de RB2160.

Toutes ces données convergent vers le fait que les études sur la famille GH57 n'en sont qu'à leurs débuts et que si RB2160 a probablement une activité liée au métabolisme de l'amidon, une approche exhaustive et systématique est encore à réaliser pour déterminer la nature exacte de son substrat. Des expériences préliminaires de dosage du glucose en solution par la glucose-oxydase ont été réalisés afin d'analyser l'éventuelle activité transférase, une mesure par les sucres réducteurs n'étant ici pas pertinente (la transglycolyse ne génère pas d'extrémités réductrices). Ces mesures laissent à penser que l'activité de RB2160 libère bien des glucoses en solution. Une des premières étapes des travaux de caractérisation sera donc de confirmer ce résultat. Il serait de plus particulièrement intéressant de tester des séries d'oligosaccharides marqués avec un fluorophore comme le p-nitrophényl-α/β-glucose, le p-nitrophényl-α/β-galactose, le p-

nitrophényl-α/β-mannose ou bien encore le p-nitrophényl-maltose, permettant de mettre en évidence un éventuel mode d'action exo.

Il faudra également élargir le champ d'investigation des chaînes polysaccharidiques aux maltodextrines courtes (quelques dizaines d'oligomères) et éventuellement cycliques. Une détermination expérimentale du pI de l'enzyme pourra également s'avérer un atout majeur dans la compréhension de son comportement en solution.

III - RB3006 : Une sialidase marine ?

III.A - La famille GH33

La famille CAZy GH33 est une famille de glycoside hydrolases tournées vers l'hydrolyse et la transglycolyse des acides sialiques (Sia). Elle est composée d'environ 300 séquences et a été assez étudiée depuis quelques décennies du fait de l'implication des acides sialiques dans divers phénomènes de pathogénicité de certains agents infectieux, en particulier chez l'homme. Les acides sialiques ne sont en effet pas des saccharides ordinaires et ils présentent une série d'originalités qui les distinguent de l'ensemble des saccharides rencontrés dans le vivant (Angata and Varki, 2002).

En premier lieu, tandis que la plupart des oligosaccharides sont constitués de sucres à cinq ou six carbones, les acides sialiques présentent la grande singularité d'être des sucres cétoacides à neuf carbones. Ils sont typiquement construits autour d'un noyau neuraminique (Neu) composé par la condensation d'une molécule de pyruvate sur le carbone réducteur d'une molécule neutre de 3-déhydroxy-3-amino-mannose ou mannosamine (ManN) (**Erreur ! Source du renvoi introuvable.**). La principale modification rapportée sur ce noyau saccharidique est l'acétylation de l'amine pour former l'acide N-acétyl-neuraminique (Neu5Ac) qui est la forme plus rencontrée dans le vivant. Beaucoup d'autres modifications du noyau neuraminique peuvent cependant survenir, voire même se cumuler. Plus de cinquante variants ont ainsi été identifiés jusqu'à aujourd'hui par O-acétylations, O-méthylations, N-glycolylations, déhydroxylations, phosphatations, ….

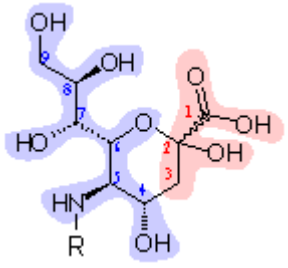


Figure III-58 : L'acide neuraminique.
Présentation de l'acide neuraminique, formé par la condensation d'un pyruvate (en rouge) et d'un mannosamine (en bleu).

La principale modification de ce composé est la forme aminoacétylée où R- = CH_3CO-

Il est en outre à noter que les bactéries gram négatives, les plantes et les algues vertes expriment également des molécules similaires mais à huit carbones (les acides 2-keto-3-

deoxy-D-manno-octulosoniques ou KDOs), selon des voies métaboliques similaires, par la condensation d'un pyruvate sur le carbone réducteur d'un arabinose.

Une deuxième grande originalité des acides sialiques réside dans leur faible et très sélective répartition dans les règnes du vivant. En effet, ils ne sont répertoriés jusqu'à présent que dans la lignée des deutérostomes pour les eucaryotes (regroupant les vertébrés et quelques invertébrés) et dans quelques phyla chez les bactéries (essentiellement les actino-bactéries) (Figure III-59). Il a été reporté dans une unique espèce d'archée (*Methanocaldococcus jannaschii*) la présence d'une putative synthase d'acides sialiques. Notons que la biosynthèse d'acides sialiques n'a pas été démontrée chez cette dernière, la reconnaissance d'une telle voie métabolique n'existant à l'heure actuelle que dans les annotations de son génome (Bult *et al.*, 1996; Graham *et al.*, 2001).

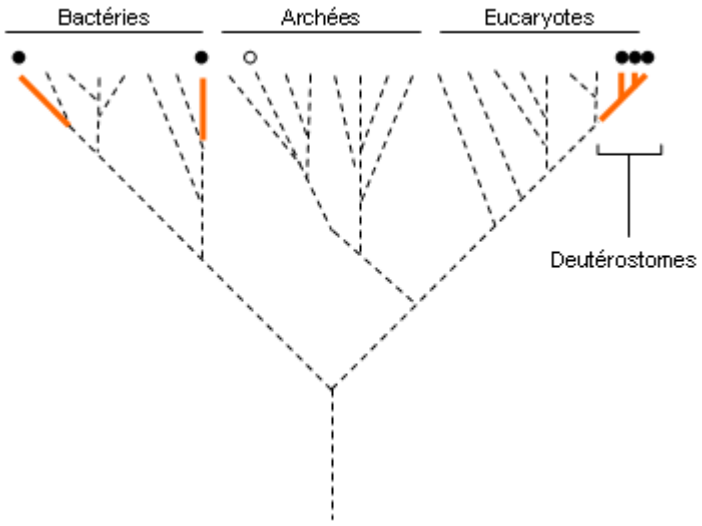


Figure III-59 : Répartition des acides sialiques dans l'arbre de la vie.
Cet arbre met en évidence les lignées utilisant les acides sialiques (en orange). Figure inspirée de (Angata and Varki, 2002).

Une troisième originalité des acides sialiques provient de la nature de leur sucre donneur sous forme de nucléotide-sucre. Ils sont en effet les seuls sucres à l'heure actuelle dans le règne vivant à être activés sous forme de cytidine monophosphate CMP-Sia.

Enfin, une dernière originalité, et non des moindres, réside dans l'utilisation des acides sialiques chez les organismes les exprimant. Ils ne sont en effet que très peu rencontrés sous forme de polysaccharides. Ils occupent en revanche typiquement la partie distale des chaînes polysaccharidiques liées aux membranes, ce qui les place en ligne de front dans la

reconnaissance intercellulaire au sein d'un organisme ainsi que dans l'interaction de l'organisme avec son environnement. C'est cette position très particulière qui explique que beaucoup d'organismes pathogènes ou commensaux des animaux ont des neuraminidases ou des lectines spécifiques aux acides sialiques. Dans le cas d'organismes pathogènes, ces protéines sont d'ailleurs considérées comme faisant partie des premiers facteurs de virulence (certaines sont mêmes obligatoires pour l'infection) voire comme toxines quand elles sont libérées sous forme soluble dans le milieu (Gaskell *et al.*, 1995; Angata and Varki, 2002). Elles sont ainsi rencontrées chez certains virus tel les virus grippaux et hépatiques où elles prennent alors le nom d'hémagglutinines, en raison de la capacité de la plupart d'entres elles à agglutiner les globules rouges, chez certaines bactéries telles *Escherichia coli* (bactérie commensale de l'intestin de l'homme) ou *Vibrio cholerae* (agent responsable du choléra) où elles prennent le nom d'adhésines quand elles permettent la reconnaissance (ou la fixation) du pathogène dans l'organisme hôte et chez certains protistes tel *Plasmodium falciparum* (agent responsable de la malaria).

La grande originalité de répartition et de structure des acides sialiques se retrouve au niveau des enzymes catalysant leur ajout ou leur hydrolyse sur des chaînes polysaccharidiques. Il est ainsi très intéressant de noter que non seulement peu de familles de la banque de données CAZy possèdent une activité sialidase, avec quatre à l'heure actuelle (familles GH-33, 34, 58 et 83), mais qu'en plus ces familles sont toutes monospécifiques et comprennent soit l'activité 3.2.1.18 (exo-α-sialidase) pour les familles GH-33, 34 et 83 (qui font d'ailleurs partie du même clan structural GH-E), soit l'activité 3.2.1.129 (endo-α-sialidase) pour la famille 58. La famille GH33 a de plus été rapportée capable de catalyser la transglycolyse d'acides sialiques (Campetella *et al.*, 1994).

Famille CAZy	Nombre de séquences	Nombre de structures (et codes PDB publiés)	Répartition
GH-33	319	10 (67)	E, B, v
GH-34	9 401	10 (83)	V
GH-83	430	5 (18)	V
GH-58	12	1 (2)	V, B

Tableau III-17 : Les sialidases dans la banque CAZy.
Répartition des sialidases chez les quatre familles CAZy présentant cette activité. La dernière colonne décrit la répartition des séquences dans l'arbre de la vie, dans l'ordre décroissant du nombre de séquences par phylum ; les phyla majoritaires sont représentés en majuscule (E : Eucaryotes, B : Bactéries, V : Virus)

Ces enzymes sont certes retrouvées dans les organismes produisant des acides sialiques mais aussi et surtout chez les virus, qui recensent 97 % des 10 200 séquences de sialidases connues. Le rôle central de ces enzymes dans l'infection de ces virus mais également dans celle de beaucoup de pathogènes en a fait les sujets de nombreuses études biomédicales et un grand nombre de structures ont été publiées au sein de ces quatre familles (Tableau III-17).

La première structure publiée dans la famille GH-33 remonte à 1994 avec la sialidase nanH de *Salmonella typhimurium*, sérotype de *Salmonella choleraesuis* (Crennell *et al.*, 1996). Notons que sur les dix protéines dont la structure a été résolue dans cette famille, huit appartiennent à des parasites ou des agents responsables de maladies chez l'homme : les bactéries *Clostridium perfringens* (responsable de gangrènes gazeuses), *Salmonella typhimurium* (responsable d'intoxication alimentaires), *Streptococcus pneumoniae* (responsable de pneumonies et de méningites) et *Vibrio cholera* (agent du choléra), ainsi que les eucaryotes *Macrobdella decora* (sangsue), *Trypanosoma cruzi* (agent de la trypanosomiase américaine) et *Trypanosoma rangeli* (parasite de l'homme non pathogène). Les séquences non issues de ce type d'organismes proviennent de l'homme avec la sialidase Neu2 et de la bactérie *Micromonospora viridifaciens* (qui synthétise naturellement plusieurs antibiotiques).

Les séquences des quatre familles présentent toutes le même repliement de *β-propeller* à 6 pales qui s'est vu historiquement attribué le nom de repliement « neuraminidase » (banque de données CATH code 2.120.10), même si un grand nombre de protéines non-sialidolytiques adoptent une conformation de ce type. La Figure III-60 présente la structure du domaine catalytique GH33 de la sialidase NedA de *Micromonospora viridifaciens* en complexe avec l'inhibiteur Neu5Ac2en (code PDB 1EUS, Gaskell *et al.*, 1995). Cette protéine est en réalité composée de trois modules : la partie N-terminale est constituée du domaine catalytique GH33 de 41 kDa, suivi par un module de liaison d'une centaine de résidus et d'un module de fixation du galactose de 150 résidus.

Le grand nombre de structures et leur cocristallisation avec divers inhibiteurs ont permis de mettre en évidence un mode d'action qui retient la configuration anomérique ainsi que les acides aminés catalytiques et les résidus impliqués dans la liaison au substrat (dont la plupart sont conservés sur l'ensemble des structures).

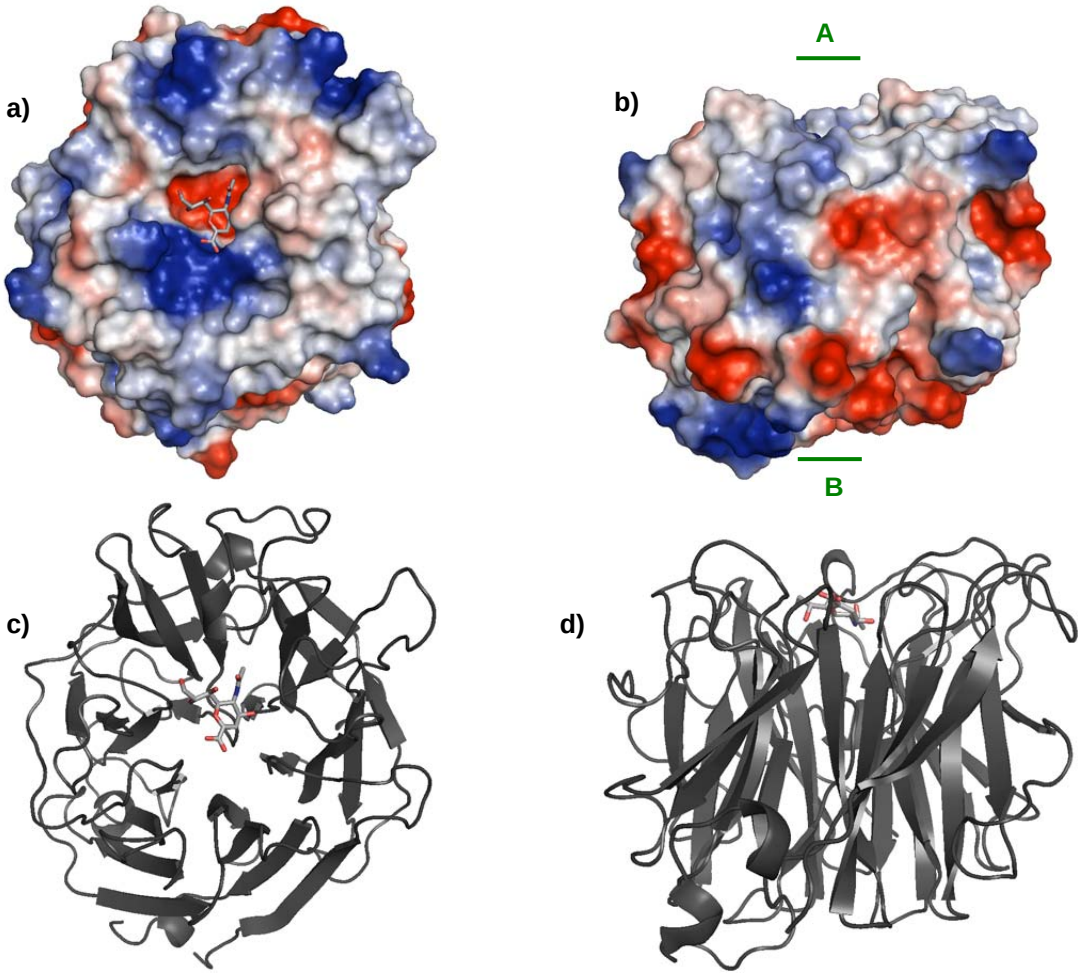


Figure III-60 : Structure de la sialidase NedA de *M. viridifaciens* (1EUS).
a-c) vue de la structure avec le site actif contenant Neu5Ac2en visible au premier plan ; b) vue en rotation de 90° perpendiculairement aux vues a et c ; les lettres **A** et **B** nomment les deux faces de la vue b

Dans le cas de 1EUS, l'attaque nucléophile est réalisée par la tyrosine Y370, stabilisée et orientée vers le carbone C2 du substrat (siège de la liaison O-glycosidique) grâce à une liaison hydrogène avec l'acide glutamique E260. Cette attaque crée un carbanion au niveau de C2 et un proton issu d'une molécule d'eau est ensuite transféré sur le C2. En effet, le donneur de proton n'est pas un résidu de la protéine, il semble que l'acide aspartique D92 (qui se situe de l'autre côté du substrat par rapport à Y370) stabilise une molécule d'eau qui se trouverait en position idéale pour effectuer le transfert (Newstead *et al.*, 2008). Ce mécanisme pourrait permettre une grande adaptabilité de l'enzyme au pH de son environnement (Burmeister *et al.*, 1993). La Figure III-61 présente une vue stéréoscopique du site actif de 1EUS avec une mise en évidence des résidus établissant des liaisons hydrogènes avec l'inhibiteur, à savoir les arginines R68, R87, R276 et R342 ainsi que les acides aspartiques D131 et D259 (Gaskell *et al.*, 1995; Newstead *et al.*, 2008).

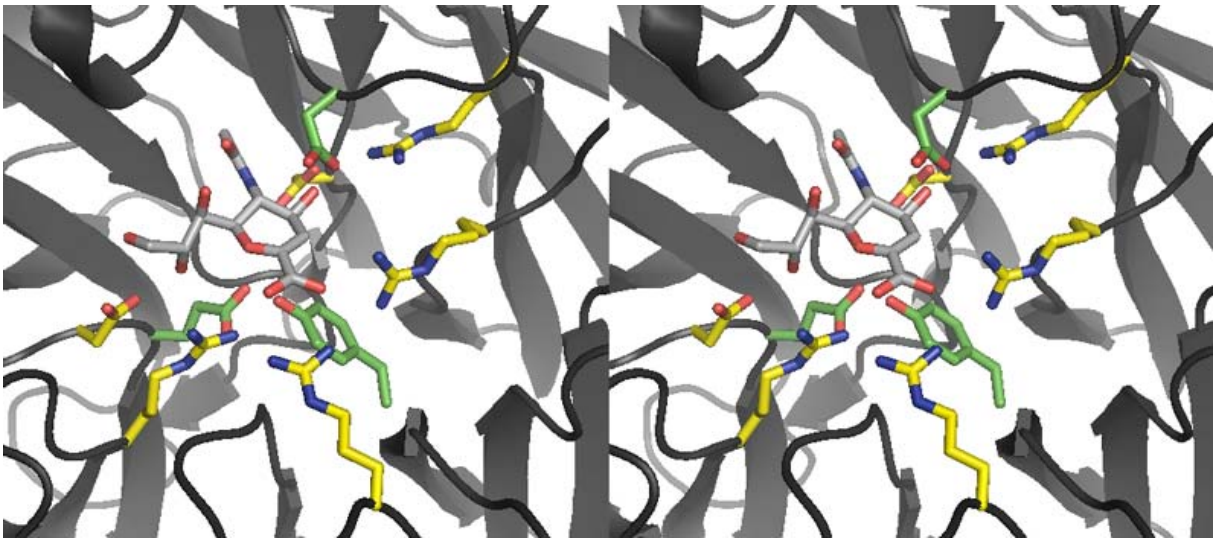


Figure III-61 : Site actif de 1EUS.
Vue détaillée du site actif de 1EUS (face **A** de la Figure III-60-b). Les acides aminés catalytiques sont colorés en vert. Les autres résidus visibles en jaune participent à la fixation du substrat, via des liaisons hydrogènes.

Un dernier élément mérite l'attention, cette fois ci au niveau des séquences de ces enzymes. En effet, les sialidases de la famille CH33 présentent plusieurs motifs d'environ huit résidus, répétés de trois à cinq fois dans les séquences, et appelés « Asp-box » dont la forme canonique est xxxYSxDxG[KR]xWxx où les « x » correspondent à des résidus équivalents dans les structures (Roggentin *et al.*, 1989; Russel, 1998; Copley *et al.*, 2001). Les pales des *β-propellers* de la famille GH33 sont formées d'un feuillet de quatre brins β antiparallèles ; les motifs Asp-box forment systématiquement une épingle β (*β-hairpin*) entre les troisième et quatrième brins de chaque feuillet en périphérie des β-propellers (Figure III-62). Les résidus aspartiques de chaque motif sont toujours exposés vers le solvant, tandis que les résidus aromatiques sont imbriqués entre deux feuillets adjacents. La fonction de ces motifs reste encore inconnue : il est proposé qu'ils pourraient simplement contribuer à la stabilité du repliement de la protéine, comme le suggère leur rencontre récurrente dans des repliements différents parmi les structures de la PDB ou qu'ils soient impliqués dans une fonction annexe de la protéine, étant donné que toutes les séquences les possédant ont des activités glycoside hydrolases ou apparentées. Copley et collaborateurs (Copley *et al.*, 2001) proposent en outre différents scénario pouvant expliquer une telle conservation entre des organismes très différents. Enfin, un éventuel rôle dans des mécanismes de sécrétion est également évoqué (Gaskell *et al.*, 1995), étant donné que l'immense majorité des protéines présentant ce motif sont également sécrétée.

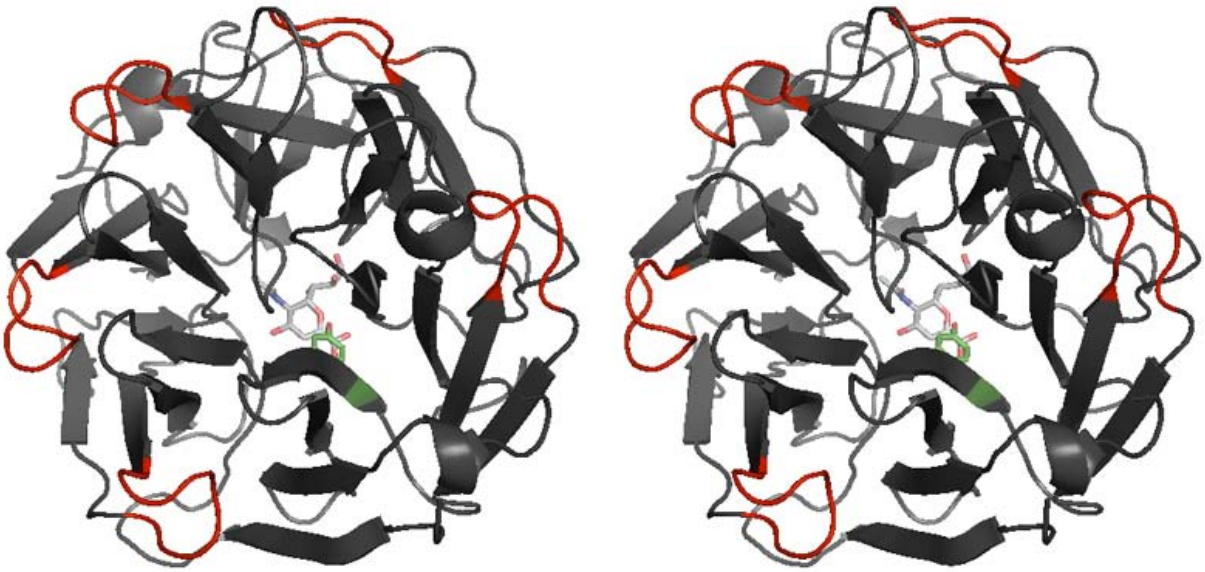


Figure III-62 : Motifs Asp-box dans la structure 1EUR.
Répartition des motifs Asp-box (en rouge) dans la structure 1EUR sur la face B de la Figure III-60-b. Ces motifs se répartissent en périphérie de chaque pale du *β-propeller* au sein d'épingles β entre deux pales adjacentes. L'inhibiteur Neu5Ac2en et la tyrosine catalytique sont observables en arrière plan.

III.B - Résultats et discussion

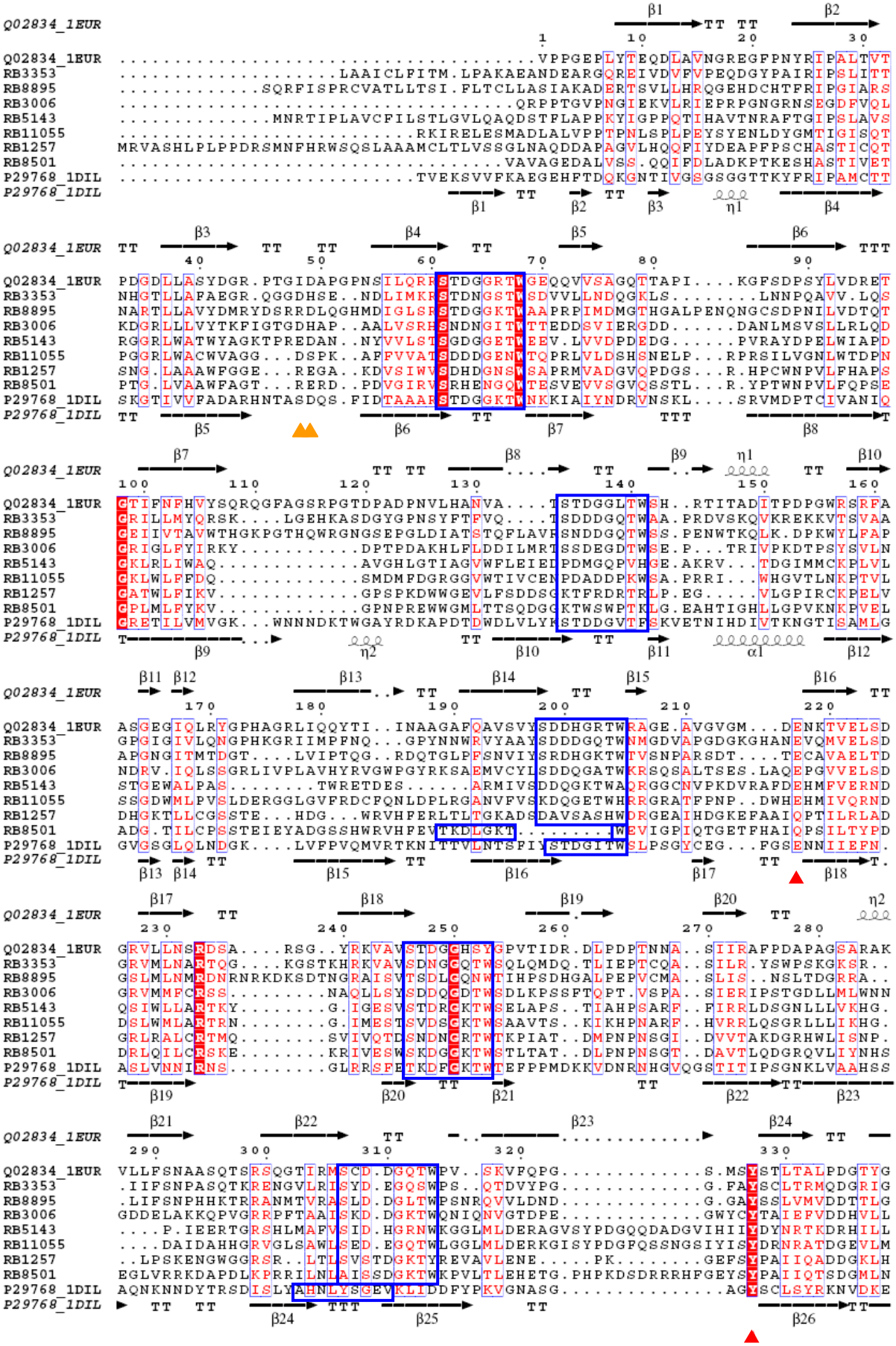
III.B.1 - Analyse bioinformatique

Rhodopirellula baltica présente un grand nombre de sialidases de la famille GH33 avec sept représentants présentés dans le **Tableau III-18**.

Protéine	Modularité	Taille résidus	Taille kDa	Annotation initiale
RB1257	GH33	375	41,4	Conserved hypothetical protein
RB3006	GH33 – UNK1 – UNK2	1153	128,8	probable sialidase (3.2.1.18)
RB3353	GH33	409	45	Sialidase precursor (3.2.1.18)
RB5143	GH33	479	53	Conserved hypothetical protein
RB8501	GH33	381	42,3	Putative uncharacterized protein
RB8895	GH33	418	45,8	Neuraminidase precursor (3.2.1.18)
RB11055	NucH – GH33	733	81	Conserved hypothetical protein

Tableau III-18 : Représentants de la famille GH33 chez *R. baltica*.
Les peptides signaux sont représentés par un module grisé.

Les sept séquences de la famille GH33 de *R. baltica* ont été alignées avec deux séquences caractérisées et dont les structures sont connues : Q02834_MICVI (sialidase de *M. viridifaciens*, code PDB 1EUR) et P29768_SALTY (sialidase de *S. typhimurium*, (Crennell *et al.*, 1996) - code PDB 1DIL). Cet alignement a été complété par les informations structurales des deux enzymes, permettant de vérifier si l'ensemble de ces enzymes présente bien le couple catalytique tyrosine + glutamate d'une part, et l'aspartate pressenti pour la stabilisation d'un donneur de proton d'autre part (Figure III-63). Cet alignement s'est révélé intéressant à plus d'un titre. Tout d'abord, il met en évidence le fait que les protéines RB1257 et RB8501 n'ont précisément pas le glutamate du couple catalytique (muté en glutamine), alors que leur séquence laisse clairement apparaître les deux autres résidus nécessaire à la catalyse, laissant ouvertes les interrogations sur leur activité. D'autre part, il souligne l'existence des motifs Asp-box au sein de toutes les séquences, avec cependant une flexibilité plus importante dans la conservation de ces résidus pour les séquences de *R. baltica*.



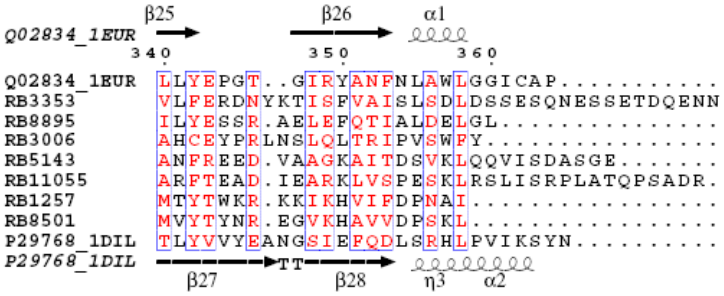


Figure III-63 : Alignement des sept sialidases de *R. baltica*.
Alignement des sept sialidases de *R. baltica* avec les séquences des structures 1EUR de *M. viridifaciens* et 1DIL de *S. typhimurium*. Les résidus catalytiques sont matérialisés par des triangles rouges (▲). L'acide aspartique stabilisant potentiellement un donneur de proton est matérialisée par un triangle orange (▲). Les motifs Asp-box sont matérialisés par des encadrés bleus en surimpression. Figure extraite de ESPript après alignement avec Multalign.

Une analyse complémentaire de phylogénie a également été menée sur l'ensemble des séquences caractérisées de la famille GH33, après alignement multiple dans MAFFT. L'arbre généré est robuste avec de bonnes valeurs de bootstrap moyennes et il semble possible d'en donner une interprétation fiable (). Les enzymes de *R. baltica* semblent s'apparier pour former des sous-groupes de noeuds assez forts mais ces paires ne s'imbriquent pas dans les groupes formés par les autres enzymes de la famille. Seule RB8895 se sépare de ses paralogues et forme un groupe robuste, mais encore une fois séparé du reste de l'arbre, avec P31206 (sialidase de *Bacteroidetes fragilis*). Il semble de plus que le fait que l'enzyme P37060 (séquence de la structure de code PDB 1KIT) forme un sous-groupe avec RB3006 et RB3353 soit artéfactuel, en accord avec les faibles valeurs de bootstrap de leur nœud, car ces séquences sont difficilement alignables tandis que RB3006 présente une similitude de 39 % (22 % d'identité) sur la longueur de son domaine catalytique GH33 avec Q02834 (séquence de la structure de code PDB 1EUR précédemment présentée), présente dans un groupe plus lointain. Enfin, les protéines RB1257 et RB8501 d'une part et RB11055 et RB5143 d'autre part forment deux couples à très forte valeur de bootstrap (100 % et 98 % respectivement). Cela semble cohérent avec les observations de l'alignement de la Figure III-63, où il est mis en évidence la non conservation chez RB1257 et RB8501 d'un des résidus catalytiques. Ces enzymes dont l'une possède un signal peptide et l'autre non pourraient donc présenter un mécanisme catalytique original, voire être utilisées pour des fonctions non catalytiques.

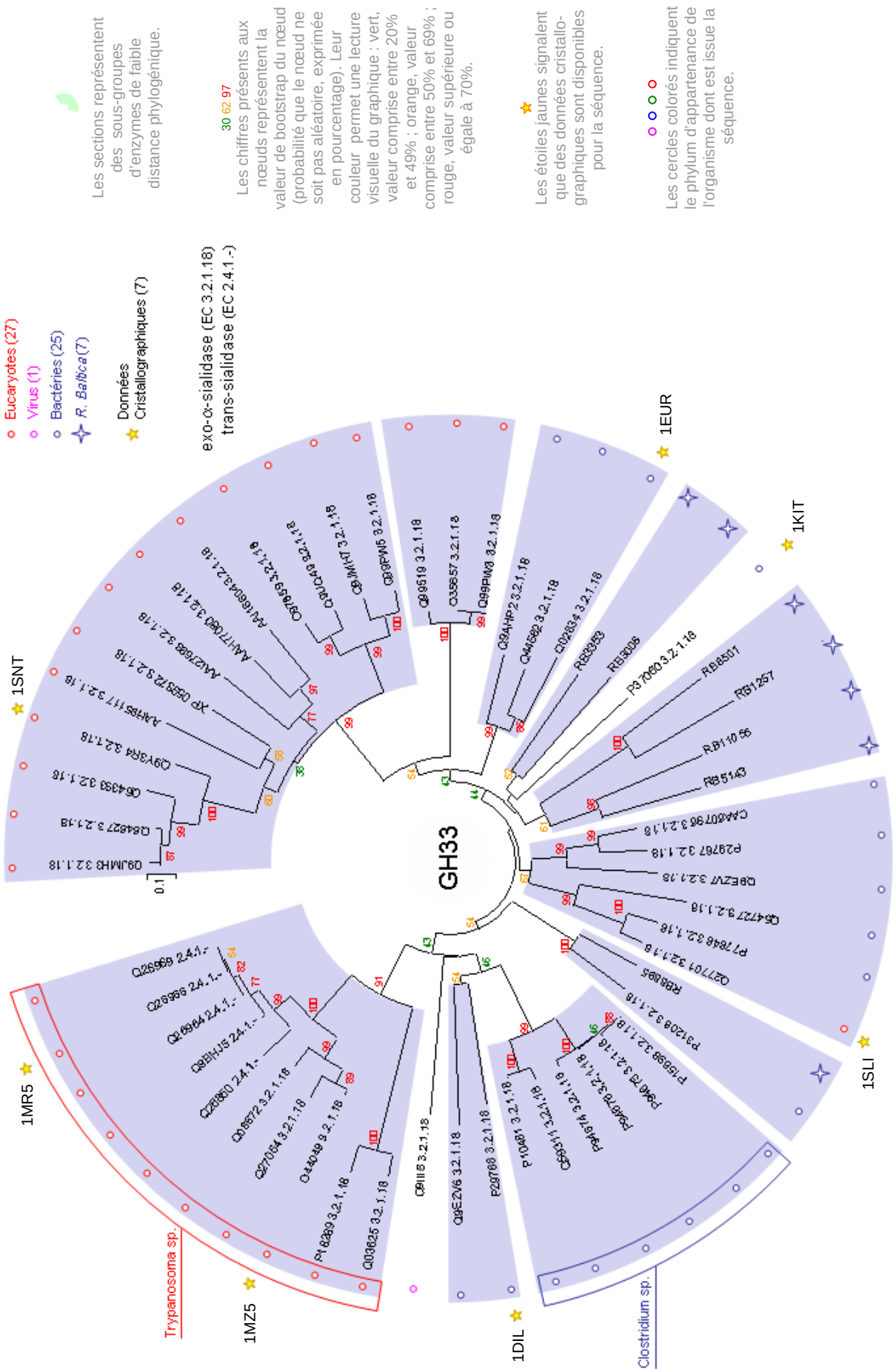


Figure III-64 : Arbre phylogénique de la famille GH33

Arbre phylogénique généré par la méthode d'évolution minimum, avec délétion des indels., à partir d'un alignement de type FFT-NS-i par MAFFT.

Enfin, il apparaît que deux des enzymes de *R. baltica* présentent une structure multimodulaire : RB11055 a ainsi un domaine ribonucleoside hydrolase de 350 résidus sur son extrémité N-terminale, tandis que RB3006 présente deux modules de fonction inconnue de tailles comparables au module catalytique GH33 (environ 400 résidus). Ces modules additionnels sont assez mystérieux quant à leur fonction potentielle. En particulier, ceux de RB3006 étaient orphelins au début de ma thèse. Au fur et à mesure que des génomes d'organismes proches ont été séquencés, ils ont tous les deux intégrés une famille de domaines conservés. Certains de leurs homologues sont de plus des protéines à part entière, ce qui ajoute encore aux interrogations sur leur fonction et ouvrent potentiellement la voie à des activités enzymatiques nouvelles reliées au métabolisme des acides sialiques. Ainsi, et de manière très intéressante, un alignement du domaine UNK1 réalisé avec ses récents homologues A6CCJ5_PLAMA (*P. maris*), B4D319_CHTFL (*C. flavus*) et A3ZZ97_BLAMA (*B. marina*), fait apparaître que ces protéines sont très conservées, avec une similitude moyenne supérieure à 50 % (Figure III-65). Des motifs ressemblant de manière distante aux Asp-box des enzymes de la famille GH33 pourraient de plus laisser penser à un module ayant évolué depuis ces sialidases.

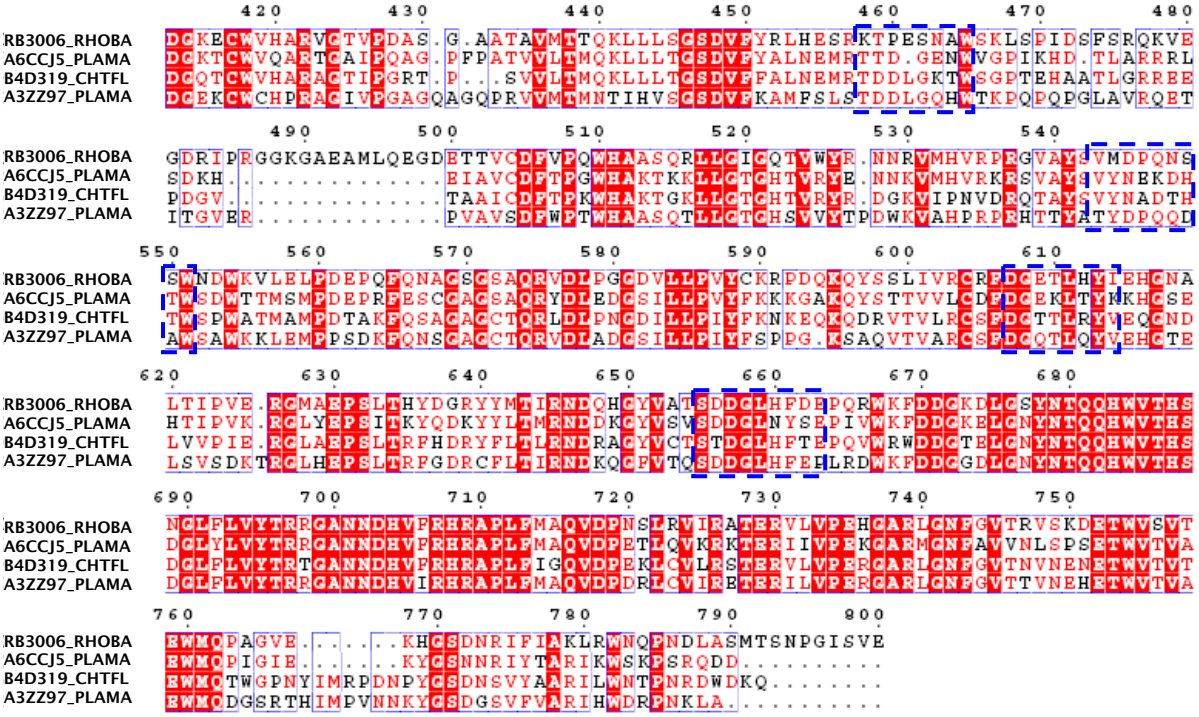


Figure III-65 : Alignement du domaine UNK1 avec trois homologues.
Alignement du domaine UNK1 de *R. baltica* avec ses homologues. La numérotation indiquées reprend la séquence de RB3006. Des motifs conservés ayant une structure similaire aux Asp-box des séquences de la famille GH33 sont encadrés en bleu.

Même si ces deux modules présentent une faible similitude entre eux, il est également possible de constater que UNK1 et GH33 de RB3006 ont quelques éléments dispersés dans leur séquence plutôt alignables suggérant effectivement une relation ancestrale entre les deux séquences. Une certaine similitude est en particulier visible au niveau du site actif du domaine GH33. Il apparaîtrait alors, dans le cas où les modules seraient effectivement apparentés, que UNK1 ne possède plus la tyrosine catalytique, suggérant un rôle non catalytique au sein de RB3006 (Figure III-66).

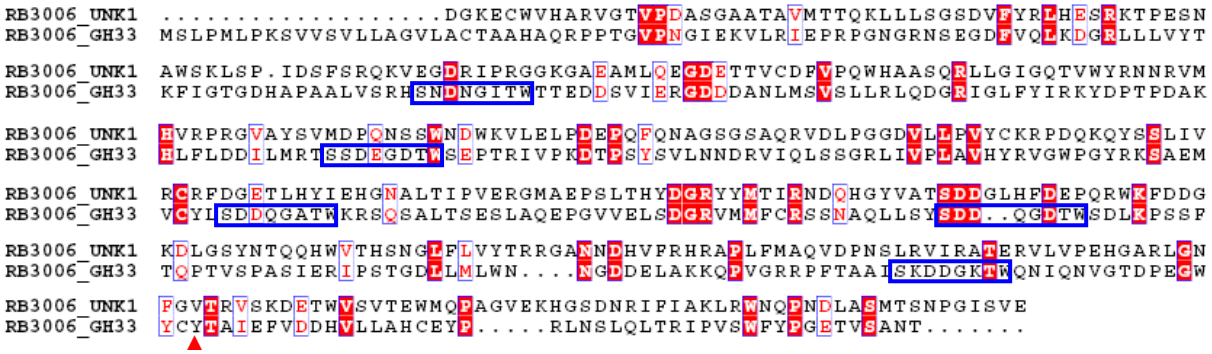


Figure III-66 : Alignement des modules UNK1 et GH33 de RB3006.
Les motifs Asp-box du module GH33 sont entourés en bleu. La tyrosine catalytique est marquée par un triangle rouge (▲)

Enfin, une dernière retombée de ces alignements a été qu'il s'est avéré que les extrémités N- et C-terminales des homologues du module UNK1 correspondaient très raisonnablement aux limites que nous avions fixées à ce module. Cela nous a confirmé *a posteriori* la validité de ce type d'analyse qui, toute éclairée soit elle de différentes analyses complémentaires (cf I-A-1), n'en reste pas moins arbitraire.

Le module UNK2 de RB3006 n'a pas présenté suffisamment de similitude avec les deux autres modules de la protéine pour émettre des hypothèses sur sa fonction potentielle.

C'est fort de ce contexte que RB3006 a été choisie pour figurer parmi les cibles de la plaque pour les tests de surexpression. Elle a été clonée sous trois formes : le module catalytique GH33 (emplacement B12), le module de fonction inconnue UNK1 (emplacement F10) et le module UNK2 (emplacement G02). Deux de ses modules présentant une forte expression soluble (modules GH33 et UNK1) ont été intégrés aux cibles finales destinées à être caractérisées biochimiquement et par cristallographie. La Figure III-67 et le Tableau III-19 présentent respectivement la séquence et les propriétés biochimiques théoriques de RB3006.

Séquence protéique de RB3123 :

```
   1 MSLPMLPKSV VSVLLAGVLA CTAAHAQRPP TGVPNGIEKV LRIEPRPGNG RNSEGDFVQL
  61 KDGRLLLVYT KFIGTGDHAP AALVSRHSND NGITWTTEDD SVIERGDDDA NLMSVSLLRL
 121 QDGRIGLFYI RKYDPTPDAK HLFLDDILMR TSSDEGDTWS EPTRIVPKDT PSYSVLNNDR
 181 VIQLSSGRLI VPLAVHYRVG WPGYRKSAEM VCYLSDDQGA TWKRSQSALT SESLAQEPGV
 241 VELSDGRVMM FCRSSNAQLL SYSDDQGDTW SDLKPSSFTQ PTVSPASIER IPSTGDLLML
 301 WNNGDDELAK KQPVGRRPFT AAISKDDGKT WQNIQNVGTD PEGWYCYTAI EFVDDHVLLA
 361 HCEYPRLNSL QLTRIPVSWF YPGETVSANT PAESQTAPLD YAVSLEVTHE GFDGKECWVH
 421 ARVGTVPDAS GAATAVMTTQ KLLLSGSDVF YRLHESRKTP ESNAWSKLSP IDSFSRQKVE
 481 GDRIPRGGKG AEAMLQEGDE TTVCDFVPQW HAASQRLLGI GQTVWYRNNR VMHVRPRGVA
 541 YSVMDPQNSS WNDWKVLELP DEPQFQNAGS GSAQRVDLPG GDVLLPVYCK RPDQKQYSSL
 601 IVRCRFDGET LHYIEHGNAL TIPVERGMAE PSLTHYDGRY YMTIRNDQHG YVATSDDGLH
 661 FDEPQRWKFD DGKDLGSYNT QQHWVTHSNG LFLVYTRRGA NNDHVFRHRA PLFMAQVDPN
 721 SLRVIRATER VLVPEHGARL GNFGVTRVSK DETWVSVTEW MQPAGVEKHG SDNRIFIAKL
 781 RWNQPNDLAS MTSNPGISVE TTAYCKPPQA MTEELGDYRS PLIFENGTRV PHASQWPQRR
 841 KEIQTRWESL LGKWPKPITD PQVTISETVH LDSVTKHTIE FQWTPNEKAT AYLLVPNTVE
 901 HADHDLPAVL SVYYEPETAI GLGKPHRDFA LQLAHRGFVT LSIGTTEATE AKTYSLYHPS
 961 IDDASVQPLS MLAYAATTAW QVLADRPEVD PNRIGVVGHS FGGKWAMFAA CLSERFACGA
1021 WSDPGIVFDE SMSGVNYWEP WYLGYHPKPW RKRGLITQDN PARGLYPRLI AQGHDLHELH
1081 ALMAPRPFLV SGGSADPIRR WTALNHSVAV NALLGHDDRV AMTNRADHSP NEDSNSVLYA
1141 FFEKHLAPSD VSL
```

Figure III-67 : Séquence protéique de RB3006.
Le module GH33 est représenté en bleu, le module UNK1 est représenté en vert, le module UNK2 est représenté en orange, les zones grises représentent respectivement le peptide signal et la zone charnière entre le module GH33 et le module UNK1.

	Protéine sauvage	Module GH33 cloné	Module UNK1 cloné	Module UNK2 cloné
Poids moléculaire	128 700 Da	40 900 Da	46 800 Da	40 000 Da
pI théorique	5,8	5,54	6,5	5,9
Nombre d'acides aminés	1153	364	413	367
Asp + Glu	141	49	50	41
Arg + Lys	113	38	45	31
Cystéines	12	4	4	3
$\varepsilon_{280\,nm}$ ($M^{-1}.cm^{-1}$)	227 800	67 600	80 120	80 000

Tableau III-19 : Récapitulatif des données biochimiques théoriques de RB3006.
Données extraites du logiciel ProtParam disponible sur le site ExPASy.

III.B.2 - Résultat d'expression, de purification et de caractérisation biophysique

Les résultats de surexpression en culture de 200 mL des modules GH33 et UNK1 de RB3006 ont confirmés ce que les tests préliminaires laissaient supposer : ces peptides ont présenté une très bonne expression soluble. Elles ont été purifiées à l'homogénéité électrophorétique par deux étapes de purification : une colonne d'affinité au nickel suivie par une colonne d'exclusion de taille (Superdex 75). Une estimation du rayon de giration et de la monodispersité en solution des deux modules a été réalisée par mesure de dispersion de la lumière (DLS) des solutions. Ces résultats ont confirmés ceux de sortie d'exclusion de taille, à savoir qu'elles sont toutes les deux monomériques. L'absence d'agrégation en solution, couplé à leur grande stabilité au cours du temps démontre le repliement correct de ces deux modules protéiques.
Les résultats respectifs de purification des enzymes sont présentés ci-après.

III.B.2.1 Purification et caractérisation du module GH33 de RB3006

Le module catalytique GH33 de RB3006 a été élué à une concentration en imidazole de 140 mM environ, sur un gradient d'imidazole allant de 50 mM à 500mM (**Figure III-68**).

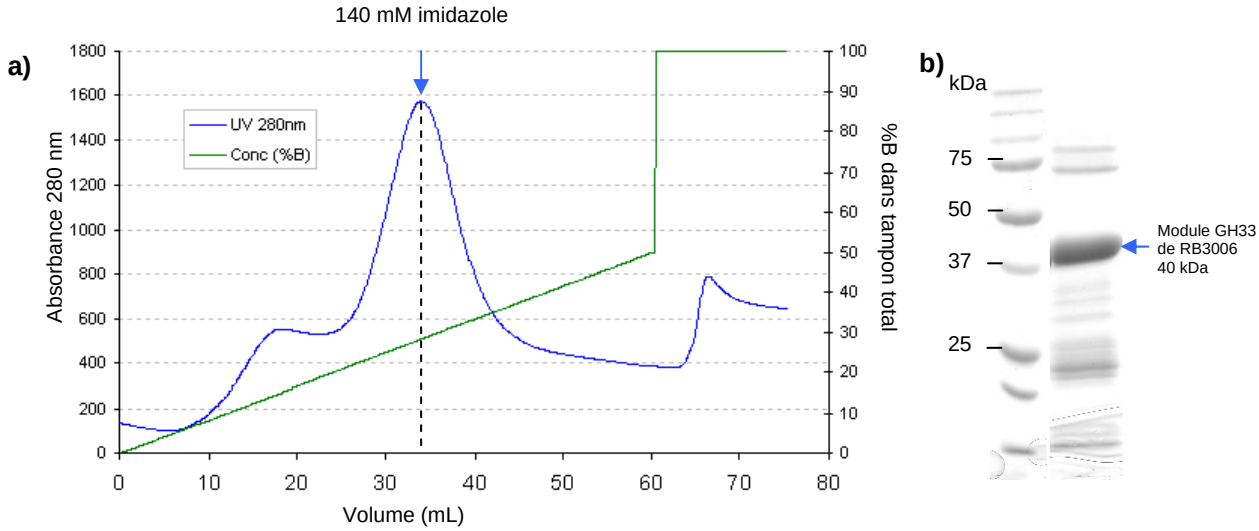


Figure III-68 : Purification de RB3006 (GH33) (Colonne d'affinité).
a) Chromatogramme de purification sur colonne d'affinité au nickel du module GH33 de RB3006 ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

L'élution sur colonne d'exclusion de taille s'est réalisée pour un volume d'environ 80 mL, ce qui est cohérent avec la taille attendue de 40 kDa pour cette protéine (**Figure III-69**).

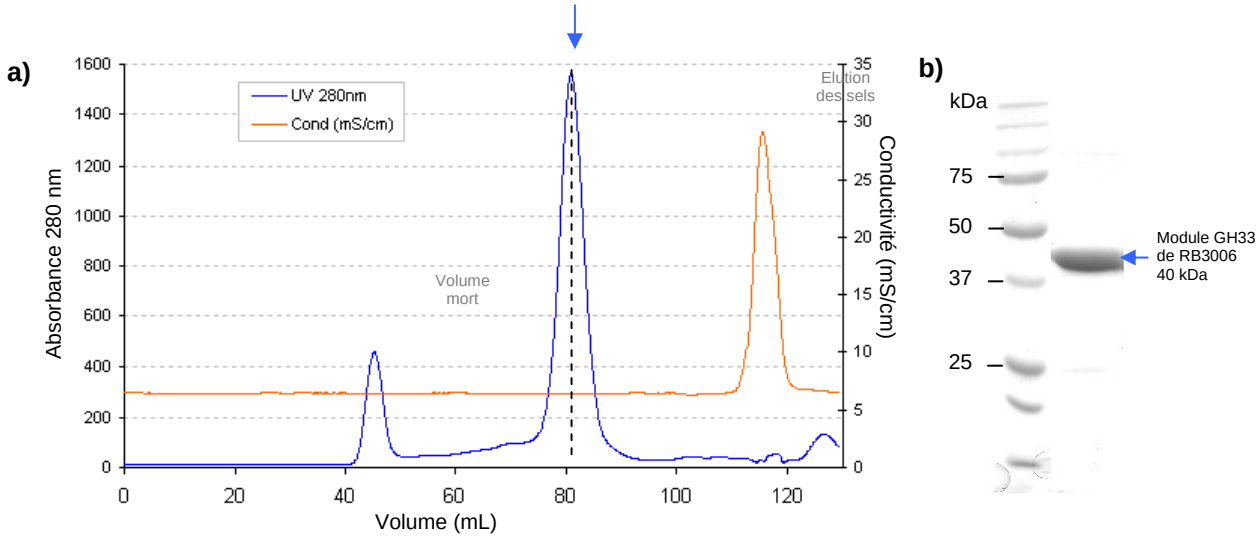


Figure III-69 : Purification de RB3006 (GH33) (Colonne d'exclusion de taille).
a) Chromatogramme de purification sur colonne d'exclusion de taille du module GH33 de RB3006. Quelques agrégats sont visibles dans le volume mort ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

Le rendement moyen des différentes productions du domaine catalytique GH33 de RB3006 a été de 20 mg de protéine purifiée / L de culture.

III.B.2.2 Purification et caractérisation du module UNK1 de RB3006

Au cours de l'étape de purification sur la colonne d'affinité au nickel, le module catalytique UNK1 de RB3006 a été élué de manière très surprenante durant le lavage de la colonne avec le tampon A (la partie Matériels et Méthodes de ce chapitre), soit à une concentration en imidazole de 65 mM environ (Figure III-70). Il est probable que cette faible fixation de la protéine à la colonne soit artéfactuelle. En effet, la protéine sort assez pure de la colonne et une petite quantité de protéine apparaît ne pas se fixer du tout à la colonne. Comme la production de cette enzyme était particulièrement bonne, il est probable que la protéine a d'une manière ou d'une autre saturée la colonne abaissant sa propre affinité au nickel par autocompétition. En perturbant la fragile fixation de la protéine à la résine, le premier lavage (qui contient 4 % de tampon B – soit 65 mM imidazole) a dû provoquer une élution massive de ces protéines. Cela est en outre confirmé par la « traîne » d'élution qui a suivie : des traces de la protéine ont été visibles jusqu'aux dernières fractions d'élution en fin de gradient (vers 500 mM imidazole).

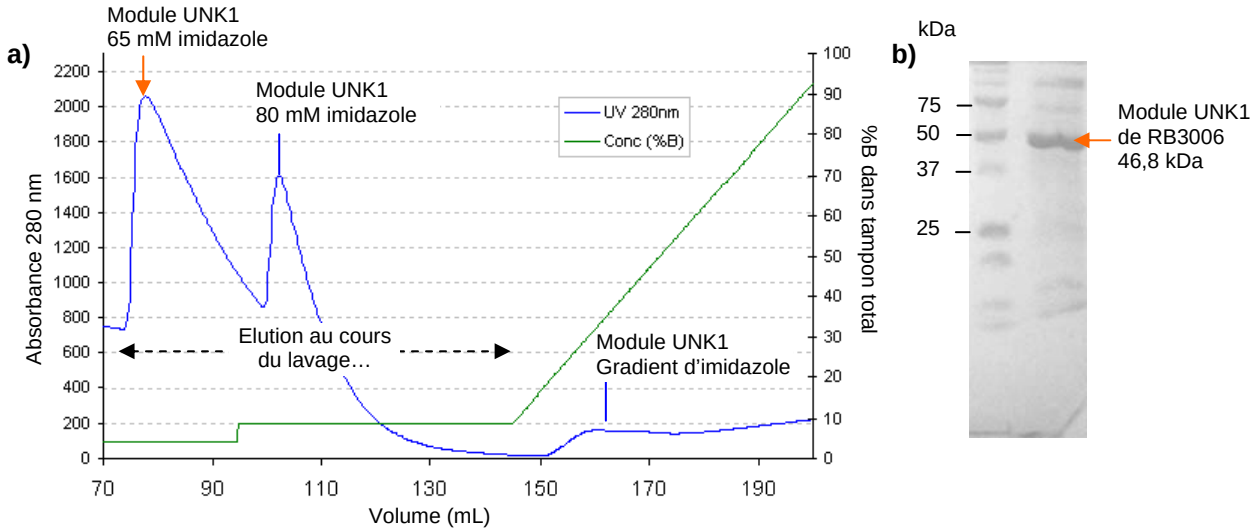


Figure III-70 : Purification de RB3006 (UNK1) (Colonne d'affinité).
a) Chromatogramme de purification sur colonne d'affinité au nickel du module UNK1 de RB3006 ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

L'élution sur colonne d'exclusion de taille s'est réalisée après 60 mL, ce qui est cohérent avec une protéine monomérique de 47 kDa. Cette étape a été par ailleurs très efficace, aboutissant à une protéine de grande pureté sur gel PAGE (Figure III-71).

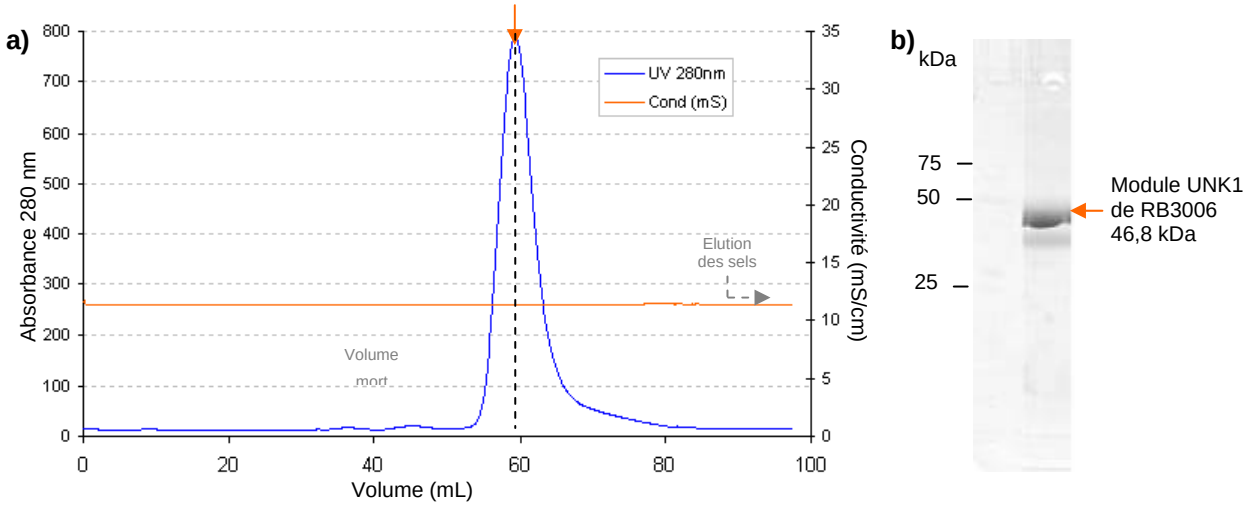


Figure III-71 : Purification de RB3006 (UNK1) (Colonne d'exclusion de taille).
a) Chromatogramme de purification sur colonne d'exclusion de taille du module UNK1 de RB3006 ; b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

Le rendement moyen des différentes productions du domaine UNK1 de RB3006 a été de 15 mg de protéine purifiée / L de culture.

III.B.3 - Tests enzymologiques

Des tests d'activités ont été réalisés sur le module catalytique GH33 de RB3006. Etant donné qu'une seule activité est recensée dans la famille, les substrats à tester se sont réduits à la dégradation d'un acide 2'-(4-méthylumbelliferyl)-N-acétyl-neuraminique (muNeu5Ac), qui est un substrat artificiel permettant de doser de l'activité exo-sialidase (Myers *et al.*, 1980). Le module catalytique de la protéine a été incubé en étuve à 37°C en présence de 5 mg/mL muNeu5Ac dans un tampon 50 mM tris-HCl pH 6,0. Aux temps 0 min, 10 min, 20 min, 50 min et 400 min, un échantillon a été prélevé et après photoexcitation à 350 nm, la fluorescence de la solution a été mesurée par un spectrofluorimètre sur un spectre de 400 nm à 500nm (Figure III-72). Il apparaît qu'aucune activité n'est décelable à ce niveau, les intensités de fluorescence étant constantes avec et sans enzymes.

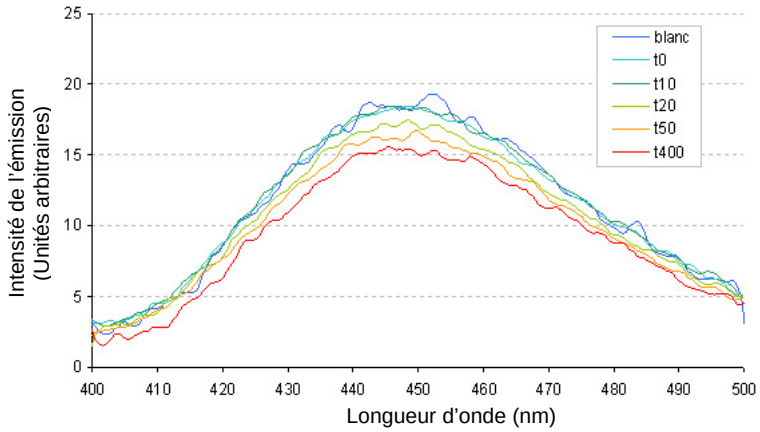


Figure III-72 : Diagramme de fluorescence : activité GH33.
Diagramme de fluorescence pour la mesure de l'activité du module GH33 de RB3006 en présence de muNeu5Ac. Le blanc correspond au substrat au temps 0 sans ajout d'enzyme.

Une mesure de la stabilité du muNeu5Ac dans les conditions expérimentales a été réalisée afin d'exclure tout artéfact de mesure. La même expérience que précédemment a été réalisée en laissant deux échantillons de muNeu5Ac en étuve à 37°C, l'un avec et l'autre sans enzyme (Figure III-73). Les résultats sont identiques, suggérant effectivement que le muNeu5Ac n'est pas stable en solution dans les conditions de l'expérience. Il est de plus connu que des conditions acides favorisent son auto-dissociation (Myers *et al.*, 1980). D'autres expériences ont été réalisées à pH neutre, sans plus de résultats. Il est possible qu'un ajustement du tampon de réaction est à effectuer. Il est également possible que d'autres facteurs interviennent, comme une activité différente, une non activité ou une

activité non conservée sous cette forme. Ceci sera discuté dans la dernière partir de cette section.

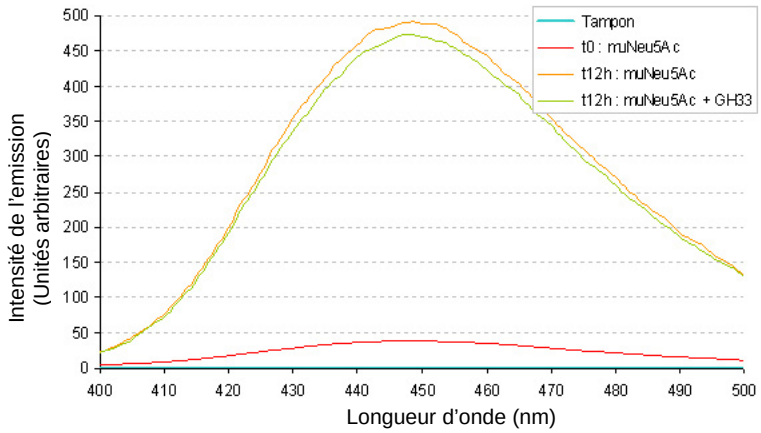


Figure III-73 : Diagramme de fluorescence : stabilité du muNeu5Ac.
Diagramme de fluorescence pour tester la stabilité du muNeu5Ac au cours du temps seul et en présence de l'enzyme.

III.B.4 - Modélisation de la structure du module GH33 de RB3006 à partir de la sialidase NedA de *M. viridifaciens*

Un modèle structural du module catalytique GH33 de RB3006 a été généré par homologie à partir de la structure de NanH de *M. viridifaciens* en complexe avec son inhibiteur Neu5Ac2en (structure de code PDB 1EUS). En effet, ces deux protéines possèdent une bonne similitude sur la longueur de leur séquence (39 % pour une identité de 25 %) et l'alignement généré à partir des séquences de ces deux protéines fait apparaître un taux assez faible d'insertions et de délétions, de plus dans des zones non adjacentes au site actif, ce qui est crucial pour l'élaboration d'un modèle valide (voir Figure III-74)

Une explication plus détaillée des critères importants dans le choix d'une bonne référence pour la modélisation des structures par homologie est donnée dans la partie I.B.4 de ce chapitre. Le modèle a été généré avec le service de modélisation du centre EBI (plateforme SwissModel) accessible via le logiciel DeepView / Swiss-PDBViewer v4.0. Ce service de modélisation peut produire des modèles structuraux par homologie à partir d'un alignement de séquence et d'un fichier de coordonnées atomiques.

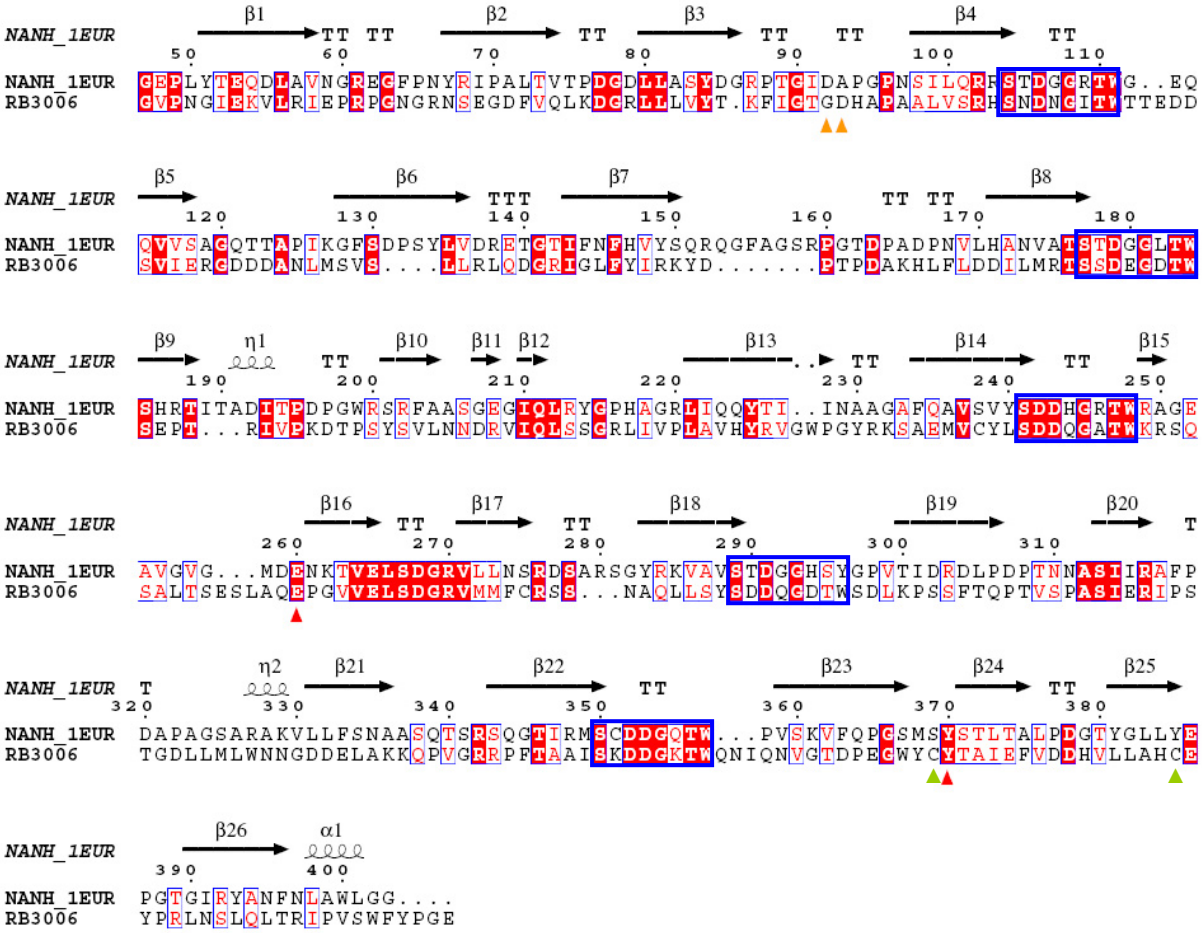


Figure III-74 : Alignement des GH33 de NanH et RB3006.

Alignement du domaine catalytique GH33 de NanH avec celui de RB3006. Les résidus catalytiques sont matérialisés par des triangles rouges (▲). L'acide aspartique stabilisant potentiellement le donneur de proton est matérialisé par un triangle orange (▲). Les triangles verts (▲) marquent un motif décrit à la fin de cette section. Les motifs Asp-box sont matérialisés par des encadrés bleus. Figure extraite de ESPript après alignement avec Multalign (et affinement à la main)

Le modèle semble raisonnablement fiable : il présente un RMSD de 0,53 Á vis-à-vis de la structure de NanH (calcul réalisé avec DeepView sur les carbones C_α). Comme expliqué dans la partie I-B-4, un modèle reste cependant une hypothèse de travail, tout ce qui suit est donc bien évidemment sujet à caution. D'autres facteurs corroborent la validité du modèle. En effet, nombre de résidus du cœur hydrophobe sont conservé, avec en particulier beaucoup de résidus aromatiques. De plus, les résidus chargés à la surface du modèle ainsi que ceux répartis le long de la séquence se superposent également à ceux de NanH. Enfin, les motifs Asp-box se positionnent sans générer de conflit stérique.

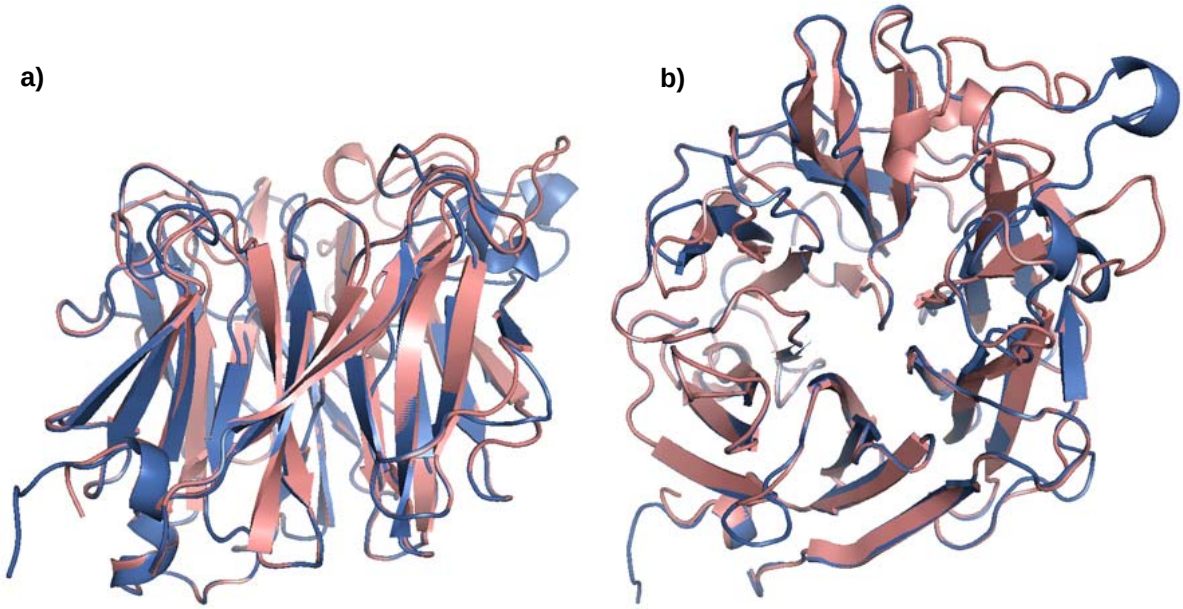


Figure III-75 : Modèle du module catalytique GH33 de RB3006.
Présentation du modèle du module catalytique GH33 de RB3006. a) et b) Vues à 90° de la superposition du modèle de RB3006 (bleu) et de 1EUR (rouge).

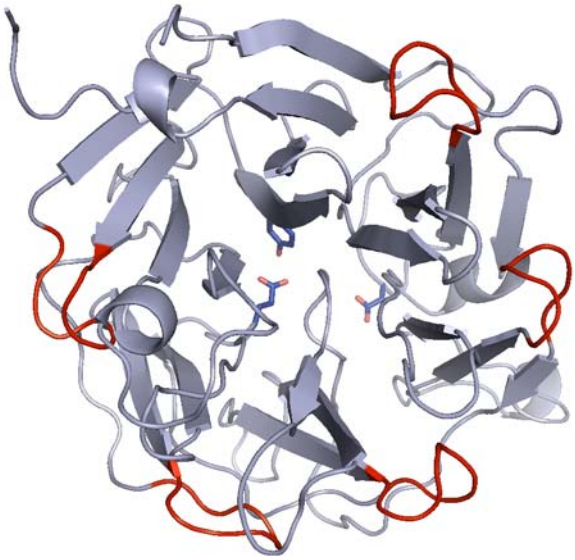


Figure III-76 : Motifs SxDxxTW.
Représentation des cinq séquences conservées SxDxxTW appartenant aux motifs Asp-box propres aux sialidases de la famille GH33.

La figure suivante met en évidence les divergences et conservations proposées pour le module catalytique de RB3006 par rapport à 1EUR. La Figure III-77 présente le site actif du module GH33 de RB3006 avec l'adoption d'un code couleur selon la conservation des résidus par rapport à 1EUR (Tableau III-20).

	RB3006	1EUR
Résidus impliqués dans la catalyse et conservés au mêmes positions[1] : Vert	E237 Y347[3]	E260 Y370[3]
Autres résidus conservés au mêmes positions[1] : Bleu foncé	R253 R316 E363	R276 R342 E386
Résidus conservés avec décalage en position[2] : Cyan	R51 D77[3]	R68 D92[3]
Nouveaux résidus proposés: Jaune	E54 R198 Q236 P285	I69 T226 D259 N311
Résidus non conservés (non représentés)	S114	D131

Tableau III-20 : Sites de fixation au substrat de RB3006 (GH33).
Présentation des acides aminés probablement impliqués dans la liaison du substrat dans le module catalytique GH33 de RB3006, classés en fonction de leur niveau d'identité avec leur référence structurale. [1] : Ces résidus se superposent quasi-parfaitement avec ceux de 1EUR ; [2] : Ces résidus ne sont pas superposés à leur homologue de 1EUR mais appartiennent à des boucles flexibles ; [3] : résidus catalytiques

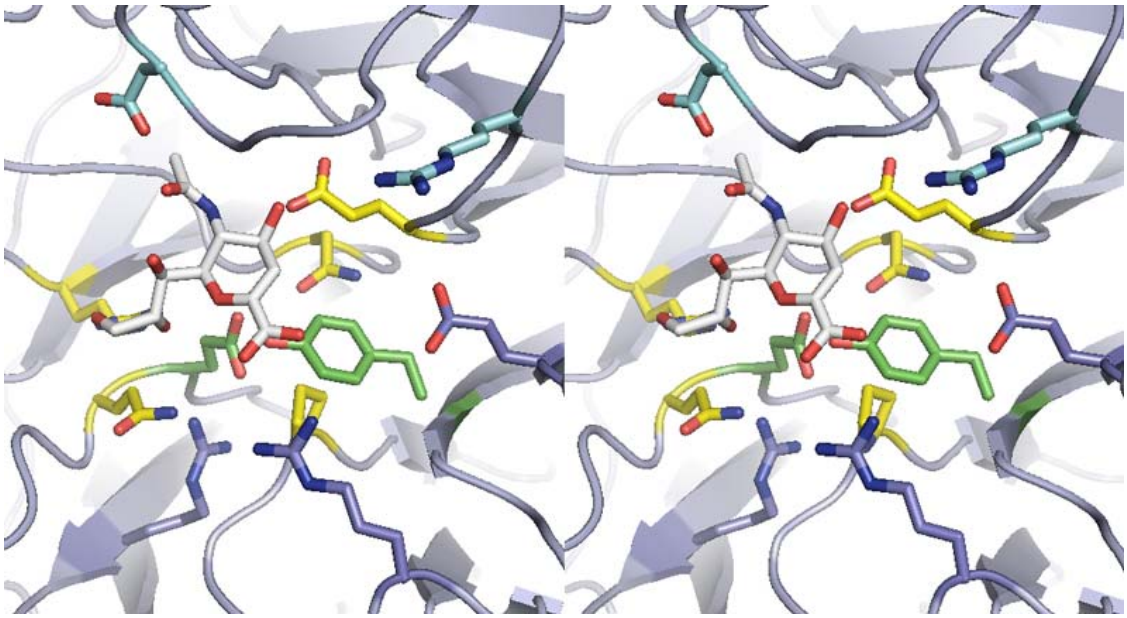


Figure III-77 : Modèle du site actif de RB3006 (GH33).
Représentation stéréoscopique des acides aminés proposés à la fixation du substrat dans le module catalytique GH33 de RB3006. Le code couleur adopté correspond à celui du Tableau III-20. L'inhibiteur Neu5Ac2en présent dans la structure de 1EUS est ajouté (en blanc) à titre indicatif.

Un dernier élément dans ce modèle semble pointer vers une singularité de la protéine de R. baltica. En effet, ce modèle a permis de constater la présence de deux cystéines (C320 et C336) relativement proches, qui ne sont pas conservées dans la famille (voir positions marquées d'un triangle vert -▲- dans la Figure III-74). Elles ne sont pas assez proches dans le modèle pour impliquer un pont disulfure mais elles le sont suffisamment dans la structure et à des positions suffisamment remarquables (extrémités de brins β adjacents) pour indiquer que la structure réelle de RB3006 possède probablement un pont disulfure à cette position. Ce pont disulfure se trouve lui-même à une position assez particulière puisqu'il serait adjacent à la tyrosine catalytique, engendrant au minimum une stabilisation de cette dernière (Figure III-78).

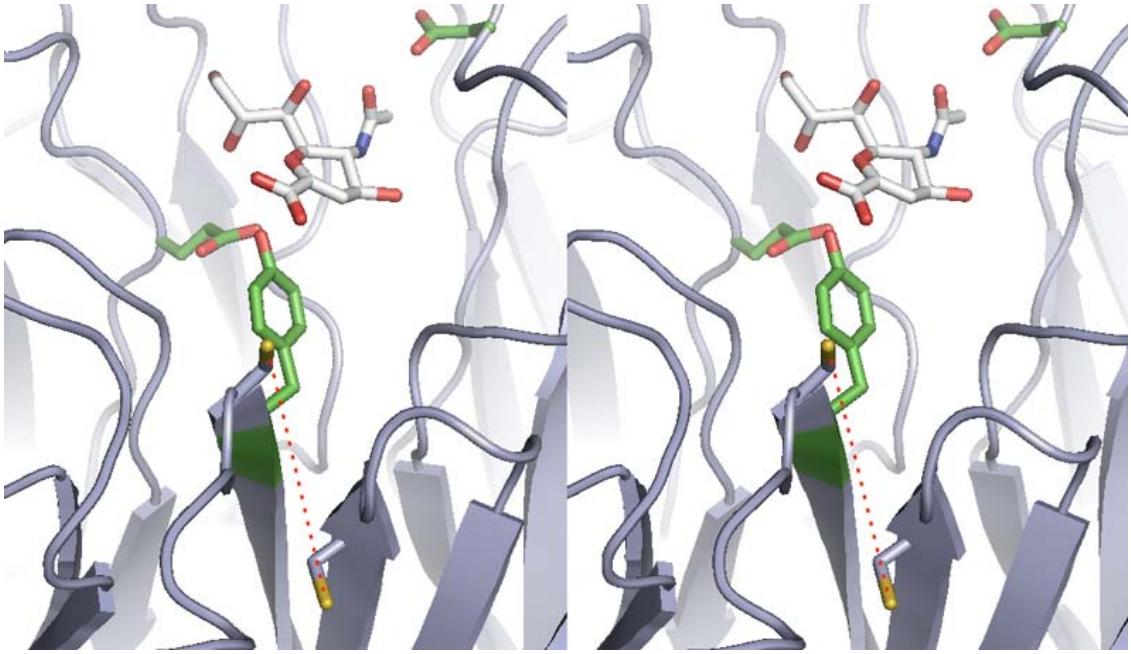


Figure III-78 : Vue stéréoscopique du modèle du site actif de RB3006 (GH33).
Vue stéréoscopique du site actif de RB3006 mettant en évidence la présence d'un pont disulfure probable entre les cystéines C320 et C336. Les trois résidus impliqués dans la catalyse sont représentés en vert. L'inhibiteur Neu5Ac2en présent dans la structure de 1EUS est ajouté (en blanc) à titre indicatif.

III.C - Cristallogénèse

Les tests de cristallisation ont été réalisés pour une concentration en enzyme d'environ 8 mg/mL pour le module catalytique et 5 mg/mL pour le module UNK1 (estimations réalisées après mesure sur NanoDrop sur 2µL d'échantillon). Comme pour RB2160, 672 conditions de cristallisation ont été testées, via 5 plaques Corning de 96 conditions issues des kits commerciaux suivants : PEG Screen I, PEG Screen II, the

Cations, PACT Suite, JCSG+, MDL I et MDL II. Les plaques ont été rempli avec un robot qui dispense des nano-gouttes assises. Ainsi 300 nL de protéines ont été mélangés avec 150 nL de solution de réservoir et ont été mis à équilibrer contre 100 µL de solution de réservoir.

Les résultats ont globalement été assez comparables à ceux obtenus avec RB2160 : sur l'ensemble des conditions, quelques unes ont données des précipités microcristallins intéressants sans cependant qu'une tendance globale ne se dégage. Des purifications avec un tampon final légèrement différent (concentration plus faible de NaCl) ont également été testées mais n'ont permis d'identifier des conditions de cristallisation plus claire. A noter cependant qu'une vérification récente des boites de cristallogénèse du module UNK1 a révélé la croissance d'un petit cristal (par observation à la loupe binoculaire avec lentille réfringente). Cet objet ayant probablement grandi au cours des douze derniers mois, ce cristal fera prochainement l'objet d'une étude par spectrométrie de masse afin de vérifier si il ne s'agirait pas d'une version tronquée de la protéine, permettant une croissance cristalline sans une zone flanquante qui perturbait l'empilement cristallin au point d'inhiber la cristallisation. Ce module est après tout le seul de l'ensemble des cibles finales dont les délimitations ont été déterminés seulement par analyse bioinformatique, et il n'est pas impensable que ce type de séquence flanquante puisse exister.

III.D - Discussion autour de l'activité de RB3006

Les modules étudiés de RB3006 auront été finalement assez faciles à manipuler parmi pour les expressions et les purifications. Les différents indicateurs de comportement en solution (DSL, chromatographie d'exclusion de taille) semblent indiquer qu'aucun des deux ne semble agrégé, multimérique, ou inhomogène en solution. Ces deux protéines sont donc très probablement bien repliées et leur fonction doit être active.

Cependant, les différents tests enzymatiques réalisés sur le module catalytique de RB3006 n'ont pas donnés les résultats souhaités. Il est pour le moins étrange que cette enzyme n'ai pas exprimé de fonction, alors qu'en plus de données biochimiques rassurantes, elle semble présenter les attributs d'une sialidase classique (résidus catalytiques, régions conservées, …). Le substrat étant peu stable, nous nous sommes retrouvés un peu limités par les conditions expérimentales pour les tests réalisés et, dans le temps imparti de ma thèse, peu de pH ont pu être testés. Il apparaît cependant que les enzymes de la famille

GH33 sont assez sensibles au pH (ref). Modifier et étendre la gamme de pH pour les tests futurs devrait donc être parmi les premières expériences réalisées pour mener à bien la caractérisation expérimentale de RB3006. Les tests de plusieurs tampons à plusieurs pH associés à un gradient de salinité en solution ont déjà été entrepris. Leur finalisation devrait se finir dans les mois qui viennent.

Si toujours aucune activité n'est observée avec les modifications des conditions expérimentales, il faudra chercher d'autres explications.

Il serait par exemple intéressant de prendre en compte sa structure modulaire. Il est possible que ses deux modules additionnels permettent par leur association de réaliser la catalyse. La sialidase NedA de *M. viridifaciens* présente une structure de ce type. Elle est composée de trois modules : le module catalytique GH33, un module de lien (structuré, et fixant des cations), et un module de fixation d'un substrat. L'activité est trouvée par l'action synergique de ces trois entités Figure III-79.

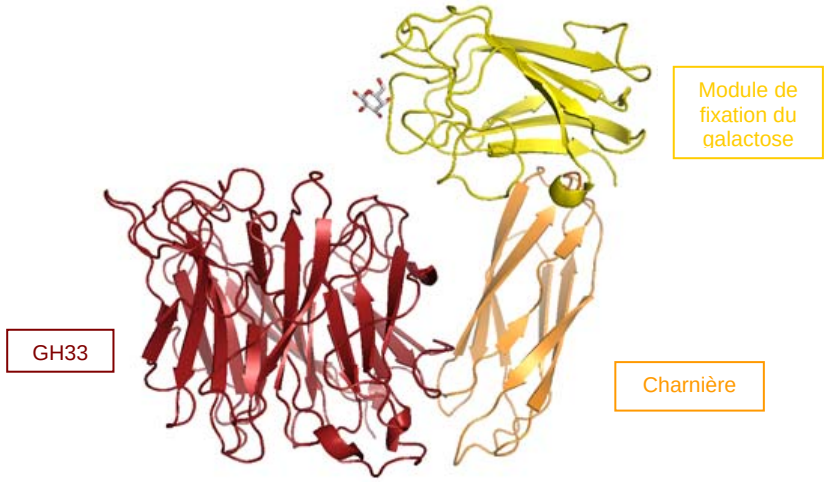


Figure III-79 : Structure de la sialidase NedA entière

Structure 1EUU de la sialidase NedA entière de *M. viridifaciens* (Gaskell et al., 1995)

Le fait que le module UNK-1 de RB3006 présente une séquence qui semble ancestralement liée aux sialidases pourrait aller dans ce sens. Peut-être lie t'elle toujours le substrat, mais a perdu la capacité de l'hydrolyser. Nous pourrions avoir plusieurs approches pour mettre cette hypothèse à l'épreuve. Il serait ainsi possible d'essayer de mettre le module catalytique de RB3006 en présence du module UNK1 et de retenter les mesures. Il faudra probablement aussi tenter d'exprimer sous forme soluble la protéine entière.

Il reste bien évidemment possible que l'enzyme puisse présenter une activité légèrement différente de celle testée avec le muNeu5Ac, ou encore que la construction actuelle ne soit tout simplement pas active sous cette forme. Il faut garder à l'esprit que

Rhodopirellula baltica ne présente pas qu'une enzyme de la famille GH33, mais sept. L'une d'entre elles est probablement une sialidase classique, vu que son gène est impliqué dans une structure qui ressemble fort à un opéron de dégradation d'acides sialiques exogènes (voir Chapitre IV). Cette même enzymes (RB3353) présente une similitude faible avec RB3006 (27% sur la longueur des domaines catalytiques), et pourtant, elles branchent toutes les deux ensemble dans l'arbre phylogénétique de la famille (Figure III-64). La bioinformatique ne tranchant pas, les caractérisations biochimique et biophysique sont les seules à pouvoir révéler les secrets de cette enzyme.

Il est à ce propos réellement dommage que les tests de cristallisation n'aient pas aboutis à la croissance de cristaux. De nombreux « précipités prometteurs » sont apparus au cours des divers tests réalisés. Aucun n'a cependant réellement évolué vers une amélioration. Il ne faut pas oublier non plus que les deux modules de RB3006 qui ont été testés en cristallisation sont les deux seuls des cibles étudiées qui sont des fragments de protéines. J'ai expliqué au cours de ce chapitre que la publication des génomes de *B. marina* et *P. maris* nous avaient fortement soulagés, le module UNK-1 s'avérant être une protéine à part entière chez ces deux *planctomycetes*. Néanmoins, elle n'est pas une protéine à part entière chez *R. baltica*. Des séquences flanquantes non repliées seraient donc à considérer. Une preuve vient malheureusement étayer cette hypothèse : l'apparition d'un cristal de UNK-1 dans une condition de cristallisation affinée, un an après la création de la boite de conditions de cristallisation. Ce phénomène arrive parfois quand une protéine présente des régions peu repliées. Les quelques protéases restant dans le mélange arrivent à dégrader ces zones, et la protéine, finalement plus compacte, arrive enfin à cristalliser. Si il s'avère que ce cristal est effectivement protéique, il faudra songer à redessiner des constructions, et à tester un ensemencement de ce cristal dans des conditions de cristallisation.

Au final de ma thèse, cette enzyme est plus que jamais intéressante et la résolution de sa structure alliée à la détermination de son activité pourraient beaucoup apporter à la connaissance de ces enzymes, ainsi qu'à la connaissance des capacités métaboliques de *R. baltica*.

IV - RB5312 : Une pectine lyase originale

IV.A - La famille PL1

Les polysaccharide lyases (ou encore transéliminases) n'ont pas un mécanisme catalytique comparable à celui des hydrolases. Elles catalysent exclusivement l'élimination des liaisons 4-*O*-glycosidiques des polysaccharides uroniques. Leur mécanisme catalytique se déroule en trois étapes. Tout d'abord, un groupe électrophile de l'enzyme subit une attaque du groupement carboxylique du substrat et une liaison est établie entre eux. Le proton du carbone C5 est ensuite éliminé par une attaque nucléophile sur ce carbone et finalement une β-élimination de la liaison 4-*O*-glycosidique est réalisée, libérant un sucre insaturé (Gacesa, 1987) (Figure III-80).

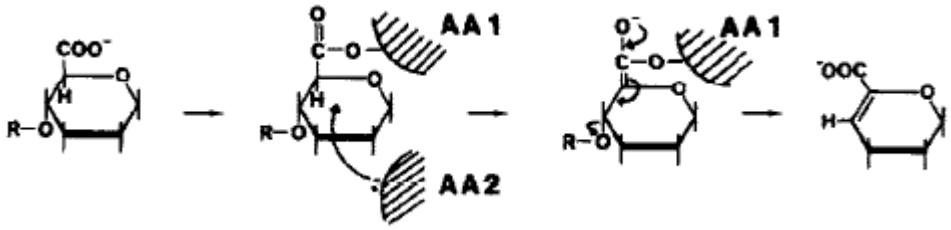


Figure III-80 : Activité lyase telle que proposée par Gacesa en 1987.
Figure extraite de Gacesa (1987).

La famille 1 des polysaccharide lyases comprend environ 400 séquences, réparties exclusivement et à égalité entre des phyla eucaryotes et bactériens. Elle est spécialisée dans la dégradation de la pectine, à travers trois activités (Tableau III-21).

Activité		Numéro EC	Nombre de structures (et codes PDB publiés)
pectate lyase	endo-α-1,4-poly galacturonate lyase	4.2.2.2	6 (32)
exo-pectate lyase	exo-α-1,4-poly galacturonate lyase	4.2.2.9	-
pectine lyase	endo-α-1,4-méthylester-polygalacturonate lyase	4.2.2.10	3 (4)

Tableau III-21 **:** Récapitulatif des activités de la famille CAZy PL1

La pectine est un polysaccharide majeur qui compose le mucilage de la paroi des plantes supérieures et est probablement une des macromolécules les plus complexes dans la nature, étant donné qu'elle peut contenir jusqu'à dix-sept monosaccharides, liés entre eux par plus de vingt types de liaisons différentes. (Mohnen, 1999; Ridley *et al.*, 2001; O'Neill *et al.*, 2004). Ses premières descriptions physico-chimiques datent de 1825 avec les travaux de Henri Braconot (Labrude and Becq, 2003) mais ses propriétés gélifiantes sont connues et ont été utilisées depuis des siècles pour la fabrication de produits de consistance gélifiée telles que les confitures. De nos jours, ce polysaccharide est toujours employé pour ses propriétés gélifiantes dans l'industrie agroalimentaire (Kashyap *et al.*, 2001), au même titre que l'agar ou les carraghénanes (décrits dans la section I-A de ce chapitre). D'un point de vue physiologique, ses fonctions au sein des plantes sont très variées. Tout d'abord, la pectine joue un rôle primordial dans la formation de leur paroi qui est une structure protéopolysaccharidique robuste et très plastique assurant le soutien de la plante (Fry, 1988). La teneur de la paroi des végétaux dicotylédones en pectine peut aller jusqu'à 35 %, faisant d'elle le principal polysaccharide devant la cellulose (30 %) et l'hémicellulose (30 %) (Fry, 1988). Elle joue également un rôle important dans la modulation des propriétés physico-chimiques de la paroi : porosité, distribution des charges de surface, pH ou encore balance ionique font d'elle un acteur majeur du transport des ions dans la paroi (McNeil *et al.*, 1984). Les oligosaccharides de pectines sont de plus connus pour induire les réactions de défense anti-microbienne chez les plantes (Hahn *et al.*, 1981; Nothnagel *et al.*, 1983; Jin and West, 1984), induire le phénomène de lignification (Robertsen, 1986) et l'accumulation d'inhibiteurs de protéases dans leurs tissus (Bishop *et al.*, 1984). Enfin, La pectine est également reconnue d'intérêt pour la santé humaine en ce qu'elle fait partie des fibres végétales non métabolisables qui favorisent la digestion (James *et al.*, 2003).

La pectine se définit comme un hétéropolysaccharide contenant essentiellement des acides galacturoniques formant une chaîne principale (GalA), certains étant sous forme de méthyl ester. Cette chaîne peut présenter des ramifications composées de divers sucres neutres (rhamnose, xylose, arabinose, mannose, …) en des proportions très variables (Kertesz, 1951).

Une chaîne de pectine est en réalité constituée de plusieurs blocs de structure différente qui se distinguent par la composition de leur chaîne principale et par le nombre et le type de leurs ramifications (Figure III-81). Dans la paroi des plantes, ces différentes chaînes de pectine sont réticulées entre elles d'une part et avec les autres polysaccharides présents d'autre part (Mohnen, 2008). On distingue quatre grands types de pectine.

Les **homogalacturonanes** (HG) représentent le type majoritairement rencontré dans la paroi des plantes et comptent pour 60 % de la quantité totale de pectine (Mohnen, 1999; O'Neill *et al.*, 2004). Il s'agit d'une chaîne linéaire de résidus GalA liés entre eux par des liaisons α-(1,4) (McNeil *et al.*, 1984). Parmi les modifications rencontrées, la méthyl-estérification du carbone *C6* est de loin la plus étudiée, en ce quelle module très fortement les propriétés physico-chimiques du polymère (Daas *et al.*, 1998). Dans la suite de ce manuscrit, un HG majoritairement non méthylé sera appelé un polygalacturonate (PGA), tandis qu'un HG majoritairement méthylé sera appelé un methylester-polygalacturonate (mePGA). Une forme de HG, nommée **xylogalacturonane** (XGA), substituée exclusivement par des β-(1,3)-D-Xylose a également été isolée dans certains tissus de plantes (Willats *et al.*, 2004)

Les **rhamnogalacturonanes de type I** (RGI) sont des polymères composés du disaccharide α-(1,2)-L-Rhamnose-α-(1,4)-D-acide galacturonique. La longueur de ces sections est assez variable et il est possible de trouver entre vingt répétitions du disaccharide idéal (betterave) et trois cents (sycomore) (Albersheim *et al.*, 1996). De plus, 20 % à 80 % des rhamnoses sont ramifiés sur leur *O4* (Albersheim *et al.*, 1996) par divers sucres, essentiellement de type galactose ou arabinose, sous forme de monosaccharides ou de structures polysaccharidiques plus ou moins longues (arabinogalacatanes ou arabinanes) (Lau J. M., 1987; Colquhoun *et al.*, 1990; Lerouge *et al.*, 1990).

Les **rhamnogalacturonanes de type II** (RGII) présentent une structure très conservée dans le règne végétal. Cette structure est caractérisée par de courtes régions du HG qui contiennent des groupes de chaînes latérales contenant des sucres très particuliers tels que le D-apiose, l'acide L-acérique (3-carboxy-5-deoxy-α-L-xylofuranose), l'acide 3-déoxy-lyxo-2-heptulosarique (Dha) ou encore l'acide 3-déoxy-manno-2-octulsonique (Kdo), assemblés en structures complexes (Mohnen, 2008).

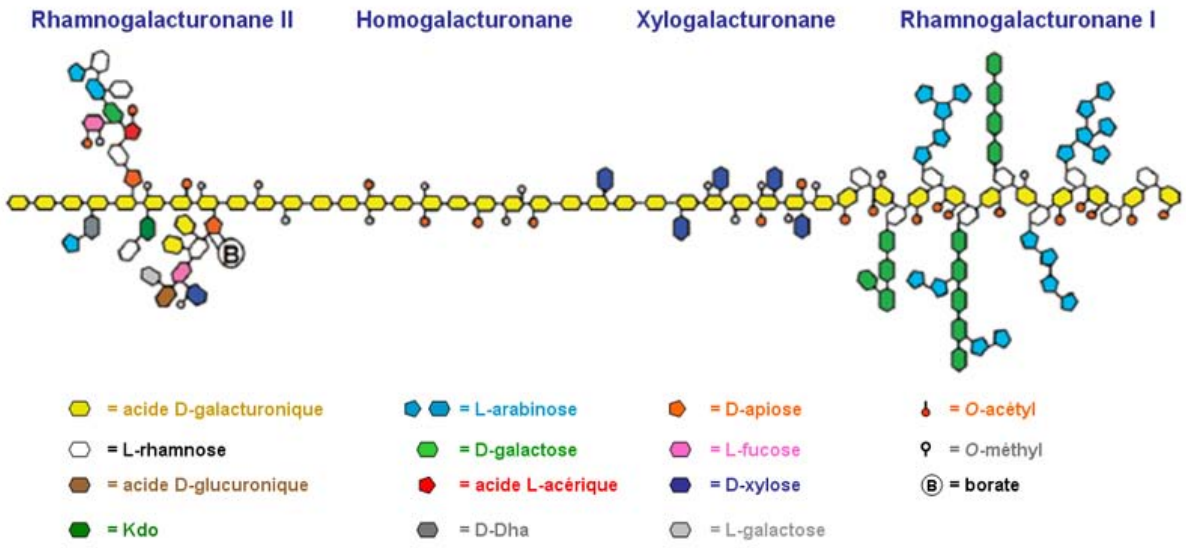


Figure III-81 : Structure de la pectine.
Présentation schématique de la structure des différents polymères composants la pectine (figure extraite de Scheller, 2007). Cette figure illustre les quatre différents domaines de la pectine qui sont retrouvés ubiquitairement dans la nature. L'abondance relative de ces différents types est variable mais les fractions HG et RGI restent les composants majeurs. A noter qu'il existe plusieurs variations des chaînes latérales, en particulier pour les RGI (Scheller *et al.*, 2007).

L'immense variété des enzymes dédiées au métabolisme de la pectine est une conséquence directe de la complexité de ce polymère. Au moins 67 glycosyl transférases sont en théorie nécessaires à sa biosynthèse, dont à l'heure actuelle seulement quatre ont été expérimentalement caractérisées (Mohnen, 2008). La connaissance de sa biodégradation est cependant plus avancée en particulier grâce à l'étude de pectinases isolées d'organismes pathogènes de plantes. Ces pectinases sont en effet reconnue pour être directement en cause dans le pourrissement de végétaux infectés, en détruisant la structure même des parois végétales (Ried and Collmer, 1986; Dinu *et al.*, 2007). Plusieurs classes d'enzymes ont ainsi été expérimentalement caractérisées : des hydrolases (rhamnosidase, exo/endo-polygalacturonase ou encore arabinogalactanase) et des transéliminases (pectate lyase et pectine lyase). Plusieurs revues détaillées sur ce sujet ont d'ailleurs été publiées ces dernières années (Pitson SM, 1996; Düsterhöft EM, 1997).

Dans le cadre de la compréhension de la pathogénicité de certains organismes particulièrement virulents contre les plantes, des études de caractérisation biochimique des activités pectinolytiques ont été menées. Les pectate lyases de la famille 1 ont fait partie des premières enzymes ainsi caractérisées, de par leur rôle central dans certains mécanismes d'infection. La première structure publiée dans la famille PL1 a été la pectate lyase PelE de

la bactérie pathogène de plante *Erwinia chrisanthemi* (code PDB 1PCL, Yoder *et al.*, 1993b). Cette structure a constitué lors de sa publication la première structure de l'ensemble des activités pectinases (lyases et hydrolases). Le repliement de cette enzyme s'est trouvé être une hélice β parallèle (également nommée hélice β « droite » du fait de l'orientation des feuillets β la composant). Ce repliement a depuis été retrouvé dans de nombreuses familles pectinolytiques puisque trois familles de lyases sur les cinq présentant cette activité suivent ce repliement, ainsi que la principale famille d'hydrolases pectinolytiques (famille GH28).

Les Figure III-82 et Figure III-83 présentent deux structures de la famille PL1 : la protéine non active Juna1 constituant l'allergène principal du pollen de *Juniperus ashei* (cèdre blanc nord américain connu pour déclencher de fortes réactions allergiques saisonnières) (code PDB 1PXZ, Czerwinski *et al.*, 2005) et la pectate lyase *bs*Pel de *Bacillus subtilis* (code PDB 2BSP, Pickersgill *et al.*, 1994). La topologie de ces structures consiste en une série de brins β arrangés en trois feuillets, stabilisés par un empilement d'asparagines et s'enroulant autour d'un cylindre, le tout formant une hélice (Yoder *et al.*, 1993a). Dans une seconde vue de ces structures, les hélices β parallèles sont mises en évidence par un code coloré (vert pour Juna1, bleu pour *bs*Pel). Les zones grisées indiquent des régions spécifiques à chacune de ces protéines, qui consistent en des protrusions très différentes en séquence et en repliement émanant de l'hélice.

a) b)

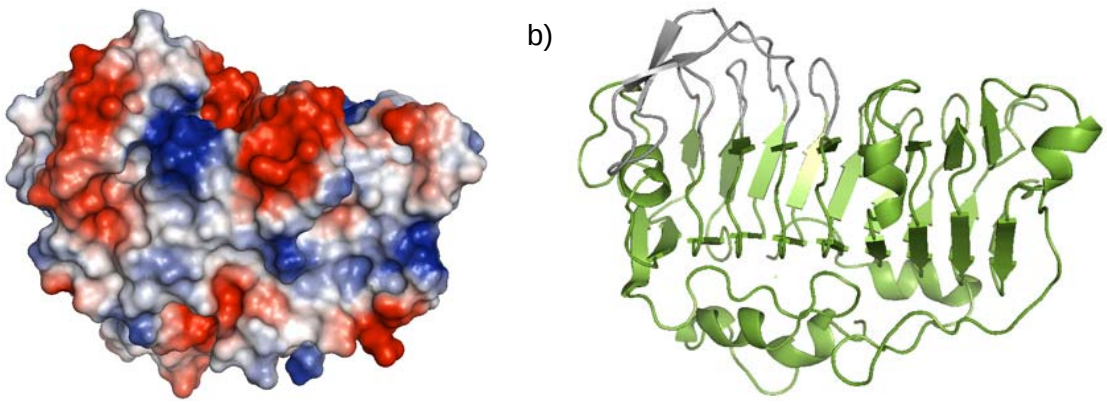


Figure III-82 : Structure de la protéine Juna1 (1PXZ).
Protéine allergène de *J. ashei* (protéine non active). a) Représentation de la protéine avec modélisation de la distribution des charges de surface. b) Représentation de la protéine en ruban avec mise en évidence des éléments de structure secondaire. L'hélice β parallèle est colorée en vert.

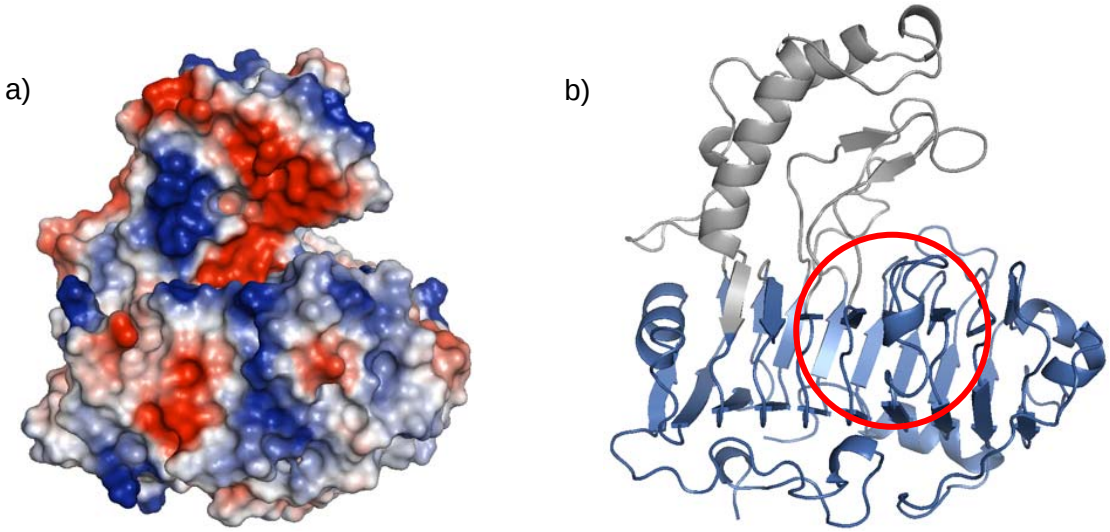


Figure III-83 : Structure de la pectate lyase *bs*Pel de *B. subtilis* (2BSP).
a) Représentation de la protéine avec modélisation de la distribution des charges de surface. b) Représentation de la protéine en ruban avec mise en évidence des éléments de structure secondaire. L'hélice β parallèle est colorée en bleu. La localisation du site actif de *bs*Pel est indiquée par un cercle rouge.

Il apparaît que le cœur de ces protéines, constitué par leur hélice β est hautement superposable, comme montré sur la Figure III-84. Cet alignement des carbones C_α du squelette protéique des hélices β des deux structures, privées de leurs éléments non communs, présente un RMSD de 3,53 Á. Sur le même type d'alignement, les enzymes PelC et PelE d'*E. chrysanthemi*, ayant une plus grande similitude (27 % entres elles contre 10% pour Juna1/*bs*Pel sur les enzymes entières) présentent un RMSD de 1,02 Á (Yoder *et al.*, 1993a; Henrissat *et al.*, 1995).

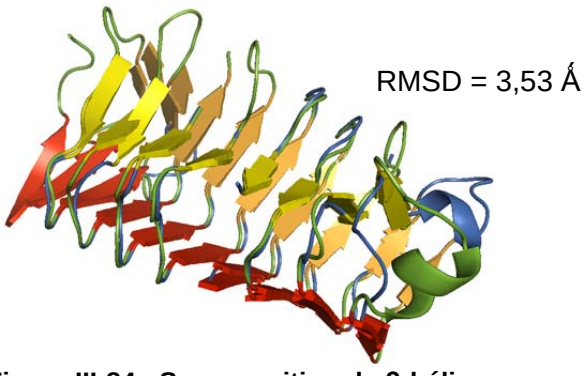


Figure III-84 : Superposition de β-hélices.
Superposition des noyaux structuraux des structures de Juna1 et *bs*Pel, à savoir leur hélice β parallèle. Les trois séries de feuillets apparaissent respectivement en rouge, orange et jaune. Les zones charnières entre les brins β sont colorées en vert pour Juna1 et en bleu pour *bs*Pel.

Peu après la résolution des premières structures de la famille, au début des années 90, plusieurs structures d'enzymes d'autres familles se sont révélées posséder le même repliement, avec toujours cette structure d'hélice β très conservée, au point qu'il s'agisse parfois de la seule chose qui lie encore les deux protéines (Yoder *et al.*, 1993a; Lietzke *et al.*, 1994; Steinbacher *et al.*, 1994; Henrissat *et al.*, 1995; Raetz and Roderick, 1995). Ces études ainsi que d'autres (Chothia and Murzin, 1993; Cohen, 1993; Sprang, 1993; Mayans *et al.*, 1997; Jenkins *et al.*, 1998) ont référencé les différentes originalités de ces repliements, soulignant leurs similitudes et mettant en perspective des alignements de séquences basés sur l'exceptionnelle conservation des hélices β. Une nomenclature a d'ailleurs été définie sur la base de cette similitude de repliement. Les feuillets jaune, rouge et orange de la Figure III-84 ont été respectivement nommés PB1, PB2 et PB3. Les régions charnières entre ces feuillets sont nommées T1 pour celle liant PB1-PB2, T2 pour PB2-PB3 et T3 pour PB3-PB1. Une description hautement détaillée de ce repliement, ainsi qu'une discussion sur le rôle des acides aminés présentant un empilement à l'intérieur de l'hélice sont présentées par Jenkins, Mayans & Pickergill (Jenkins *et al.*, 1998)

La grande conservation de certains résidus dans la famille a permis de donner quelques pistes sur la nature des acides aminés impliqués dans la catalyse. Il est à noter que l'observation de cette grande conservation de motifs dans les séquences datant de 1995 avec 14 séquences de pectate lyases, est toujours d'actualité avec les ajouts de nouvelles séquences dans la famille (360 séquences au total dans la base de données CAZy à l'heure actuelle, dont des pectine lyases, pourtant plus divergentes). Il apparaît qu'il existe plusieurs régions dans les différentes séquences qui présentent une grande conservation. Cependant, une région en particulier présente des sites strictement conservés au sein des séquences possédant une activité pectate lyase. Cette région ayant été démontrée capable de fixer un ion calcium crucial pour l'activité pectinolytique de *bs*Pel (Pickersgill *et al.*, 1994), il a été proposé que ce site corresponde au site actif de l'enzyme (Lietzke *et al.*, 1994; Pickersgill *et al.*, 1994; Henrissat *et al.*, 1995; Yoder and Jurnak, 1995; Jenkins *et al.*, 1998). En particulier une arginine (R279 chez *bs*Pel, R176 chez *an*PelA – voir Figure III-86), ainsi que quelques acides aspartiques et glutamiques (Figure III-86) ont été proposés pour être impliqués dans la catalyse. La cocristallisation du mutant inactif R218K de la PelC d'*E. chrysanthemi* (*ec*PelC) avec un pentasaccharide de PGA (pPGA) a permis d'avancer encore dans la compréhension de la composition des acides aminés du site actif (Scavetta *et al.*, 1999). L'arginine R218 pressentie pour le rôle de la base attaquant le proton du *C5* du sucre catalysé (voir début de cette section) a en effet été confirmée dans ce rôle. Les résidus

acides tapissant la gorge du site actif se sont révélés fixer une série d'ions Ca^{2+} permettant une interaction forte avec la molécule de pPGA (Figure III-86).

Il est frappant de constater qu'il s'agit très précisément de ce site qui subit un grand remaniement entre les activités pectate lyases (catalyse réalisée au niveau d'un acide galacturonique, au sein d'une chaîne HG faiblement méthylée) et les pectine lyases (catalyse cette fois réalisée au niveau d'un galacturonate de méthyle, au sein d'une chaîne HG hautement méthylée) (Mayans *et al.*, 1997).

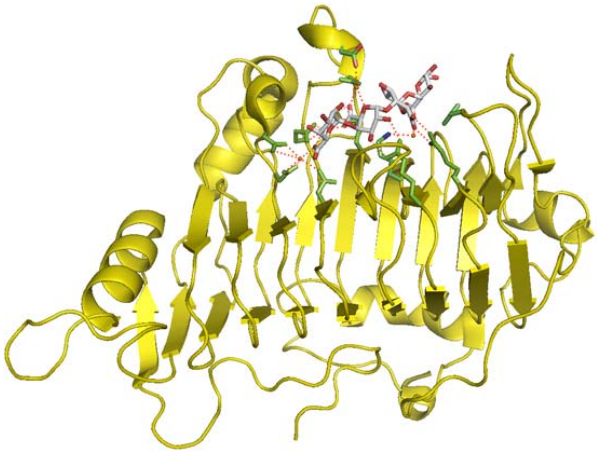


Figure III-85 : Structure du mutant R218K de la PelC de *E. chrysanthemi.* *Enzyme* en complexe avec un pentaPGA (code PDB 2EWE, Scavetta *et al.*, 1999).

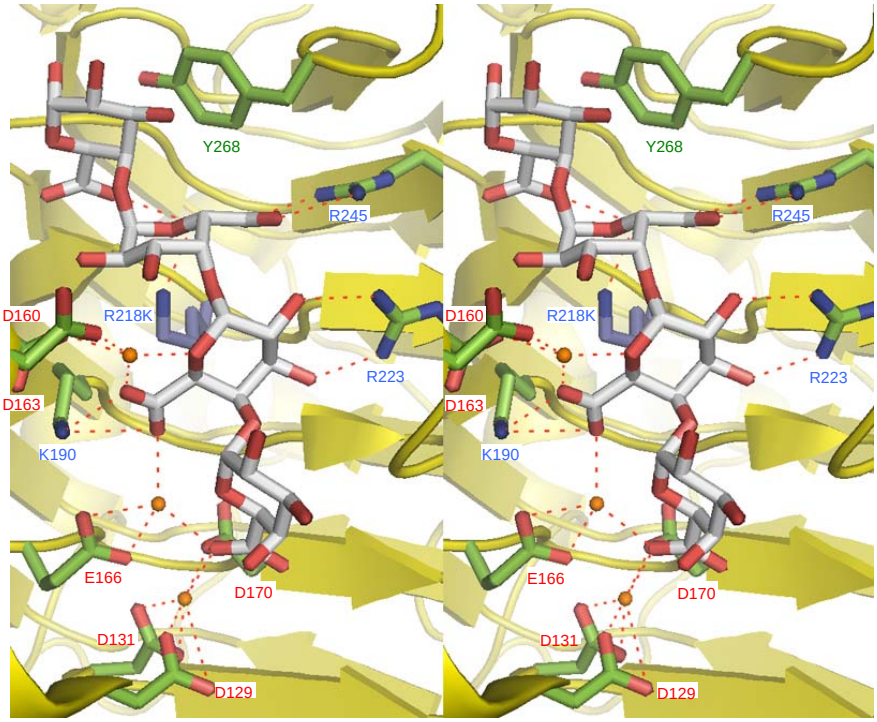


Figure III-86 : Site actif de la pectate lyase C d'*E. chrysanthemi*. Présentation des acides aminés composant le site actif du mutant R218K de *ec*PelC, représentant les pectate lyases de la famille PL1. Quatre ions calcium sont chélatés dans le site actif et participent à la fixation du substrat (d'après Scavetta *et al.*, 1999).

Le site actif des pectine lyases semble drastiquement différent. Ce site actif proposé y est en effet tapissé de résidus hydrophobes, en particulier de résidus aromatiques, connus pour participer à la fixation de polysaccharides neutres dans d'autres familles de polysaccharidases (Maenaka *et al.*, 1994). Des expériences de mutagenèse dirigée sur la pectine lyase PelA d'*Aspergillus niger* (*an*PelA) ont permis de caractériser des résidus cruciaux pour l'activité de cette enzyme (Sanchez-Torres *et al.*, 2003). Les mutants D154E/A, R176A/D/K, R236A/K ainsi que K239N, D186N, D221N et D186N-D221N ont été réalisés. Il apparaît que l'arginine R236 est essentielle à l'activité, tandis que les autres mutations affectent plus ou moins le rendement catalytique sans inactiver totalement l'enzyme (Figure III-87). Les acides aromatiques proposés pour être impliqués dans la fixation du substrat comprennent quatre résidus tryptophanes (W66, W81, W151 et W212) et trois tyrosines (Y85, Y211 et Y215). Ces résidus ne forment pas d'empilement mais semblent plutôt former des paires (W81-W151, W66-W212 et Y211-212) arrangées en interactions proches face contre côté (Mayans et al., 1997).

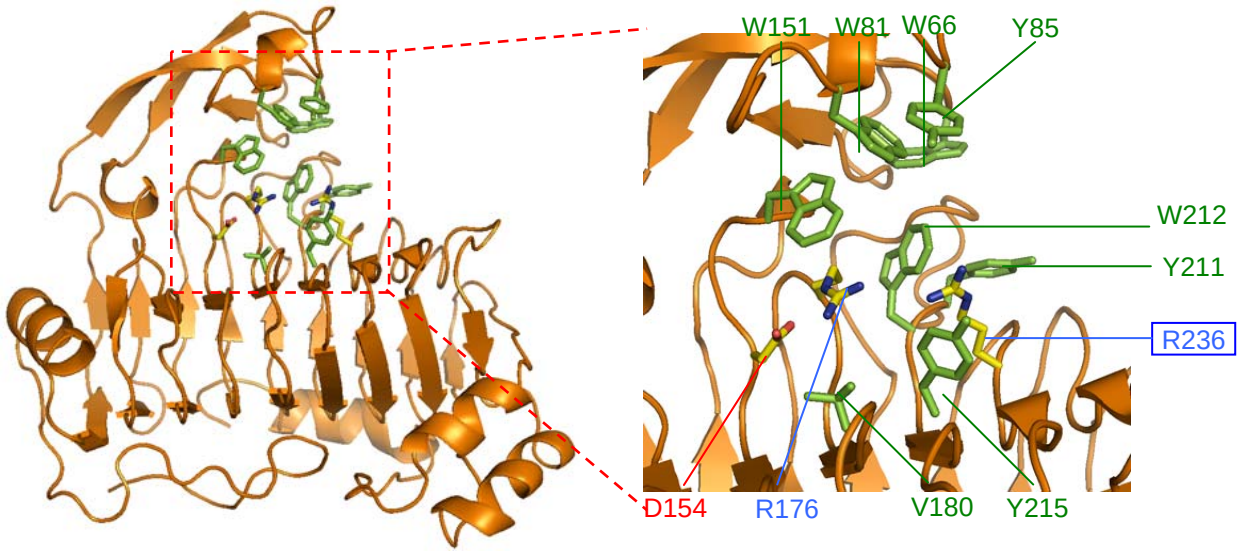


Figure III-87 : Structure de la pectine lyase A d'*A. niger* (1QCX).
Présentation des acides aminés proposés pour composer le site actif de *an*PelA, représentant les pectine lyases de la famille PL1. L'arginine R236, base catalytique de la lyase, est encadrée en bleu.

Enfin, parmi les motifs conservés de la famille PL1, deux en particulier présentent une structure remarquablement conservée à travers l'ensemble des enzymes de la famille. La première, commençant en position 151 chez *bs*Pel, est de type *phhhbph* (*p* : polaire, *h* : hydrophobe, *b* : basique avec *phhhbph* = [ND][VI][IV][IVL][RQ]N[VLI] ; quelques enzymes présentent cependant localement des variantes, souvent propres à une espèce), au sein d'une région elle-même conservée sur le feuillet PB2 de la protéine en vis-à-vis du site actif. Une autre région est également remarquable pour sa quasi-stricte conservation

dans la famille. Cette région, située en position 194 chez *bs*Pel est constituée du motif [VI]W[IVLF]DH. Ces deux régions se retrouvent en réalité dans des brins β adjacents dans les structures (Figure III-88 et Figure III-89). Ces motifs ont été proposés pour constituer un site catalytique secondaire (Henrissat *et al.*, 1995) mais de récentes analyses semblent indiquer qu'ils serviraient plutôt au maintien de l'intégrité de la structure de l'hélice β (Mayans *et al.*, 1997).

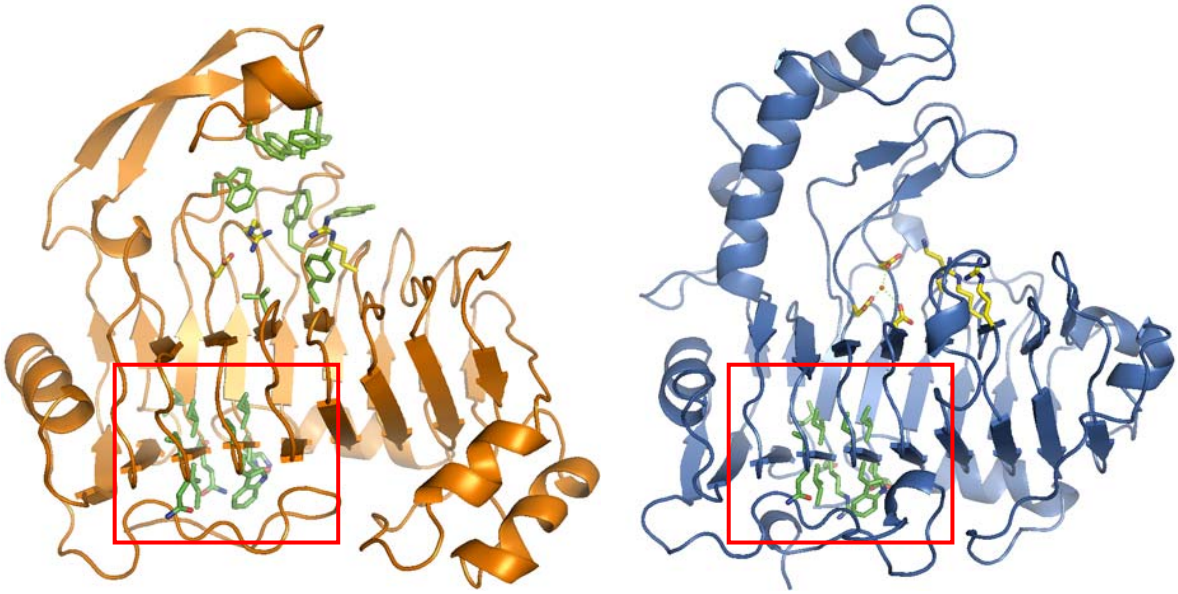


Figure III-88 : Comparaison des structures de PL1 1QCX et 2BSP
Présentation du site conservé présent sur la face opposée au site actif de la pectate lyase *bs*Pel et de la pectine lyase *an*PelA. Vues globales des structures de *an*PelA (orange) et *bs*Pel (bleu). La région conservée est matérialisée par des résidus en vert encadrés en rouge.

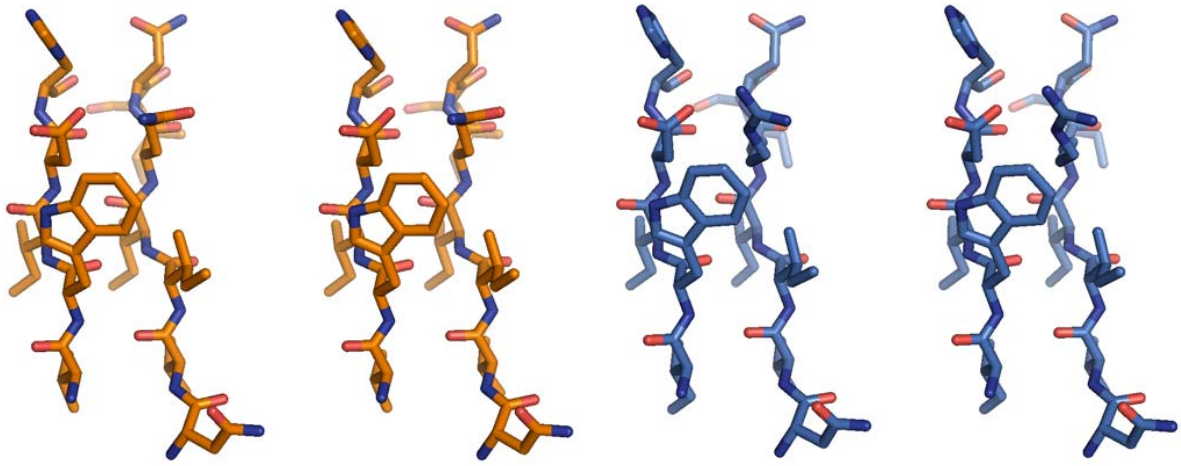


Figure III-89 : Détail des structures de PL1.
Présentation du site conservé présent sur la face opposée au site actif de la pectate lyase *bs*Pel et de la pectine lyase *an*PelA. Vues stéréoscopiques rapprochées des sites respectifs de *an*PelA (orange) et *bs*Pel (bleu). Noter la conservation des orientations entre les deux structures.

IV.B - Résultats et discussion

IV.B.1 - Analyse bioinformatique

R. baltica présente deux séquences appartenant à la famille des PL1, toutes deux monomodulaires, représentées dans le Tableau III-22.

Protéine	Modularité	Taille résidus	Taille kDa	Annotation initiale
RB5312	PL1	455	50	Pectate lyase (EC 4.2.2.2)
RB5316	PL1	353	38,9	Pectate lyase (EC 4.2.2.2)

Tableau III-22 : Protéines membres de la famille PL1 chez *R. baltica*.
Les peptides signaux sont représentés par un module grisé.

Un alignement multiple de ces deux séquences a été généré avec les séquences caractérisées suivantes de la famille PL1: P39116_BACSU (*bs*Pel), P81294_JUNAS (Juna1), P0C1A2_ERWCH (pectate lyase PelA de *E. chrysanthemi*) et P11073_ERWCH (pectate lyase PelC de *E. chrysanthemi*). Cet alignement permet de constater que les enzymes de *R. baltica* présentent toutes les deux un site actif de pectate lyase et non de pectine lyase, ainsi que beaucoup des régions conservées de la famille, corroborant cette étonnante conservation propre à ce type de repliement (Figure III-90).

Cet alignement met également en évidence que les deux paralogues de *R. baltica* sont assez divergents l'un de l'autre. RB5316 semble présenter les attributs typiques de la famille. Elle présente l'ensemble des résidus acides considérés comme catalytiques chez *bs*Pel (D184, D223 et D227), l'arginine R279 qui est considérée cruciale à l'activité et la lysine K247, mais pas l'arginine R284. Elle présente également des zones communes avec ses homologues, en particulier les sites en vis-à-vis du site actif *phhhbph* et xWxDH en quasi-stricte conservation avec *bs*Pel. De plus, RB5316 se trouve présenter une identité moyenne envers ses homologues de l'ordre de 25 % (son meilleur score étant 32 % d'identité avec la protéine non caractérisée Q9KF01 de *Bacillus halodurans*).

RB5312 présente beaucoup plus de singularité au sein de la famille PL1. Tout d'abord, l'acide catalytique D227 supposé être en interaction avec l'ion calcium indispensable à l'activité se trouve muté en sérine, ce qui constitue une mutation majeure au

sein du site actif. De plus, si RB5312 présente également une mutation au niveau de la lysine K284, elle possède cependant les autres résidus proposé pour l'activité, à savoir les acides D184 et D223, la lysine K247 et l'arginine R279.

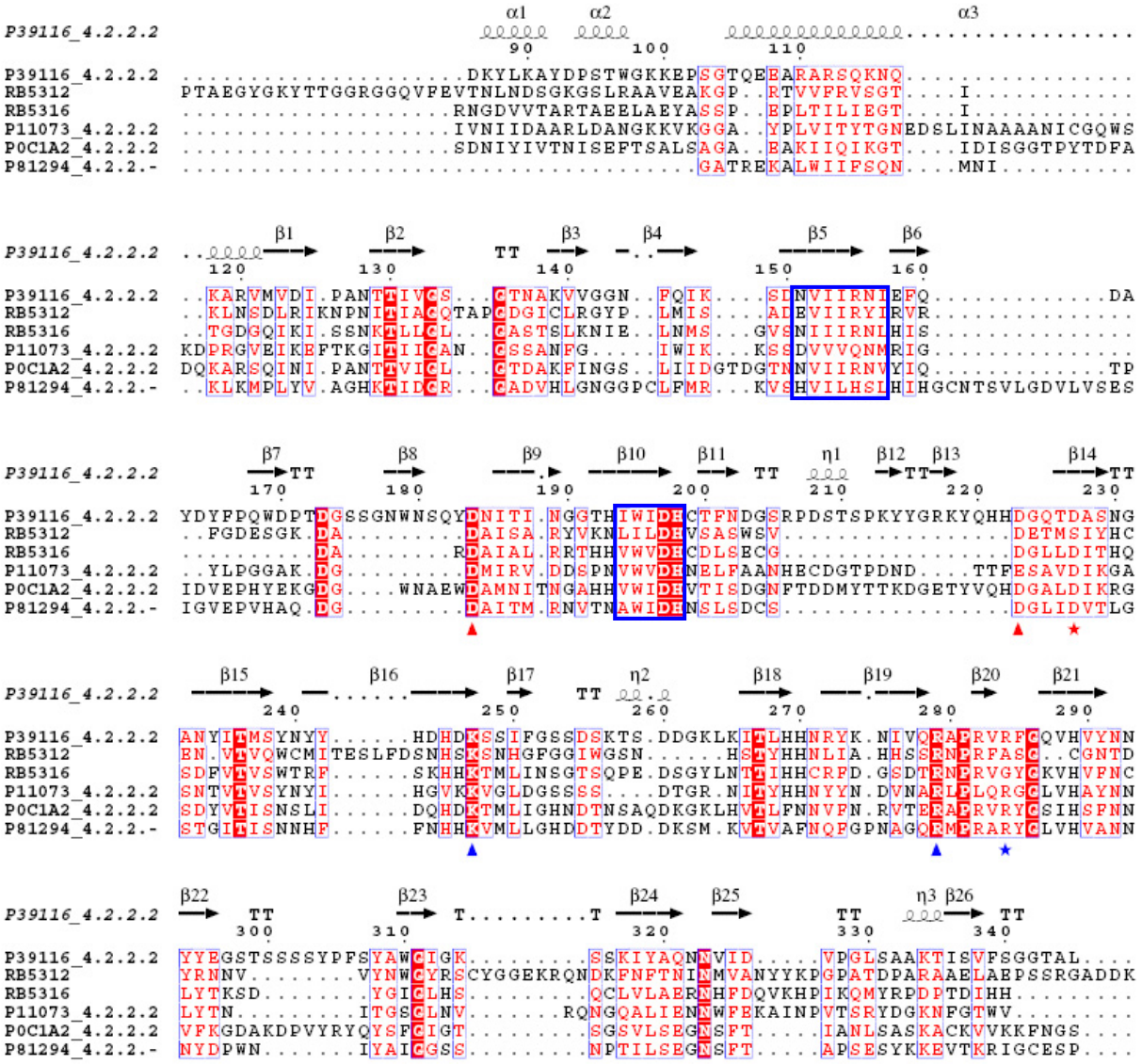


Figure III-90 : Alignement dans la famille PL1.

Alignement de séquences représentatives des domaines catalytiques pectate lyase de la famille PL1 avec les deux séquences de *R. baltica*. Les résidus considérés comme catalytiques sont indiqués par un motif de couleur : un triangle rouge (▲) ou bleu (▲) indique que le résidu respectivement acide ou basique est strictement conservé dans cet alignement ; une étoile rouge (★) ou bleue (★) indique que le résidu n'est pas strictement conservé dans cet alignement. Les motifs *phhhbph* et xWxDH présentés dans la partie IV-A sont encadrés en bleu.

Il apparaît également que RB5312 est la seule séquence de l'alignement à ne pas posséder strictement les mêmes résidus que ses homologues dans les régions phhhbhp et xWxGH. Elle se trouve notamment la seule à avoir le tryptophane de la séquence xWxDH muté en isoleucine. Afin de valider ces différences de séquence, un alignement de plus

grande ampleur a été réalisé sur les séquences entières de 87 protéines caractérisées de la famille avec celles de *R. baltica* avec MAFFT (données non montrées). Tout d'abord, cet alignement confirme les résultats du premier : RB5312 reste la seule séquence à présenter un site xWxDH muté sur le tryptophane. Il confirme ainsi que RB5312 est une protéine à part dans cette famille. Elle présente en particulier une alternance entre des régions très alignables et d'autres beaucoup moins. De manière très intéressante, son site actif fait très précisément partie des régions peu alignables et semble avoir subi de nombreuses modifications, en particulier des séries d'insertions et de délétions. Ces éléments semblent confirmer que si les deux enzymes de *R. baltica* sont des pectates lyases probablement fonctionnelles, elles doivent présenter des activités plus originales que l'essentiel de la famille (en particulier pour RB5312).

Une analyse complémentaire de phylogénie a été réalisée avec 89 séquences de la famille PL1 (dont celles de *R. baltica*) ayant été caractérisées expérimentalement (Figure III-91). L'arbre phylogénique qui en sort est cohérent avec ce à quoi l'on pouvait s'attendre. En effet, les pectine lyases s'excluent des pectate lyases et forment deux groupes très robustes au sein de l'arbre (avec des valeurs de bootstrap de 99 % pour chacun), l'un contenant les enzymes de champignons et l'autres celles de bactéries. De la même manière, les pectate lyases d'organismes eucaryotes forment des groupes robustes (un groupe de plantes et un de champignons) qui s'excluent des pectate lyases de bactéries. Il est important de rappeler à ce niveau que l'immense majorité des pectate lyases bactériennes de la famille PL1 qui ont été caractérisées appartiennent à des phyla bactériens très récurrents. En particulier, les genres *Erwinia* et *Bacillus* y sont surreprésentés. Ils représentent à eux deux 30 séquences sur les 48 bactériennes actuellement caractérisées. Il ne serait donc pas surprenant que des enzymes d'une planctomycètes soit un peu exclues des différents sous-groupes de l'arbre. Cependant, il apparaît que RB5316 appartient à un sous groupe, sans que son nœud ne soit très robuste (23 %). RB5312, elle, est totalement exclue et forme de manière très surprenante un groupe à la racine de l'arbre avec la PelZ d'*E. chrysanthemi* (ces deux séquences présentant une similitude non négligeable de 29 %).

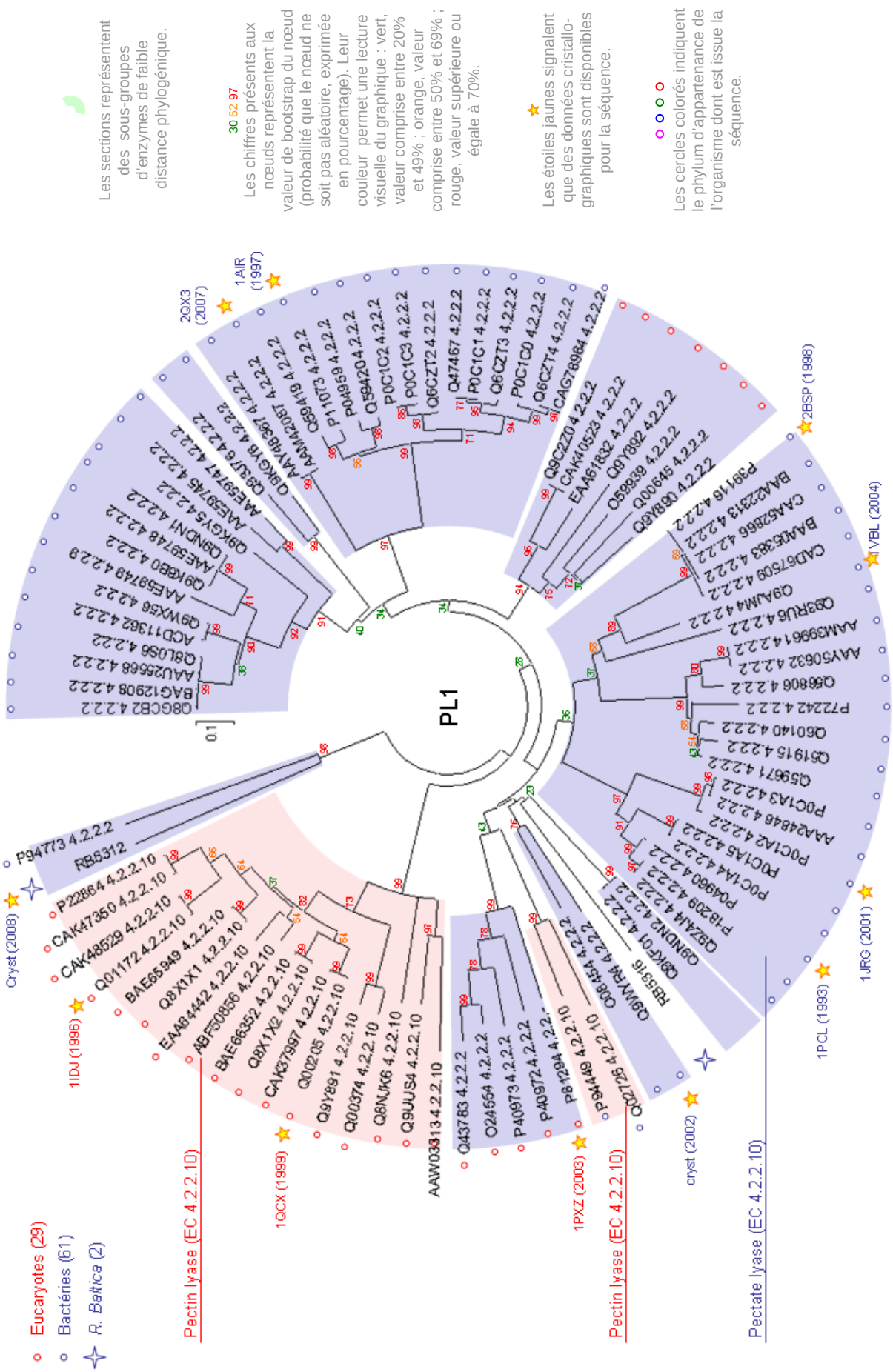


Figure III-91 : Arbre phylogénique présentant la famille PL1

Arbre phylogénique généré par la méthode d'évolution minimum, avec délétion des indels., à partir d'un alignement de type FFT-NS-i par MAFFT.

Lors de la conception de la plaque de cibles, RB5312 semblait ainsi une enzyme particulièrement intéressante. Tout d'abord, l'activité pectate lyase est en soit intrigante dans une bactérie marine. De plus, *R. baltica* possède quatre enzymes annotées pectate lyase (deux PL1 : RB5312 et RB5316, et deux PL10 : RB3417 et RB9973) ainsi que d'autres enzymes ayant des activités potentiellement liées à la dégradation de la pectine comme des α-L-rhamnosidases de la famille GH78, laissant suggérer que ce métabolisme n'est pas anodin pour cette bactérie. Ensuite, la forte divergence de RB5312 laisse présager une activité potentiellement plus originale que les homologues de cette famille. Enfin, les gènes codant pour RB5312 et RB5316 se trouvent être dans une zone génique intéressante (Figure III-92). Ils ne sont séparés que par seul gène, codant une protéine de fonction inconnue mais conservée et définie par le DUF1680, lointain descendant de glycosyl hydrolases. De plus, ils sont encadrés par des gènes codant des enzymes potentiellement impliquées dans la dégradation de polysaccharides : une autre lyase prédite pour dégrader un D-arabino-3-hexulose-6-phosphate formaldéhyde, qui se trouve être un maillon de la voie métabolique d'interconversion des pentoses et de la dégradation des acides galacturoniques (voir base de données KEGG code : cac00040) et une protéine de fonction inconnue mais également conservée et définie par le DUF1593, rencontré quasi-exclusivement chez *R. baltica*.

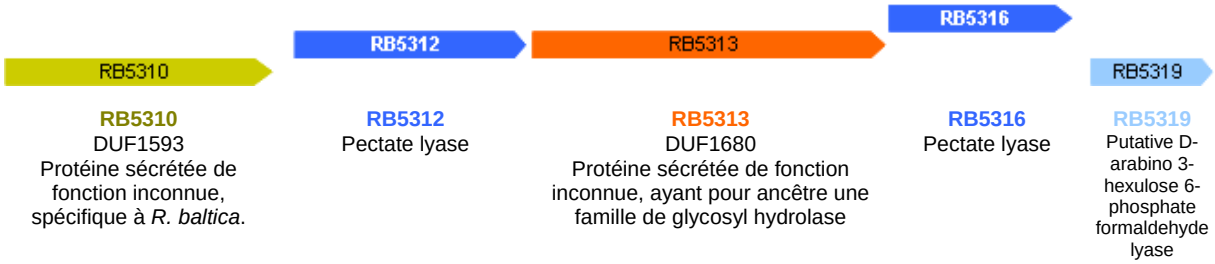


Figure III-92 : Locus des gènes *rb5312* et *rb5316.*
Ces deux gènes codent les deux PL1 de *R. baltica*. Figure tirée de Glöckner *et al.*, (2003).

Il serait donc envisageable que ces deux pectinases agissent en synergie avec d'autres enzymes au sein d'un opéron pour une dégradation optimisée d'un polysaccharide pectinique visé et qu'elles possèdent pour ce faire des activités différentes sur le même polysaccharide.

Pour toutes ces raisons RB5312 a intégré le choix final des cibles à caractériser dès qu'il a été avéré qu'elle présentait une forte expression soluble. Elle a été clonée sous sa forme entière privée de son peptide signal, cette protéine n'étant pas modulaire. La Figure III-93 présente sa séquence et le Tableau III-23 dresse une liste de ses propriétés physico-chimiques telles que prédites par ProtParam sur le site internet ExPASy.

Séquence protéique de RB5312 :

```
  1 MHHHHHHRSQ KPLAFPTAEG YGKYTTGGRG GQVFEVTNLN DSGKGSLRAA
 51 VEAKGPRTVV FRVSGTIKLN SDLRIKNPNI TIAGQTAPGD GICLRGYPLM
101 ISADEVIIRY IRVRFGDESG KDADAISARY VKNLILDHVS ASWSVDETMS
151 IYHCENVTVQ WCMITESLFD SNHSKSNHGF GGIWGSNHST YHHNLIAHHS
201 SRNPRFASGC GNTDYRNNVV YNWGYRSCYG GEKRQNDKFN FTNINMVANY
251 YKPGPATDPA RAAELAEPSS RGADDKGHWY VAENVIEGSP TVSADNWSGV
301 RGANYIQLDQ PWEAMPINQQ TAEEAFEDVL QHAGASWPKR DPIDTRIIQE
351 VRDGTATYEG VYKTKKRVSS DTQITGIIDS QQDVGGWPEL KNAAAAPDTD
401 HDGIPDAWEA EHGMDPNDAS DGNRTGNDGY TMLEQYINSI P
```

Figure III-93 : Séquence protéique de RB5312

	Protéine sauvage	Module PL1 cloné
Poids moléculaire	47 600 Da	48 700 Da
pI théorique	5,4	5,8
Nombre d'acides aminés	433	441
Asp + Glu	55	55
Arg + Lys	41	42
Cystéines	5	5
Méthionines	8	8
$\varepsilon_{280\,nm}$ ($M^{-1}.cm^{-1}$)	83 500	83 500 $M^{-1}.cm^{-1}$

Tableau III-23 : Récapitulatif des données biochimiques théoriques de RB5312.
Données extraites du logiciel ProtParam disponible en ligne sur le site ExPASy.

IV.B.2 - Résultats d'expression, de purification et de caractérisation biophysique

Tout comme les autres enzymes sélectionnées pour leur bonne expression soluble dans la plaque initiale, RB5312 a présenté une très bonne expression soluble et les purifications se sont dans leur ensemble plutôt bien passées. Elle a été purifiée à homogénéité électrophorétique par deux étapes de chromatographie : une colonne d'affinité au nickel et une colonne d'exclusion de taille Superdex 75. Ci-après sont présentés les chromatogrammes issus de ces purifications.

Le passage sur colonne de nickel a permis d'obtenir d'importantes quantités de protéine RB5312 relativement pure. Le même phénomène que décrit pour la purification du module UNK1 de RB3006 s'est initialement produit : une quantité trop importante de protéine a été probablement injectée sur la résine, provoquant une saturation des sites de fixation et finalement engageant une élution à faible concentration d'imidazole. Ce problème a été résolu en diminuant la quantité d'extrait protéique total à charger. Les purifications de routine ont été ainsi typiquement réalisées sur la moitié d'un culot bactérien issu d'une culture de 200 mL de milieu ZYP-5052. L'enzyme a été éluée à une concentration de l'ordre de 155 mM d'imidazole (Figure III-94).

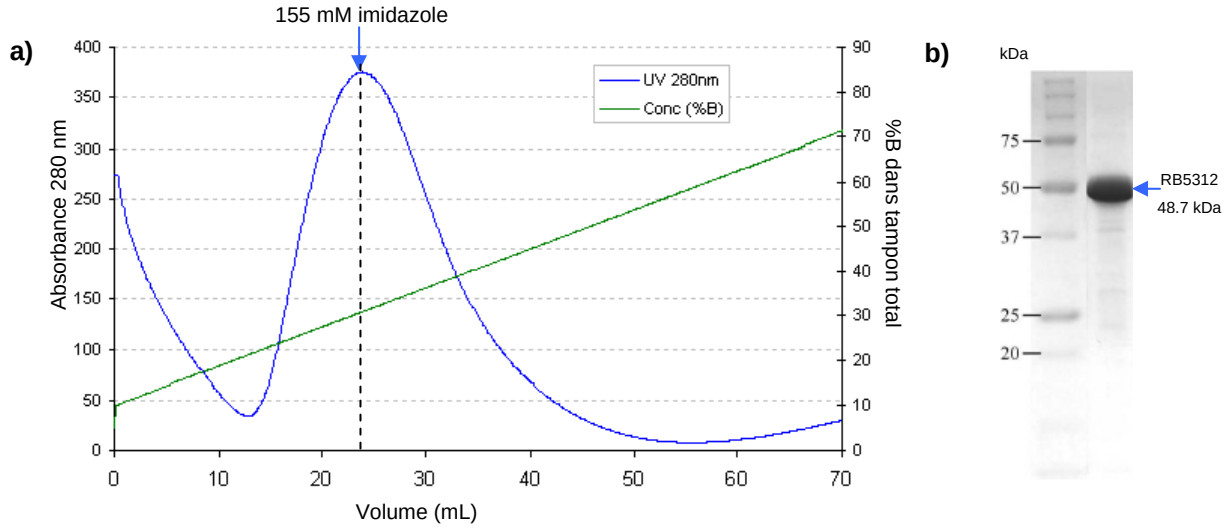


Figure III-94 : Purification de RB5312 (Colonne d'affinité).
a) Chromatogramme de purification sur colonne d'affinité au nickel du de RB5312 ;b) Gel SDS-PAGE de la fraction la plus intense en sortie de colonne.

Cette protéine s'est révélée moins sensible à la concentration en sel de la solution que les autres enzymes de *R. baltica* de ce chapitre. Le tampon initial de sortie de colonne d'exclusion de taille (100 mM tris-HCl pH 8,0 ; 100 mM NaCl ; 2 % glycérol) a ainsi pu être

changé en un autre moins concentré en sel et sans glycérol (50 mM HEPES-HCl pH 7,5 ; 50 mM NaCl), ces deux éléments pouvant être incompatibles avec la croissance cristalline. Néanmoins, la protéine a présenté une stabilité moins grande dans ce tampon que les autres enzymes de *R. baltica* car elle précipite au bout d'une à deux semaines. Le volume d'élution de RB5312 sur colonne d'exclusion de taille Superdex 75 a été d'environ 60 mL, correspondant bien à une protéine monomérique de 48,7 kDa (Figure III-95). Cela a été corroboré par les mesures de DLS sur la protéine purifiée.

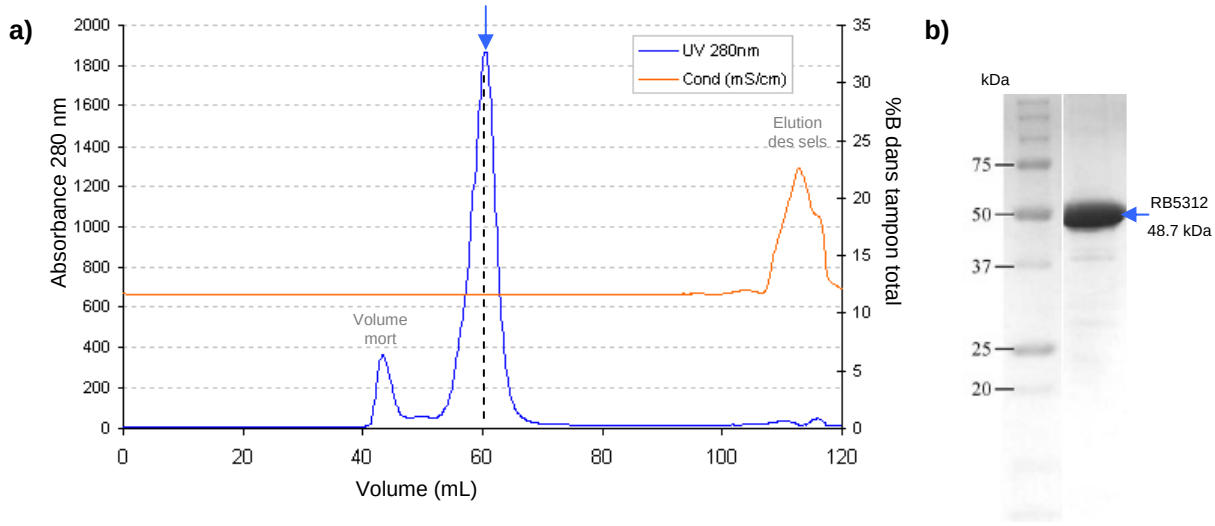


Figure III-95 : Purification de RB5312 (Colonne d'exclusion de taille).
a) Chromatogramme de purification sur colonne d'exclusion de taille de RB5312 ; b) Gel SDS-PAGE de la fraction les plus intense en sortie de colonne.

Le rendement moyen des différentes productions de RB5312 a été de 12 mg de protéine purifiée / L de culture.

IV.B.3 - Tests enzymologiques

La potentielle activité pectinolytique de la protéine RB5312 a été testée en suivant l'apparition d'oses insaturés en C4-C5 par mesure de l'absorbance à 235 nm (Albersheim *et al.*, 1996). Plusieurs sources de pectines disponibles au laboratoire ont été testées: un PGA de citron (Sigma P-9135), un mePGA de pomme (Sigma P-8471) et une pectine nommée SkW de la compagnie d'hydrocolloïdes Degussa. Un test préliminaire réalisé sur ces trois pectines a confirmé l'activité pectinolytique de RB5312, avec une meilleure efficacité sur la pectine SkW et dans une moindre mesure sur le mePGA de pomme (Figure III-96). Cette activité pectinolytique a été également confirmée par C-PAGE, avec l'apparition d'oligosaccharides après action de l'enzyme. Cette technique montre également que la

dégradation de la pectine SKW est plus complète que celle des pectines fournies par SIGMA. La structure exacte de ces pectines n'étant pas connue, il n'est pas possible à ce stade, de distinguer entre les activités pectate lyase (EC 4.2.2.2) et pectine lyase (4.2.2.10). De plus, la pureté de ces substrats était limitée, ce qui rendait difficilement reproductible la mesure de l'activité de RB5312. Pour améliorer la mesure de cette activité, ces pectines ont été partiellement purifiées d'un part par filtration directe sur un filtre Amicon à 0,22 µm et d'autre part par précipitation à l'éthanol 96 % (avec resuspension dans l'eau). La filtration à 0,22 µm a donné de bons résultats sur la pectine SkW et a aussi grandement amélioré le signal de dégradation du mePGA de pomme.

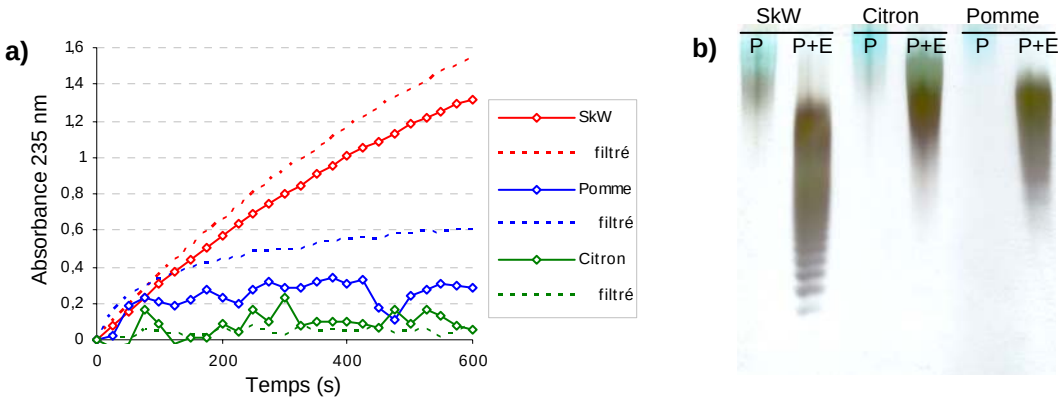


Figure III-96 : Premiers résultats de dégradation de pectines par RB5312.
a) Absorbances à 235 nm montrant la dégradation des pectines par RB5312 ; b) Gel de C-PAGE avec les trois profils de dégradation après 600 s (de gauche à droite : SkW, citron et pomme). P : Polysaccharide seul ; P+E : Polysaccharide avec RB5312.

Pour la détermination de la température optimale d'activité de RB5312, le mélange réactionnel a été incubé au bain marie à des températures variant de 5°C en 5°C de 20°C à 65°C et une mesure a été réalisée au spectrophotomètre toutes les minutes (Figure III-97-a). L'activité pectinolytique est maximale entre 40°C et 55°C (Figure III-97-b). Au-delà de ces températures, l'enzyme précipite et plus aucune activité n'est détectable.

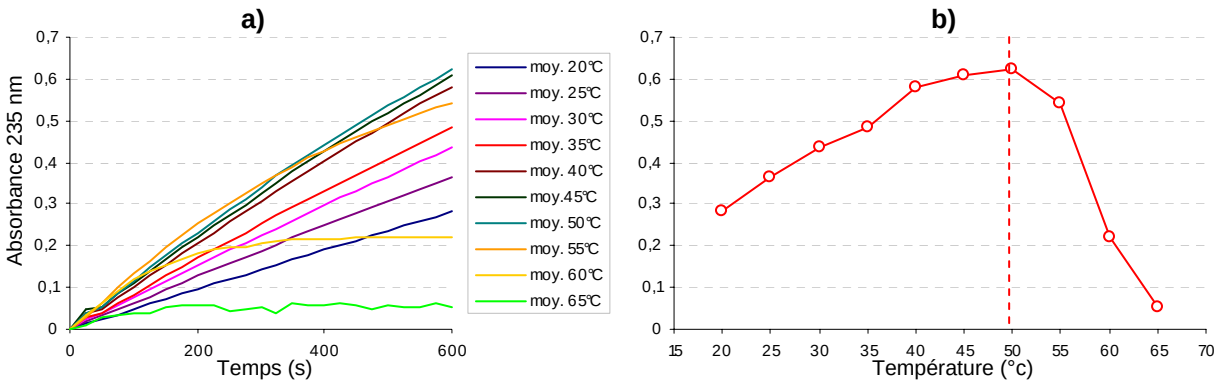


Figure III-97 : Mesure de température optimale de l'activité de RB5312.
a) Absorbances à 235 nm du volume réactionnel à plusieurs températures en fonction du temps (s) ; b) Absorbances à 235 nm à 600 s en fonction de la température. Trait rouge vertical : température optimale.

Pour la détermination du pH optimum, le tampon de la solution à été adapté au pH : 50 mM tris-HCl pour la gamme de pH 7,0 à 9,0 et 50 mM glycine pour la gamme de pH de 9,0 à 10,5. Le mélange réactionnel à été incubé au bain marie à 40°C et une mesure d'absorbance à 235 nm a été réalisée au spectrophotomètre toutes les 1 min (Figure III-98-a). Le maximum d'activité est observé à pH basique à 9,5. Cet optimum de pH est relativement strict car l'activité décroît rapidement dès qu'on s'éloigne de pH 9,5 (Figure III-98-b). L'activité est en particulier quasiment abolie à pH 7,0. Il est à noter que les mesures au-delà de pH 10,5 sont difficiles car la pectine a tendance à gélifier aux pH basiques, rendant les mesures inopérantes.

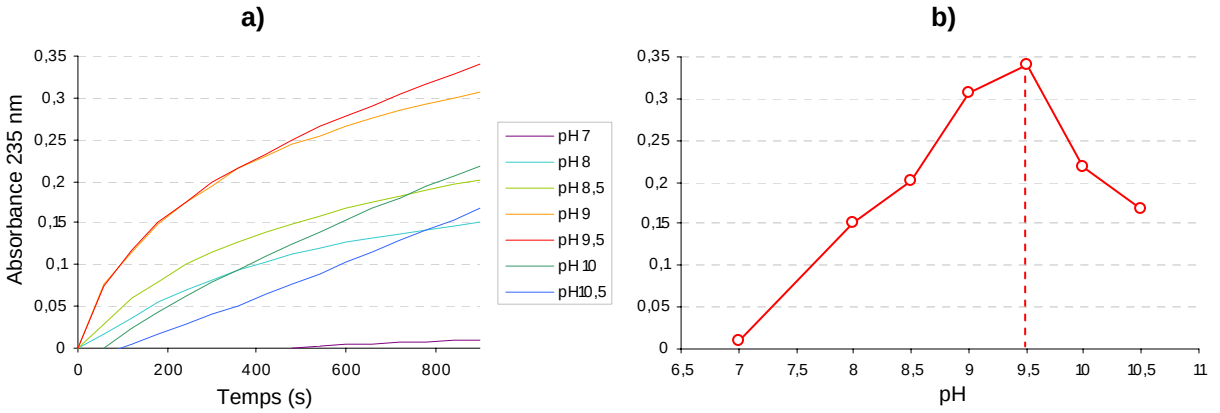


Figure III-98 : Mesure de pH optimal de l'activité de RB5312.
a) Absorbances à 235 nm du volume réactionnel à plusieurs pH en fonction du temps (s) ; b) Absorbances à 235 nm à 15 min en fonction du pH. Trait rouge vertical : pH optimal.

Pour distinguer entre les activités pectate lyase et pectine lyase, nous avons testé si l'activité enzymatique de RB5312 est dépendante de la présence de cations divalents, en

particulier le calcium. En effet, les ions calcium participent au mécanisme catalytique des pectate lyases de la famille PL1, ainsi qu'à la fixation du substrat (Scavetta *et al.*, 1999). Par contre l'activité des pectine lyases est indépendante de la présence d'ions (Matthews, 1968; Mayans *et al.*, 1997). Cela a été testé en incubant RB5312 en sortie de purification en présence ou en absence de 0,5 mM EDTA, un chélateur de cations divalents. Cette expérience a montré que l'EDTA abolissait complètement l'activité pectinolytique de RB5312. (Figure III-99-b). Parmi les cations que nous avons testés, seuls les ions calcium et cobalt ont restauré l'activité de RB5312 (Figure III-99-a). Par conséquent, la protéine RB5312 est bien une pectate lyase, en dépit de la faible similitude de séquence de cette protéine avec des pectate lyases caractérisées.

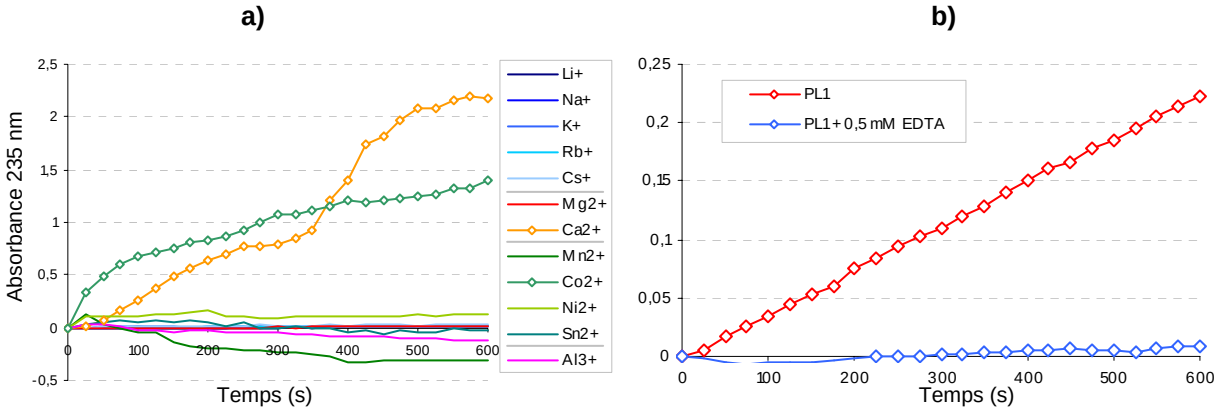


Figure III-99 : Mesure de dépendance ionique de l'activité de RB5312.
a) Absorbances à 235 nm du volume réactionnel en fonction du temps en présence de 5 mM de plusieurs ions métalliques, après incubation avec 0,5 mM EDTA ; b) Absorbances à 235 nm en fonction du temps de RB5312 en sortie de purification et avec 0,5 mM EDTA.

Le Tableau III-24 résume les paramètres biochimiques d'activité de RB5312.

	Optima	Figure de référence
Température	40°C – 55°C	Figure III-97
pH	9,5	Figure III-98
Dépendance ionique	Ca^{2+} et Co^{2+}	Figure III-99

Tableau III-24 : Résumé des caractéristiques biochimiques de RB5312

L'étude des produits limites de dégradation de RB5312 a été également entreprise. Pour cela, nous avons utilisé un acide polygalacturonique issu de citron et qui a été déméthylé par une pectine méthylestérase (Mégazyme). La Figure III-100 présente une

cinétique de dégradation de ce PGA par la protéine recombinante RB5312. Cette pectate lyase a probablement un mode d'action endo avec la libération initiale d'oligosaccharides de grande taille. Au fur et à mesure de la réaction, les fragments de haute masse moléculaire sont clivés en oligosaccharides de taille inférieure. La dégradation semble à complétude après 24h, avec deux produits terminaux majoritaires, bien qu'une fraction non négligeable de cette pectine ne soit pas dégradée par RB5312.

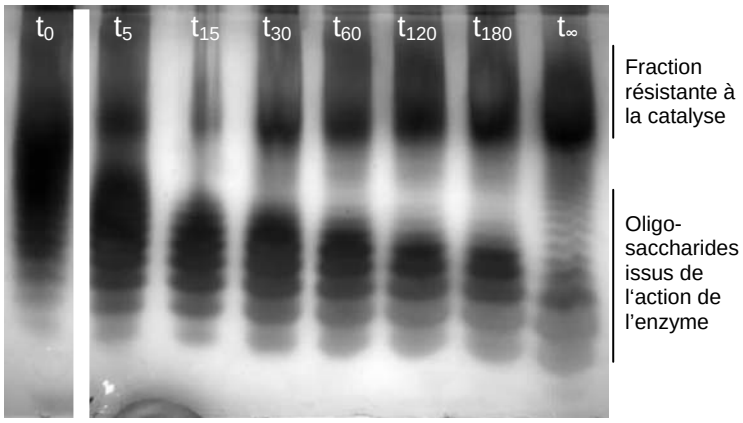


Figure III-100 : Dégradation d'un PGA par RB5312
Gel de C-PAGE après dégradation du PGA Mégazyme par RB5312 en fonction du temps (min). t_0 correspond au polymère avant action de RB5312.

Il est intéressant de constater que le polymère initial contient déjà une quantité significative d'oligosaccharides. Avant d'entamer une étude structurale par RMN des produits terminaux de RB5312, le PGA de Mégazyme a été dialysé en utilisant une membrane Amicon de seuil de coupure 8000 Da, afin d'éliminer les oligosaccharides présents initialement. Ensuite, 20 mL de PGA dialysée (0,2 % w/v) ont été dégradés à 40°C en présence de 5 ng/mL de RB5312 purifiée. Les oligosaccharides libérés ont été partiellement purifiés par chromatographie d'exclusion de taille sur une colonne Superdex S30 26/60 (GE HealthCare) (Figure III-101). Après lyophilisation et échange contre D_2O, ce mélange de produits de dégradation a été envoyé au service RMN de l'Université de Bretagne Occidentale. Une analyse RMN 1H à une dimension est actuellement en cours.

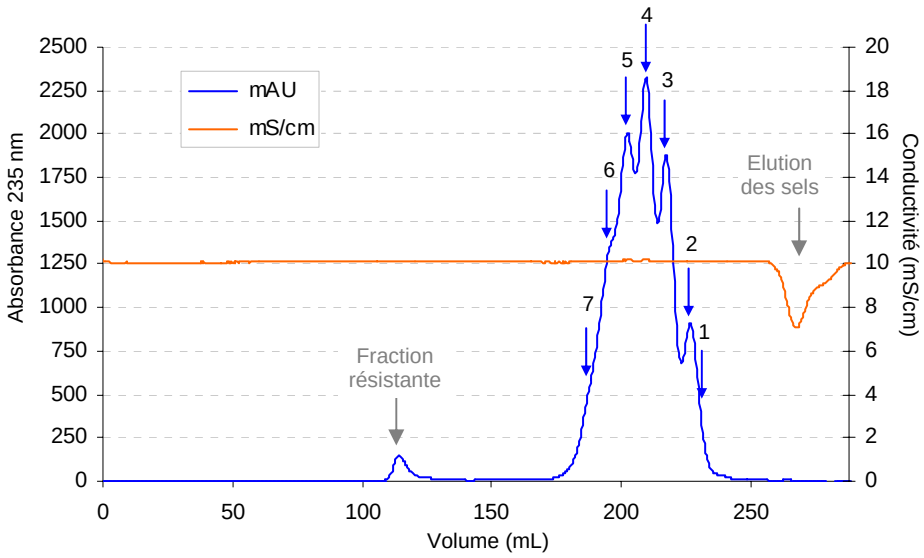


Figure III-101 : Purification d'oligoPGA.
Chromatogramme de sortie de colonne de la purification des oligoPGA issus de la dégradation à complétion du PGA de Mégazyme. Les flèches bleues se situent aux pics d'absorbance à 235 nm correspondant aux oligosaccharides de taille minimum non dégradés par l'enzyme.

IV.B.4 - Cristallogénèse

Les tests de cristallisation de RB5312 ont été réalisés après avoir concentré cette protéine à 9,7 mg/mL. En utilisant un robot dispensant des nano-gouttes (Honeybee, Cartesian), 192 conditions de cristallisation ont été testées à 20°C avec les kits commerciaux Wizard I & II (Emerald BioStructures, Inc.) et JCSG+ Suite (Qiagen). Les gouttes assises ont été réalisées en mélangeant 300 nL de protéine pure (9,7 mg/ML) à 100 nL de solution réservoir. Cette protéine a donné des cristaux dans plusieurs conditions contenant du MPD et du polyéthylène glycol (PEG). Les conditions de cristallisation ont été optimisées manuellement par la technique des gouttes suspendues en mélangeant 2 µl de protéine pure et 2 µl de solution réservoir. Une condition a été optimisée pour produire des cristaux en forme d'aiguilles : 40% 2-méthyl-pentane-2,4-diol (MPD), 4% polyéthylèneglycol (PEG) 8000, 150 mM cacodylate de sodium pH 6.0. Les cristaux présentent une forme qui apparaît tubulaire (certains ont l'air « creux ») et aux extrémités fracturées. Ils ont poussé en l'espace de quelques jours (Figure III-102).

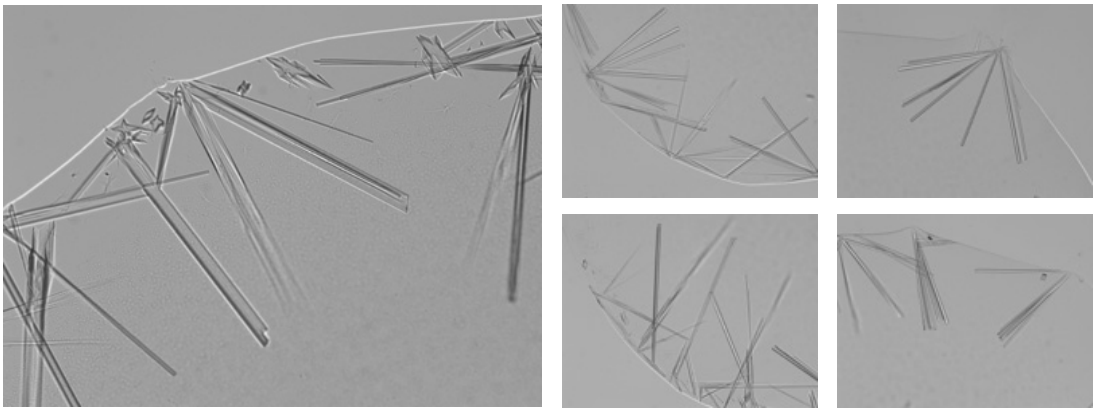


Figure III-102 : Cristaux en forme d'aiguilles de RB5312.

IV.B.5 - Cristallographie

Les expériences de cristallographie par diffraction aux rayons X ont été réalisées sur les cristaux de RB5312. Deux sources de rayons X ont été utilisées : une source classique (anode tournante) au laboratoire et une source synchrotron en partenariat avec le laboratoire européen de rayonnement synchrotron (European Synchrotron Radiation Facility, ESRF).

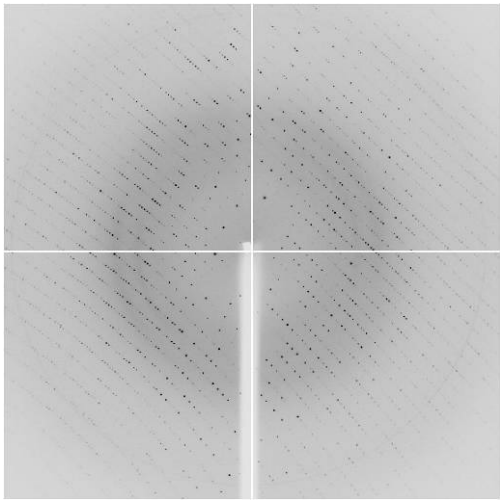
IV.B.5.1 Résultats de cristallographie

Les résultats de cristallogénèse ainsi que les résultats préliminaires de cristallographie sur les cristaux de RB5312 ont fait l'objet d'un article au cours de ma thèse dans Acta Crystallographica Section F (Dabin et al., 2008).

Du fait que les conditions de cristallisation contenaient 40 % de MPD, il n'a pas été nécessaire de tremper les cristaux dans une solution de cryoprotectant avant congélation à l'azote liquide.

Les premières mesures de clichés de diffraction sur les cristaux natifs de RB5312 ont montré que le pouvoir diffractant des cristaux était souvent limité à une résolution autour de 2,5-3,0 Å. En revanche, les statistiques d'analyse des données de diffraction ont toujours été correctes. Cependant, parmi la dizaine de cristaux testés, l'un d'entre eux (dimensions 0,3 x 0,05 x 0,05 mm) a donné des résultats avec une diffraction allant jusqu'à 1,8 Å. En conséquence, un jeu de données complet a été collecté sur ce cristal dont les statistiques de collecte sont résumées dans le **Tableau III-25**. La collection des données a été réalisée à 100 K sur la ligne ESRF ID14-eh2 (longueur d'onde des rayons X fixée à 0,933 Å). Le groupe d'espace a été déterminé comme étant de système de Bravais orthorhombique primitif, $P2_12_12_1$, avec pour paramètres de maille a = 39,05 Å, b = 144,05 Å, c = 153,97 Å. L'analyse

du coefficient de Matthews nous a permis de prédire la présence de deux molécules dans l'unité asymétrique, pour un volume de maille par masse moléculaire de protéine V_M = 2,17 $Å^3.Da^{-1}$ et une proportion de solvant dans le cristal de 43 % du volume (Matthews, 1968).



Première image du jeu de données

Groupe d'espace	P $2_1 2_1 2_1$
Paramètres de maille (Å, °)	a = 39,05, b = 144,05, c = 153,97 α = β = γ = 90
Gamme de résolution (Å)	52,63 – 1,8 (1,90 – 1,80)[1]
Nombre d'observations	304 224 (14 846)
Nombre de réflexions uniques	76 671 (7 858)
Complétude (%)	93,3 (67,0)
$\langle I/\sigma(I)\rangle$	17,3 (3,9)
Redondance	4,0 (1,9)
R_{merge}[2] (%)	5,4 (16,7)

[1] Entre parenthèses : Données correspondant à la plus haute sphère de résolution

[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$

Tableau III-25 : Statistiques de la collection de données des cristaux natifs de RB5312.

Les données brutes ont été intégrées et mises à l'échelle respectivement par les logiciels *MOSFLM* (Leslie, 1992) et *SCALA*, qui font partie de la suite intégrée de logiciels CCP4 (Collaborative Computational Project Number 4, 1994).

IV.B.5.2 Phasage des jeux de données

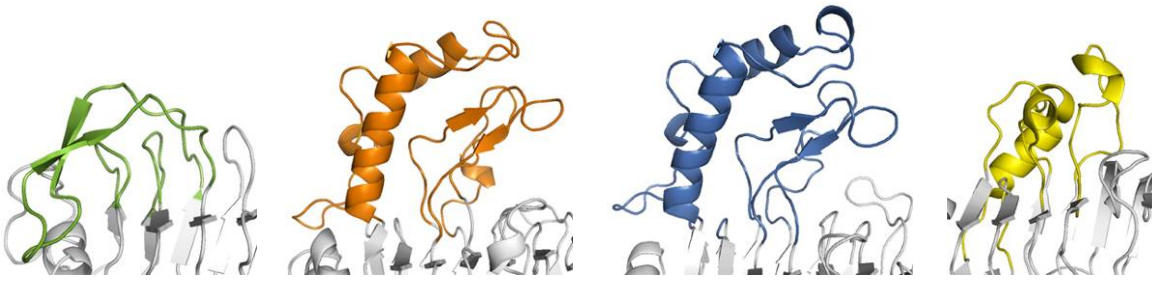
Plusieurs tentatives de phasage des données ont été réalisées. Nous avons tout d'abord utilisé le **remplacement moléculaire**, qui consiste à exploiter l'information structurale de protéines similaires et permet un phasage rapide en cas de succès. Nous nous sommes également tournés vers des techniques expérimentales d'incorporation d'atomes lourds. Nous avons ainsi pu tester la **méthode MIR**, en profitant de la présence probable de calcium dans la structure de RB5312 pour le remplacer par de l'erbium et de l'ytterbium. Enfin, nous avons testé la **méthode MAD**, en réalisant des productions de protéines sélénométhionylées. Les différents résultats obtenus avec ces techniques sont exposés dans cette section.

IV.B.5.2.1 Remplacement moléculaire

Avec un jeu de données natif à haute résolution et devant le nombre raisonnable de structures disponibles dans la famille PL1, une tentative de phasage par remplacement moléculaire a l'aide du logiciel *AMoRe* (Navaza, 2001) a semblé une première méthode à essayer. Nous nous attendions cependant à rencontrer quelques écueils, ne serait-ce que du fait que RB5312 présente une réelle divergence par rapport aux séquences de sa famille, y compris au sein de régions pourtant très conservées. Un autre écueil dont nous avions conscience est la structure même des hélices β, qui ont grossièrement la forme d'une spirale droite de section triangulaire. Dans ce type de structures, il est possible de générer de nombreuses solutions similaires à une fraction de tour d'hélice près, bruitant les solutions de la fonction de rotation.

RB5312 s'est avérée présenter de très faibles similitudes avec toutes les séquences de la banque PDB. Le meilleur score par BLAST, avec seulement 14% de similitude, s'est trouvé être la protéine non fonctionnelle Juna1 de *J. ashei* (code PDB 1PXZ), présentée dans l'introduction sur la famille PL1. La zone alignable est essentiellement localisée dans les parties N-terminales des deux protéines. Les alignements générés entre ces protéines intègrent de plus de nombreuses zones d'insertions chez RB5312.

Aucune solution avec contraste n'ayant été obtenue avec les premiers tests utilisant la structure 1PXZ de Juna1, une série de structures chimériques a été réalisée à partir des différentes structures connues. En effet, il s'avère que si le cœur des hélices β est très conservé, en revanche de larges insertions très divergentes (tant en taille qu'en structure) apparaissent dans l'ensemble de la famille, typiquement au niveau de la région N-terminale du feuillet PB1 des protéines. Quelques exemples de ces zones divergentes sont donnés dans la Figure III-103.



Juna1 - **1PXZ** *J. ashei* | Pl47 - **1VBL** *Bacillus sp.* | Pel - **1BN8** *B. subtilis* | PelC - **2EWE** *E. chrysanthemi*

Figure III-103 : Extensions dans les structures de la famille PL1.
Présentation des extensions du feuillet PB1 de quatre structures de la famille PL1, leur hélice β étant alignée.

Avec les différentes chimères générées, j'ai essayé de couvrir l'ensemble de ces extensions, sachant qu'il n'est pas possible de prédire laquelle est la plus probable, étant donnée encore une fois la forte divergence de RB5312. Trois versions des hélices β ont été également générées : une plus courte d'un tour d'hélice et deux plus longues respectivement d'un et de deux tours d'hélices. Je ne vais pas présenter exhaustivement l'ensemble de ces chimères structurales, la Figure III-104 se propose d'en donner deux exemples.

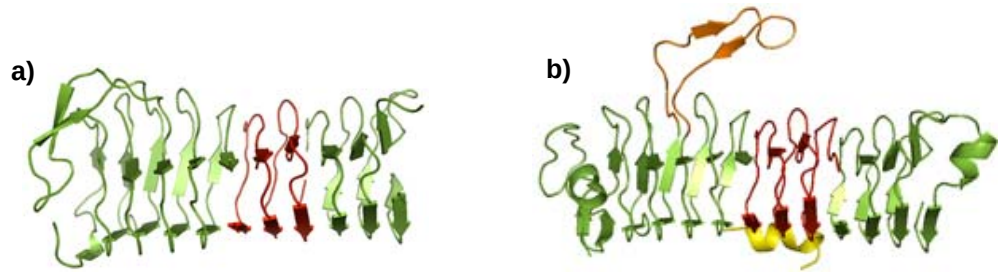


Figure III-104 : Chimères de structures de la famille PL1.
Présentation de deux chimères structurales basées sur l'hélice β de 1PXZ et possédant des insertions provenant de plusieurs structures de la famille PL1. a) Chimère reprenant le repliement de 1PXZ avec son extension mais privée de ses extrémités C- et N-terminales. Une insertion de deux tours d'hélice dans l'hélice β est visible (en rouge) ; b) Chimère reprenant le repliement de 1PXZ, avec une partie de ses extrémités C- et N-terminales avec trois modifications : une insertion de deux tours dans l'hélice (en rouge), une hélice α commune à de nombreuses structures, se repliant traditionnellement sur la face opposée au site actif (jaune) et enfin, une des extensions de 1VBL (en orange).

Ces chimères n'ont malheureusement pas non plus abouti à la génération de solutions exploitables. Nous sommes donc rapidement passés à la détermination de la phase par des méthodes expérimentales.

IV.B.5.2.2 Phasage expérimental par la méthode MIR

Afin de produire des phases par la technique MIR, un trempage des cristaux de RB5312 native dans des solutions contenant 5 mM de nitrate d'ytterbium et de nitrate d'erbium a été réalisé.

Ces terres rares ont été choisies d'une part pour leur facilité à échanger les ions calcium, que nous savions être fixés dans le site actif de la protéine (Pickersgill *et al.*, 1994), et d'autre part car elles possèdent une bande très intense d'absorption à des longueurs d'onde accessible en routine sur les lignes de lumière de l'ESRF. Il était donc envisageable d'effectuer des mesures MAD avec les mêmes jeux. La présence ainsi que la quantité des

métaux lourds ont été estimées grâce à l'intensité des spectres de fluorescence X réalisés sur les cristaux trempés. Il est ainsi apparu que seul l'ytterbium avait été incorporé dans les cristaux, cependant en faible quantité.

Ne possédant qu'un unique dérivé lourd, nous avons choisi de tirer parti du signal anomal de l'ytterbium. Trois jeux complets de données ont donc été collectés sur le cristal ayant incorporé le plus de métal aux longueurs d'onde correspondant respectivement au maximum de la composante imaginaire du coefficient d'absorption de l'ytterbium, au point d'inflexion de sa composante réelle et loin du seuil d'absorption soit respectivement à 1,18 Å, 1,19 Å et 1,21 Å.

Il est apparu dès l'intégration des tâches de diffraction que ces cristaux supportaient bien plus difficilement l'irradiation que les cristaux natifs. Sur les cinq cristaux présentant une incorporation suffisante d'ytterbium, seulement deux ont pu produire des jeux de données de bonne qualité. De plus, seule la première collecte de ces cristaux (au pic d'absorption à 1,18 Å) a été exploitable. En effet, la qualité des données s'est malheureusement détériorée au fil des irradiations et les collectes aux autres longueurs d'ondes ont révélé une forte dégradation des données. La sélection d'images correspondant à des sections angulaires des collectes présentant de meilleures statistiques n'a pas permis d'amélioration notable de la qualité de ces données.

L'analyse des données a révélé que le signal anomal était très faible. Le R_{ano} s'est ainsi trouvé systématiquement inférieur au R_{merge} du jeu de données, ce qui signifie que l'intensité du signal anomal est inférieure au bruit de fond moyen du jeu de données. Il semble en particulier que soit apparue une dérive des statistiques au cours des collectes successives des jeux de données, en particulier au niveau de la mosaicité des clichés de diffraction, signe que le rayonnement X a particulièrement affaibli le cristal au cours des collectes.

L'expérience de remplacement du calcium par une terre rare n'ayant pas abouti, nous avons abandonné la recherche de dérivés par trempage dans des solutions de sels d'atomes lourds, cette méthode pouvant être parfois assez hasardeuse. Nous avons entrepris de marquer de manière covalente la protéine avec un diffuseur anomal, technique qui présente l'avantage d'être moins aléatoire. Je suis donc passé à la production de la protéine RB5312 avec incorporation de sélénométhionine, pour résoudre la phase par mesures de MAD sur le sélénium.

IV.B.5.2.3 Tentative de phasage expérimental par la méthode MAD au sélénium

Une production de la protéine RB5312 avec incorporation de sélénométhionine a été réalisée. L'efficacité du remplacement a été mesurée par estimation de la masse des protéines par spectrométrie de masse. En réalité, trois protocoles de marquage ont été utilisés au cours de ma thèse.

Le premier a utilisé le milieu de culture PASM-5052 (Studier, 2005). Ce protocole, basé sur celui en ZYP-5052 déjà utilisé pour les productions régulières des protéines, nous a séduit par sa très grande simplicité : il était compatible avec l'utilisation d'une souche non auxotrophe à la méthionine et permettait d'obtenir de grandes quantités d'enzymes marquées, ce qui n'est pas toujours évident avec d'autres milieux. Il n'a cependant pas donné les résultats escomptés en produisant notamment des protéines dont le remplacement des méthionines n'a pas été systématique. En particulier deux populations ont été produites : une avec trois remplacements sur les huit méthionines de la protéine native, et l'autre avec quatre remplacements.

Un second protocole a donc été utilisé, cette fois-ci plus traditionnel. Il a tout d'abord fallu retransformer une souche bactérienne auxotrophe à la méthionine, pour nous assurer que le remplacement serait total. La souche B834(DE3) a été utilisée. Il s'agit d'une souche mutée de la souche BL21, utilisée pour la production de la protéine native, et également lysogénisée par le bactériophage λ(DE3). La production de la protéine séléniée a consisté à cultiver cette souche en milieu LB jusqu'à atteindre une absorbance de la solution d'environ 1 unité, puis à transférer les bactéries dans un milieu frais (après centrifugation), contenant de la sélénométhionine à la place de la méthionine. La production a été alors induite par ajout d'IPTG au milieu. Les résultats ont été analysés par spectrométrie de masse (figure X).

Une production avec un troisième protocole a été tentée pendant que les mesures de spectrométries de masse étaient en cours en utilisant cette fois le protocole décrit par Doublié et collaborateurs en 1997 (Doublié, 1997). Il s'agit d'une autre philosophie : cette fois-ci, une souche non auxotrophe est envisageable. La culture se réalise dans un milieu M9, supplémenté avec de la sélénométhionine à la place de la méthionine. L'ajout d'acides aminés connus pour inhiber la voie de biosynthèse de la méthionine quand ils sont en excès nous assure que la bactérie ne va pas produire de méthionine endogène. Ces deux dernières méthodes ont en réalité donné un remplacement uniforme de tous les sites. Les résultats suivants ne concerneront donc que la production réalisée par le second protocole.

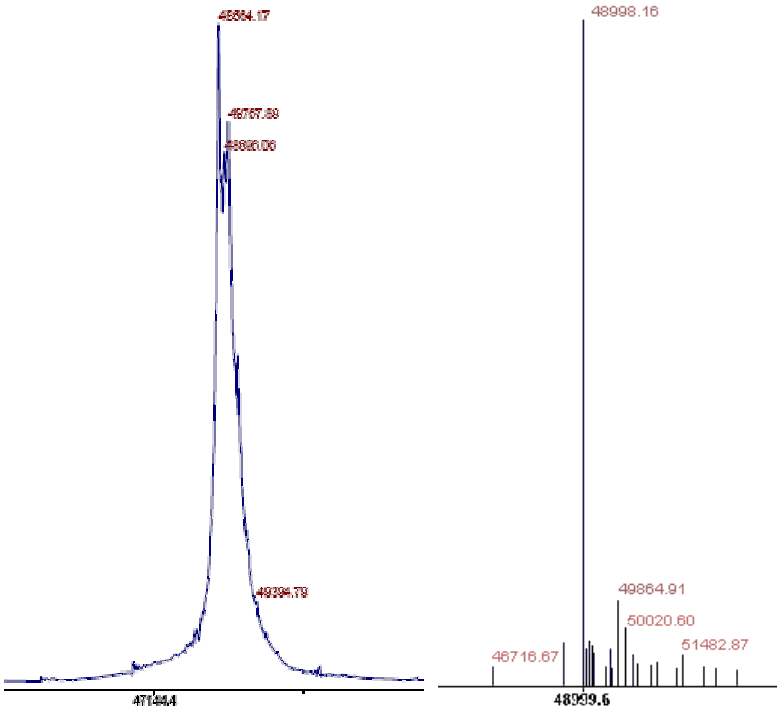


Figure III-105 : Diagramme de spectrométrie de masse.
Le diagramme de gauche présente la protéine native (pic à 48 554 Da). Le diagramme de droite présente la protéine séléniée (pic à 48 997 Da). La masse de la protéine séléniée correspond à un écart de 433 Da, soit, aux incertitudes de mesures près à 8 incorporations de sélénium.

Les rendements de la production de cette protéine ont été beaucoup plus faibles que ceux de RB5312 native, autour de 1 mg de protéine purifiée, par litre de culture. La concentration a été amenée à 4,4 mg/mL pour les tests de cristallisation. En utilisant le robot de cristallisation, 192 conditions ont été testées à 20°C en utilisant les kits commerciaux Wizard I & II (Emerald BioStructures, Inc.) et JCSG+ Suite (Qiagen). Des gouttes assises ont été réalisées en mélangeant 300 nL de protéine à 100 nL de solution réservoir. La protéine séléniée n'a pas présenté les mêmes conditions de cristallisation que la protéine native, avec une préférence pour les conditions en sels. Elles ont été optimisées manuellement par la technique des gouttes suspendues en mélangeant 2 µl de protéine pure et 2 µl de solution réservoir. La condition optimisée a permis de produire des cristaux en 1 M LiCl, 18% PEG 6000, 0,1 M Citrate de sodium pH 6,0. Elle a présenté un profil cristallin également différent : des petits oursins de plaquettes, qui ont poussé en quelques jours.

Trois jeux de données ont été collectés à trois longueurs d'onde autour du pic d'absorption du sélénium, à l'ESRF, sur la ligne ID23-EH1. Le spectre de fluorescence X mesuré sur le cristal est présenté Figure III-106. Les statistiques des jeux de données sont montrées dans le Tableau III-26.

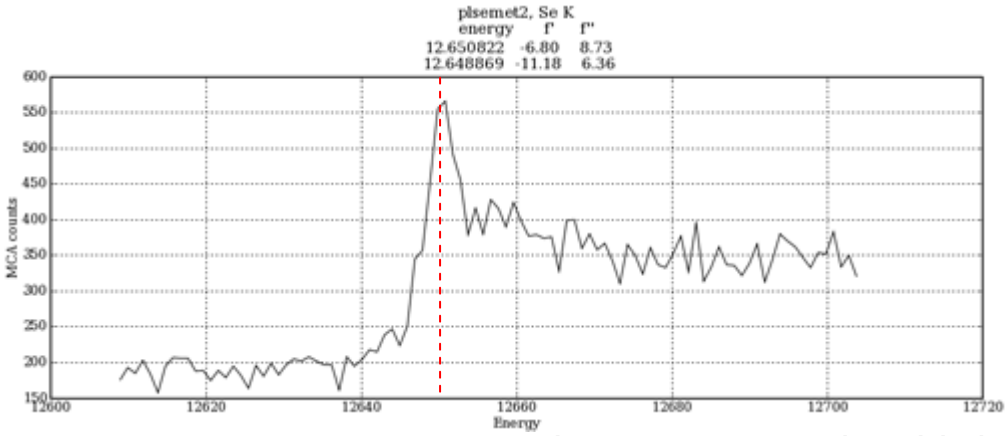


Figure III-106 : Spectre d'absorption mesuré sur un cristal de protéine séléniée.
Spectre mesuré à la longueur d'onde d'absorption de la raie K_α du Se. Le trait matérialise le pic d'absorption.

Groupe d'espace		$P\,2_1\,2_1\,2_1$	
Longueur d'onde (Å)	0,93	0,94	0,92
Paramètres de maille (Å, °)	a = 39,04 b = 146,82 c = 154.50 α = β = γ = 90		
Gamme de résolution (Å)	40-3.16 (3.33-3.16)	40-3.16 (3.33-3.16)	40-3.16 (3.33-3.16)
Nombre d'observations	110160	110686	110819
Nombre de réflexions uniques	16121	16171	16328
Complétude (%)	99.8 (99.8)	99.8 (99.8)	99.8 (99.8)
⟨I/ σ(I)⟩	23.9 (16.9)	22.8 (15.7)	25.4 (7.0)
Redondance	6.8 (7.1)	6.8 (7.1)	6.8 (7.0)
R_{merge}[2] *(%)*	8.6 (10.3)	8.3 (10.6)	7.7 (10.6)
Redondance Ano	3.7	3.6	3.6
Complétude Ano (%)	99.9	99.9	99.8
R_{ano} *(%)*	3.1 (2.5)	3.2 (2.4)	2.9 (2.2)

[1] Entre parenthèses : Données correspondant à la plus haute sphère de résolution
[2] $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$

Tableau III-26 : Statistiques de collecte de données avec RB5312 séléniée

Les données ont été intégrées et mises à l'échelle entre elles, par rapport au jeu sur le pic d'absorption. L'intensité du signal anomal a été estimée grâce au logiciel XPREP. Un extrait des résultats de ce logiciel est montré dans le Tableau III-27 suivant.

Resl.	Inf - 8.0	- 6.0	- 5.6	- 5.4	- 5.2	- 5.0	- 4.8	- 4.6	- 4.4	- 4.2	- 3.98
N(data)	1545	2072	808	512	578	699	803	925	1138	1338	1623
<I/sig>	39.5	26.1	22.0	21.7	21.3	20.8	21.8	22.3	21.5	18.1	15.6
%Compl	98.0	99.1	99.5	99.6	99.8	99.3	99.4	99.5	99.3	99.3	99.1
<d"/sig>	2.40	1.65	1.41	1.15	1.15	1.11	1.12	1.08	1.00	0.96	0.94

Resl : Sphères de résolution
N(data) : nombre de taches de diffraction par sphère de résolution
<I/sig> : intensité signal sur bruit ;
%Compl : complétude la sphère de résolution ;
<d"/sig> : intensité du signal anomal sur le bruit

Tableau III-27 : Résultats de XPREP

Il apparaît à la lecture du Tableau III-27 que l'intensité du signal anomal est faible. Le rapport <d''/sig> traduit la moyenne de l'amplitude du signal anomal, calculé par la différence d'intensité moyenne sur les taches de Friedel séparée par de la diffraction anomale, par rapport à la différence d'intensité des taches sans signal anomal. En dessous d'une valeur de 1,2, les variations sont considérées comme aléatoires. Il apparaît que seules les données supérieures à 5,6 Å de résolution (colonne jaune du Tableau III-27) portent un signal anomal, relativement faible.

Des tentatives de phasage à partir de ces données, et à la résolution maximum de 5,6 Å ont été tentées en utilisant plusieurs logiciels : Solve/Resolve et Sharp. Aucun n'a convergé vers une solution. Sur les différentes cartes de densité électronique calculées quelques zones sont apparues, correspondant probablement d'après leur nombre (six), aux sites des séléniums. Il apparaît que ces sites se situeraient dans des zones adjacentes, et il n'a pas été possible d'aller plus loin dans l'utilisation des données.

J'ai tenté une déconvolution à la main des cartes de Patterson de différence anomale, qui traduisent l'information anomale brute des jeux de données. Elles sont calculées pour chaque longueur d'onde et pour chaque section de Harker. Ces sections représentent des positions spéciales de la maille cristalline dans lesquelles toute l'information de diffraction du cristal apparaît. L'exercice consiste à repérer les intensités de diffraction à la fois communes aux cartes de différence anomales, et non représentées dans les cartes natives. Le groupe d'espace des cristaux ($P2_12_12_1$) génère trois sections de Harker : x = 0,5 ; y = 0,5 ; z = 0,5. Les intensités observables sur ces diagrammes correspondent à des zones de la maille présentant un signal de diffraction fort, et en l'occurrence, à une diffraction anomale forte. La Figure III-107 présente l'un de ces diagrammes, comparé au même diagramme, cette fois sans tenir compte du signal anomal. La différence des deux est sensée refléter ce signal.

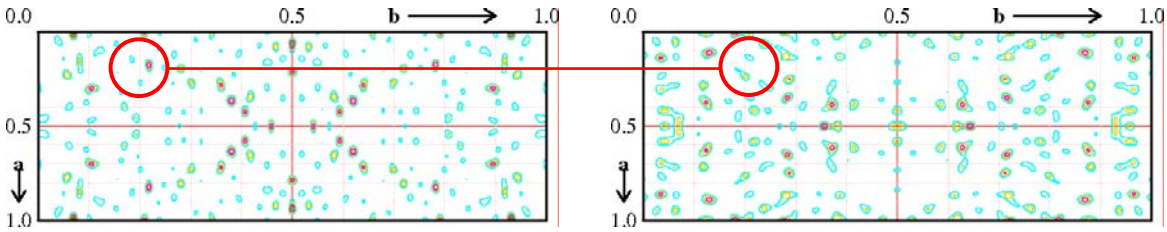


Figure III-107 : Cartes de Patterson des section de Harker z=0,5.
Carte de gauche : carte de différence anomale ; carte de droite : carte native. Les dimensions sont exprimées en pourcentage des dimensions atomiques de la maille. Un pic de la carte anomale, non présent dans la carte native est matérialisé par un cercle rouge.

Il suffit, pour faire correspondre les différentes cartes entre elles, de regénérer « à la main » les positions symétriques de chaque intensité intéressante. Ces positions équivalentes générées par la symétrie du cristal sont accessibles dans les tables de cristallographie. Si les données sont cohérentes, les taches, après application des symétries, doivent se superposer.

Il apparaît que le signal anomal est réellement faible et que les données ne sont pas exploitables, même en cherchant les signaux les plus faibles…

IV.B.6 - Discussion

RB5312 aura vraiment été une cible tout à la fois passionnante et frustrante à plus d'un titre. Ses purifications se sont plutôt bien passées, et n'ont pas présentées de difficultés majeures. Il a fallu un peu ajuster les tampons de la forme séléniée de cette protéine, qui était plus hydrophobe et tolérait moins bien les faibles salinités. Nous l'avons finalement purifiée en 200 mM de LiCl dans les tampons de colonne d'exclusion de taille, vu que les conditions de cristallisation comportaient 1 M de LiCl.

Sur le plan biochimique, l'activité pectate lyase pose plus de questions qu'elle n'apporte de réponses. En effet, Shlechner et al, 2003 ont montré que *Rhodopirellula baltica* ne croissait pas avec la pectine comme unique source de carbone. Effectivement, dans son génome, la voie métabolique de la dégradation de la pectine n'apparaît pas de manière évidente (ceci sera discuté dans le chapitre IV). Nous nous attendions donc à ce que cette enzyme présente une activité sur un substrat peut-être potentiellement proche sans être pectique, de type alginique. Il est certain que le polymère dégradé par *R. baltica* dans son environnement naturel peut être d'une autre nature, des analyses transcriptomiques pourraient typiquement répondre à ce genre de question. Biochimiquement, il reste que l'activité de RB5312 sur la pectine est très intense et il ne fait aucun doute que le substrat

naturel de cette enzyme est, sinon de la pectine elle-même, du moins très proche. Enfin, il devrait être possible sous peu de caractériser les oligosaccharides limites, vu que les expériences ^{1}H RMN ont eu lieu et que les spectres mesurés sont en attente d'exploitation.

Il est possible d'imaginer plusieurs scénarios expliquant pourquoi *R. baltica* pourrait posséder des enzymes de dégradation de la pectine qui *a priori* ne lui apporteraient aucun avantage puisqu'elle ne peut pas utiliser les produits formés. Par exemple, il pourrait être imaginée une association microbienne pour la dégradation du mélange de substrats que constituent les neiges marines. Elles sont en effet un agrégat informe de plusieurs polymères, où chaque population microbienne apporterait un élément à la dégradation globale du macroagrégat. Il est également possible d'envisager que *R. baltica* ait acquis cette enzyme pour permettre la déliquescence de la paroi de sa cible pour accéder aux substrats sur lesquels elle peut croître.

Sur le plan structural, cette enzyme aura été très frustrante. Les résultats de cristallisation de la protéine native étaient pourtant de bon augure. Un jeu de données avec de très bonnes statistiques, et une très bonne résolution a été collecté. Cependant, les tentatives de phasages se seront révélées sans issues jusqu'au bout. Quelque chose reste très surprenant dans ces résultats. Il est clair que les jeux de données, sauf dégradation due aux rayonnement X, sont plutôt bons. La redondance des données, la complétude, le rapport signal sur bruit, la mosaicité, tous ces indicateurs sont plutôt positifs. Seul le signal anomal est faible. Un phénomène est à considérer : il semble qu'il y ait deux molécules dans l'unité asymétrique. Nous pensons probable que l'une des deux présente un certain désordre, qui « tue » le signal anomal, du fait qu'il doit y avoir peu de diffuseurs anomaux contribuant au signal. Cela n'affecterait pas le signal non anomal car il est beaucoup plus fort.

Quoi qu'il en soit, nous sommes retournés à une stratégie MIR et des cristaux trempés dans beaucoup plus de métaux ont été réalisés. Le crible des sels métalliques s'incorporant au cristal a été réalisé d'une autre manière, avec un gel retard. Cette technique consiste à faire migrer la protéine mélangée à la solution métallique dans un gel PAGE natif ; les échantillons fixant l'ion lourd migrent plus lentement. Les sels de mercure semblent ainsi présenter une piste intéressante. Des tests de cristallisation avec du mercure seront tentés prochainement. D'autres métaux seront criblés et nous espérons pouvoir en avoir suffisamment pour réaliser un phasage cette fois de qualité.

Un point très positif de toutes ces expériences restera que j'ai vu les différentes procédures du traitement du signal quand les jeux de données ne sont pas bons. Même si j'espère pouvoir un jour réussir une résolution de structure de protéine sans encombre, cette expérience me sera je pense très profitable dans les cas difficiles.

V - Matériels et méthodes

V.A - Production des protéines

V.A.1 - Production de protéines recombinantes natives

Le milieu ZYP-5052 est un milieu de culture bactérienne développé pour la production de protéines recombinantes ayant pour base le milieu ZY, un mélange d'extrait de levure à 5 g/L et d'acides aminés produits par digestion tryptique de caséine (de type tryptone) à 10 g/L. A cela sont ajoutés dans l'ordre 1 mM $MgSO_4$; 0,5% glycérol ; 0,05% glucose, 0,2% α-lactose ; 25 mM $(NH_4)_2SO_4$; 50 mM KH_2PO_4 et 50 mM Na_2HPO_4.

Une préculture de 2 mL de milieu LB additionné de 100 µg/mL d'ampicilline a été inoculée depuis le stock glycérol constitué à partir de la plaque de plasmides initiale et laissée 12 h à 37 °C sous agitation à 250 rpm.

Une culture de 200 mL de milieu ZYP-5052 additionnée de 100 µg/mL d'ampicilline a ensuite été inoculée avec les 2 mL de préculture et l'ensemble a été laissé en enceinte thermostatée à 37 °C sous agitation à 250 rpm. Un prélèvement de 1 mL a été effectué deux fois par jour pour suivre la croissance bactérienne par mesure de la turbidité de la solution (lecture spectrophotométrique de l'absorbance de la solution à 600 nm). La culture a été arrêtée lorsque la phase stationnaire est observée après la phase de croissance exponentielle (typiquement 3 à 4 jours). L'absorbance à 600 nm observée doit être alors supérieure à 4 pour que l'expression soit considérée comme ayant été induite.

Après arrêt de la culture, les cellules ont été centrifugées à 6000 rpm et 4 °C pendant 30 min et les culots stockés en congélateur -80 °C (où ils peuvent rester jusqu'à plusieurs mois).

V.A.2 - Production de protéines recombinantes séléniées

Au cours de ma thèse, deux protocoles de marquage aux sélénométhionines des protéines recombinantes ont été utilisés pour exprimer la protéine RB5312 (pectate lyase,

famille PL1), le premier n'ayant pas permis une sélénométhionylation uniforme de la protéine engendrant quelques difficultés lors des expériences de diffusion anomale multiple (Multiple Anomalous Diffraction, MAD).

Le premier protocole s'utilise indifféremment avec une souche auxotrophe à la méthionine (telle que la souche *E. coli* B834) ou avec une souche non auxotrophe (telle que la souche *E. coli* BL21). Il est basé sur l'utilisation du milieu autoinductif PASM-5052, développé par Studier (Studier, 2005). Ce milieu est une version modifiée du milieu autoinductif ZYP-5052 permettant la production de protéines recombinantes sélénométhionylées. La différence entre ces deux milieux réside dans leur source de carbone et d'acides aminés. En effet, le milieu PASM-5052 incorpore l'ensemble des additifs du milieu ZYP-5052 dans de l'eau stérile, sans l'extrait de levure ni la tryptone. En remplacement est ajouté extemporanément une solution de 17 acides aminés (tous sauf cystéine, méthionine et tyrosine) pour une concentration finale de 200 mg/L par acide aminé, suivi de l'ajout de 10 mg/L de méthionine, 100 nM de vitamine B12 et 125 mg/L de D-sélénométhionine (ou 250 mg/L d'un mélange racémique de DL-sélénométhionine). La culture est ensuite réalisée dans les mêmes conditions que décrites précédemment (cf I-A-1).

Le second protocole est assez différent dans son principe. Il se base sur l'utilisation d'un milieu minimum de type M9, avec induction de l'expression en présence d'isopropyl β-D-1-thiogalactopyranoside (IPTG). Ce protocole est de plus réservé à l'utilisation d'une souche bactérienne auxotrophe à la méthionine (de type *E. coli* B834). Tout d'abord une préculture de 10 mL de milieu LB additionné de 100 µg/mL d'ampicilline a été inoculée à partir du stock glycérol et laissée 12 h à 37 °C sous agitation à 250 rpm. Un volume de 1 L de milieu LB-ampicilline (100 µg/mL) a ensuite été inoculé avec la préculture et laissé sous agitation à 250 rpm et 37 °C jusqu'à ce que le milieu atteigne une absorbance à 600 nm entre 0,8 et 1,0. L'ensemble des étapes suivantes a été réalisée au maximum à basse température (sur glace, centrifugations à 4 °C) et stérilement (sous hotte avec stérilisation de tous les ustensiles par trempage dans l'éthanol 20%). La culture a été centrifugée stérilement 15 min à 4000 rpm et 4 °C, le surnageant a été enlevé et le culot bactérien transféré dans 500 mL de la solution saline servant de base au milieu M9 (22 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.5 mM NaCl, 19 mM NH_4Cl). Le culot a été resuspendu jusqu'à homogénéité, centrifugé (15 min, 4000 rpm, 4 °C) et le surnageant a été une nouvelle fois enlevé. Finalement, le culot a été retransféré dans 1 L d'une solution de sels du milieu M9 supplémenté pour la culture avec 100 nM vitamine B12, 50 µM $CaCl_2$, 500 µM $FeCl_3$, 2 mM $MgSO_4$, 8 mM glucose, ½ concentration de la solution de métaux du milieu ZYP-5052, une solution de 19 aminoacides (tous sauf la méthionine) pour une concentration finale de 42

mg/L par aminoacides et 10 mg/L de D-sélénométhionine (ou 20 mg/L d'un mélange racémique de DL-sélénométhionine). Une fois l'ensemble bien homogénéisé, 1 mM d'IPTG a été ajouté pour induire l'expression.

La culture a alors été placée 18 h en incubateur sous agitation à 250 rpm et à 20 °C. Après arrêt de la culture, les cellules ont été centrifugées à 6000 rpm et 4 °C pendant 30 min et les culots stockés en congélateur -80 °C.

V.B - Purification des protéines recombinantes

Le culot bactérien congelé contenant la protéine surexprimée a été resuspendu dans 50 mL de **Tampon A** (50mM Na_2HPO_4 pH 8,0 ; 500mM NaCl ; 50mM imidazole ; 5% glycérol), additionné d'inhibiteur de protéases Complete Protease Inhibitor Cocktail (Roche) et de benzonase (Merck). Les cellules ont été lysées par presse de French et le lysat a été ultracentrifugé à 20 000 ***g*** et 4 °C pendant 2 h.

Le surnageant a ensuite été chargé sur une colonne d'affinité au nickel HisTrap Fast Flow (GE HealthCare), elle-même préchargée avec 100 mM $NiSO_4$ et prééquilibrée en tampon A, sur un système Äkta Purifier FPLC (Amersham Pharmacia Biotech) au débit de 1 mL/min.

La colonne a ensuite été lavée avec du tampon A jusqu'au retour à la ligne de base de l'absorbance à 280 nm en sortie de colonne. Une première étape dans la purification a alors été un palier de lavage avec un mélange tampon A et 4% de **Tampon B** (50mM Na_2HPO_4 pH 8,0 ; 500mM NaCl ; 400mM imidazole ; 5% glycérol), pour une concentration en imidazole en entrée de colonne de 65 mM. Lorsque l'absorbance à 280 nm retourne une nouvelle fois à la ligne de base, un gradient entre les tampons A et B allant de 4% à 100% de tampon B est appliqué à la colonne, au débit de 1 mL/min. La sortie de colonne a été divisée en fractions de 1 mL tout au long de la purification. En fin de gradient, les fractions ont été analysées en SDS-PAGE et celles présentant la protéine d'intérêt en bonne pureté sur gel ont été regroupées. Le volume de cette solution a été réduit à 2 mL par ultrafiltration sur un système Amycon (membrane en polyethersulfone, de taille d'exclusion dépendant de la protéine purifiée, généralement choisie pour correspondre à la moitié de la taille de la protéine d'intérêt) sous pression de diazote.

La purification s'est achevée par une dernière étape de purification sur une colonne d'exclusion de taille de type Superdex 75 HiLoad 16/60 prep grade (S75 - GE HealthCare),

ou Superdex 200 HiLoad 16/60 prep grade (S200 - GE HealthCare). Ces colonnes se distinguent par le maillage de leur gel respectif (elles sont toutes deux de même composition et ont un même volume total de 120 mL), permettant une gamme de séparation en fonction de la masse moléculaire des échantillons drastiquement différente l'une de l'autre : la S75 peut séparer des protéines dans la gamme 10 kDa – 70 kDa et la S200 pour dans la gamme 20 kDa – 200 kDa. Après équilibration de la colonne en **Tampon C** (50mM tris-HCl pH 7.5, 100mM NaCl), l'échantillon a été chargé et élué sur un volume de 120 mL de tampon C au débit de 1 mL/min, avec fractionnement de l'élution tous les 1 mL. La protéine est attendue pour un volume supérieur au volume d'exclusion des colonnes (environ 36 mL), et, après analyse par SDS-PAGE, les fractions contenant la protéine les plus pures ont été mélangées.

Les protéines destinées à la cristallogénèse ont de plus été concentrées par centrifugation à 6000 rpm et 4 °C en système Amicon Bioseparation Centricon (Millipore) sur membrane Ultracell YM (cellulose) jusqu'à atteinte de leur concentration critique de solubilité par mesure spectrophotométrique à 280 nm sur un appareil Nanodrop Spectrophotometer ND-1000 (LabTech), le facteur d'extinction molaire théorique ε_{280nm} a été calculé à partir de la séquence des protéines recombinantes (Gill and von Hippel, 1989).

Des modifications ont été appliquées à ce protocole standard selon les besoins de chaque protéine. Ainsi, la protéine RB2160 (module catalytique GH57) a été purifiée dans un tampon C modifié constitué de 50mM tris-HCl pH 7,5 ; 200 mM NaCl ; 2% glycérol pour des raisons d'instabilité au cours de la phase de concentration entraînant une forte précipitation de la protéine. La version sélénométhionylée de la protéine RB5312 (module catalytique PL1) a, quant à elle, été purifiée dans un tampon C également modifié en 50mM tris-HCl pH 7,5 ; 200 mM LiCl afin d'augmenter sa solubilité (ce tampon a été déterminé après des expériences préliminaires en cristallogénèse).

V.C - Caractérisation biophysique

V.C.1 - Chromatographie analytique

En plus de permettre la séparation des éventuelles impuretés de l'échantillon purifié, les chromatographies en colonne d'exclusion de taille avec les colonnes S75 et S200 ont permis d'estimer la masse moléculaire des cibles purifiées grâce au calibrage des volumes

d'élution des colonnes (sous réserve de modulation par le volume hydrodynamique des protéines, pouvant présenter une masse apparente plus grande ou plus petite).

La calibration consiste à injecter sur chaque colonne une série de protéines pures et de masses moléculaires définies, à une température et un débit donné. Les comparaisons des volumes d'élution ne seront alors pertinentes que si les purifications sont opérées dans les mêmes conditions. Les Tableau III-28 et Tableau III-29 présentent les différentes séries de protéines utilisées pour la calibration et fournies sous forme de kits de calibration (GE Healthcare). Un volume de 1 mL a été injecté par protéine, à température ambiante et au débit de 1 mL/min. Le bleu dextran est totalement exclu de la colonne de par sa masse moléculaire (2 000 kDa) et sert à calibrer le volume mort des colonnes (environ 40 mL pour les deux).

	Protéine	Masse mol. (Da)	Concentration (mg/mL)
Superdex 75 HiLoad 16/60 prep grade	*Résolution de la colonne : 10 kDa – 70 kDa*		
	Aprotinine	6 500	3
	Cytochrome C	12 400	2
	Anhydrase carbonique	29 000	2
	Albumine de sérum de bœuf	66 000	5
	Dextran bleu	2 000 000	1

Tableau III-28 : Protéines de calibration de la colonne d'exclusion de taille Superdex75.

	Protéine	Masse mol. (Da)	Concentration (mg/mL)
Superdex 200 HiLoad 16/60 prep grade	*Résolution de la colonne : 10 kDa – 180 kDa*		
	Cytochrome C	12 400	8
	Anhydrase carbonique	29 000	3
	Albumine de sérum de bœuf	66 000	10
	Alcool déshydrogénase	150 000	5
	β-amylase	200 000	4
	Dextran bleu	2 000 000	1

Tableau III-29 : Protéines de calibration de la colonne d'exclusion de taille Superdex200.

V.C.2 - Mesure de diffusion de la lumière

La présence éventuelle d'isoformes multimériques des échantillons ou d'agrégats solubles dans les solutions a été estimée en ayant recours à une mesure de la dispersion de la lumière des solutions (Dynamic Light Scattering, DLS). Cette technique consiste à faire passer un faisceau lumineux cohérent de longueur d'onde 633 nm à travers la solution et d'étudier la déviation de la lumière à son contact à certains angles. Les grosses particules de la solution présenteront une déviation de la lumière différente des particules plus petites et il est ainsi possible d'estimer la répartition des populations moléculaires de la solution en fonction de leur taille. Par exemple, il est possible de voir si la protéine a tendance à former des agrégats solubles (invisibles sur SDS-PAGE du fait de la dénaturation des protéines) et si d'une manière générale, la solution présente une ou plusieurs populations de particules. Cette technique est très complémentaire du SDS-PAGE et des colonnes d'exclusion de taille. Elle présente l'avantage d'être très rapide (une mesure dure 15 min, quand un passage sur colonne prend une journée et un dépôt sur gel deux heures), non destructrice, quantitative et un bon indicateur de la structure quaternaire de la protéine (monomérique, multimérique, mono- ou poly-disperse, …). Elle peut cependant être facilement illisible si de grosses particules telles des traces de poussières sont présentes dans la solution, masquant de par leur taille toute autre particule plus petite.

Les échantillons ont été préalablement filtrés à 0.20 µm sur Millex-LG (Millipore) pour éviter toute trace de macroparticules. Les mesures ont été réalisées sur un volume de 1 mL de solution protéique après sortie de purification en colonne d'exclusion de taille S75 ou S200 (et souvent après concentration) dans des cuves en quartz, sur un appareil ZetaSizer Nano-S (Malverne Instruments) calibré sur une solution de toluène pur fournie par le fabricant.

V.C.3 - Spectrométrie de masse

La masse moléculaire des protéines recombinantes sélénométhionylées a été estimée par comparaison avec les protéines natives (i.e. non sélénométhionylées) en spectrométrie de masse. Toutes les protéines ont été dessalées par dialyse contre un tampon 25 mM tris-HCl pH7,0 ; 25 mM NaCl et la concentration des échantillons a été amenée autour de 1 mg/mL. Deux types de mesures de spectrométrie de masse ont été réalisées.

Le remplacement des sélénométhionines de la première production des protéines marquées a été analysé par une spectrométrie de type Electrospray-TOF sur une source Q-TOF 2 (Waters), étalonnée par la myoglobine (1 mg/mL). Les échantillons ont été chargés directement sur la source. La protéine native a été préparée en additionnant 1 µL de solution protéique ; 25 µL CH3CN 100% ; 12 µL H2O 100% ; et 12 µL HCOOH 1%. La protéine sélénométhionylée a été préparée en additionnant 1 µl de solution protéique ; 5 µL CH3CN 100% ; 2 µL H2O 100% ; et 2 µL HCOOH 1%. Un volume de 10 µL pour chaque échantillon a été chargé et une tension de 45 V et 50 V respectivement pour les protéines native et séléniée a été appliquée sur le cône d'introduction.

Le remplacement des sélénométhionines de la seconde production des protéines marquées a été analysé par une spectrométrie de type MALDI-TOF sur une source Voyager DE STR (Applied Biosystem), étalonnée par l'albumine de sérum de bœuf (BSA) (1 mg/mL, pour un ratio BSA:matrice de 1:29). Les mesures ont été réalisées en mode linéaire avec une tension d'accélération de 25 kV. La matrice utilisée pour la protéine native a été l'acide sinapinique (10 mg/ml dans CH3CN 50%). La matrice utilisée pour la protéine séléniée a été l'acide 2,5-dihydroxybenzoïque (10 mg/ml dans CH3CN 50%). La protéine séléniée a été de plus dessalée en CH3CN 50% par élution sur une colonne ZipTip C18. Pour les deux protéines, 1 µL de solution protéique a été mélangé à 1 µL de matrice et 1 µL du mélange a été déposé sur la plaque d'expérimentation.

V.D - Tests enzymatiques

V.D.1 - Dosage des sucres réducteurs

Ce dosage est basé sur la réduction du réactif ferricyanure de potassium (hexacyanoferrate III de potassium) en ferrocyanure de potassium (hexacyanoferrate II de potassium) par les molécules réductrices du mélange réactionnel, en l'occurrence les extrémités réductrices des oses de la solution. Le dosage consiste à déterminer la concentration en ions ferricyanures par lecture spectrophotométrique de la solution à 420 nm. Plus leur concentration est basse (et donc l'absorbance à 420 nm faible), plus le mélange contient de sucres réducteurs (Kidby et al., 1973). Cette technique a été utilisée pour les tentatives de caractérisation des modules catalytiques GH16 de RB3123, PL1 de RB5312 et GH57 de RB2160.

La solution de réactif est préparée en ajoutant 300 mg de ferricyanure de potassium à 28 g de carbonate de sodium monohydrate et 1 mL de soude concentrée (5 M). La solution est complétée à 1L avec de l'eau bidistillée. Une gamme étalon au glucose est réalisée en préparant 4 solutions contenant 100 mM NaCl, 5 mM $CaCl_2$, 50 mM MES pH 7 et respectivement 0 µg/mL, 100 µg/mL, 200 µg/mL et 300 µg/mL de glucose.

La solution de substrat est préparée en ajoutant 100 mM NaCl, 5 mM $CaCl_2$, 50 mM MES pH 7 et le polysaccharide à doser pour une concentration finale de ce dernier entre 0,1% et 1% (voir cas suivants). Pour les tests d'activités du domaine catalytique GH16 de la protéine RB3123, onze substrats polysaccharidiques ont été utilisés : agarose 0,1% (EuroGenTec), κ-carraghénane 0,1% (Degussa), ι-carraghénane 0,1% (Degussa), λ-carraghénane 0.1% (Degussa), laminarine 0.1% (Goëmar), laminarine 0.1% (Sigma), galactane 0,1% (Aldrich), xylane 1% (Sigma), cellulose soluble 1% (Merck), chitine 1% (Sigma), chitosane 1% (Sigma). Enfin, pour le domaine catalytique GH57 de la protéine RB2160, plusieurs polysaccharides ont été testés : amylose recristallisé (don gracieux de W. Helbert) à 0,3%, amidon soluble (Sigma) à 0,5%, glycogène (Sigma) à 0,5%, pullulane (Sigma) à 0,1%, α-galactane (Sigma) à 0,1%.

Un volume de 900 µL de la solution de substrat a été thermalisée à 37 °C. A ensuite été ajouté un volume de 100 µL de la solution enzymatique et l'ensemble a été mélangé par vortex. Un premier échantillon de 100 µL a directement été prélevé à l'instant du dépôt de la solution enzymatique et déposé sur glace. Il sera utilisé pour l'estimation des paramètres initiaux de la réaction. Un échantillon de 100 µL a ensuite été périodiquement prélevé du mélange réactionnel (typiquement aux temps 15 min, 30 min, 60 min, 120 min et le lendemain de l'expérience) et conservé dans la glace. Une fois l'ensemble des échantillons prélevés, il est possible de les congeler pour une mesure ultérieure, l'ajout du réactif impliquant une mesure dans les plus brefs délais (le ferricyanure étant assez instable).

Pour effectuer la mesure, chaque échantillon de la réaction a été ajouté à 1 mL de la solution de réactif. La gamme étalon a été préparée au même moment, en ajoutant pour chaque solution standard, 100 µL de la solution de glucose et 900 µL de réactif. L'ensemble a été bouilli au bain marie pendant 7 min et refroidi à température ambiante. Une mesure spectrophotométrique de l'absorbance à 420 nm a alors été réalisée pour chaque échantillon.

V.D.2 - Caractérisation de l'activité pectinolytique

V.D.2.1 Dosage enzymatique

Les polysaccharide lyases agissent par un mécanisme de β-élimination et libèrent des disaccharides dont l'ose non réducteur est insaturé en C_4-C_5. La libération des oligosaccharides insaturés a été suivie par mesure spectrophotométrique de l'absorbance de la double liaison en C_4-C_5 à 235 nm (Collmer *et al.*, 1988). Pour le domaine catalytique PL1 de la protéine RB5312, les tests ont été réalisés sur plusieurs types de pectines disponibles au laboratoire : homogalacturonane de citron (Sigma), méthylester-homogalacturonane de pomme (Sigma), pectine SKW (Degussa).

Un prélévement de 10 µL de la solution d'enzyme à environ 1 mg/mL a été ajouté à 1 mL d'une solution tamponnée contenant 50 mM tris-HCl pH 9.5, 100 mM NaCl, 0.5 mM $CaCl_2$ et 0.2 % pectine SKW (Degussa). Le mélange réactionnel a été laissé en enceinte thermostatée à 40°C et un prélèvement de 100µL a été effectué toutes les 30 min pour les études de cinétique enzymatique. La mesure spectrophotométrique de l'absorbance à 235 nm de l'échantillon a ensuite été réalisée sur un spectrophotomètre UV-2450PC (Shimadzu).

L'optimum de température de RB5312 a été déterminé en effectuant des dosages d'activité similaires en variant la température de réaction de 20°C à 65°C, par pas de 5°C. De même, le pH optimum a été déterminé à 40°C en utilisant différents tampons de réaction : 50 mM tris-HCl pour la gamme de pH 7,0 à 9,0 et 50 mM glycine pour la gamme de pH de 9,0 à 10,5. Enfin, la dépendance de l'activité de RB5312 à des cations divalents a été testée en réalisant le dosage en présence ou en absence de 5 mM d'EDTA. En cas de perte d'activité, différents cations à 5mM ont été ajouté au milieu réactionnel pour tenter de restaurer l'activité (Li^+, Na^+, K^+, Rb^+, Cs^+ ; Mg^{2+}, Ca^{2+}, Mn^{2+}, Co^{2+} Ni^{2+}, Sn^{2+}, et Al^{3+})

V.D.2.2 Analyse des produits de dégradation de RB5312

Pour l'étude des produits de dégradation de RB5312, un acide polygalacturonique de citron déméthylé par une pectine méthylestérase a été utilisé (MEGAZYME). Ce substrat a été préalablement ultrafiltré sur une membrane amicon de seuil de coupure 8000 Da pour éliminer d'éventuels oligosaccharides déjà présents. Vingt millilitres d'une solution de cette pectine (0,2 % w/v) ont été incubé pendant une nuit à 40°C en présence de 5 ng/mL RB5312 (100 mM tris-HCl pH 9,0, 1 mM $CaCl_{2)}$. Les produits de dégradation de RB5312 étant supposés contenir des résidus acides galacturoniques chargés négativement, ils ont donc

été analysés par électrophorèse sur gel de polyacrylamide (Carbohydrate-PolyAcrylamide Gel Electrophoresis - C-PAGE, Zablackis and Perez, 1990). Cinq microlitres de produits de dégradation sont déposés sur un gel comprenant une phase de concentration (« stacking ») de 6% de polyacrylamide et une phase de séparation (« running ») de 27% de polyacrylamide dans un tampon 50 mM Tris–HCl pH 8,7, 2 mM EDTA. Après migration, le gel est coloré au bleu alcian suivie d'une coloration au nitrate d'argent (Min and Cowman, 1986).

Les oligosaccharides libérés par RB5312 ont été partiellement purifiée par une chromatographie d'exclusion de taille sur une colonne Superdex S30 26/60 (GE HealthCare). L'élution a eu lieu dans un tampon 100 mM tris-HCl pH 9,0, 1 mM $CaCl_2$ à 1,5 mL/min et a été suivie par une mesure de l'absorbance à 235 nM. Les fractions contenant des oligosaccharides ont été regroupées, puis lyophilisées et resolubilisées trois fois dans du D^2O afin de les analyser par RMN 1H (service RMN de l'université de Bretagne Occidentale, Brest).

V.D.3 - Test de l'activité 4-α-glucanotransférase

Le domaine catalytique GH57 de RB2160 ayant potentiellement une activité 4-α-glucanotransférase, plusieurs tests ont été entrepris pour tenter de la caractériser, sinon quantitativement du moins qualitativement.

V.D.3.1 Zymographie en conditions non dénaturantes

La zymographie des sucres en conditions non dénaturantes consiste à réaliser sur un échantillon protéique une électrophorèse en gel de polyacrylamide non dénaturant (PAGE natif i.e. sans SDS ni agent réducteur, voir Chap II - II.C.1), en présence du substrat supposé de l'enzyme (en l'occurrence le glycogène et l'amidon). En fin d'électrophorèse, le gel est incubé à température ambiante en présence d' α-glucanes de faibles degrés de polymérisation (de maltose, i.e. DP2, à maltoheptaose, i.e. DP7). Le gel est ensuite transféré dans une solution de Lugol (5 % I_2, 10 % KI) dont l'iode se fixe au glycogène et à l'amidon pour donner une coloration brun foncé. L'intensité de coloration du gel est uniforme sauf aux endroits où une enzyme a modifié le polysaccharide : une activité de dégradation est visible par une bande plus claire (i.e. substrat moins dense), une activité de synthèse est visible par une bande plus foncée (i.e. substrat plus dense), tandis que les activités de modification (branchement, cyclisation, …) sont visibles par une bande d'une teinte différente pouvant être assez variable (Lantz et al., 1994).

Les gels ont été polymérisés en utilisant un mélange 70 % acrylamide – 30 % bisacrylamide à la concentration standard de 7 % et en ajoutant de 0.05 % à 0.3 % de glycogène ou d'amidon (Sigma), en plaque de 0.5 mm d'épaisseur avec 10 puits de dépôts, pour un volume de 20 µL par échantillon. Les échantillons ont été colorés en 0.2 % bleu de bromophénol avant dépôt, conservés sur glace et déposés sur gel extemporanément.

Les électrophorèses natives ont été réalisées en chambre réfrigérée à 4 °C et au voltage constant de 100 V par gel. La séparation électrophorétique a été considérée réalisée au bout d'environ 1 h 30 min (lorsque le front de migration, matérialisé par le bleu de bromophénol, atteint la limite inférieure du gel). Les gels ont ensuite été découpés suivant leur longueur et les bandes des gels ont été analysées. Les bandes destinées au repérage de la protéine dans les gels ont été colorées au bleu de Coomassie (20% éthanol, 10% acide acétique, 0,25% bleu de Coomassie avec décoloration dans un mélange 20% éthanol, 10% acide acétique). Les bandes destinées à l'analyse de l'activité ont été incubées 1 h minimum sous agitation douce dans une solution 50 mM tris-HCl ; 200 mM NaCl ; 5 mM $CaCl_2$; 1 mM $MnCl_2$; 1 mM $MgCl_2$ et 1 mg/mL de chaque oligosaccharide suivant : maltose, maltotriose, maltopentose, maltohexaose et maltoheptaose. Plusieurs conditions d'incubation ont alors été testées : température ambiante ou 40 °C et pH 5, pH 7 ou pH 9. Finalement, les bandes de gel ont été colorées par trempage dans une solution de lugol (5 % I_2, 10 % KI) pour estimer l'activité enzymatique. La décoloration a dans ce cas été réalisée en présence d'eau. Même si les deux colorations sont réalisables sur un même gel, elles ont été dans la plupart des cas réalisées sur des dépôts gémellaires.

V.D.3.2 Zymographie en conditions dénaturantes

Cette méthode est très proche de la zymographie en conditions non dénaturantes. Il s'agit essentiellement de réaliser cette dernière en présence de 0,1% SDS (ajouté dans les tampons de migration, de polymérisation du gel et de dépôt des échantillons). Avant analyse, le gel est incubé dans une solution 400 mM tris-HCl pH 7,5 et 0,1% nonyl phénoxylpolyéthoxyléthanol (NP-40) renouvelée toutes les 15 min 4 fois, puis en solution 400 mM tris-HCl pH 7,5 renouvelée toutes les 15 min 4 fois. Les incubations et colorations (bleu de Coomassie et Lugol) précédentes sont ensuite appliquées.

V.D.4 - Test de l'activité sialidase

L'activité exo-sialidase a été estimée par mesure de la dégradation du substrat générique de l'enzyme, l'acide N-acétyl-neuraminique, lié par une liaison glycosidique au fluorophore 4-méthylumbelliferol. La mesure consiste à exciter la solution à 365 nm et à lire l'intensité de la fluorescence à 450 nm de la 4-méthylumbelliferone libérée. En effet, sous sa forme liée au substrat le 4-méthylumbelliferol ne présente pas de fluorescence, celle-ci n'apparaissant qu'une fois clivé sous forme de 4-méthylumbelliferone (Potier *et al.*, 1979).

A 1 mL d'une solution à 50 mM tris-HCl pH 7,0 on été ajoutés 5 mg d'acide 2'-(4-méthylumbelliferyl)-N-acétyl-neuraminique ainsi que 10 µL de la solution protéique à environ 1 mg/mL, et le mélange réactionnel a été placé en étuve à 37 °C. Un échantillon de 200 µL a été prélevé du milieu réactionnel aux temps 0 min, 10 min, 20 min, 50 min et 400 min et photoexcité à la longueur d'onde de 365 nm et la fluorescence induite à 450 nm a été mesurée sur un spectrofluoromètre LS-50B (Perkin Elmer) dans des cuves en quartz à trois fenêtres (disposées en « T ») fournies par le fabricant.

V.E - Cristallogenèse

Les protéines ont été concentrées en fin de purification jusqu'à leur limite de solubilité dans un tampon le moins salé possible, les sels interférant fortement avec le processus de cristallisation.

Les techniques de la cristallogenèse ont pour but la production de cristaux de protéines. Celles utilisées au cours de cette thèse sont basées sur la diffusion de vapeur entre une goutte contenant la protéine, mélangée à une solution d'additifs, et un réservoir de volume infiniment plus grand que la goutte (d'un facteur au moins cent) ne contenant que la solution d'additifs (« condition de cristallisation »). Un gradient de diffusion de vapeur va progressivement s'établir entre les deux solutions et va tendre à un équilibre entre leurs concentrations respectives en solutés, pondérées par leur volume. Le volume de la goutte contenant la protéine étant infiniment plus petit que celui de la solution réservoir, l'équilibre va être dominé par cette dernière et va aboutir au dessèchement de la goutte qui va perdre de l'eau sous forme de vapeur, permettant finalement la concentration de la protéine (**Erreur ! Source du renvoi introuvable.**).

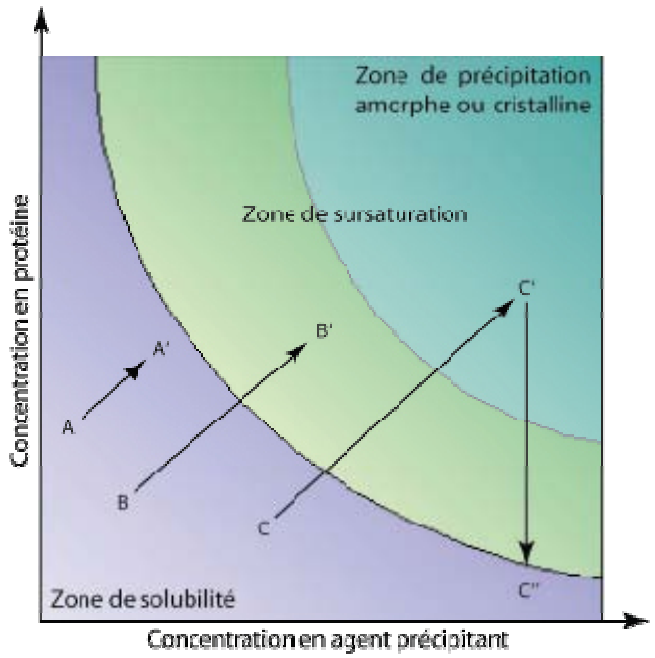


Figure III-108 : Cristallogenèse
Diagramme de phase présentant les conditions nécessaires à la cristallogenèse. Les flèches tracent l'évaporation de la goutte entre son état initial et final.

Cas A : la protéine reste soluble ; *Cas B :* la protéine va entrer en zone de sursaturation, et peut même commencer à nucléer mais la croissance cristalline sera faible ; *Cas C :* la protéine passe par sa concentration de nucléation et continue à se concentrer. Si un cristal croit la concentration en protéine diminue alors virtuellement et permet de ne pas dépasser un seuil qui provoquerait sa précipitation.

C'est précisément au cours de ce processus de concentration que la protéine peut entrer en phase de nucléation. Cela est réalisé spontanément lorsque la protéine atteint sa concentration limite de précipitation ET que le tampon dans lequel elle se trouve favorise les interactions intermoléculaires.

L'ensemble du crible des conditions de cristallisation a été réalisé en utilisant les techniques dites « de la goutte suspendue » pour les affinements de conditions et « de la goutte assise » pour le crible des kits de cristallogenèse. Dans le cas de la goutte suspendue, une goutte (de 0,5 µL à 4 µL) d'un mélange protéine – solution de cristallisation est déposée sur une lamelle de verre siliconée et scellée avec de la graisse au silicone au dessus d'un puits contenant entre 200 µL et 1 mL de la solution de cristallisation. (Figure III-109).

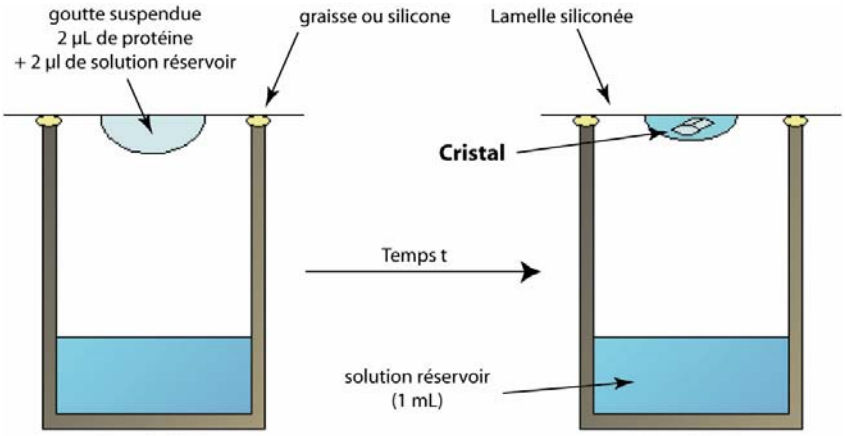


Figure III-109 : Technique de la goutte suspendue.

Les affinements de conditions avérées ou prometteuses d'après les tests avec les kits commerciaux ont été réalisés en utilisant cette technique dans des boites de 24 conditions.

La technique de la goutte assise est conceptuellement très proche. La goutte contenant la protéine mélangée à la solution de cristallogenèse est simplement déposée cette fois-ci sur un support dans une enceinte fermée contenant une solution réservoir. Le criblage des conditions de cristallisation en kits a été réalisé selon cette technique en plaque 96 conditions en utilisant un robot de remplissage HoneyBee 963 (Cartesian). Plusieurs kits commerciaux ont été utilisés pour la détermination des conditions initiales (couvrant une large gamme de tampons et d'additifs) : les suites PEG Screen I et II, The Cations, PACT, Wizard I & II (Emerald BioStructures, Inc.) et JCSG+ Suite (Qiagen). Deux types de plaques ont été utilisés : un avec une goutte par condition et un avec 3 gouttes (de ratio différents protéine/condition du crible) par condition.

V.F - Cristallographie

V.F.1 - Des rayons X au service de la biologie

La source de rayons X classique de notre laboratoire consiste en un appareil Nonius Proteum (Brücker). La source de rayonnement est constituée d'un tube à anode tournante en cuivre dont la puissance nominale atteint 44 kW. Les rayons X produits par ce dispositif sont monochromatiques à la longueur d'onde de la raie de transition électronique K_α du cuivre (environ 1,54 Å) et le cristal est fixé à une tête goniométrique à travers un montage quatre cercles. La détection des rayons X est réalisée par un détecteur CCD Brücker.

Les rayons X produits par la source synchrotron de l'ESRF ont des propriétés radicalement différentes. Ils sont émis par des électrons accélérés à des vitesses relativistes, en rotation dans un accélérateur de particules de 800 m de circonférence. Le fait que ces électrons soient hautement énergétiques leur confère entre autres la propriété de produire un faisceau lumineux de plusieurs ordres de grandeurs plus intense que les sources classiques et couvrant un large spectre de rayons X allant du rayonnement infrarouge (~10 nm) aux rayons X durs (~0,01 nm). Dans le cadre de la résolution de structure de macromolécules biologiques (protéines, acides nucléiques,…), des rayons X de longueur d'onde de l'ordre de l'angström (~0,1 nm) sont utilisés.

Le partenariat avec l'ESRF m'a permis de réaliser les expériences de diffraction sur les lignes de lumière ID14-EH2, ID14-EH4, BM30A et ID23-EH1 (**Figure III-110**), qui présentent des caractéristiques différentes. Ainsi, ID14--EH2 est une ligne délivrant un faisceau monochromatique fixé à 0,93 Å, tandis que ID14-EH4, ID23-EH1 et BM30A sont des lignes adaptées à des expériences à longueur d'onde modulable. Ces différents sites ont de plus développés une politique d'interface utilisateur commune ces dernières années qui permet une gestion très automatisée des différentes collectes.

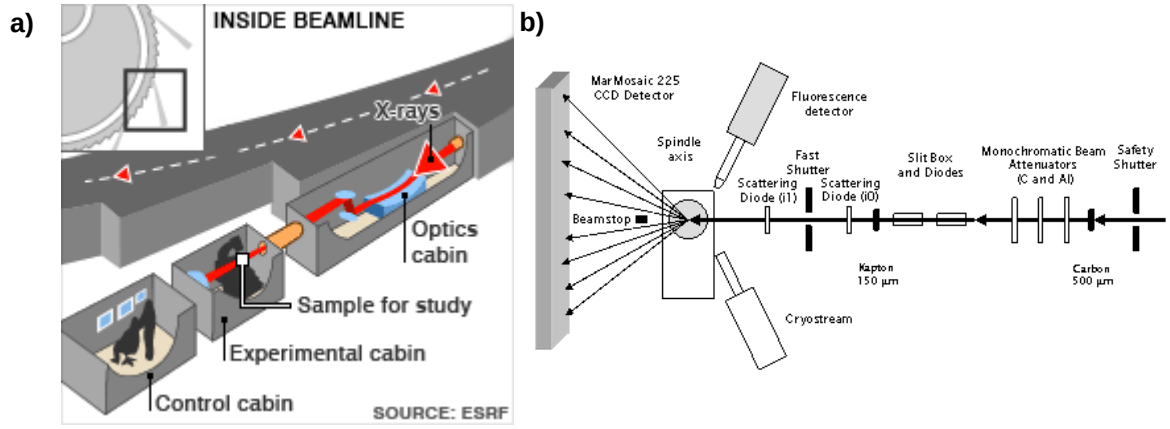


Figure III-110 : Présentation de l'ESRF
a) Présentation de l'organisation générale d'une ligne de lumière à l'ESRF ; b) Présentation de l'organisation de la cabine expérimentale de la ligne ID23-1. Source : ESRF

V.F.2 - Préparation des expériences de cristallographie

Les cristaux ont été systématiquement montés sur microboucle magnétique de 22 mm (Hampton Research). En effet, l'ensemble des lignes de lumière de l'ESRF destinées à l'étude de macromolécules biologiques a été optimisé pour ce type de montage de cristaux. Les robots de montage d'échantillons sont en particulier conçus pour tirer parti de ces capsules magnétiques.

Les cristaux ont été congelés à 100 K par trempage des boucles dans un bain d'azote liquide et stockage en Dewar réfrigéré à 100 K également par azote liquide. Afin d'éviter la formation de glace dans les boucles lors de leur refroidissement, les cristaux ont été trempés lors du montage en boucle dans une solution contenant la solution mère de cristallisation additionnée d'un cryoprotectant en faible pourcentage (typiquement du glycérol ou du MPD). Des tests de diffraction ont été systématiquement réalisés sur ces solutions pour estimer la formation de glace lors de leur congélation.

V.F.3 - Collecte des données

Avant chaque expérience à l'ESRF, des mesures du groupe d'espace, des paramètres de maille ainsi que du pouvoir diffractant des cristaux ont été réalisés sur le diffractomètre du laboratoire. Cela nous a permis d'initier les collectes en ayant une préconnaissance des différents cristaux et de ce que l'on peut en attendre, augmentant d'autant notre efficacité sur place.

La stratégie de collecte des jeux de données à l'ESRF a consisté à enregistrer deux images séparées de 90° du cristal testé. Ces images ont été traitées par MOSFLM et la stratégie de collecte angulaire proposée a été très souvent suivie. D'une manière générale, nous avons évité de collecter en aveugle afin de maximiser les chances d'avoir un jeu de données avec une bonne redondance et des statistiques invariantes dans le temps (nous avons en particulier été très conservateurs dans le choix du temps d'exposition des cristaux aux rayons X).

V.F.4 - Traitement des données collectées

Le traitement des données se décompose en plusieurs étapes. Tout d'abord l'**indexation** des taches de diffraction. Elle a été réalisée en utilisant en routine le logiciel MOSFLM. Elle consiste en la détermination des paramètres de maille du cristal (module des vecteurs directeurs de la maille **a**, **b** et **c**, ainsi que les angles qui les relient α, β et γ). Les taches de diffraction sont ensuite nommées par leur position dans le réseau réciproque avec leurs indices de Miller *hkl*. Vient ensuite l'**intégration** des intensités. La mosaicité du cristal, ainsi que la largeur spectrale et la divergence du faisceau incident, ont pour conséquence d'étaler plus ou moins les taches sur le cliché de diffraction. Les programmes de traitement permettent d'établir les profils moyens de ces taches et de calculer l'intensité de chaque réflexion par intégration numérique suivant les profils calculés. Enfin, la **mise à l'échelle** permet d'effectuer des comparaisons entre les taches équivalentes issues de plusieurs images, voire de plusieurs jeux de données. En effet, lorsqu'une réflexion et/ou ses symétriques sont mesurées plus d'une fois sur deux images différentes, leur intensité doit être théoriquement identique. En pratique, cela est difficilement réalisable en raison de plusieurs paramètres tels la dégradation du cristal sous l'exposition du flux de rayons X ou encore la forme du cristal et la décroissance continue du rayonnement synchrotron. Une mise à l'échelle globale des intensités d'un jeu de données est réalisée en déterminant pour chaque image un facteur de mise à l'échelle k linéaire et un facteur d'agitation thermique B en fonction de la résolution.

Par la suite, la **réduction** des données a été réalisée par les logiciels SCALA et TRUNCATE. Ces logiciels permettent le tri des taches en fonction de leurs indices de Miller, la moyennation des intensités des réflexions équivalentes en respectant leurs facteurs d'échelle respectifs et le calcul du module des facteurs de structure. Un fichier contenant l'ensemble des réflexions unique est finalement généré.

V.F.5 - Estimation de la qualité d'un jeu de données

La qualité d'un jeu de données est déterminée par différents critères statistiques calculés au cours du traitement. Les paramètres les plus significatifs comprennent la résolution, le rapport signal sur bruit I/σ(I), la complétude, la redondance, le facteur R_{sym}, et quelques autres.

La **résolution** détermine la séparation des détails dans l'image de la densité électronique. La limite haute de la résolution des données mesurées peut être imposée par la configuration de l'expérience (taille du détecteur, distance cristal-détecteur, longueur d'onde…) ou par la qualité de diffraction du cristal. Lors de cette thèse, la limite de résolution des jeux de données a été fixée à la tranche de plus haute résolution vérifiant au moins l'un de ces deux critères : $I/\sigma(I) \leq 2$ et/ou $R_{sym} \geq 30\%$. Le **rapport signal sur bruit I/σ(I)** permet d'avoir une estimation de l'intensité moyenne des réflexions mesurées en fonction de la résolution (où I est l'intensité des réflexions et σ(I), l'écart type des intensités des réflexions – il correspond au bruit de fond). La **complétude** donne le pourcentage de réflexions mesurées par rapport au nombre total de taches mesurables pour une résolution donnée. Sa valeur doit être la plus grande possible (supérieure à 90%). La **redondance** établit le nombre moyen de mesures d'une réflexion et de ses symétriques. Plus cette valeur est grande, meilleure est l'estimation de l'intensité moyenne d'une réflexion unique (en diminuant l'influence d'erreur de mesure). Le **facteur R_{sym}** correspond à :

$$R_{sym} = \frac{\sum_{hkl} \left| I_{hkl} - \langle I \rangle \right|}{\sum_{hkl} I_{hkl}}$$

Il permet de comparer l'intensité I_{hkl} de chaque réflexion équivalente par la symétrie du cristal à la valeur moyenne <I> de ces réflexions. Plus ce facteur est bas, meilleure est la cohérence du jeu de données. D'autres paramètres peuvent également être pris en compte pour évaluer la qualité des données de diffraction comme par exemple la mosaicité du cristal, exprimée en degrés, qui correspond à l'écart à l'idéalité des paramètres des mailles

dans le cristal par rapport à la maille moyenne du cristal ou encore le graphique de Wilson qui, grâce à l'étude de la distribution des intensités en fonction de la résolution, permet d'estimer un facteur de mise à l'échelle et un facteur de température, mais aussi d'évaluer la présence de macles dans les cristaux (Yeates, 1997).

V.F.6 - Phasage des données

Les détecteurs de rayons X mesurent l'intensité des rayons diffractés par le cristal et l'information de phase qui existe dans ces rayons est perdue, ce qui rend impossible le calcul de la densité électronique. Il est possible de surmonter ce problème dans le cas des petites molécules grâce aux méthodes dites « directes » de calcul de la phase à partir des données collectées. Cependant, dans le cas des molécules biologiques, il faut passer par des méthodes nécessitant dans la plupart des cas l'enregistrement de données de diffraction provenant d'un nouveau cristal. Ces méthodes indirectes peuvent être classées en deux groupes : celles dites expérimentales nécessitant l'incorporation d'un atome lourd dans le cristal et dont les propriétés de diffuseur anomal pourront éventuellement être utilisées, et celles nécessitant l'utilisation d'un modèle de structure connue présentant une certaine similitude.

V.F.6.1 Phasage par remplacement moléculaire

La technique de phasage par remplacement moléculaire est employée lorsque l'on dispose d'un modèle d'une protéine supposée structuralement proche de la protéine d'intérêt.

La comparaison des séquences en acides aminés est une bonne indication de l'homologie structurale. Deux molécules ayant plus de 30% d'identité de séquence ont beaucoup de chances d'avoir un repliement similaire. Afin d'améliorer artificiellement la corrélation entre le modèle et la protéine d'intérêt, il peut aussi être préférable de modifier le modèle en supprimant les chaînes latérales ou certaines parties (boucles, domaines peu structurés, …) lorsqu'on le suppose trop éloigné de la structure de la protéine d'intérêt. Le remplacement moléculaire va chercher les six paramètres de la transformation géométrique (trois pour la rotation et trois pour la translation) permettant de passer du modèle à la structure de la protéine d'intérêt. Les procédures du remplacement moléculaire ont été initiées en 1962 (Rossmann and Blow, 1962) et, depuis lors, largement améliorées par le développement des techniques de calcul numérique (Navaza, 2001; Read, 2001).

Les essais de remplacement moléculaires ont été réalisés en utilisant le logiciel AmoRe (Automatic Molecular Replacement - Navaza, 2001).

V.F.6.2 Phasage expérimental

Le phasage expérimental consiste en une série de techniques impliquant toutes le marquage du cristal par un métal lourd et une voire plusieurs collectes de jeux de données sur les cristaux marqués. Deux paramètres sont alors cruciaux : l'incorporation au sein des mailles doit avoir eu lieu en des positions périodiques du cristal, et les paramètres de maille des cristaux doivent être rigoureusement identiques si plusieurs collectes sont réalisées. Au cours de ma thèse, une seule de ces techniques a été utilisée : la diffraction anomale multiple (Multiple-wavelength Anomalous Diffraction, MAD).

La technique du MAD consiste donc à incorporer un métal au sein des cristaux. Ce métal doit présenter des propriétés de diffuseur anomal (phénomène d'absorption) à une longueur d'onde accessible aux différentes sources de rayons X utilisées. Le sélénium est un diffuseur anomal particulièrement prisé en cristallographie des protéines en raison de ses propriétés chimiques proches de celles du soufre et de son seuil d'absorption assez intense à une longueur d'onde facilement accessible dans les accélérateurs de particules (raie K_α à 0,98 Å). Il est ainsi possible de produire des protéines dans lesquelles les acides aminés soufrés telles que les méthionines sont remplacés par des équivalents sélénométhionines. Lors de la cristallisation de ces protéines, les séléniums se retrouveront dans des positions cristallographiques, les assurant de contribuer à la diffraction du cristal et les mesures seront toutes réalisées sur le même cristal, permettant de conserver une isotropie stricte entre les jeux de données (sauf cas de dégradation majeure du cristal sous l'effet des rayonnements ionisants). Trois collectes sont ensuite effectuées à trois longueurs d'onde différentes:

- au maximum de la composante imaginaire du coefficient d'absorption du métal ;
- au point d'inflexion de la composante réelle du coefficient d'absorption du métal ;
- loin du seuil d'absorption.

L'intégration de ces différentes données permet de séparer le signal du diffuseur anomal des données non diffusives du jeu de données. Cela est réalisé en découplant les taches de diffraction liées par la relation de Friedel. Un facteur de qualité de ce signal anomal (R_{ano}) est alors déduit. Le pouvoir phasant du diffuseur anomal dépend de l'écart entre le facteur R_{sym} du jeu de données (qui correspond au bruit moyen des données) et le facteur R_{ano}. Plus le facteur R_{ano} est grand par rapport au facteur R_{sym}, plus le pouvoir de phasage du jeu de données est important.

Chapitre IV

-

Vers la reconstruction du métabolisme des sucres de *Rhodopirellula baltica*

I - Révision des annotations des enzymes du métabolisme des sucres

Le but de mon travail de thèse était d'avoir une meilleure compréhension du métabolisme des sucres dans la planctomycète marine *Rhodopirellula baltica*, en particulier celui des polysaccharides complexes. Le point de départ de l'étude a été le recensement des CAZymes (Carbohydrate Active enZymes) de *R. baltica*, qui est effectué régulièrement sur tous les nouveaux génomes par les équipes maintenant la banque de données CAZy (Coutinho and Henrissat, 1999) (http://www.cazy.org/).

Cent vingt-huit enzymes de ce métabolisme ont été identifiées dans le génome de *R. baltica* : 39 GH, 5 PL, 23 CE ainsi que 61 GT. Dans le chapitre I, une analyse bioinformatique approfondie de ces protéines a été réalisée et a permis de déterminer la présence de 165 modules au sein de diverses architectures modulaires. Ces résultats ont pu être utilisés afin de cloner et exprimer 96 modules choisis sur l'ensemble. Cette première analyse a montré que ces architectures modulaires avait rarement été prises en compte dans l'annotation initiale du génome de *R. baltica* (Glöckner *et al.*, 2003). Dans le chapitre II, en complément des expériences de caractérisations biochimiques, j'ai également étudié par une approche phylogénétique quatre familles de polysaccharidases qui m'ont permis d'affiner voire de corriger la prédiction initiale de la spécificité de substrats de ces enzymes. Il est apparu qu'il pourrait être très intéressant de réaliser ce travail sur l'ensemble des enzymes du métabolisme des sucres de *R. baltica*, dans le but d'améliorer l'annotation de ces enzymes.

La procédure d'annotation fonctionnelle s'est divisée en trois étapes :

- Analyse de l'architecture modulaire (Cf. chapitre II) ;
- Réalisation de comparaisons de séquences par BLAST uniquement avec des protéines de fonction expérimentalement démontrée ;
- Enfin, analyse phylogénétique incluant des enzymes des familles concernées de *R. baltica* et des homologues de fonction expérimentalement démontrée (Cf. chapitre III).

Le Tableau III-30 présente l'ensemble des résultats de ce travail de réannotation.

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB4024	Q7UT87	3.2.1.8	probable endo-1,4-beta-xylanase Z [precursor]	-CE1	3.1.1.73	Feruloyl esterase, family CE1
RB6145	Q7UQR5	-	conserved hypothetical protein	-CE1	3.1.1.-	Carbohydrate esterase, family CE1
RB8823	Q7UMH7	3.2.1.8	probable endo-1,4-beta-xylanase 1 precursor	-CE1	3.1.1.73	Feruloyl esterase, family CE1
RB9546	Q7ULE9	-	conserved hypothetical protein-putative a hydrolase	-CE1	3.1.1.-	Carbohydrate esterase, family CE1
RB9732	Q7UL52	-	probable gluconolactonase precursor-hypothetical secreted or membrane associated protein	-CE1-UNK1	3.1.1.-	Carbohydrate esterase, family CE1
RB11024	Q7UJV7	-	enterochelin esterase	-UNK1-CE1	3.1.1.-	Carbohydrate esterase, family CE1
RB11670	Q7TTY3	-	probable ferric enterobactin esterase-related protein Fes	-CE1	3.1.1.-	Carbohydrate esterase, family CE1
RB11978	Q7UJC7	-	conserved hypothetical protein-putative hydrolase	-CE1a-CE1fes-GlucL	3.1.1.-/3.1.1.-/3.1.1.17	Carbohydrate esterase, family CE1 / Carbohydrate esterase, family CE1 / Gluconolactonase
RB7144	Q7UP58	-	probable predicted xylanase/chitin deacetylase	-CE4	3.-.-.-	Carbohydrate esterase, family CE4
RB12316	Q7UIU9	-	similar to chitooligosaccharide deacetylase	-CE4	3.-.-.-	Carbohydrate esterase, family CE4
RB2091	Q7UWE4	-	similar to chitooligosaccharide deacetylase	-CE4	3.-.-.-	Carbohydrate esterase, family CE4
RB5007	Q7UGU5	-	probable acetyl xylan esterase AxeA	-CE6	3.1.1.-	Carbohydrate esterase, family CE6
RB763	Q7UYA8	3.1.6.13	iduronate-2-sulfatase	-CE6-Sulf	3.1.1.-/3.1.6.13	Carbohydrate esterase, family CE6 / Iduronate-2-sulfatase
RB977	Q7UXZ7	3.5.1.25	probable N-acetylglucosamine-6-phosphate deacetylase	-CE9a-CE9b	3.5.1.-	Carbohydrate esterase, family CE9

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB981	Q7UXZ6	-	conserved hypothetical protein-putative imidazolonepropionase or related amidohydrolase	-CE9	3.5.1.-	Carbohydrate esterase, family CE9
RB1356	Q7UXF7	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase NAGA	-CE9	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase, family CE9
RB12559	Q7UIF8	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase	-CE9	3.5.1.25	N-acetylglucosamine-6-phosphate deacetylase, family CE9
RB4934	Q7UGZ2	3.5.1.-	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	-CE11	3.5.1.-	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase
RB3639	Q7UTX2	-	conserved hypothetical protein	-CE12	3.1.1.-	Carbohydrate esterase, family CE12
RB5540	Q7URP5	-	conserved hypothetical protein	-CE14	3.5.1.-	Carbohydrate esterase, family CE14
RB10036	Q7UKN9	-	conserved hypothetical protein	-CE14	3.5.1.-	Carbohydrate esterase, family CE14
RB4014	Q7UT93	-	probable acetyl xylan esterase	-CE15	3.1.1.-	Carbohydrate esterase, family CE15
RB3405	Q7UUA7	3.2.1.23	putative hydrolase	-UNK1-GH2	3.2.1.-	Glycoside hydrolase, family GH2
RB5256	Q7UGF4	3.2.1.4	cellulase	-UNK1-GH5	3.2.1.4	Endo-1,4-beta-D-glucanase, family GH5
RB9136	Q7UM13	3.2.1.8	probable endo-1,4-beta-xylanase homolog T27I17	-UNK1-GH10	3.2.1.8	Endo-1,4-beta-xylanase, family GH10
RB9911	Q7UKV6	3.2.1.8	endo-1,4-beta-xylanase	-GH10-UNK2	3.2.1.8	Endo-1,4-beta-xylanase, family GH10
RB10416	Q7UF11	3.2.1.8	similar to xylanase	-DOCK1-UNK1-PA14a-PA14b-CalxB-UNK2-GH10-UNK3	3.2.1.8	Endo-1,4-beta-xylanase, family GH10
RB548	Q7UYJ7	2.4.1.18	1,4-alpha-glucan branching enzyme	-GH13-CBM48	2.4.1.18	1,4-Alpha-glucan branching enzyme, family GH13

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB2638	Q7UVH1	2.4.1.18	1,4-alpha-glucan branching enzyme (glycogen branching enzyme)	-GH13	2.4.1.18	1,4-Alpha-glucan branching enzyme, family GH13
RB2986	Q7UUY3	3.2.1.10	oligo-1,6-glucosidase	-GH13	3.2.1.20	oligo-1,4-glucosidase, family GH13
RB4894	Q7UH19	3.2.1.-	glycogen operon protein glgX-2	-GH13-CBM48	3.2.1.-	Glycogen debranching enzyme, family GH13
RB5196	Q7UGI7	2.4.1.4	alpha-amylase, amylosucrase	-GH13	2.4.1.4	Amylosucrase, family GH13
RB5200	Q7UGI4	3.2.1.1	alpha-amylase	-HAD-GH13	3.1.3.-/3.2.1.141	Phosphatase, sucrose-6P phosphatase family / Malto-oligosyltrehalose trehalohydrolase, family GH13
RB9292	Q7ULT9	3.2.1.-	glycogen operon protein glgX-2	-GH13-CBM48	3.2.1.-	Glycogen debranching enzyme, family GH13
RB12343	Q7UIS9	2.4.1.7	sucrose phosphorylase	-GH13	2.4.1.7	Sucrose phosphorylase, family GH13
RB2702	Q7UVD8	3.2.1.83	Kappa-carrageenase [precursor]	-GH16	3.2.1.83	Kappa-carrageenase, family GH16
RB3123	Q7UUR7	3.2.1.-	probable glycosyl hydrolases-putative kappa-carrageenase precursor	-GH16-CBM16	3.2.1.-	Glycoside hydrolase, family GH16
RB4561	Q7USD8	3.2.1.52	beta-hexosaminidase	-GH20-UNK1	3.2.1.52	beta-N-acetyl-hexosaminidase, family GH20
RB5816	Q7UR90	3.2.1.51	alpha-L-fucosidase	-UNK1-GH29	3.2.1.51	Alpha-L-fucosidase, family GH29
RB12360	Q7UIS2	3.2.1.65	levanase precursor	-GH32	3.2.1.80	Fructan beta-fructosidase, family GH32
RB3006	Q7UUX1	3.2.1.18	probable sialidase	-GH33-UNK1-UNK2	3.2.1.18	Sialidase, family GH33
RB3353	Q7UUD9	3.2.1.18	Sialidase [Precursor]	-GH33	3.2.1.18	Sialidase, family GH33

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB11055	Q7UJU4	-	conserved hypothetical protein	-NucH-GH33	3.2.-.-/3.2.1.18	Ribonucleoside hydrolase / Sialidase, family GH33
RB5143	Q7UGL7	-	conserved hypothetical protein	-GH33	3.2.1.18	Sialidase, family GH33
RB8895	Q7UMD4	3.2.1.18	neuraminidase precursor	-GH33	3.2.1.18	Sialidase, family GH33
RB1257	Q7UXL3	-	conserved hypothetical protein	-GH33	3.2.1.18	Sialidase, family GH33
RB2377	Q7UVX7	-	arylsulfatase	-SufCa-LamG-SufCb-GH43	3.6.1.-/3.6.1.-/3.2.1.-	Formglycine-dependent sulfatase / Formglycine-dependent sulfatase / Glycoside hydrolase, family GH43
RB8073	Q7UG75	3.2.1.55	alpha-L-arabinofuranosidase II	-GH43	3.2.1.55	Alpha-L-arabinofuranosidase, family GH43
RB2160	Q7UWA6	3.2.1.1	alpha-amylase	-GH57	-	Glycoside hydrolase, family GH57
RB4161	Q7UT23	2.4.1.25	4-alpha-glucanotransferase	-GH77	2.4.1.25	4-alpha-Glucanotransferase, family GH77
RB700	Q7UYD5	3.2.1.40	alfa-L-rhamnosidase	-UNK1-GH78-UNK2	3.2.1.40	alpha-L-Rhamnosidase, family GH78
RB736	Q7UYB9	-	conserved hypothetical protein-putative rhamnosidase	-DUF1080-GH78	3.2.1.40	alpha-L-Rhamnosidase, family GH78
RB3421	Q7TU03	3.2.1.81	probable Beta-agarase [Precursor]	-UNK1-GH86	3.2.1.-	Glycoside hydrolase, family GH86
RB4699	Q7US58	-	conserved hypothetical protein	-GH95	3.2.1.-	Glycoside hydrolase, family GH95
RB10507	Q7UEW8	3.2.1.20	alpha-glucosidase	-GH97	3.2.1.20	alpha-Glucosidase, family GH97
RB10996	Q7UJX3	-	putative UDP-glucose:sterol glucosyltransferase	-GT1	2.4.1.173	Sterol 3-beta-glucosyltransferase, family GT1

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB289	Q7UYZ8	2.4.1.83	dolichol-phosphate mannosyltransferase	-GT2	2.4.1.-	Dolichol-phosphate mannosyltransferase homolog, family GT2
RB1637	Q7UX10	-	putative glycosyl transferase	-GT2	2.4.1.-	Dolichol-phosphate mannosyltransferase homolog, family GT2
RB2206	Q7UW80	2.4.1.-	conserved hypothetical protein	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB2469	Q7UVS5	2.4.1.-	glycosyltransferase-like protein	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB5846	Q7UR75	-	probable glycosyltransferase involved in cell wall biogenesis	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB6305	Q7UQI5	2.4.1.83	dolichol-phosphate mannosyltransferase	-GT2	2.7.8.-	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase, family GT2
RB6377	Q7UQE3	-	periplasmic glucans biosynthesis protein MdoH	-GT2-UNK1	2.4.1.-	Glucans biosynthesis glucosyltransferase H, family GT2
RB7643	Q7UND3	2.4.1.83	conserved hypothetical protein-putative dolichyl-phosphate hexose synthase	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB8905	Q7UMC7	2.4.1.83	putative polyprenol phosphate mannosyl transferase 1 (Ppm1)	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB9237	Q7ULW4	2.4.1.-	probable two-domain glycosyltransferase	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB9623	Q7ULA8	-	sugar transferase-putative a glycosyl transferase	-GT2	2.4.1.-	Glycosyltransferase, family GT2
RB9637	Q7ULA0	-	glycosyl transferase	-GT2-UNK1	2.4.1.-	Glycosyltransferase, family GT2
RB11941	Q7UJE8	-	conserved hypothetical protein-putative glycosyl transferase	-GT2	2.4.1.-	4,4'-diaponeurosporenoate glycosyltransferase homolog, family GT2
RB12831	Q7UI07	2.4.1.83	probable dolichyl-phosphate mannose synthase	-GT2	2.4.1.-	Glycosyltransferase, family GT2

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB13211	Q7UHG9	2.4.1.83	probable dolichol-phosphate mannosyltransferase-putative membrane bound sugar transferase involved in LPS biosynthesis	-GT2-GT83-UNK1	2.7.8.-/2.-.-.-	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase, family GT2 / Undecaprenyl phosphate-alpha-4-amino-4-deoxy-L-arabinose arabinosyl transferase, familyGT83
RB1027	Q7UXY0	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2001	Q7UWJ4	2.4.1.-	probable glycosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2471	Q7UVS3	2.4.1.52	conserved hypothetical protein-putative poly(glycerol-phosphate) alpha-glucosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2473	Q7UVS1	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2474	Q7UVS0	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2482	Q7UVR5	-	similar to hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2484	Q7UVR4	2.4.1.-	putative glycosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2485	Q7UVR3	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Lipopolysaccharide core biosynthesis, family GT4
RB2493	Q7UVQ7	2.4.1.-	LPS biosynthesis RfbU related protein	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2499	Q7UVQ1	-	glycosyl transferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2985	Q7UUY4	-	conserved hypothetical protein-putative glycosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB2990	Q7UUY1	-	conserved hypothetical protein-putative glycosyltransferase	-GT4-UNK	2.4.1.-	Glycosyltransferase, family GT4
RB3591	Q7UU03	2.-.-.-	putative transferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB4333	Q7USS4	2.4.1.-	probable galactosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB5197	Q7UGI6	2.4.1.14	sucrose-phosphate synthase 1	-GT4-Phos	2.4.1.14/3.1.3.24	Sucrose-phosphate synthase, family GT4 / Sucrose-6-phosphate phosphatase
RB5991	Q7UQZ2	-	capsular polysaccharide biosynthesis protein	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB6241	Q7UQL8	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB6520	Q7UQ48	-	putative glycosyl transferase	-GT4	2.4.1.-	Lipopolysaccharide core biosynthesis, family GT4
RB7146	Q7UP57	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Lipopolysaccharide core biosynthesis, family GT4
RB8313	Q7UFV7	-	putative glycosyl transferase (WbnE)	-GT4	2.4.1.-	Lipopolysaccharide core biosynthesis, family GT4
RB9617	Q7ULB1	5.-.-.-	mannosyltransferase B	-GT4	2.4.1.-	LPS mannosyltransferase, family GT4
RB10434	Q7UF02	-	mannosyltransferase	-GT4	2.4.1.-	LPS mannosyltransferase, family GT4
RB10436	Q7UF01	2.4.1.-	mannosyl transferase	-GT4	2.4.1.-	LPS mannosyltransferase, family GT4
RB10614	Q7UKI5	2.-.-.-	lipopolysaccharide core biosynthesis glycosyl transferase lpsD	-GT4	2.4.1.-	LPS mannosyltransferase, family GT4
RB12344	Q7UIS8	2.4.1.-	predicted glycosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB12590	Q7UIE3	2.4.1.-	probable hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB12604	Q7UID6	2.4.1.-	similar to hexosyltransferase	-GT4	2.4.1.-	Glycosyltransferase, family GT4

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB9627	Q7ULA4	-	probable mannosyltransferase A (mtfA)	-GT4	2.4.1.-	Glycosyltransferase, family GT4
RB6654	Q7UPY2	2.4.1.21	glycogen synthase	-GT5	2.4.1.21	Glycogen synthase, family GT5
RB9245	Q7ULV9	-	probable heptosyl III transferase	-GT9	2.4.-.-	Lipopolysaccharide heptosyltransferase, family GT9
RB9238	Q7ULW3	2.4.1.-	putative fucosyl transferase	-GT11	2.4.1.69	O-antigen alpha-1,2-fucosyltransferase, family GT11
RB11356	Q7UEG0	2.4.1.-	probable N-acetylgalactosaminyltransferase	-GT12	2.4.1.-	Beta-1,4 N-acetylgalactosaminyltransferase, family GT12
RB11527	Q7UE72	2.4.1.-	probable N-acetylgalactosaminyltransferase	-GT12	2.4.1.-	Beta-1,4 N-acetylgalactosaminyltransferase, family GT12
RB11529	Q7UE71	2.4.1.-	probable beta-1,4 N-acetylgalactosaminyltransferase	-GT12	2.4.1.-	Beta-1,4 N-acetylgalactosaminyltransferase, family GT12
RB2725	Q7UVC4	-	lipid-A-disaccharide synthetase	-GT19	2.4.1.182	Lipid-A-disaccharide synthase, family GT19
RB5016	Q7UGT9	-	similar to beta-1,4-galactosyltransferase waaX	-GT25	2.4.1.-	LPS beta-1,4-galactosyltransferase, family GT25
RB11533	Q7UE70	-	conserved hypothetical protein	-GT25-SulfT	2.4.1.-	Glycosyltransferase, family GT25 / Sulfotransferase family protein
RB1367	Q7UXE9	2.4.1.-	UDP-N-acetyl-D-mannosaminuronic acid transferase	-GT26	2.4.1.-	EPS glycosyltransferase, family GT26
RB12601	Q7UID8	2.4.1.-	UDP-N-acetyl-D-mannosaminuronic acid transferase	-GT26	2.4.1.-	EPS glycosyltransferase, family GT26
RB11688	Q7UDY7	2.-.-.-	3-deoxy-D-manno-octulosonic-acid transferase	-GT30	2.4.-.-	3-deoxy-D-manno-octulosonic-acid transferase, family GT30
RB9614	Q7ULB2	-	similar to surface protein Sur1	-GT32	2.4.1.-	Ceramide glucosyltransferase, family GT32

Protéine	Code UniProt	Numéro EC Initial	Annotation Initiale	Modules	Numéro EC Révisé	Annotation révisée
RB8383	Q7UFR8	2.4.1.1	phosphorylase 2	-GT35	2.4.1.1	Glycogen phosphorylase, GT35
RB1005	Q7UXY8	-	conserved hypothetical protein-putative glycosyltransferase	-GT81	2.4.1.-	Glucosyl-3-phosphoglycerate synthase, family GT81
RB13211	Q7UHG9	2.4.1.83	probable dolichol-phosphate mannosyltransferase-putative membrane bound sugar transferase involved in LPS biosynthesis	-GT2-GT83-UNK1	2.7.8.-/2.-.-.-	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase, family GT2 / Undecaprenyl phosphate-alpha-4-amino-4-deoxy-L-arabinose arabinosyl transferase, familyGT83
RB9640	Q7UL99	-	probable ceramide glucosyltransferase	-GTNC	2.4.1.-	Glycosyltransferase
RB9648	Q7UL94	-	probable ceramide glucosyltransferase	-GTNC	2.4.1.-	Glycosyltransferase
RB5312	Q7US17	4.2.2.2	pectate lyase	-PL1	4.2.2.2	Pectate Lyase, family PL1
RB5316	Q7US15	4.2.2.2	pectate lyase	-PL1	4.2.2.2	Pectate Lyase, family PL1
RB3601	Q7UTZ6	4.2.2.3	probable alginate lyase [Precursor]	-PL7	4.2.2.11	Polyguluronate lyase, family PL7
RB3417	Q7UUA0	4.2.2.2	pectate lyase	-PL10	4.2.2.2	Pectate Lyase, family PL10
RB9973	Q7UKS6	-	hypothetical protein-transmembrane prediction	-PL10	4.2.2.2	Pectate Lyase, family PL10

Tableau III-30 : Révision de l'annotation des polysaccharidases de *Rhodopirellula baltica*.

Les modules recensés par la banque CAZY sont représentés par les codes suivant : GH, glycoside hydrolase ; PL, polysaccharide lyase, CE, carbohydrate esterase ; GT, glycosyltransferase, CBM : carbohydrate binding module. Les autres types de modules sont représentés par les codes suivants : UNK, module de fonction inconnue ; GlucL, Gluconolactonase ; Sulf, sulfatase, DOCK, dockerin ; PA14, famille Pfam, interactions protéines-protéines, CalxB, famille Pfam, fixation d'ions calcium ; HAD, famille Pfam incluant sucrose-6P phosphatases ; NucH, famille Pfam, Ribonucleoside hydrolase ; LamG, famille Pfam, Laminin G ; DUF1080, famille Pfam, fonction inconnue ; Phos, Sucrose-6-phosphate phosphatase ; SulfT, famille des sulfotransferases.

Il est tout d'abord intéressant de noter que plusieurs polysaccharidases de *R. baltica* présentent quelques modules, non directement impliqués avec les sucres, qui ne sont pas répertoriés par la banque CAZY : des modules catalytiques de type sulfatases (RB763, RB2377), phosphatases (RB5200, RB5197), glucanolactonase (RB11978), etc ; des modules d'interaction protéine-protéine (Dockerin, RB10416), ainsi que de nombreux modules de fonction inconnue mais conservés (DUF1080, PA14), ou limités aux planctomycètes, voire même uniquement présents chez *R. baltica*. Une fonction précise n'a été attribuée que dans le cas où le niveau d'identité était suffisant avec la plus proche protéine caractérisée (> 30%), ou quand la prédiction était robustement cohérente avec l'analyse phylogénétique. Dans les autres cas, dans un souci de limiter les surprédictions au maximum, l'inférence est restée à un niveau plus général comme « glycoside hydrolase » ou « sulfatase ». Pour le degré de précision des numéros EC, la même règle a été appliquée. Enfin, je propose d'indiquer systématiquement dans l'annotation finale la famille à laquelle appartiennent les différents modules composant la protéine.

Cette étude a permis de corriger de nombreuses erreurs d'annotation plus ou moins graves. En particulier, certaines erreurs se sont révélées dues à une absence d'analyse modulaire de la protéine. Par exemple, les protéines RB4024 et RB8823 étaient annotées initialement des « endo-1,4-beta-xylanases ». Ces protéines présentent bien des fortes similitudes de séquence avec l'endo-1,4-beta-xylanase XynZ de *Clostridium thermocellum* (P10478), mais cette enzyme est bifonctionnelle et la zone de similitude correspond en réalité à un module N-terminal feruloyl esterase appartenant à la famille CE1. L'absence d'analyse modulaire peut aussi conduire à des annotations qui, sans être fausses, n'en sont pas moins incomplètes. C'est par exemple le cas de la protéine RB2377, qui avait été annotée « arylsulfatase ». Cette protéine est en réalité composée de quatre modules : deux modules sulfatases séparés par un module de type laminin G, qui est connu pour interagir avec des polysaccharides sulfatés comme l'héparine (Harrison *et al.*, 2007), et d'un module glycoside hydrolase de la famille GH43. Si cette protéine n'est clairement pas une sulfatase en soi, elle semble néanmoins impliquée dans la dégradation d'un polysaccharide sulfaté.

Fort heureusement, la majorité des erreurs n'ont pas présenté ce niveau de gravité. Les analyses phylogénétiques et les comparaisons avec les protéines de fonctions expérimentalement démontrées ont surtout permis d'affiner des prédictions et d'éviter les surinterprétations sur certaines autres.

II - Vers la reconstruction du métabolisme des sucres de Rhodopirellula baltica

Dans ce paragraphe, une approche intégrative des analyses bioinformatiques précédentes avec les données de la littérature et les résultats expérimentaux obtenus est proposée, afin de mieux décrire certaines parties du métabolisme des sucres de *R. baltica*. Quatre métabolismes ont été choisis, pour leur cohérence avec le sujet de thèse : la dégradation des polysaccharides de la paroi des plantes supérieures, la dégradation des polysaccharides de la paroi des algues marines, le métabolisme du glycogène et de l'amidon, et le métabolisme des acides sialiques.

II.A - La dégradation des polysaccharides de la paroi de plantes supérieures

Rhodopirellula baltica possède plusieurs enzymes qui pourraient être impliquées dans la dégradation de polysaccharides pariétaux des plantes supérieures, en particulier ceux constituant la phase matricielle. Le génome de *R. baltica* contient en effet trois xylanases de la famille GH10 (RB9136, RB9911, RB10416) et une glycoside hydrolase de la famille GH43, qui est une xylosidase ou une arabinofuranosidase (RB8073). Plusieurs carbohydrate esterases des familles CE6 et CE15 avaient été initialement annotées « acetyl xylan esterases », mais elles présentent une grande divergence et l'attribution de cette spécificité reste pour l'instant trop spéculative. J'ai déjà mentionné le cas des protéines RB4024 et RB8823, deux feruloyl esterases qui sont responsables de l'hydrolyse des liaisons esters liant des composés phénoliques à des arabinoxylanes et des pectines (Christov and Prior, 1993).

Plusieurs enzymes participeraient de plus à la dégradation des pectines : deux pectate lyases de la famille PL1 (RB5312 et RB5316), et deux pectate lyases de la famille PL10 (RB3417 et RB9973). J'ai démontré expérimentalement que la protéine recombinante RB5312 est bien une pectate lyase dont l'activité dépend des ions calcium (Chapitre III IV.A). On peut également supposer que les deux rhamnosidases de la famille GH78 (RB700 et RB736) pourraient jouer un rôle dans l'élimination des régions branchées des pectines (rhamnogalacturonanes).

La capacité d'une bactérie marine comme *R. baltica* à utiliser des polysaccharides de plantes supérieures terrestres peut paraître étrange au premier abord, mais il ne faut pas oublier la très forte influence fluviale dans la mer Baltique. Il est donc envisageable que des

débris de végétaux terrestres participent à la formation des neiges marines. Ce potentiel de dégradation est néanmoins à nuancer. En effet, selon les travaux de Schlesner et collaborateurs sur la taxonomie de *R. baltica*, cette bactérie ne serait pas capable d'utiliser la pectine comme source de carbone (Schlesner *et al.*, 2004). Ce résultat est très surprenant, puisque l'activité pectinolytique de RB5312 a été démontrée. Il faudra donc reproduire les expériences de Schlesner pour confirmer ou non cette incapacité de *R. baltica* à utiliser la pectine pour croître en solution. En cas de confirmation, plusieurs hypothèses sont à envisager. Tout d'abord, il est possible que *R. baltica* ne possède qu'une voie de dégradation incomplète de la pectine. Elle pourrait dans ce cas utiliser la pectine en consortium avec d'autres populations microbiennes qui complèteraient la voie catabolique. Cette hypothèse est assez plausible, étant données les nombreuses espèces bactériennes qui ont été isolées sur les neiges marines (Glöckner *et al.*, 2003). Il pourrait être également envisagé que ces gènes de dégradation de la pectine ne soient qu'un moyen pour *R. baltica* d'accéder à des substrats qu'elle pourrait alors prendre en charge. Il est en effet démontré que le pourrissement des végétaux suite à une infection par une bactérie pathogène est du à l'action de pectine lyases (Mohnen, 2008). *R. baltica* pourrait avoir une stratégie de dégradation basée sur la déliquescence de la paroi, pour en faciliter l'extraction des autres polysaccharides. Enfin, il est également à considéré que la forme et la nature du substrat puisse jouer un rôle majeur dans la capacité de la bactérie à le dégrader. En effet, la paroi des algues, celle des végétaux supérieurs, et même les neiges marines, ne sont pas des structures dissoutes, mais de véritables gels denses. De plus, chaque espèce, voire chaque portion du cycle de vie de chaque espèce, présente des polysaccharides pariétaux différents (nature, quantité, taille, forme, ramifications, …). Cette bactérie peut ne pas tous les dégrader. Il pourrait donc être envisagé une étude plus systématique sur un grand nombre des polysaccharides accessibles en milieu marin.

II.B - La dégradation des polysaccharides de la paroi de macroalgues marines

L'une des particularités les plus surprenantes de *R. baltica* est que son génome contient plus d'une centaine de sulfatases à formylglycine (Glöckner *et al.*, 2003). Comme l'eau de mer est riche en sulfate (autour de 20 mM), il est peu probable que ces nombreuses sulfatases servent à récupérer du sulfate. Il est en revanche plus probable qu'elles servent plutôt à éliminer les groupements sulfates des polysaccharides chargés, présents chez toutes les algues marines (vertes, rouges ou brunes) (Kloareg and Quatrano, 1988).

Il apparaît que plusieurs enzymes dans le génome de *R. baltica* pourraient dégrader des polysaccharides sulfatés d'algues. Ainsi, RB2702 (famille GH16), annotée « kappa-carraghénase » semble bien présenter cette activité, d'après l'analyse bioinformatique approfondie des deux enzymes de la famille GH16 de *R. baltica* (voir Chapitre II). Il s'est cependant avéré que RB3123, également annotée « kappa-carraghénase », n'en était pas une. La modélisation de la structure de cette enzyme suggère cependant que cette nouvelle enzyme de la famille GH16 hydrolyse bien un polysaccharide chargé négativement.

RB3421 appartient à la famille GH86, qui ne contient pour l'instant que des beta-agarases. Cette protéine est cependant très divergente dans sa famille. Deux CAZymes présentent, enfin, des modules additionnels sulfatases : la carbohydrate esterase RB763 (famille CE6) et la glycoside hydrolase RB2377 (famille GH43). Ces protéines modulaires pourraient donc être de bons candidats pour agir sur des polysaccharides sulfatés. Comme la famille GH43 comprend des xylosidases, il est possible que RB2377 soit une enzyme spécifique des glucuronoxylanes sulfatés qui ont été isolés chez les algues vertes (Kloareg and Quatrano, 1988). Enfin, RB3601 présente une forte similitude de séquence (41% d'identité) avec l'alginate lyase de *Corynebacterium* sp. (Osawa *et al.*, 2005). Cette enzyme appartient à la famille PL7 et est spécifique des acides L-guluroniques (les C5-epimères des acides D-mannuroniques).

Pour vérifier certaines de ces hypothèses, nous avons fait croitre *R. baltica* sur un milieu M40 (Schlesner *et al.*, 2004) gélifié respectivement par 1% d'agar, 1% de kappa-carraghénane et 2% de iota-carraghénane (Figure IV-111).

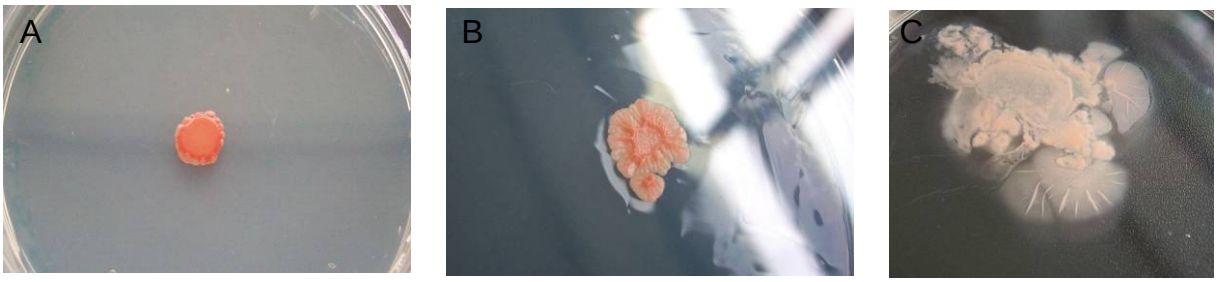


Figure IV-111 : Colonies de *R. baltica*.
Colonies de *R. baltica* près une semaine de culture sur gel d'agar (A), de κ-carraghénane (B) et de ι-carraghénane (C).

Après une semaine de culture on constate qu'une dépression dans le gel est visible pour le kappa-carraghénane, tandis que le gel de iota-carraghénane est complètement liquéfié. Par contre, le gel d'agar ne semble pas avoir été dégradé. Par conséquent, cette simple expérience confirme que *R. baltica* a bien une activité kappa-carraghénolytique, qui est certainement due à la protéine RB2702.

La forte dégradation du iota-carraghénane est par contre totalement inattendue. Le génome de *R. baltica* ne contient en effet pas de glycoside hydrolases de la famille GH82, la seule famille de iota-carraghénases connue à ce jour (Barbeyron *et al.*, 2000). Deux hypothèses sont ici envisageables : (i) une sulfatase pourrait être spécifique du 2-sulfate des résidus 3,6-anhydro-D-galactose 2-sulfate du iota-carraghénane. Cette enzyme convertirait alors le iota-carraghénane en kappa-carraghénane, qui pourrait être hydrolysé par RB2702. (ii) *R. Baltica* possèderait une iota-carraghénase appartenant à une nouvelle famille de glycoside hydrolases.

L'incapacité de *R. baltica* à dégrader l'agar est également une surprise. Ceci semble indiquer que la protéine RB3421 n'est pas une β-agarase, mais présenterait une spécificité de substrat qui n'a pas encore été décrite dans la famille GH86.

Enfin d'après Schlesner et collaborateurs (Schlesner *et al.*, 2004), *R. baltica* ne serait pas non plus capable de dégrader l'alginate. Ce résultat, qui demande à être confirmé, peut cependant se comprendre. En effet, la polyguluronate lyase RB3601 présenterait une spécificité assez étroite. L'alginate est initialement produit comme un polymère d'acides D-mannuronique qui sont ensuite partiellement convertis en acides L-guluronique par des mannuronate C5-epimerases (Nyvall *et al.*, 2003). En absence de polymannuronate lyases, la seule présence d'une polyguluronate lyase pourrait être insuffisante pour cliver l'alginate.

Plusieurs scénarios possibles sont à envisager pour expliquer cette putative guluronate lyase dans l'arsenal enzymatique de *R. baltica*. A vrai dire, il s'agit plus ou moins des mêmes scénarios que ceux envisagés que ceux de la pectine lyase. *R. baltica* pourrait ainsi utiliser ce polysaccharide comme source de carbone en synergie avec d'autres bactéries marines. Elle pourrait également utiliser cette enzyme pour casser la paroi des algues brunes, puis dégrader les autres composants polysaccharidiques rendus accéssibles.

II.C - Métabolisme du glycogène et de l'amidon

Au cours de la réannotation, j'ai remarqué que le génome de *R. baltica* comprenait de nombreuses protéines liées au métabolisme du glycogène et de l'amidon. L'étude taxonomique de *R. baltica* avait de plus révélé que cette bactérie marine pouvait utiliser de l'amidon exogène comme source de carbone (Schlesner *et al.*, 2004). En me basant sur la littérature (Ball and Morell, 2003), j'ai entrepris de vérifier si *R. baltica* possédait bien une voie complète de biosynthèse et de dégradation du glycogène endogène, ainsi qu'une voie de dégradation de l'amidon exogène. Ceci a nécessité de chercher par BLAST certaines enzymes qui ne sont pas répertoriées dans la banque de données CAZy, en particulier les glucose-1-phosphate adenylyltransférases. Cette analyse m'a permis de vérifier si *R. baltica*

possède toutes les enzymes nécessaires au stockage du carbone sous forme de glycogène et à son recyclage. Ces différentes voies sont détaillées ci-dessous :

II.C.1 - La biosynthèse du glycogène

La première étape est la production d'ADP-glusose (réaction 1) qui est catalysée par la glucose-1-phosphate adenylyltransférase (GlgC, EC 2.7.7.27). Le génome contient deux gènes codant cette enzyme : *rb1358* et *rb10465*.

(1) *ATP* + *α-D-glucose 1-phosphate* ↔ *diphosphate* + *ADP-glucose*

La deuxième étape est catalysée par la glycogène synthase (GlgA, EC 2.4.1.21) qui polymérise la chaine de α-(1,4)-glucose (2). Cette enzyme appartient à la famille des glycosyltransférases GT5 et est codée chez *R. baltica* par le gène *rb6654*.

(2) *α-(1,4)-(glycosyl)$_n$ linéaire* + *ADP-glucose* → *α-(1,4)-(glycosyl)$_{n+1}$ linéaire* + *ADP*

La dernière étape de synthèse du glycogène est la formation des branchements α-(1,6). Cette réaction (3) est catalysée par une glycoside hydrolase de la famille GH13 (1,4-alpha-glucan-branching enzyme, GlgB, EC 2.4.1.18) qui clive une liaison glycosidique α-(1,4) de la chaine linéaire et transfère l'α-(1,4)-oligoglucane libéré à une position α-(1,6). Le génome de R. baltica comprend deux paralogues de GlgB : *rb548* et *rb2638*.

(3) *α-(1,4)-(glycosyl) linéaire* → *α-(1,4)-(glycosyl) linéaire avec un branchement α-(1,6)*

Il est ainsi intéressant de noter que *Rhodopirellula baltica* diffère de la bactérie modèle *Escherichia coli*, par la présence de deux paralogues de GlgC (RB1358 et RB10465) et de GlgB (RB548 et RB2638), alors qu'*E. coli* ne possède qu'une seule glucose-1-phosphate adenylyltransférase et une seule 1,4-alpha-glucan-branching enzyme(Ball and Morell, 2003).

II.C.2 - Le catabolisme du glycogène

Le recyclage du glycogène nécessite au moins quatre enzymes, la glycogène phosphorylase (GlgP, EC 2.4.1.1, famille GT35), une enzyme de débranchement (glycogen debranching enzyme, glgX, EC 3.2.1.-, famille GH13), une alpha-(1,4)-glucanotransférase (MalQ, EC 2.4.1.25, famille GH77) et une maltodextrin-phosphorylase (MalP, EC 2.4.1.1, famille GT35).

La glycogène phosphorylase est l'enzyme majeure du catabolisme du glycogène. Elle transfère un phosphate inorganique sur l'extrémité réductrice d'une chaîne externe (***ec***) d'α-(1,4)-glucane, libérant ainsi un glucose-1-phosphate (réaction 1). Toutes les glycogène phosphorylases connues arrêtent leur action au niveau d'un branchement α-(1,6) (Ball and Morell, 2003). *R. baltica* possède une glycogène phosphorylase : RB8383.

(1) ***ec***$(glycosyl)_n$ + *Pi* $\longrightarrow$ $(glycosyl)_{n-1}$ + *glucose-1-P*

L'étape suivante est l'élimination des chaînes extérieures α-(1,6). Elle est catalysée par l'enzyme de débranchement GlgX (2). Le génome de *R. baltica* contient deux paralogues de GlgX, RB4894 et RB9292. Cette dernière protéine a la particularité d'avoir chez *R. baltica* un module additionnel de fixation de sucres de la famille CBM48.

(2) ***ec***$[\alpha\text{-}(1,4)\text{-}glucose]_4$*-α-(1,6)-glycogène* $\longrightarrow$ *glycogène* + *maltotétraose*

Les maltooligosaccharides qui sont ainsi libérés sont ensuite pris en charge par une alpha-(1,4)-glucanotransférase qui va allonger ces courtes chaînes pour qu'elles redeviennent substrats de la maltodextrine phosphorylase. A chaque réaction catalysée par l' alpha-(1,4)-glucanotransférase un résidu glucose est également libéré (réaction 3). *R. baltica* possède une alpha-(1,4)-glucanotransférase appartenant à la famille GH77 : RB4161.

(3) *maltotriose* + *maltotriose* $\longrightarrow$ *maltopentaose* + *glucose*

La dernière enzyme majeure est la maltodextrine phosphorylase qui dégrade les maltooligosaccharides jusqu'au maltotetraose (réaction 4).

(4) *maltooligosaccharide* $(glucosyl)_n$ + *Pi* ⟶ *maltooligosaccharide* $(glucosyl)_{n-1}$ + *glucose-1-P*

Cette enzyme, qui a été caractérisée chez *E. coli* (Boos and Shuman, 1998), appartient aussi à la famille GT35 comme la glycogène phosphorylase, mais s'en distingue par une forte activité sur les oligosaccharides alors que la glycogène phosphorylase n'agit que sur le glycogène. Le génome de *R. baltica* n'a cependant révélé qu'une seule protéine appartenant à la famille GT35, la protéine RB8383. Il est donc probable que RB8383 a une spécificité de substrat plus large que les deux enzymes spécialisées d'*E. coli* et qu'elle assume à la fois les rôles de glycogène et de maltodextrine phosphorylases. De telles enzymes ont déjà été identifiées chez plusieurs bactéries (Schinzel and Nidetzky, 1999).

Enfin, le génome de *R. baltica* comprend aussi une α-1,4-glucosidase de la famille GH97 (RB10507) et une alpha-(1,6)-glucosidase de la famille GH13 (RB2986), qui sont sûrement impliquées dans la dégradation finale des maltooligosaccharides, qu'ils soient issus du recyclage du glycogène endogène ou de la dégradation de l'amidon exogène.

II.C.3 - Catabolisme de l'amidon

L'annotation initiale de *R. baltica* avait identifié trois α-amylases potentielles : deux enzymes de la famille GH13 (RB5196 et RB5200) et une protéine de la famille GH57 (RB2160). Mon analyse biochimique de RB2160 démontre que cette protéine n'a pas d'activité amylolytique, bien que je n'aie pas réussi à identifier exactement son activité (Cf. chapitre II). De la même manière, l'analyse modulaire et phylogénétique menée cette fois sur les protéines RB5196 et RB5200 ne supporte pas la prédiction que ces enzymes soient des α-amylases. RB5196 est très proche de l'amylosucrase de *Neisseria polysaccharea* (52% d'identité de séquence), une enzyme impliquée dans la synthèse d'un α-(1,4)-glucane linéaire (Buttcher *et al.*, 1997). Quant à RB5200, c'est une protéine bimodulaire. Le module N-terminal appartient à une famille de phosphatases incluant les sucrose-6-phosphate phosphatases et les trehalose-6-phosphate phosphatases. Le module C-terminal appartient à la famille GH13 et son homologue le plus proche est la malto-oligosyltrehalose trehalohydrolase (TreZ, EC 3.2.1.141) d'*Arthrobacter ramosus* (35% d'identité de séquence, 49% de similitude). Cette enzyme participe avec la malto-oligosyltrehalose synthase TreY à la production de trehalose à partir de glycogène (Yamamoto *et al.*, 2001). Bien que je n'aie pas trouvé d'homologue de TreY dans le génome de *R. baltica*, les deux modules de RB5200 ont en commun un lien avec le tréhalose ou du moins un disaccharide 6-phosphate.

Il est en tout cas peu probable que RB5200 soit une α-amylase. Au total, il n'y a pas actuellement de candidat clair pour assumer une activité amylolytique expliquant la capacité de *R. baltica* à utiliser l'amidon comme source de carbone (Schlesner *et al.*, 2004).

II.D - **Le métabolisme des acides sialiques**

Comme décrit au chapitre II, *R. baltica* présente plusieurs représentants d'une famille multigénique de sialidases (famille GH33). Pour essayer de comprendre le rôle de ces enzymes, j'ai tout d'abord vérifié si cette bactérie marine possédait les enzymes clés de la synthèse des acides sialiques (Angata and Varki, 2002) : (i) la N-acetylneuraminate synthase (neuB, EC 2.5.1.56) qui catalyse la formation de l'acide N-acetyl neuraminique (réaction 1) ; (ii) la N-acylneuraminate cytidylyltransferase (neuA, EC 2.7.7.43) qui active l'acide N-acetyl neuraminique (réaction 2) :

(1) *Phosphoenol pyruvate* + *N-acetyl-D-Mannosamine* + *H2O* ↔ *phosphate* + *N-acetylneuraminate*

(2) *CTP* + *N-acylneuraminate* ↔ *diphosphate* + *CMP-N-acylneuraminate*

J'ai donc effectué une recherche BLAST sur le génome de *R. baltica* en utilisant les protéines caractérisées NeuA de *Neisseria meningitidis* (Edwards and Frosch, 1992) et NeuB d'*E. coli* (Edwards and Frosch, 1992). Aucun homologue de ces deux enzymes n'a cependant été trouvé. Elles ont été cependant trouvées (par cette méthode, et avec les mêmes sondes) dans les génomes d'autres *Planctomycetes* (*B. marina*, *C. Kuenenia stuttgartiensis* et *P. maris*). Une recherche BLAST sur le génome de *R. baltica* en utilisant comme sondes les enzymes de ces trois planctomycètes a également été réalisée, sans plus de succès. Par conséquent, *R. baltica* ne synthétise pas d'acides sialiques et utilisent certainement ses sialidases pour dégrader des polysaccharides ou des glycoconjugés exogènes contenant des acides neuraminiques.

Pour renforcer encore cette idée, j'ai recherché dans le génome de R. baltica d'autres enzymes du catabolisme des acides sialiques, en m'appuyant cette fois sur la base de voies métaboliques KEGG. Deux protéines clés ont été identifiées : (i) RB3352, une N-acetylneuraminate lyase (EC 4.1.3.3) qui catalyse le clivage des acides N-acetyl-neuraminiques (réaction 3) ; (ii) RB3348, une N-acetyl-D-glucosamine 2-epimerase (EC

5.1.3.8) qui interconvertit le N-acetyl-D-mannosamine en N-acetyl-D-glucosamine (réaction 4).

(3) *N-acetyl neuraminate* ↔ *pyruvate* + *N-acetyl-D-mannosamine*

(4) *N-acetyl-D-mannosamine* ↔ *N-acetyl-D-glucosamine*

Il est particulièrement intéressant de noter que les gènes de ces deux protéines se trouvent justement en amont du gène d'une des sialidases de R. baltica (RB3353), ainsi qu'un gène codant un symporteur (RB3349). Ces quatre gènes forment probablement un opéron qui permet à R. baltica de libérer des acides N-acetyl-neuraminiques de biomolécules exogènes et de les convertir en N-Acetyl-D-glucosamine (Figure IV-112).

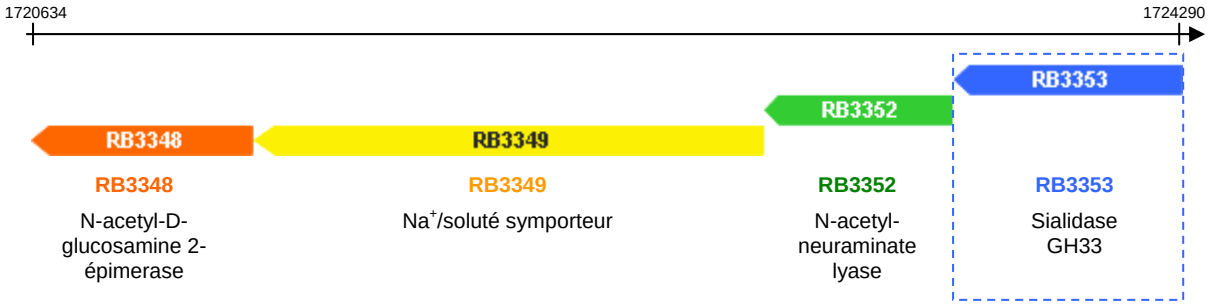


Figure IV-112 : Potentiel opéron de dégradation d'acides sialiques.
Locus des gènes *rb3348* à *rb3353* codant pour un potentiel opéron de dégradation d'acides sialiques exogènes chez *R. baltica*.

Conclusions
et
Perspectives

I - Conclusions

Le génome de la *Planctomycetes* marine *Rhodopirellula baltica* a été séquencé il y a quelques années (Glöckner *et al.*, 2003). Ce génome a révélé contenir un grand nombre de sulfatases et de polysaccharidases. Etant donné la large répartition du phylum des *Planctomycetes* dans les écosystèmes marins, terrestres et d'eau douce, ces bactéries sont pressenties pour jouer un rôle majeur dans le recyclage des polysaccharides environnementaux (provenant de débris de végétaux ou d'animaux). *Rhodopirellula baltica* a été identifiées au cours d'un programme de métagénomique dans la mer Baltique, elle est considérée comme un acteur potentiellement important, participant au cycle du carbone dans cet écosystème.

Afin de comprendre les bases moléculaires de son rôle écologique, la détermination expérimentale des voies métaboliques de *Rhodopirellula baltica* impliquant des polysaccharides reste encore à être réalisée. En attendant, la fiabilité de la prédiction de fonction des enzymes agissant au sein de ces voies est essentielle.

Le but de mon travail de thèse était de permettre la mise en place d'un processus de validation de l'annotation des gènes impliqués dans le métabolisme des polysaccharides complexes chez *Rhodopirellula baltica*, ainsi que l'amorçage de l'effort de validation expérimentale. Plusieurs approches ont été suivies afin de réaliser cet objectif :

(i) Une approche à moyen-débit afin permettre l'analyse d'un grand nombre de ces polysaccharidases, choisies pour leur pertinence dans un cadre d'écologie marine ;

(ii) Une approche à plus bas débit de caractérisation biochimique, menée en parallèle d'une caractérisation cristallographique de quelques unes de ces enzymes ;

(iii) Une approche bioinformatique afin de réévaluer l'annotation initiale par une méthode d'annotation experte.

J'ai ainsi participé au développement d'une méthodologie à moyen débit, basée sur la philosophie de traitement multicible des grands programmes de génomique structurale, pour étudier un nombre conséquent de protéines en parallèle. Le but de ce type d'analyse est de permettre la surexpression des protéines choisies afin de faciliter leur étude, par exemple de caractérisation biochimique. Un programme de plateforme de surexpression à été créé pendant ma thèse, en se basant sur cette approche.

J'ai pu utiliser les développements réalisés pour cette méthodologie à moyen débit pour lancer un programme de surexpression de modules isolés des polysaccharidases de *R. baltica*. Le but de cette étude étant de permettre la caractérisation biochimique et biophysique d'un maximum de ces modules.

J'ai choisi quatre cibles parmi l'ensemble des enzymes clonées pour procéder à leur caractérisation complète. Les quatre cibles ont été choisies dans un souci de compromis entre la pertinence scientifique de la protéine (appartenant de préférence à une voie originale de dégradation de *R. baltica*) et une faisabilité à court terme de sa purification (meilleures surexpression solubles). C'est ainsi qu'une étude structurale et fonctionnelle a été entamée sur :

(i) RB3123 : enzyme annotée kappa-carraghénase de la famille GH16 ;
(ii) RB2160 : enzyme annotée alpha-amylase de la famille GH57 ;
(iii) RB3006 : enzyme annotée sialidase de la famille GH33 ;
(iv) RB5312 : enzymes annotée pectate lyase de la famille PL1 ;

Enfin, en se basant sur mes travaux de caractérisation expérimentale des activités des quatre protéines sélectionnées, j'ai également pu procéder à la réévaluation des annotations des gènes impliqués dans des voies métaboliques de polysaccharides ou de sucres plus simples. Une reconstruction de ces voies métaboliques a permis d'affiner notre connaissance du potentiel de dégradation de *R. baltica*.

I.A - Etude des polysaccharidases de *R. baltica*

L'étude modulaire systématique de l'ensemble des polysaccharidases de *R. baltica* montre qu'un grand nombre d'entre elles sont modulaires. Cette étude a également montré qu'un grand nombre de ces enzymes présentaient un peptide signal, confirmant leur probable sécrétion dans l'environnement de la bactérie.

J'ai pu cloner 92 des 165 modules que contient *R. baltica*, répartis au sein de 128 enzymes. Les profils de surexpression de chacun de ces modules ont été estimés par les techniques analytiques complémentaires que sont l'analyse de gels SDS-PAGE, et l'hybridation sur membrane de type Dot-Blot avec marquage par anticorps.

Fort du succès de clonage des différents modules, j'ai choisi quatre enzymes parmi celles présentant le meilleur compromis entre une bonne expression sous forme soluble et une originalité métabolique pertinente dans le cadre d'un écosystème marin.

I.B - RB3123 : Une nouvelle activité GH16 ?

I.B.1 - Conclusions

RB3123 est une enzyme bi-modulaire, présentant un peptide signal. Elle est constituée d'un module catalytique GH16 et d'un module CBM16. Cette enzyme est assez divergente dans sa famille et les analyses phylogéniques ne tranchent pas sur son appartenance à un sous-groupe de l'arbre de sa famille.

L'étude réalisée sur RB3123 montre que celle-ci est encore une cible récalcitrante d'un point de vue biochimique. Néanmoins, à l'heure où ma thèse se termine, j'ai pu prouver qu'elle ne présente pas d'activité kappa-carraghénase, qui est pourtant celle prédite par son annotation. Il semble de plus qu'elle ne présente aucune activité non plus parmi celles décrites dans la famille GH16. Une modélisation de sa structure par homologie a été réalisée. Il semble que le modèle généré soit de bonne qualité, d'après des critères objectifs tels que sa distribution de charge ou encore la répartition de ses résidus hydrophobes. Cette étude a permis de mettre en évidence que sa gorge de fixation du substrat serait nettement occupée par des résidus basiques. Cela irait dans le sens d'un rôle potentiel dans la dégradation de polysaccharides anioniques, qui sont des molécules très abondantes dans les écosystèmes marins.

I.B.2 - Perspectives

La clarification de son comportement en solution sera primordiale pour pouvoir fiablement décrire son activité. Il est probable qu'il faille passer par l'utilisation de la construction de la protéine entière (qui fait également partie des cibles de l'étude à moyen-débit), la création de nouvelles constructions, ou encore le clonage dans d'autres systèmes hétérologues. Des cribles de substrats marins pourraient également être des pistes prometteuses. Parallèlement à ces études, l'autre protéine de la famille GH16 de *R. baltica* va être également étudiée afin de constater si elle présente une vraie activité kappa-carraghénase.

I.C - RB2160 : Une nouvelle activité GH57 ?

I.C.1 - Conclusions

RB2160 est une enzyme bi-modulaire, ne présentant pas de peptide signal. Elle est constituée d'un module catalytique GH57 et d'un module de fonction non connue exclusivement rencontré dans cette famille, et qui n'existe de plus que chez certains de ses représentants. Par analyse phylogénique, elle groupe avec ses homologues présentant ce module additionnel, cependant, les quelques uns caractérisés ont des fonctions différentes.

Clonée sous sa forme entière, cette enzyme a été assez facile à manipuler, et son comportement en solution indique qu'elle est très probablement repliée correctement (monomérique, faible précipitation avec le temps).

J'ai cependant pu prouver au cours de l'étude biochimique de RB2160 qu'elle ne présentait pas d'activité alpha-amylase, qui était pourtant l'activité prédite par son annotation. L'investigation approfondie des autres activités connues de la famille GH57, une famille essentiellement amylolytique, indique que RB2160 n'exprime aucune activité référencée dans cette famille. Il est toujours possible que les substrats choisis (amylose et glycogène animal) soient effectivement substrats de l'enzyme, mais ne soient pas reconnus dans les conditions expérimentales réalisées. Néanmoins, RB2160 étant relativement divergente dans sa famille, il n'est pas exclu qu'elle présente une fonction différente de ce qui a déjà été décrit. De plus, de nouvelles activités ont été récemment découvertes, qui ouvrent encore le champ d'action de cette famille.

I.C.2 - Perspectives

La biochimie de cette enzyme étant peu problématique, il sera nécessaire de procéder à de nouvelles analyses pour déterminer son activité. Cette enzyme semblant un peu déconnectée du contexte enzymatique de *R. baltica*, la détermination de son activité pourrait se révéler très informative sur un métabolisme original impliquant éventuellement un nouveau substrat. Un problème devra également être résolu : malgré un pI théorique de d'environ 5 elle semble présenter un bilan de charges neutre à sa surface, ce qui pourrait avoir beaucoup de conséquences sur son activité, et son comportement en solution. Une mesure de pI réel devra donc être également conduite.

I.D - RB3006 : Une sialidase marine ?

I.D.1 - Conclusions

RB3006 est une enzyme trimodulaire, présentant un peptide signal. Le module catalytique appartient à la famille GH33, qui est une famille exclusivement sialolytique. La grande interrogation autour de cette enzyme *est en quoi peut-elle diverger de ses paralogues ?* En effet, *R. baltica* présente sept enzymes de cette famille. Même si la famille ne présente que l'activité « sialidase », il est très probable que les sept sialidases de *R. baltica* agissent sur des substrats sensiblement différents, ou dans des contextes différents.

Les trois modules de RB3006 ont été clonés séparément et ses modules GH33 et UNK-1 présentent une bonne expression soluble. Les différentes caractérisations biophysiques réalisées indiquent que ces deux modules sont solubles, stables, et monomériques.

J'ai réalisé plusieurs tests enzymatiques et aucune activité n'a pu être détectée, ce qui est assez surprenant. En effet, le substrat utilisé est un substrat générique (fluorophore-acide sialique), traditionnellement utilisé pour décrire l'activité de cette famille. Plusieurs hypothèses sont envisageables : (i) soit l'activité est nouvelle ; (ii) soit les conditions expérimentales ne sont pas réunies ; (iii) soit l'enzyme n'est active que sous sa forme entière. Je ne pense pas que cette enzyme présente une nouvelle activité. En effet, elle n'est pas spécialement divergente, elle possède les résidus catalytiques, ainsi qu'une grande partie des régions conservées de la famille, et enfin, cette famille est étudiée depuis longtemps et est plutôt bien caractérisée. Le second point peut véritablement être sensible. Seuls deux conditions ont été testées, il est possible que cela ne soit pas assez et qu'il faille élargir le champ d'analyse. Enfin, le point trois n'est pas à négliger. Il est très possible que les modules de RB3006 participent d'une manière ou d'une autre à l'activité.

I.D.2 - Perspectives

Les prochaines expériences à planifier pour cette protéine devront s'assurer que les conditions expérimentales ne peuvent pas révéler l'activité. Des tests sont en préparation pour varier le tampon, pH, salinité, température… Il est probable que la clé de son activité réside dans ces variations. De plus, des constructions de la protéine entière ou au moins des deux premiers modules seront à réaliser, afin de tester le possible effet synergique de la présence des autres modules.

I.E - RB5312 : Une pectate lyase originale

I.E.1 - Conclusions

RB5312 est une enzyme non modulaire, présentant un peptide signal. Cette enzyme appartient à la famille des PL1, dont les activités tournent autour de la dégradation des pectines (au sens large : les polysaccharides de la phase amorphe des plantes). *R. baltica* semble présenter de plus quatre activités pectine lyases.

Clonée sous forme entière, privée de son peptide signal, RB5312 a été également assez facile à purifier. Elle présente un profil stable et se trouve sous forme monomérique en solution. Cette enzyme a été la seule à présenter une activité catalytique et à donner des cristaux.

J'ai ainsi prouvé que RB5312 présentait bien une activité pectine lyase. Il semble de plus que cette enzyme ait une préférence pour les pectines déméthylées, ce qui suggère plus précisément une activité pectate lyase. J'ai également pu déterminer les caractéristiques biochimiques de cette enzyme et les mesures de constante catalytique et de vitesse maximum sont en cours de réalisation. Enfin, j'ai également commencé l'étude des oligosaccharides limites produits par son action, dont des analyses ^{1}H RMN sont en cours.

Cette enzyme a également donné de nombreux cristaux et j'ai pu mesurer plusieurs jeux de données de cristaux natifs de la protéine et l'un d'entre eux à donné un jeu à 1,8 Å. Trois méthodes de phasage ont été testées : le remplacement moléculaire avec un parent structural, le MAD sur l'Ytterbium, et la production de protéine avec incorporation de sélénométhionines pour réaliser des expériences MAD sur le Sélénium. Aucune de ces méthodes n'a cependant permis de résoudre la structure de RB5312.

I.E.2 - Perspectives

Des trempages de cristaux de la protéine native dans un panel plus large de métaux lourds ont été réalisés et des expériences de MAD voire de MIR, selon l'intensité du signal anomal accessible avec chaque métal lourd sera tenté d'ici peu. La résolution de la structure de RB5312 devrait permettre de comprendre en quoi la grande divergence de cette enzyme influe sur sa fonction. J'espère sincèrement pouvoir participer à cette aventure, qui me permettrai de clore ce chapitre de ma thèse.

Il reste bien sûr également à comprendre le rôle de cette enzyme dans l'action de *R. baltica* sur la paroi des végétaux, vu que la bactérie ne semble pas capable de croître avec la pectine comme source de carbone. La caractérisation de son paralogue, qui est moins divergent, pourrait peut-être apporter des éléments de réponse.

I.F - Reconstruction des voies métaboliques

L'analyse bioinformatique de l'ensemble des polysaccharidases de *Rhodopirellula baltica* aura été très riche en enseignements. Il s'agit d'une bactérie vraiment divergente et pour produire des arbres significatifs, j'ai dû parfois creuser les banques de données pour trouver les informations que je cherchais. Le résultat en vaut la peine, l'analyse a permis de réannoter une grande partie de ces enzymes. J'espère que ces résultats pourront participer à une meilleure compréhension de l'étendue des voies de dégradation accessibles aux populations bactériennes marines, et à *Rhodopirellula baltica* en particulier.

La première application de cette réannotation aura permis la reconstruction de quatre voies métaboliques dont j'ai tenté de caractériser des enzymes au cours de ma thèse, et qui renvoient à des originalités de substrat de *R. baltica*.

Elle semble ainsi disposer d'une large game d'enzymes dédiées à la dégradation de la paroi des algues, en particulier de leurs polysaccharides sulfatés. Ses nombreuses sulfatases sont probablement impliquées dans cette dégradation. De plus, des activités originales restent à découvrir : le rôle de RB3123 par exemple, ou encore les bases enzymatiques de sa dégradation du iota-carraghénane, que la bactérie peut dégrader bien qu'aucune iota-carraghénase n'ait été découverte.

J'ai pu identifier l'ensemble des enzymes probablement impliquées dans le métabolisme du glycogène bactérien. Les annotations initiales prédisaient, de plus, trois α-amylases de la famille GH13. Mes analyses ne sont pas aussi tranchées, et leur rôle respectif n'est pas clair. Enfin, le rôle de la GH57 apparaît réellement obscur dans ce contexte. Je pense que la détermination de sa fonction apportera une première réponse

Le métabolisme des acides sialiques est également intrigant. Ces saccharides étant quasi-exclusivement rencontrés dans l'ordre des *Deuterostomes* et chez quelques bactéries, cela pourrait suggérer que *R. baltica* ne se nourrit pas que de plante, mais également d'animaux. Cela serait de plus cohérent avec le fait qu'elle est également capable de croitre sur une source de chondroïtine sulfate, un polysaccharide de la matrice extracellulaire des animaux.

Le fait que *R. baltica* ne puisse pas croitre sur la pectine (Schlesner *et al.*, 2004), mais qu'elle possède au moins une pectate lyase fonctionnelle est assez intrigant. Cela renvoie au

rôle biologique de cette enzyme, un rôle peut-être d'assistance dans la dégradation des parois végétales.

Enfin, beaucoup de voies métaboliques ont été révélées par les études de l'ensemble des polysaccharidases de *R. baltica*. Il semble par exemple cette bactérie puisse dégrader des fructanes. Des voies de biosynthèses se sont également apparues confortées et ont confirmé certaines capacités de *R. baltica*, comme par exemple la synthèse de lipide A dont *R. baltica* possède les enzymes centrales, de nombreuses enzymes qui semblent impliquées dans la glycosylation protéique, la synthèse potentielle de différents osmoprotectants ou encore la synthèse de lipopolysaccharides (LPS) et d'exopolysaccharides (EPS).

Des expériences préliminaires de croissance de *R. baltica* en présence de nombreux polysaccharides ont été menées dans notre laboratoire. Ces résultats seront à confirmer mais permettent de confronter les estimations bioinformatiques avec la réalité de la dégradation. La plupart des résultats présentés par Schlesner et collaborateurs (Schlesner *et al.*, 2004) ont ainsi pu être confirmé, cependant certains résultats ne sont pas aussi tranchés.

Il semblerait par exemple que, contrairement aux résultats publiés, *R. baltica* puisse croitre en présence de pectine, ce qui indique que le type de pectine utilisé (origine, pureté, méthylation, …) pourrait être critique d'une part, et (surtout) que la voie métabolique de dégradation de la pectine est probablement très originale, au point d'être imprédictible à l'heure actuelle. De la même manière, la bactérie est capable de croitre en présence d'amidon mais pas en présence de glycogène. Mes analyses bioinformatiques semblaient pourtant indiquer que les enzymes de la famille GH13 initialement prédites pour être des α-amylases n'en étaient probablement pas. Deux hypothèses seraient donc à envisager : le fait que au moins l'une de ces enzymes soit réellement une α-amylase ou que la ou les α-amylase de *R. baltica* n'ait pas encore été trouvées.

II - Perspectives générales

Des études expérimentales basées sur ma réannotation seront à mener. Je pense avoir permis un éclaircissement du potentiel catalytique de *Rhodopirellula baltica*, tout en ayant fait émerger de nombreuses questions sur les buts de ces voies métaboliques.

Par exemple, la question de la synthèse de ce qui semble être de nombreux osmoprotectants disaccharidiques (tréhalose, fructose-mannose et saccharose) est peut-être à relier à des variations salines importantes du fait que *R. baltica* vit à la frontière des mondes marin et d'eau douce. Il a en effet été montré que les neiges marines étaient des structures complexes et extrêmement étanches, permettant la concentration en leur sein de nombreux composés (Alldredge, 2000). Il est probable que les populations bactériennes croissant dessus doivent être réactives dans leur adaptation à des variations d'osmolarité.

La présence des différents EPS que *R. baltica* semble capable de synthétiser est peut-être à relier, quant à elle, à la constitution des populations microbiennes sur les neiges marines.

Enfin, le lien entre la dégradation effective des polymères par la bactérie et la prédiction des enzymes impliquées dans cette dégradation n'est pas complètement établi pour la plupart des voies métaboliques.

Ces questions n'ont pas de réponses simples. Leur étude nécessitera des analyses variées, mélangeant caractérisation biochimique d'activités enzymatiques, identification claire des polymères utilisables comme source de carbone, et analyses de transcriptomique pour comprendre les contextes d'induction de ces enzymes.

Il apparaît plus que jamais à la fin ma thèse, que les études sur *Rhodopirellula baltica* n'en sont qu'à leurs débuts.

Annexes

Annexe 1

Amorces utilisées pour le clonage de l'ensemble des cibles

Groupe 1 : Séquences compatibles avec BamHI / EcoRI			
Pos.	Nom	Amorces aller (5'->3')	Amorces retour (5'->3')
A01	*rb3405*	ggggggggATCCTgTTCCgTCgCAACCgggCAg	CCCCCCgAATTCTTAATCgACCACgTgAACATCgATgAA
B01	*rb5256*	ggggggggATCCTCCgCTCAAACCgCgTTgggTg	CCCCCCgAATTCTTAgTTgCgTgCCgTCAgCACgAC
C01	*rb10416*	ggggggggATCCAgCAACgCCTTCgAACATCTTgAT	CCCCCCgAATTCTTAgTCCTCATCATTgACAAAgTTAATgA
D01	*rb10416*	ggggggggATCCACACgTgTCATCAACCAgTCTTTg	CCCCCCgAATTCTTAggCTCggTCgACgATgAACgg
E01	*rb10416*	ggggggggATCCACgCCgCTggTTCCTgCAAATC	CCCCCCgAATTCTTACTgTggTCCggAAgCAAgCTgg
F01	*rb9136*	ggggggggATCCCAAgAATCgCAAACTCCggCgg	CCCCCCgAATTCTTATgggTAgAACAACCgCgTgAACg
G01	*rb9136*	ggggggggATCCggCgCggACAACgATgCCTgg	CCCCCCgAATTCTTACTCACCgCTAACTCCAATTTgCAA
H01	*rb9911*	ggggggggATCCggCgATCgTTTgAAAgAgTTggC	CCCCCCgAATTCTTACgAgATCggAACgATgCCATCg
A02	*rb9911*	ggggggggATCCAAgTTggCgTATTTggAgATgCAg	CCCCCCgAATTCTTATTggCCgATgCggTggAAgTAC
B02	*rb3123*	ggggggggATCCgCACAAgAgCATCCTTTTTCCgAT	CCCCCCgAATTCTTATggCTgTTCggCCTCTACTAACg
C02	*rb3123*	ggggggggATCCgCACAAgAgCATCCTTTTTCCgAT	CCCCCCgAATTCTTAgCTgCTCTgCCAgATCCggCA
D02	*rb3123*	ggggggggATCCAACCTCATTAgCAATCCTggATTTg	CCCCCCgAATTCTTATggCTgTTCggCCTCTACTAACg
E02	*rb4561*	ggggggggATCCgAACTTgCTCgCATTgCCgCgT	CCCCCCgAATTCTTAgAAgACTTgCACTTCgACgATTTTT
F02	*rb5816*	ggggggggATCCTCCggCgCCgCCCCgAgTTT	CCCCCCgAATTCTTACTCggCgATggTCgCgTgAATA
G02	*rb3006*	ggggggggATCCCCgggAATCAgCgTggAgACg	CCCCCCgAATTCTTACAgAgAAACgTCggATggAgCC
H02	*rb2377*	ggggggggATCCTCTCAggAACggCCgAACgTTAT	CCCCCCgAATTCTTAgCCAgCgTACggCTCCCACg
A03	*rb2377*	ggggggggATCCCCgTTgATTgCgAATCgAACgCT	CCCCCCgAATTCTTATCTTgTTgCCTCgACATTgTggTg
B03	*rb2377*	ggggggggATCCAACTTTgTgTTTgTgCTCACCgATg	CCCCCCgAATTCTTAATCAgCCgTCggATgCTTgggA
C03	*rb2377*	ggggggggATCCgCATTgTTCATCAAggCAggCTg	CCCCCCgAATTCTTAgAACTTCTgAgACTCgTCTggCC
D03	*rb8073*	ggggggggATCCCAggAATCCTCgCCCgCCgAg	CCCCCCgAATTCTTACCTCTCTTCACCgAgTgCTTgg
E03	*rb736*	ggggggggATCCAACgAgACATTggCAACCATTgCg	CCCCCCgAATTCTTACTTTTCACCgTgATgCTgAATgCC
F03	*rb3421*	ggggggggATCCCgAACCgCgAAATggTgggAAg	CCCCCCgAATTCTTATTgACTgTgTTTgCTATCgAgggC
G03	*rb2206*	ggggggggATCCAAAgTCAgTATCATCgTgCCgACC	CCCCCCgAATTCTTATCggTCgTAgATTCgACgCAAATC
H03	*rb2469*	ggggggggATCCTCCgAAgAATCAACCATgAAgATTTC	CCCCCCgAATTCTTACCgTTTCAgAAACTggggCAgC
A04	*rb6305*	ggggggggATCCAATCgCCTCgAAgAAATTTCggTg	CCCCCCgAATTCTTATCggCgACCgTATCCggTCAC
B04	*rb6377*	ggggggggATCCACggCggTgTTggTACCggTC	CCCCCCgAATTCTTAATgAAAgCCTTCACTgAgCAACAAT
C04	*rb6377*	ggggggggATCCgAAATTggCAAgACACTggCAAAg	CCCCCCgAATTCTTAACCTgCTgCTgTAgCCAgACgT
D04	*rb7643*	ggggggggATCCgATgCgTgTCgTgCACTTTCgAT	CCCCCCgAATTCTTAACggCgAAgCAATCCgTAgCgg
E04	*rb8905*	ggggggggATCCCCCgAAgAACAAAAAgAACAAgCg	CCCCCCgAATTCTTACgACTTCTCTgCTTTCACATAgTC
F04	*rb9623*	ggggggggATCCACAgAAACCATCAgTgTTgTgATCC	CCCCCCgAATTCTTACAgAgCgCgTCTCCggTCCAA
G04	*rb9637*	ggggggggATCCTgCATCAgCgTTgTgATCCCCg	CCCCCCgAATTCTTATTCgATgTggAACAAgTgCCAAgC
H04	*rb9637*	ggggggggATCCCgACgAATCgCTCACgATTgCTT	CCCCCCgAATTCTTATTTCgAgTTTTgAAgCTCAACCAgA
A05	*rb10614*	ggggggggATCCAACATTgACCTgATCATCACCgAg	CCCCCCgAATTCTTATgCCACTATTTTggCgTAgAgTTC
B05	*rb12590*	ggggggggATCCTCggCTCTgAATgTCCCTCgCg	CCCCCCgAATTCTTAACCgACggTTgAATgACCCATTg
C05	*rb2471*	ggggggggATCCgTggCAggTATTAATTgTTTTgACTg	CCCCCCgAATTCTTAATCCAATCTgACgCAATTTggACg

D05	*rb2473*	gggggggATCCACTgCTCATATCgAAAAgCgACTg	CCCCCCgAATTCTTACCCTTCgggTAgTAggCAAgTTg
E05	*rb2990*	gggggggATCCTCgTCCTTgAAgCCACTggCCg	CCCCCCgAATTCTTAgTgAgTTggATTggCgACgACAC
F05	*rb5197*	gggggggATCCAgCATgCCCgCCAgCgggAC	CCCCCCgAATTCTTAggACTgCTCAATgATCTCTgTgAC
G05	*rb9245*	gggggggATCCCCCATgCggATTCTgCTCAgCC	CCCCCCgAATTCTTAAgCCgCCTTggCgAACgCgg
H05	*rb9238*	gggggggATCCgCCACATCAgCTCACCTCCATC	CCCCCCgAATTCTTAAATTCgATgCCAACTTgATgggAC
A06	*rb11356*	gggggggATCCCgTAATCAAgCAACTTTCTgCATCA	CCCCCCgAATTCTTATTTCCATACAACTTTCTCCAgTCCg
B06	*rb11527*	gggggggATCCCgTAATCAAgCCACCTTCTgTATC	CCCCCCgAATTCTTATTTCCATACgACTTTCTCCAATCCg
C06	*rb11533*	gggggggATCCgACgCACCgTTCTTTgCCgTgC	CCCCCCgAATTCTTACCCAAACTTCCggAgATCAggAT
D06	*rb12601*	gggggggATCCAACAAAACCATgTTggACTACggAC	CCCCCCgAATTCTTAgCCAAACATCAgTTCTTTgCTggg
E06	*rb9614*	gggggggATCCAgCATTCCgAAgATCCTgCATCAA	CCCCCCgAATTCTTACgAgCgTggATCgCCCAACCA
F06	*rb9640*	gggggggATCCAATCTgCCTCgggTCgCTgTTTT	CCCCCCgAATTCTTACCATTggCTgCTCATgTACgTCT
G06	*rb3417*	gggggggATCCgCCgAAAAAgTCgCgAAgCAAAC	CCCCCCgAATTCTTACTTCATCCATTTTgCgTgAgCTTTg
H06	*rb4024*	gggggggATCCCgAACTCgCAAACCgATCACCTg	CCCCCCgAATTCTTAACggTTTCgTAggAgCTgTTTgC
A07	*rb6145*	gggggggATCCgAAgCCTACACATggTCggCCg	CCCCCCgAATTCTTACggCgTCAACTCTTTCCATCCAC
B07	*rb8823*	gggggggATCCggATCACCTCAgACgCAgTCCg	CCCCCCgAATTCTTACTCCTgCTCgATCgAggTATCg
C07	*rb9546*	gggggggATCCTgTCCCTCCAgCAACgACTTgTC	CCCCCCgAATTCTTACTTTTCgTTTTggTgACgCAACAAC
D07	*rb12316*	gggggggATCCgAAggTTACTggCCCgACggAA	CCCCCCgAATTCTTATTCAgggATACTCgTgACggATg
E07	*rb977*	gggggggATCCgATgTgCTgATTCggTCCggTC	CCCCCCgAATTCTTACgTCCTgCCAATCAACTCCgCC
F07	*rb981*	gggggggATCCCATgACATCgTTCCTggTgCACC	CCCCCCgAATTCTTAgggCAACTgTTCgTATTTgCggT
G07	*rb4934*	gggggggATCCATggggATTCgCAACgAACATACg	CCCCCCgAATTCTTATgCTgCTCggTTTCTCCgCTgg

Groupe 2 : Séquences compatibles avec BamHI / MfeI			
Pos.	Nom	Amorces aller (5'->3')	Amorces retour (5'->3')
H07	*rb5256*	gggggggATCCggAgTCATCCgTgTCgCgTTTg	CCCCCCCAATTgTTACTTTCCgAggATCgCgTCgCg
A08	*rb10416*	gggggggATCCCAggTCAgTTTgCATgCCgAgAT	CCCCCCCAATTgTTAAATCgTTCCggTCACCTgCCgA
B08	*rb10416*	gggggggATCCACggAACTgCTTCCCgTCgATC	CCCCCCCAATTgTTAggCCTgggCTCgCCAggCg
C08	*rb2702*	gggggggATCCCAgTCCAACgAAAATTCgAAACCC	CCCCCCCAATTgTTACTCATCCAgTTCCCACACCCgg
D08	*rb3353*	gggggggATCCgAAgCCAACgACgAAgCCAggg	CCCCCCCAATTgTTAATTgTTCTCCTggTCggTTTCAgA
E08	*rb2160*	gggggggATCCTCgCCTCACgTTCACCTTTgCTT	CCCCCCCAATTgTTACgCgTTCACATgATCgTgggATT
F08	*rb700*	gggggggATCCCAACgATTgCggTgCgAgTACTT	CCCCCCCAATTgTTAgTTgCggAACTCgTAgAACgCg
G08	*rb700*	gggggggATCCTACgAgTTTTTggCgTCCAgTgg	CCCCCCCAATTgTTAATCgACCAgTCgATAAATCgAgATC
H08	*rb3421*	gggggggATCCCAggCTCCgACggCCCCCAA	CCCCCCCAATTgTTAgTCgCCAAAgTgCTCgATCAAAg
A09	*rb13211*	gggggggATCCCACgCTATCgAAACCAgTCATgC	CCCCCCCAATTgTTATgAggATCTTgTTTCTgAATTCCgg
B09	*rb13211*	gggggggATCCgCgTACAgCCgATCCCAAACAC	CCCCCCCAATTgTTATCTCgAACgCgTTTCTCgTTgAg
C09	*rb10434*	gggggggATCCTggTTgAgCgCATgCAAACACCg	CCCCCCCAATTgTTAATgACTgCATTTggggAgAATTgC
D09	*rb2484*	gggggggATCCCCTgACgCAgCCAAACCgAATTg	CCCCCCCAATTgTTAAgTgACCTTCCTTAgTTTCgATTCT
E09	*rb2485*	gggggggATCCTTTAAATTgAATTCCgAACAACTCATC	CCCCCCCAATTgTTAggCACCggTCggAAgCCCCA
F09	*rb2499*	gggggggATCCTgCACTCTTCCgCACTgCCgC	CCCCCCCAATTgTTACCCgTAATgATTgTgTTgCgATTg
G09	*rb5197*	gggggggATCCCgATTgCCggAATTCgACCggA	CCCCCCCAATTgTTAgTTCggAATCCggATATgATTCAg
H09	*rb11529*	gggggggATCCAgCATCATggCgAAAAgTATTTCCA	CCCCCCCAATTgTTACgCACggACgATTTgACCTCCT
A10	*rb11533*	gggggggATCCATgAATCTCgATCgCCgCgATgA	CCCCCCCAATTgTTAgTTgAAgAACCgCgTTTgACTCC
B10	*rb1367*	gggggggATCCTgCgTgCgTgACACCgTgAgC	CCCCCCCAATTgTTACgAggTCATgCCCCAgCTTTTg

Groupe 3 : Séquences compatibles avec BglII / EcoRI			
Pos.	Nom	Amorces aller (5'->3')	Amorces retour (5'->3')
C10	*rb10416*	gggggg**AgATCT**AATgAgggACTCACCgCTgAgg	CCCCCC**gAATTC**TTAACTCggCAATgCCgTTTTAAAggA
D10	*rb10416*	gggggg**AgATCT**TCTgACgAgCAAACgCTTgTTCg	CCCCCC**gAATTC**TTAggCggTCTCAggCAATTCAATCg
E10	*rb12360*	gggggg**AgATCT**ggAATCgACTACCgAgAAgACTAC	CCCCCC**gAATTC**TTACCAgATgCTTTTgACTCggTTgAg
F10	*rb3006*	gggggg**AgATCT**CCggCTgAATCTCAAACAgCACC	CCCCCC**gAATTC**TTAATTgCTggTCATCgATgCgAggT
G10	*rb8895*	gggggg**AgATCT**gACgAACgAACATCggTgCTCC	CCCCCC**gAATTC**TTAAAgCCCCAgTTCATCCAgTgCg
H10	*rb5846*	gggggg**AgATCT**gCATCgCCTgTTAAAAAgATTgATg	CCCCCC**gAATTC**TTATTTTCCgTTgCggTgAgCgATCA
A11	*rb4333*	gggggg**AgATCT**AAAATCggCATCATAggTCATCTCA	CCCCCC**gAATTC**TTACgCTgCgACCTCgTTTggAAAg
B11	*rb8313*	gggggg**AgATCT**CgTATCgCCCATgTgATCACCC	CCCCCC**gAATTC**TTAgCgAgCAgAAATCAgATCCgCAT
C11	*rb11688*	gggggg**AgATCT**CgCTATCgCCgTggCATCggC	CCCCCC**gAATTC**TTAAgCggCTTTTTgACgTgACTgAAC
D11	*rb9648*	gggggg**AgATCT**AAAgCCgCggTCATCCTgTgTC	CCCCCC**gAATTC**TTACCAACTCgAATgATAAAgCTTCgC
E11	*rb5312*	gggggg**AgATCT**CAgAAgCCATTggCCTTTCCgAC	CCCCCC**gAATTC**TTACgggATgCTgTTgATgTATTgTTC
F11	*rb2091*	gggggg**AgATCT**ggCgAACCgCAgTCACTgACAA	CCCCCC**gAATTC**TTAgggCAggTCgTTgAAggTgTAg
G11	*rb5007*	gggggg**AgATCT**gCACAgCTTCCACCAACgggAC	CCCCCC**gAATTC**TTATTCATTggATgAgCCAgTgAgTTg

Groupe 4 : Séquences compatibles avec BglII / MfeI			
Pos.	Nom	Amorces aller (5'->3')	Amorces retour (5'->3')
H11	*rb10416*	gggggg**AgATCT**CCggTCCAAATCCTCgTCCAgg	CCCCCC**CAATTg**TTATgTTTggTCCAACAACCATTggCC
A12	*rb4561*	gggggg**AgATCT**gAggTTTCCATTgTTCCATTgCCg	CCCCCC**CAATTg**TTAgCgTTgTTgAAAgTTCTggTAATCg
B12	*rb3006*	gggggg**AgATCT**CAgCgTCCACCgACTggCgTg	CCCCCC**CAATTg**TTAATAgAACCAAgAgACgggAATTCg
C12	*rb700*	gggggg**AgATCT**CAACgATTgCggTgCgAgTACTT	CCCCCC**CAATTg**TTAATCgACCAgTCgATAAATCgAgATC
D12	*rb700*	gggggg**AgATCT**ggCgAgCAACgTgACTTTggTTT	CCCCCC**CAATTg**TTAgCTgCCCgAATgAACCgTTAATg
E12	*rb2990*	gggggg**AgATCT**TCAATgATATCAAATgACCTTCTgAAg	CCCCCC**CAATTg**TTACCgATTgAACgAATCAATgTACAAC
F12	*rb8383*	gggggg**AgATCT**TCCAACACgCTTTCCgCTCCgC	CCCCCC**CAATTg**TTACgACAgTggTCgCACgTCCCA
G12	*rb3601*	gggggg**AgATCT**gAAACgCCCgCCgACgTTCTC	CCCCCC**CAATTg**TTACTggTCTTCggTATTgATCTCgTC
H12	*rb3639*	gggggg**AgATCT**AgCgACAAAATCACCATCgCATTg	CCCCCC**CAATTg**TTATTgTgCgTCTggggAggAAggC

Annexe 2

Résumé des résultats des tests de surexpression, par analyse par SDS-PAGE et Dot-Blot

Cible			Expression SDS PAGE		Expression Dot Blot	
position	Protéine		Fraction totale	Fraction soluble	Fraction totale	Fraction soluble
A01	RB3405	UNK	+	+	+	+
B01	RB5256	UNK	+	+	+	+
C01	RB10416	Dock1	-	-	+	+
D01	RB10416	UNK-1	-	-	+	+
E01	RB10416	PA14-2	-	-	+	+
F01	RB9136	UNK	+	+	++	++
G01	RB9136	GH10	-	-	++	++
H01	RB9911	GH10	-	-	-	-
A02	RB9911	UNK	+	+	+	+
B02	RB3123	ALL	+	+	+	+
C02	RB3123	GH16	+	+	+	+
D02	RB3123	CBM16	-	-	+	+
E02	RB4561	UNK	+	-	+	+
F02	RB5816	UNK	+	-	+	+
G02	RB3006	UNK-2	+	-	+	+
H02	RB2377	Sulf-1				
A03	RB2377	LamG	+	+	-	-
B03	RB2377	Sulf-2	-	-	+	+
C03	RB2377	GH43	+	+	+	+
D03	RB8073	GH43	++	++	+	+
E03	RB736	DUF1080	-	-	+	+
F03	RB3421	GH86	-	-	+	+
G03	RB2206	GT2	-	-	+	+

Cible			Expression SDS PAGE		Expression Dot Blot	
position	**Protéine**		**Fraction totale**	**Fraction soluble**	**Fraction totale**	**Fraction soluble**
H03	RB2469	GT2	-	-	+	+
A04	RB6305	GT2	nd	nd	+	+
B04	RB6377	GT2	-	-	+	+
C04	RB6377	UNK	-	-	+	+
D04	RB7643	GT2	-	-	+	+
E04	RB8905	GT2	-	-	++	++
F04	RB9623	GT2	nd	nd	+	+
G04	RB9637	GT2	-	-	+	+
H04	RB9637	UNK	+	-	+	+
A05	RB10614	GT4	-	-	+	+
B05	RB12590	GT4	-	-	+	+
C05	RB2471	GT4	-	-	+	+
D05	RB2473	GT4	++	+	+	+
E05	RB2990	UNK	+	+	+	+
F05	RB5197	GT4	-	-	+	+
G05	RB9245	GT9	-	-	+	+
H05	RB9238	GT11	-	-	+	+
A06	RB11356	GT12	-	-	+	+
B06	RB11527	GT12	+	-	+	+
C06	RB11533	SulfT	-	-	-	-
D06	RB12601	GT26	++	-	+	+
E06	RB9614	GT32	+	+	+	+
F06	RB9640	GTNC	+	+	+	+
G06	RB3417	PL10	+	-	-	-
H06	RB4024	CE1	-	-	-	-
A07	RB6145	CE1	-	-	+	+
B07	RB8823	CE1	+	-	++	++
C07	RB9546	CE1	+	-	++	++

Annexe 2

Cible			Expression SDS PAGE		Expression Dot Blot	
position	Protéine		Fraction totale	Fraction soluble	Fraction totale	Fraction soluble
D07	RB12316	CE4	+	-	+	+
E07	RB977	CE9-1	-	-	+	+
F07	RB981	CE9				
G07	RB4934	CE11	-	-	+	+
H07	RB5256	GH5	-	-	+	+
A08	RB10416	CalxB	-	-	+	+
B08	RB10416	UNK-2	+	+	++	++
C08	RB2702	GH16	+	+	+	+
D08	RB3353	GH33	+	-	+	+
E08	RB2160	GH57	+	-	++	++
F08	RB700	UNK-1	-	-	+	+
G08	RB700	UNK-2	+	-	++	++
H08	RB3421	UNK	+	-	+	+
A09	RB13211	GT2	-	-	+	+
B09	RB13211	UNK	-	-	+	+
C09	RB10434	GT4				
D09	RB2484	GT4	-	-	+	+
E09	RB2485	GT4	-	-	-	-
F09	RB2499	GT4	-	-	+	+
G09	RB5197	SPP	-	-	+	+
H09	RB11529	GT12	-	-	+	+
A10	RB11533	GT25	+	+	+	+
B10	RB1367	GT26	-	-	-	-
C10	RB10416	PA14-1	+	+	++	++
D10	RB10416	UNK-3	+	-	+	+
E10	RB12360	GH32	+	-	++	++
F10	RB3006	UNK-1	+	+	++	++
G10	RB8895	GH33	++	+	+	+

Cible			Expression SDS PAGE		Expression Dot Blot	
position	**Protéine**		**Fraction totale**	**Fraction soluble**	**Fraction totale**	**Fraction soluble**
H10	RB5846	GT2	-	-	+	+
A11	RB4333	GT4	-	-	-	-
B11	RB8313	GT4	-	-	-	-
C11	RB11688	GT30				
D11	RB9648	GTNC	-	-	+	+
E11	RB5312	PL1	+ +	+	++	++
F11	RB2091	CE6	+	-	++	++
G11	RB5007	CE6	-	-	+	+
H11	RB10416	GH10	-	-	-	-
A12	RB4561	GH20	-	-	-	-
B12	RB3006	GH33	+	-	++	++
C12	RB700	ALL	+	+	++	++
D12	RB700	GH78	+	-	-	-
E12	RB2990	GT4	-	-	+	+
F12	RB8383	GT35	-	-	-	-
G12	RB3601	PL7	-	-	++	++
H12	RB3639	CE12	-	-	-	-
Pourcentage de réussite			38%	23%	82%	82%

Publications

Publications

J'ai publié mes résultats de cristallisation dans une note de cristallisation au courant de cette année. L'article a été publié dans *Acta Crystallographica Section F*. Cet article est joint dans les pages qui suivent.

Un article décrivant l'ensemble des résultats présentés dans ce manuscrit, ainsi que mes analyses sur les annotations du génome de *Rhodopirellula baltica* est en cours de rédaction et sera soumis idéalement avant décembre, dans un journal à sensibilité écologique. Son titre ainsi que la liste des auteurs sont présentés ci-après :

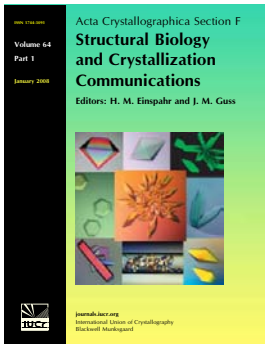
> Jérôme DABIN, Agnès GROISILLIER, Murielle JAM, Tristan BARBEYRON, Margarette BAUER, Franck-Oliver GLÖCKNER, Mirjam CZJZEK and Gurvan MICHEL. Exploring the carbohydrate metabolism of *Rhodopirellula baltica*: a post-genomic validation of the enzymatic potential of a key polymer degrader in the sea.

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091
Editors: **H. M. Einspahr and J. M. Guss**

Expression, purification, crystallization and preliminary X-ray analysis of the polysaccharide lyase RB5312 from the marine planctomycete *Rhodopirellula baltica*

Jérôme Dabin, Murielle Jam, Mirjam Czjzek and Gurvan Michel

Acta Cryst. (2008). F**64**, 224–227



Acta Crystallographica Section F: Structural Biology and Crystallization Communications is a rapid all-electronic journal, which provides a home for short communications on the crystallization and structure of biological macromolecules. It includes four categories of publication: protein structure communications; nucleic acid structure communications; structural genomics communications; and crystallization communications. Structures determined through structural genomics initiatives or from iterative studies such as those used in the pharmaceutical industry are particularly welcomed. *Section F* is essential for all those interested in structural biology including molecular biologists, biochemists, crystallization specialists, structural biologists, biophysicists, pharmacologists and other life scientists.

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Expression, purification, crystallization and preliminary X-ray analysis of the polysaccharide lyase RB5312 from the marine planctomycete *Rhodopirellula baltica*

Jérôme Dabin,[a,b]‡ Murielle Jam,[a,b]‡ Mirjam Czjzek[a,b]* and Gurvan Michel[a,b]

[a]UPMC University Paris 06, UMR 7139, Marine Plants and Biomolecules, Station Biologique de Roscoff, F-29682 Roscoff, Bretagne, France, and [b]CNRS, UMR 7139, Marine Plants and Biomolecules, Station Biologique de Roscoff, F-29682 Roscoff, Bretagne, France

‡ These authors contributed equally to the work described in this article.

Correspondence e-mail: czjzek@sb-roscoff.fr

Received 14 November 2007
Accepted 13 February 2008

Polysaccharide lyases belonging to family PL1 act on pectins. These anionic polymers are usually produced by terrestrial plants and therefore pectinolytic enzymes are not frequently observed in marine microorganisms. The protein RB5312 from the marine bacterium *Rhodopirellula baltica* is distantly related to family PL1 pectate lyases, but its exact function is unclear. In this study, the expression and purification of a recombinant form of RB5312 are described. This protein was crystallized using the hanging-drop vapour-diffusion method. The crystals belongs to space group $P2_12_12_1$, with unit-cell parameters $a = 39.05$, $b = 144.05$, $c = 153.97$ Å, $\alpha = \beta = \gamma = 90^\circ$. A complete data set was collected to 1.8 Å resolution from a native crystal.

1. Introduction

The cell walls of marine algae and seagrasses are characterized by an abundance of anionic polysaccharides formed from carboxylic or sulfated sugars (Aquino *et al.*, 2005; Kloareg & Quatrano, 1988). These polymers constitute a crucial carbon source for a number of marine bacteria that secrete specific glycoside hydrolases or polysaccharide lyases (Michel *et al.*, 2006). Among these microorganisms, the Planctomycetes are recognized as one of the key phyla catalyzing important transformations in the global carbon cycle in the sea (Alldredge, 2000; DeLong *et al.*, 1993). Planctomycetes frequently inhabit phytodetrial macroaggregates in marine environments where they mineralize this organic matter, which is mainly composed of polysaccharides (Miskin *et al.*, 1999; Neef *et al.*, 1998; Wang *et al.*, 2002). *Rhodopirellula baltica* is a marine representative of this important bacterial phylum and it was the first to have its genome completely sequenced (Glöckner *et al.*, 2003). Interestingly, its genome revealed an unprecedented large number of sulfatases, accompanied by numerous polysaccharidases. These observations suggest that *R. baltica* is able to degrade a wide range of anionic polysaccharides from marine algae and seagrasses.

Among the various enzymes found in the *R. baltica* genome, two have been annotated as putative pectate lyases: RB5312 (GenBank accession No. CAD74167) and RB5316 (CAD74169) (Glöckner *et al.*, 2003). Both proteins have been assigned to family PL1 of the polysaccharide lyases by the Carbohydrate Active Enzyme website (http://www.cazy.org/; Coutinho & Henrissat, 1999). Pectins are a group of plant polysaccharides that are mainly constituted of D-galacturonic acid (GalA) and three polysaccharide domains are often found: homogalacturonan (HGA), rhamnogalacturonan-I (RGI) and rhamnogalacturonan-II (RGII) (Mohnen, 1999; O'Neill *et al.*, 1990). HGA is a linear homopolymer of (1,4)-α-linked D-galacturonic acid and is deposited in the cell wall of land plants in a form that has 70–80% of GalA residues methyl-esterified at the C-6 carboxyl (Mohnen, 1999; O'Neill *et al.*, 1990). Pectate and pectin lyases cleave the α-(1,4)-linkages in HGA domains by a β-elimination mechanism, releasing an unsaturated C4—C5 bond at the non-reducing end of the cleaved polysaccharide. Pectate lyases (EC 4.2.2.2) are specific for demethylated forms of HGA and require Ca^{2+} for activity (Pilnik & Rombouts, 1981; Scavetta *et al.*, 1999). In contrast, pectin lyases (EC 4.2.2.10) cleave methylated HGA

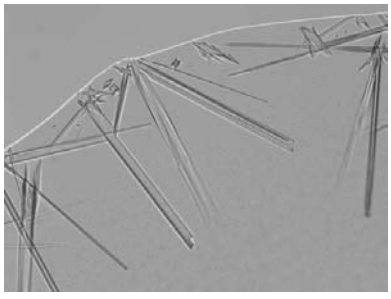


Table 1
Oligonucleotides used for the cloning of *rb5312*.

The primers were designed for a half-hybridization temperature of 345 K.

Forward	GGGGGGAGATCTCAGAAGCCATTGGCCTTTCCGAC
Reverse	CCCCCCGAATTCTTACGGGATGCTGTTGATGTATTGTTC

according to a calcium-independent mechanism (Mayans *et al.*, 1997; Pilnik & Rombouts, 1981).

The presence of putative pectinolytic enzymes in a marine bacterium is surprising since pectins are typical of terrestrial higher plants. Nevertheless, pectin-like polymers have been identified in some marine green algae (Kloareg & Quatrano, 1988) and seagrasses (Ovodov *et al.*, 1975). To evaluate the quality of the functional predictions of Glöckner *et al.* (2003), we have performed our own sequence analyses using *BLAST* searches against the SwissProt database. RB5316 displays 37% and 29% sequence identity to the pectate lyases PelA from *Emericella nidulans* and PelE from *Erwinia chrysanthemi*, respectively (Ho *et al.*, 1995; Lietzke *et al.*, 1994). Therefore, RB5316 is indeed homologous to characterized members of the PL1 family and is likely to be a pectate lyase. In contrast, for the second protein RB5312 only the N-terminal region shares some similarities to characterized proteins: the pectate lyase PelA from *Em. nidulans* (29% sequence identity for Ala22–Ser222, RB5312 numbering) and the noncatalytic pollen allergen AgE from *Ambrosia artemisiifolia* (27% sequence identity for Ala19–Asp160; Rafnar *et al.*, 1991). The C-terminal region of RB5312 (Gly225–Pro455) displays no significant similarity to any characterized protein. At this low level of sequence identity, the exact function of RB5312 is unclear and the prediction that it is a pectate lyase is difficult to substantiate.

Therefore, the focus of our study is to unravel the biochemical function of the marine polysaccharide lyase RB5312. Determination of its crystal structure will also help to establish the role of this divergent member of the PL1 family. We have thus cloned the *rb5312* gene from *R. baltica* into *Escherichia coli* and produced pure recombinant protein. We also report the crystallization of RB5312 and the preliminary X-ray analysis of the crystals.

2. Experimental

2.1. Overexpression and purification of RB5312

According to the program *SignalP* (Bendtsen *et al.*, 2004), RB5312 features an N-terminal signal peptide that is cleaved between residues Ala23 and Gln24. The nucleotide sequence corresponding to the mature RB5312 protein (Gln24–Pro455) was amplified by PCR from *R. baltica* genomic DNA using a set of primers (see Table 1) at a half-hybridization temperature of 345 K. The obtained PCR product was then purified using the Qiagen QIAquick system and digested with *Bgl*II/*Eco*RI (5′/3′ ends) in NEB2 buffer (BioLabs) at 310 K for 3 h. Ligation was performed overnight at 293 K using T4 ligase (Sigma) in a pFO4 vector (derivative of pET15) pre-digested with *Bam*HI/*Eco*RI. This resulted in a gene coding for a recombinant protein (441 residues in total, 48 701 Da) flanked by an N-terminal tag encompassing a methionine followed by six histidine residues and two residues, arginine and serine, corresponding to the *Bgl*II restriction site. The ligation mixture was transformed into *E. coli* DH5α strain. The recombinant plasmid was extracted with the Wizard Plus SV Minipreps kit (Promega) and used to transform *E. coli* BL21 (DE3). The recombinant cells were incubated at 293 K in ZYP-5052 medium (Studier, 2005) with 100 µg ml^{-1} ampicillin until saturation of the culture (final $OD_{600\,nm} \simeq 15$). The cells were harvested by centrifugation (4000*g*, 277 K, 20 min). The cell pellet was resuspended in buffer *A* (50 m*M* NaH_2PO_4 pH 8.0, 500 m*M* NaCl, 50 m*M* imidazole, 5% glycerol). The cells were disrupted using a French press and the lysate was cleared by centrifugation (50 000*g*, 277 K, 30 min). The protein supernatant was applied onto a 10 ml column consisting of IMAC HyperCell resin (Pall Corporation) charged with 100 m*M* $NiSO_4$ and pre-equilibrated with buffer *A*. After a step of washing with buffer *A* (two column volumes), the protein was eluted with a 60 ml linear gradient from buffer *A* to buffer *B* (50 m*M* NaH_2PO_4 pH 8.0, 500 m*M* NaCl, 400 m*M* imidazole, 5% glycerol) at a flow rate of 1 ml min^{-1}. The protein eluted at 160 m*M* imidazole. The protein peak was analyzed by SDS–PAGE (Fig. 1) and the purest fractions were pooled (10 ml final volume). The volume was reduced to 4 ml by ultrafiltration on an Amicon membrane (polyethersulfone, 30 kDa cutoff). The protein sample was injected onto a Superdex 75 HiLoad (GE Healthcare) column pre-equilibated with buffer *C* (50 m*M* HEPES–HCl pH 7.5, 50 m*M* NaCl). Elution was performed with 60 ml buffer *C* at a flow rate of 1 ml min^{-1}. The final protein yield was 2 mg per 500 ml of culture medium. The purified protein was concentrated to 9.7 mg ml^{-1} by ultrafiltration on an Amicon membrane.

2.2. Preliminary pectinolytic activity tests

The pectinolytic activity of RB5312 was tested spectrophotometrically at 235 nm with a UV-2041 PC spectrophotometer (Shimadzu, Japan) by measuring the increase in absorbance caused by the formation of C4–C5 unsaturated sugars. Measurements were taken every 25 s over 10 min at 313 K. The standard assay (1 ml) contained 5 µl purified enzyme at 1 mg ml^{-1}, 995 µl 100 m*M* Tris–HCl pH 8.0 and 1%(*w*/*v*) polygalacturonic acid (PGA) from citrus fruit (Sigma). This activity test was also performed with 995 µl 100 m*M* Tris–HCl pH 8.0, 1%(*w*/*v*) PGA and 5 m*M* EGTA as a control.

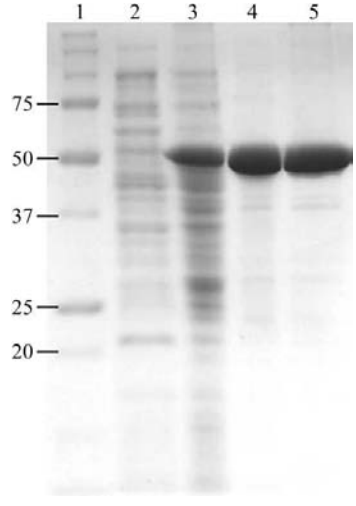


Figure 1
SDS–PAGE gel (12.5%) stained with Coomassie Blue showing the different steps in the heterologous expression and purification of the polysaccharide lyase RB5132. Lane 1, molecular-weight markers (kDa); lane 2, soluble fraction of nontransformed *E. coli* BL21 lysate; lane 3, soluble fraction of *E. coli* lysate with expression; lane 4, fraction with maximum intensity (absorption at 280 nm) after Ni–IMAC column chromatography; lane 5, fraction with maximum intensity after size-exclusion column chromatography.

Table 2
Data-collection statistics of RB5312 crystals.

Values in parentheses are for the highest resolution shell.

Space group	$P2_12_12_1$
Unit-cell parameters (Å, °)	a = 39.05, b = 144.05, c = 153.97, $\alpha = \beta = \gamma = 90$
Resolution range (Å)	52.63–1.8 (1.90–1.80)
No. of observations	304224 (14846)
No. of unique reflections	76 671 (7858)
Completeness (%)	93.3 (67.0)
$\langle I/\sigma(I)\rangle$	17.3 (3.9)
Redundancy	4.0 (1.9)
$R_{\rm merge}$† (%)	5.4 (16.7)

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$.

2.3. Protein crystallization

All crystallization experiments were carried out at 292 K. Initial crystallization trials were performed with the Wizard I and II (Emerald BioStructures, Inc.) and JCSG+ Suite (Qiagen) kits; that is, a total of 192 conditions in two 96-well plates from Corning. The trials were set up using a Cartesian crystallization robot and the sitting drops were made up by mixing 300 nl protein solution (9.7 mg ml^{-1} in 50 m*M* HEPES buffer pH 7.5, 50 m*M* NaCl) with 150 nl reservoir solution. Subsequently, the best conditions were optimized in 24-well Linbro plates using the hanging-drop vapour-diffusion method. These drops were prepared on siliconized cover slips by mixing 2 µl protein solution with 1 µl well solution. The drops were equilibrated against 0.5 ml reservoir solution.

2.4. Data collection and X-ray diffraction analysis

X-ray diffraction data were collected from a crystal of recombinant RB5312 at 100 K on beamline ID14-EH2 at the ESRF (Grenoble, France) using an ADSC Quantum 4R CCD detector. All crystals were flash-cooled in a liquid-nitrogen stream. Since the crystallization condition contained 40% 2-methyl-2,4-pentanediol (MPD), it was not necessary to add a cryoprotectant. The wavelength of the synchrotron X-rays was 0.933 Å. The crystal was rotated through 100° with 0.4° oscillation per frame. Further data-collection statistics are given in Table 2. All raw data were processed using the program *MOSFLM* (Leslie, 1992). The resultant data were merged and scaled using the program *SCALA*, which is part of the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). An attempt to solve the structure of RB5312 by molecular replacement was performed using the program *AMoRe* (Navaza, 2001). The model was selected by a sequence search using *BLAST* against the PDB sequence database.

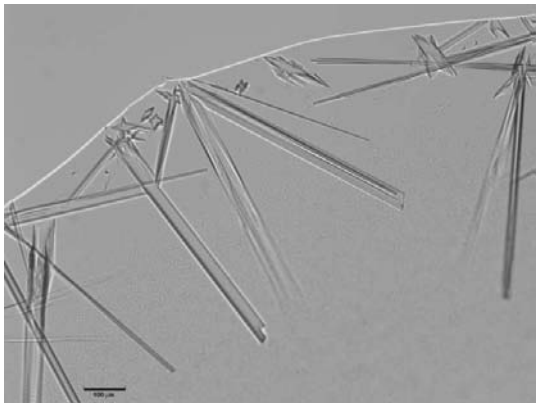


Figure 2
Needle-shaped crystals of polysaccharide lyase RB5132.

3. Results

RB5312 was expressed in *E. coli* without its signal peptide as a soluble protein. This recombinant protein was purified by a combination of IMAC and size-exclusion chromatography in sufficient quantities for crystallization (Fig. 1). A linear increase in $A_{235\,\rm nm}$ was observed when RB5312 was added to a reaction assay containing polygalacturonic acid from citrus fruit. This phenomenon is completely suppressed by the addition of 5 m*M* EGTA. Therefore, RB5312 displays pectinolytic activity and its mechanism is calcium-dependent, suggesting a pectate-lyase specificity. To confirm this substrate specificity, purification of the terminal products of RB5312 is under way. The purified oligosaccharide will be characterized by ^{1}H NMR (Dabin *et al.*, in preparation).

The crystallization screening of RB5312 resulted in the identification of several successful conditions containing MPD and polyethylene glycol (PEG). The optimized conditions for crystal growth are 40% MPD, 4% PEG 8000, 150 m*M* sodium cacodylate pH 6.0 at 292 K. Needle-shaped crystals grew within several days (Fig. 2). Most crystals of the recombinant native protein diffracted to 2.5 Å resolution. However, by screening numerous crystals (∼15) one single crystal (of dimensions 0.3 × 0.05 × 0.05 mm) was found to diffract to 1.8 Å resolution. A complete data set was collected from this crystal and the data-collection quality is reported in Table 2. The space group was determined to be $P2_12_12_1$, with unit-cell parameters a = 39.05, b = 144.05, c = 153.97 Å, $\alpha = \beta = \gamma = 90°$. The asymmetric unit most probably contains two molecules, giving a crystal volume per protein weight ($V_{\rm M}$) of 2.17 Å^3 Da^{-1} and a solvent content of 43% by volume (Matthews, 1968).

The closest sequence match to RB5312 in the PDB is Jun a 1 (PDB code 1pxz), the major cedar pollen allergen from *Juniperus ashei* (Czerwinski *et al.*, 2005). Jun a 1 adopts a β-helix fold and clearly belongs to the PL1 family, but its sequence identity to RB5312 is hardly significant (only 14%). Although the number of insertions and deletions between the two sequences is low, we nevertheless attempted to solve the structure of RB5312 by molecular replacement using the atomic coordinates of 1pxz. However, as expected, these attempts failed. We have therefore decided to produce selenomethionine-labelled protein and the first crystallization trials are under way.

This research was supported by grants to GM from the GIS 'Génomique Marine' and the French Research Ministry (ACI Young Researcher). Additional support was provided by the Region Bretagne (Marine3D program to MC). We are grateful to Drs Rudolph Amann and Frank-Oliver Glöckner for the gift of the *R. baltica* strain and genomic DNA. We thank M. Robert Larocque and Dr Miroslaw Cygler for the gift of the pFO4 vector (BRI, Montreal, Canada). We are indebted to the staff of the European Synchrotron Radiation Facility (ESRF, Grenoble, France) beamline ID14-EH2 for technical support during data collection and processing.

References

Alldredge, A. L. (2000). *Limnol. Oceanogr.* **45**, 1245–1253.
Aquino, R. S., Landeira-Fernandez, A. M., Valente, A. P., Andrade, L. R. & Mourao, P. A. (2005). *Glycobiology*, **15**, 11–20.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). *J. Mol. Biol.* **340**, 783–795.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Coutinho, P. M. & Henrissat, B. (1999). *Recent Advances in Carbohydrate Bioengineering*, edited by H. J. Gilbert, G. Davies, B. Henrissat & B. Svensson, pp. 3–12. Cambridge: The Royal Society of Chemistry.
Czerwinski, E. W., Midoro-Horiuti, T., White, M. A., Brooks, E. G. & Goldblum, R. M. (2005). *J. Biol. Chem.* **280**, 3740–3746.
DeLong, E. F., Franks, D. G. & Alldredge, A. L. (1993). *Limnol. Oceanogr.* **38**, 924–934.
Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R. & Reinhardt, R. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 8298–8303.
Ho, M. C., Whitehead, M. P., Cleveland, T. E. & Dean, R. A. (1995). *Curr. Genet.* **27**, 142–149.
Kloareg, B. & Quatrano, R. (1988). *Oceanogr. Mar. Biol. Annu. Rev.* **26**, 259–315.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
Lietzke, S. E., Yoder, M. D., Keen, N. T. & Jurnak, F. (1994). *Plant Physiol.* **106**, 849–862.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Mayans, O., Scott, M., Connerton, I., Gravesen, T., Benen, J., Visser, J., Pickersgill, R. & Jenkins, J. (1997). *Structure*, **5**, 677–689.
Michel, G., Nyval-Collen, P., Barbeyron, T., Czjzek, M. & Helbert, W. (2006). *Appl. Microbiol. Biotechnol.* **71**, 23–33.
Miskin, I. P., Farrimond, P. & Head, I. M. (1999). *Microbiology*, **145**, 1977–1987.
Mohnen, D. (1999). *Comprehensive Natural Products Chemistry*, edited by D. Baron & K. Nakanishi, Vol. 3, pp. 497–527. Amsterdam: Elsevier.
Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.
Neef, A., Amann, R., Schlesner, H. & Schleifer, K. H. (1998). *Microbiology*, **144**, 3257–3266.
O'Neill, M. A., Albersheim, P. & Darvill, A. G. (1990). *Methods in Plant Biochemistry*, edited by P. M. Dey, Vol. 2, pp. 415–441. London: Academic Press.
Ovodov, Y. S., Ovodova, R. G., Shibaeva, V. I. & Mikheyskaya, L. V. (1975). *Carbohydr. Res.* 42, 197–199.
Pilnik, W. & Rombouts, F. M. (1981). *Enzymes and Food Processing*, edited by G. G. Birch, N. Blakerough & K. J. Parker, pp. 105–128. London: Applied Science Publishers.
Rafnar, T., Griffith, I. J., Kuo, M. C., Bond, J. F., Rogers, B. L. & Klapper, D. G. (1991). *J. Biol. Chem.* **266**, 1229–1236.
Scavetta, R. D., Herron, S. R., Hotchkiss, A. T., Kita, N., Keen, N. T., Benen, J. A., Kester, H. C., Visser, J. & Jurnak, F. (1999). *Plant Cell*, **11**, 1081–1092.
Studier, F. W. (2005). *Protein Expr. Purif.* **41**, 207–234.
Wang, J., Jenkins, C., Webb, R. I. & Fuerst, J. A. (2002). *Appl. Environ. Microbiol.* **68**, 417–422.

Bibliographie

Bibliographie

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF*, et al.* (2000). "The genome sequence of Drosophila melanogaster." *Science* **287**(5461): 2185-2195.

Albersheim, P., Darvill, A. G., O'Neill, M. A., Schols, H. A. and Voragen, A. G. J. (1996). An hypothesis: The same six polysaccharides are components of the primary cell walls of all higher plants. *Pectins and Pectinases*. A. G. J. Voragen. Amsterdam, Elsevier Science BV**:** 47–55.

Alldredge, A. L. (2000). "Interstitial dissolved organic carbon (DOC) concentrations within sinking marine aggregates and their potential contribution to carbon flux." *Limnology and Oceanography* **45**: 1245-1253.

Allouch, J., Helbert, W., Henrissat, B. and Czjzek, M. (2004). "Parallel substrate binding sites in a beta-agarase suggest a novel mode of action on double-helical agarose." *Structure* **12**: 623-632.

Allouch, J., Jam, M., Helbert, W., Barbeyron, T., Kloareg, B., Henrissat, B. and Czjzek, M. (2003). "The three-dimensional structures of two beta-agarases." *Journal of Biological Chemistry* **278**: 47171-47180.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). "Basic local alignment search tool." *Journal of Molecular Biology* **215**: 403-410.

Amann, R., Binder, B., Olson, R., Chisholm, S., Devereux, R. and Stahl, D. (1990). "Combination of 16S rRNA targeted oligonucleotide probes with flow-cytometry for analyzing mixed microbial populations." *Applied and Environmental Microbiology* **56**: 1919-1925.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." *Nucleic Acids Research* **32**(Database issue): 226-229.

Angata, T. and Varki, A. (2002). "Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective." *Chem Rev* **102**(2): 439-69.

Armisen, R. (1991). "Agar and agarose biotechnological applications." *Hydrobiologia* **221**: 157-166.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.*, et al.* (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* **25**(1): 25-9.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P.*, et al.* (2003). "PRINTS and its automatic supplement, prePRINTS." *Nucleic Acids Res* **31**(1): 400-2.

Avery, O. T., MacLeod, C. M. and McCarty, M. (1995). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944." *Mol Med* **1**(4): 344-65.

Bae, B., Ohene-Adjei, S., Kocherginskaya, S., Mackie, R. I., Spies, M. A., Cann, I. K. and Nair, S. K. (2008). "Molecular basis for the selectivity and specificity of ligand recognition by the family 16 carbohydrate-binding modules from Thermoanaerobacterium polysaccharolyticum ManA." *J Biol Chem* **283**(18): 12415-25.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004). "Swiss-Prot: juggling between evolution and stability." *Brief Bioinform* **5**(1): 39-55.

Ball, S. G. and Morell, M. K. (2003). "From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule." *Annu Rev Plant Biol* **54**: 207-33.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., Bedell, J. A., McPherson, J. D. and Johnson, S. L. (2000). "The syntenic relationship of the zebrafish and human genomes." *Genome Res* **10**(9): 1351-8.

Barbeyron, T., Michel, G., Potin, P., Henrissat, B. and Kloareg, B. (2000). "iota-Carrageenases constitute a novel family of glycoside hydrolases, unrelated to that of kappa-carrageenases." *Journal of Biological Chemistry* **275**(45): 35499-35505.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I. and Schomburg, D. (2007). "BRENDA, AMENDA and FRENDA: the enzyme information system in 2007." *Nucleic Acids Research* **35**: D511-514.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L.*, et al.* (2004). "The Pfam protein families database." *Nucleic Acids Res.* **32**(Database issue): 138-141.

Baumann, M. J., Eklof, J. M., Michel, G., Kallas, A. M., Teeri, T. T., Czjzek, M. and Brumer, H., 3rd (2007). "Structural evidence for the evolution of xyloglucanase activity from xyloglucan endo-transglycosylases: biological implications for cell wall metabolism." *Plant Cell* **19**(6): 1947-1963.

Belshaw, R. and Bensasson, D. (2006). "The rise and falls of introns." *Heredity* **96**(3): 208-13.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2008). "GenBank." *Nucleic Acids Research* **36**: D25-30.

Berman, H. M., Henrick, K. and Nakamura, H. (2003). "Announcing the worldwide Protein Data Bank." *Nature Structural Biology* **10**(12): 980.

Bishop, P. D., Pearce, G., Bryant, J. E. and Ryan, C. A. (1984). "Isolation and characterization of the proteinase inhibitor-inducing factor from tomato leaves. Identity and activity of poly- and oligogalacturonide fragments." *J Biol Chem* **259**(21): 13172-7.

Blow, N. (2008). "Structural genomics: inside a protein structure initiative center." *Nature Methods* **5**: 203-207.

Bochkarev, A. and Tempel, W. (2008). "High throughput crystallography at SGC Toronto: an overview." *Methods Mol Biol* **426**: 515-21.

Boneca, I. G., de Reuse, H., Epinat, J. C., Pupin, M., Labigne, A. and Moszer, I. (2003). "A revised annotation and comparative analysis of Helicobacter pylori genomes." *Nucleic Acids Res* **31**(6): 1704-14.

Boos, W. and Shuman, H. (1998). "Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation." *Microbiol Mol Biol Rev* **62**(1): 204-29.

Brenner, S. E. (1999). "Errors in genome annotation." *Trends Genet* **15**(4): 132-3.

Brent, M. R. (2005). "Genome annotation past, present, and future: how to define an ORF at each locus." *Genome Res* **15**(12): 1777-86.

Brent, M. R. (2008). "Steady progress and recent breakthroughs in the accuracy of automated genome annotation." *Nat Rev Genet* **9**(1): 62-73.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D.*, et al.* (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." *Science* **273**(5278): 1058-1073.

Burmeister, W. P., Henrissat, B., Bosso, C., Cusack, S. and Ruigrok, R. W. (1993). "Influenza B virus neuraminidase can synthesize its own inhibitor." *Structure* **1**(1): 19-26.

Busso, D., Thierry, J. C. and Moras, D. (2008). "The structural biology and genomics platform in strasbourg: an overview." *Methods Mol Biol* **426**: 523-36.

Buttcher, V., Welsh, T., Willmitzer, L. and Kossmann, J. (1997). "Cloning and characterization of the gene for amylosucrase from Neisseria polysaccharea: production of a linear alpha-1,4-glucan." *J Bacteriol* **179**(10): 3324-30.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B. and Mornon, J. P. (1997). "Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives." *Cell Mol Life Sci* **53**(8): 621-45.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A.*, et al.* (2003). "The Gene Ontology Annotation (GOA)

project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro." *Genome Res* **13**(4): 662-72.
Campetella, O. E., Uttaro, A. D., Parodi, A. J. and Frasch, A. C. (1994). "A recombinant Trypanosoma cruzi trans-sialidase lacking the amino acid repeats retains the enzymatic activity." *Mol Biochem Parasitol* **64**(2): 337-40.
Carpita, N. and McCann, M. (2000). The Cell Wall. *Biochemistry & Molecular Biologie of Plants*. R. Jones, American Society of Plant Physiologists**:** 52-108.
Carrette, O., Burkhard, P. R., Sanchez, J. C. and Hochstrasser, D. F. (2006). "State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research." *Nature Protocols* **1**: 812-823.
Check, E. (2002). "Venter aims for maximum impact with minimal genome." *Nature* **420**(6914): 350.
Chothia, C. and Murzin, A. G. (1993). "New folds for all-beta proteins." *Structure* **1**(4): 217-22.
Christov, L. P. and Prior, B. A. (1993). "Esterases of xylan-degrading microorganisms: production, properties, and significance." *Enzyme Microb Technol* **15**(6): 460-75.
Chua, T. K., Bujnicki, J. M., Tan, T. C., Huynh, F., Patel, B. K. and Sivaraman, J. (2008). "The structure of sucrose phosphate synthase from Halothermothrix orenii reveals its mechanism of action and binding mode." *Plant Cell* **20**(4): 1059-72.
Cohen, F. E. (1993). "The parallel beta helix of pectate lyase C: something to sneeze at." *Science* **260**(5113): 1444-5.
Collaborative Computational Project Number 4 (1994). "The CCP4 suite: programs for protein crystallography." *Acta Crystallographica. Section D, Biological Crystallography* **50**(Pt 5): 760-763.
Collmer, A., Ried, J. L. and Mount, M. S. (1988). "Assay Methods for Pectic Enzymes." *Methods In Enzymology* **161**: 329-335.
Colquhoun, I. J., de Ruiter, G. A., Schols, H. A. and Voragen, A. G. (1990). "Identification by n.m.r. spectroscopy of oligosaccharides obtained by treatment of the hairy regions of apple pectin with rhamnogalacturonase." *Carbohydr Res* **206**(1): 131-44.
Comfort, D. A., Chou, C. J., Conners, S. B., VanFossen, A. L. and Kelly, R. M. (2008). "Functional-genomics-based identification and characterization of open reading frames encoding alpha-glucoside-processing enzymes in the hyperthermophilic archaeon Pyrococcus furiosus." *Appl Environ Microbiol* **74**(4): 1281-3.
Consortium, A. G. I. (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." *Nature* **408**(6814): 796-815.
Consortium, C. e. S. (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." *Science* **282**(5396): 2012-2018.
Consortium, H. g. P. (2004). "Finishing the euchromatic sequence of the human genome." *Nature* **431**(7011): 931-45.
Consortium, T. U. (2008). "The Universal Protein Resource (UniProt) 2009." *Nucleic Acids Res*.
Copley, R. R., Russell, R. B. and Ponting, C. P. (2001). "Sialidase-like Asp-boxes: sequence-similar structures within different protein folds." *Protein Sci* **10**(2): 285-92.
Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." *Nucleic Acids Research* **16**: 10881-10890.
Coutinho, P. M. and Henrissat, B. (1999). Carbohydrate-active enzymes: an integrated database approach. *Recent Advances in Carbohydrate Bioengineering*. B. Svensson. Cambridge, The Royal Society of Chemistry**:** 3-12.
Crennell, S., Garman, E., Laver, G., Vimr, E. and Taylor, G. (1994). "Crystal structure of Vibrio cholerae neuraminidase reveals dual lectin-like domains in addition to the catalytic domain." *Structure* **2**(6): 535-44.
Crennell, S. J., Garman, E. F., Philippon, C., Vasella, A., Laver, W. G., Vimr, E. R. and Taylor, G. L. (1996). "The structures of Salmonella typhimurium LT2 neuraminidase and its complexes with three inhibitors at high resolution." *J Mol Biol* **259**(2): 264-80.

Czerwinski, E. W., Midoro-Horiuti, T., White, M. A., Brooks, E. G. and Goldblum, R. M. (2005). "Crystal structure of Jun a 1, the major cedar pollen allergen from *Juniperus ashei*, reveals a parallel beta-helical core." *Journal of Biological Chemistry* **280**(5): 3740-3746.

Daas, P. J., Arisz, P. W., Schols, H. A., De Ruiter, G. A. and Voragen, A. G. (1998). "Analysis of partially methyl-esterified galacturonic acid oligomers by high-performance anion-exchange chromatography and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." *Anal Biochem* **257**(2): 195-202.

Dahm, R. (2008). "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research." *Hum Genet* **122**(6): 565-81.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998). "Conservation of gene order: a fingerprint of proteins that physically interact." *Trends Biochem Sci* **23**(9): 324-8.

Davies, G. J. and Henrissat, B. (1995). "Structures and mechanisms of glycosyl hydrolases." *Structure* **3**(9): 853-859.

Dayhoff, M. O., Schwartz, R. M., Orcutt, B. C. and In Dayhoff, M. O. (1978). Atlas of protein sequence and structure. Washington DC.

de Reviers, B. (2002). Biologie et Phylogénie des Algues. Paris, Belin.

de Souza, N. (2007). "From structure to function." *Nature Methods* **4**: 771.

Devos, D. and Valencia, A. (2001). "Intrinsic errors in genome annotation." *Trends Genet* **17**(8): 429-31.

Dickmanns, A., Ballschmiter, M., Liebl, W. and Ficner, R. (2006). "Structure of the novel alpha-amylase AmyC from Thermotoga maritima." *Acta Crystallogr D Biol Crystallogr* **62**(Pt 3): 262-70.

Dinu, D., Nechifor, M. T., Stoian, G., Costache, M. and Dinischiotu, A. (2007). "Enzymes with new biochemical properties in the pectinolytic complex produced by Aspergillus niger MIUG 16." *J Biotechnol* **131**(2): 128-37.

Discala, C., Benigni, X., Barillot, E. and Vaysseix, G. (2000). "DBcat: a catalog of 500 biological databases." *Nucleic Acids Research* **28**: 8-9.

Doublié, S. (1997). "Preparation of selenomethionyl proteins for phase determination." *Methods Enzymol* **276**: 523-30.

Düsterhöft EM, L. V., Voragen A, Beldman G (1997). "Purification, characterization, and properties of two xylanases from Humicola insolens." *Enzyme and Microbial Technology* **20**(6): 437-445.

Editorial (2008). "Structural genomics in the spotlight." *Nature Methods* **5**: 115.

Edwards, U. and Frosch, M. (1992). "Sequence and functional analysis of the cloned Neisseria meningitidis CMP-NeuNAc synthetase." *FEMS Microbiol Lett* **75**(2-3): 161-6.

Eisenstein, M. (2007). "The shape of things." *Nature Methods* **4**: 95-102.

Etzold, T., Ulyanov, A. and Argos, P. (1996). "SRS: information retrieval system for molecular biology data banks." *Methods in Enzymology* **266**: 114-128.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, J. S., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L.*, et al.* (2008). "The Pfam protein families database." *Nucleic Acids Research* **36**: D281-288.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." *Trends in Genetics* **16**: 227-231.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A. and Merrick, J. M. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." *Science* **269**(5223): 496-512.

Fowler, S. W. and Knauer, G. A. (1986). "Role of large particles in the transport of elements and organic compounds through the oceanic water column." *Progress In Oceanography* **16**(3): 147-194.

Fox, B. G., Goulding, C., Malkowski, M. G., Stewart, L. and Deacon, A. (2008). "Structural genomics: from genes to structures with valuable materials and many questions in between." *Nature Methods* **5**: 129-132.

Fry, S. C. (1988). Wall polymers : chemical characterization. *The growing plant cell wall*. H. L. S. Technical**:** pp. 103-185.

Fuerst, J. A., Gwilliam, H. G., Lindsay, M., Lichanska, A., Belcher, C., Vickers, J. E. and Hugenholtz, P. (1997). "Isolation and molecular identification of planctomycete bacteria from postlarvae of the giant tiger prawn, Penaeus monodon." *Appl Environ Microbiol* **63**(1): 254-62.

Fukusumi, S., Kamizono, A., Horinouchi, S. and Beppu, T. (1988). "Cloning and nucleotide sequence of a heat-stable amylase gene from an anaerobic thermophile, Dictyoglomus thermophilum." *Eur J Biochem* **174**(1): 15-21.

Fulton, D. L., Li, Y. Y., Laird, M. R., Horsman, B. G., Roche, F. M. and Brinkman, F. S. (2006). "Improving the specificity of high-throughput ortholog prediction." *BMC Bioinformatics* **7**: 270.

Funderburgh, J. L. (2000). "Keratan sulfate: structure, biosynthesis, and function." *Glycobiology* **10**(10): 951-8.

Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J. P. (1987). "Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences." *FEBS Letters* **224**(1): 149-155.

Gacesa, P. (1987). "Alginate-modifying enzymes. A proposed unified mechanism of action for the lyases and epimerases." *FEBS Letters* **212**: 199–202.

Gade, D., Gobom, J. and Rabus, R. (2005a). "Proteomic analysis of carbohydrate catabolism and regulation in the marine bacterium Rhodopirellula baltica." *Proteomics* **5**(14): 3672-83.

Gade, D., Theiss, D., Lange, D., Mirgorodskaya, E., Lombardot, T., Glockner, F. O., Kube, M., Reinhardt, R., Amann, R., Lehrach, H.*, et al.* (2005b). "Towards the proteome of the marine bacterium Rhodopirellula baltica: mapping the soluble proteins." *Proteomics* **5**(14): 3654-71.

Galperin, M. Y. and Kolker, E. (2006). "New metrics for comparative genomics." *Current Opinion in Biotechnology* **17**(5): 440-447.

Galperin, M. Y. and Koonin, E. V. (1998). "Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption." *In Silico Biol* **1**(1): 55-67.

Galperin, M. Y. and Koonin, E. V. (1999). "Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes." *Genetica* **106**: 159-170.

Gaskell, A., Crennell, S. and Taylor, G. (1995). "The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll." *Structure* **3**(11): 1197-205.

Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C. J., Lachaize, C.*, et al.* (2003). "Automated annotation of microbial proteomes in SWISS-PROT." *Comput Biol Chem* **27**(1): 49-58.

Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., Stockwell, T. B., Brownley, A., Thomas, D. W. A., M.A., Merryman, C.*, et al.* (2008). "Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome." *Science* **319**(5867): 1215-1220.

Gill, S. C. and von Hippel, P. H. (1989). "Calculation of protein extinction coefficients from amino acid sequence data." *Anal Biochem* **182**(2): 319-26.

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K.*, et al.* (2003). "Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1." *Proceedings of the National Academy of Sciences of the United States of America* **100**(14): 8298-8303.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M.*, et al.* (1996). "Life with 6000 genes." *Science* **274**(5287): 546,563-567.

Gomase, V. S. and Tagore, S. (2008). "Transcriptomics." *Current Drug Metabolism* **9**(3): 245-249.

Graham, D. E., Kyrpides, N., Anderson, I. J., Overbeek, R. and Whitman, W. B. (2001). "Genome of Methanocaldococcus (Methanococcus) jannaschii." *Methods Enzymol* **330**: 40-123.

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A.*, et al.* (2007). "The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution." *Nucleic Acids Research* **35**: D291-297.

Gross, R., Hauer, B., Otto, K. and Schmid, A. (2007). "Microbial biofilms: new catalysts for maximizing productivity of long-term biotransformations." *Biotechnol Bioeng* **98**(6): 1123-34.

Guerin, M. E., Kordulakova, J., Schaeffer, F., Svetlikova, Z., Buschiazzo, A., Giganti, D., Gicquel, B., Mikusova, K., Jackson, M. and Alzari, P. M. (2007). "Molecular recognition and interfacial catalysis by the essential phosphatidylinositol mannosyltransferase PimA from mycobacteria." *J Biol Chem* **282**(28): 20705-14.

Hahn, M. G., Darvill, A. G. and Albersheim, P. (1981). "Host-Pathogen Interactions : XIX. THE ENDOGENOUS ELICITOR, A FRAGMENT OF A PLANT CELL WALL POLYSACCHARIDE THAT ELICITS PHYTOALEXIN ACCUMULATION IN SOYBEANS." *Plant Physiology* **68**(5): 1161-1169.

Hanahan, D. (1983). "Studies on transformation of Escherichia coli with plasmids." *J Mol Biol* **166**(4): 557-80.

Haquin, S., Oeuillet, E., Pajon, A., Harris, M., Jones, A. T., van Tilbeurgh, H., Markley, J. L., Zolnai, Z. and Poupon, A. (2008). "Data management in structural genomics: an overview." *Methods Mol Biol* **426**: 49-79.

Harrison, D., Hussain, S. A., Combs, A. C., Ervasti, J. M., Yurchenco, P. D. and Hohenester, E. (2007). "Crystal structure and cell surface anchorage sites of laminin alpha1LG4-5." *J Biol Chem* **282**(15): 11573-81.

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

Henrissat, B. (1998). "Glycosidase families." *Biochem Soc Trans* **26**(2): 153-6.

Henrissat, B. and Davies, G. J. (2000). "Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics." *Plant Physiology* **124**(4): 1515-1519.

Henrissat, B., Heffron, S. E., Yoder, M. D., Lietzke, S. E. and Jurnak, F. (1995). "Functional implications of structure-based sequence alignment of proteins in the extracellular pectate lyase superfamily." *Plant Physiol* **107**(3): 963-76.

Hertz-Fowler, C., Peacock, C. S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K.*, et al.* (2004). "GeneDB: a resource for prokaryotic and eukaryotic organisms." *Nucleic Acids Res* **32**(Database issue): D339-43.

Hieu, C. X., Voigt, B., Albrecht, D., Becher, D., Lombardot, T., Glockner, F. O., Amann, R., Hecker, M. and Schweder, T. (2008). "Detailed proteome analysis of growing cells of the planctomycete Rhodopirellula baltica SH1T." *Proteomics* **8**(8): 1608-23.

Hueber, S. D. and Lohmann, I. (2008). "Shaping segments: Hox gene function in the genomic age." *Bioessays* **30**(10): 965-79.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S. and Sigrist, C. J. (2008). "The 20 years of PROSITE." *Nucleic Acids Res* **36**(Database issue): D245-249.

Idaka, M., Terada, T., Murayama, K., Yamaguchi, H., Nureki, O., Ishitani, R., Kuramitsu, S., Shirouzu, M., Yokoyama, S., RIKEN Structural Genomics/Proteomics Initiative

(RSGI) (2003). "Crystal structure of TT1467 from Thermus thermophilus HB8." *To Be Published*.

Iliopoulos, I., Tsoka, S., Andrade, M. A., Enright, A. J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C.*, et al.* (2003). "Evaluation of annotation strategies using an entire genome sequence." *Bioinformatics* **19**(6): 717-26.

Imamura, H., Fushinobu, S., Yamamoto, M., Kumasaka, T., Jeon, B. S., Wakagi, T. and Matsuzawa, H. (2003). "Crystal structures of 4-alpha-glucanotransferase from Thermococcus litoralis and its complex with an inhibitor." *J Biol Chem* **278**(21): 19378-86.

James, S. L., Muir, J. G., Curtis, S. L. and Gibson, P. R. (2003). "Dietary fibre: a roughage guide." *Intern Med J* **33**(7): 291-6.

Janecek, S. (2002). "A motif of a microbial starch-binding domain found in human genethonin." *Bioinformatics* **18**(11): 1534-7.

Jenkins, J., Mayans, O. and Pickersgill, R. (1998). "Structure and evolution of parallel beta-helix proteins." *J Struct Biol* **122**(1-2): 236-46.

Jin, D. F. and West, C. A. (1984). "Characteristics of Galacturonic Acid Oligomers as Elicitors of Casbene Synthetase Activity in Castor Bean Seedlings." *Plant Physiol* **74**(4): 989-992.

Jones, C. E., Brown, A. L. and Baumann, U. (2007). "Estimating the annotation error rate of curated GO database sequence annotations." *BMC Bioinformatics* **8**: 170.

Juncosa, M., Pons, J., Dot, T., Querol, E. and Planas, A. (1994). "Identification of active site carboxylic residues in Bacillus licheniformis 1,3-1,4-beta-D-glucan 4-glucanohydrolase by site-directed mutagenesis." *J Biol Chem* **269**(20): 14530-5.

Kalnins, A., Otto, K., Ruther, U. and Muller-Hill, B. (1983). "Sequence of the lacZ gene of Escherichia coli." *Embo J* **2**(4): 593-7.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T.*, et al.* (2008). "KEGG for linking genomes to life and the environment." *Nucleic Acids Research* **36**: D480-484.

Kashyap, D. R., Vohra, P. K., Chopra, S. and Tewari, R. (2001). "Applications of pectinases in the commercial sector: a review." *Bioresour Technol* **77**(3): 215-27.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic Acids Research* **30**(14): 3059-3066.

Keitel, T., Simon, O., Borriss, R. and Heinemann, U. (1993). "Molecular and active-site structure of a *Bacillus* 1,3-1,4-beta-glucanase." *Proceedings of the National Academy of Sciences of the United States of America* **90**(11): 5287-5291.

Kertesz, M. A. (2000). "Riding the sulfur cycle--metabolism of sulfonates and sulfate esters in gram-negative bacteria." *FEMS Microbiology Reviews* **24**(2): 135-175.

Kertesz, Z. I. (1951). The pectic substances. New York-London.

Khoroshko, L. O., Petrova, V. N., Takhistov, V. V., Viktorovskii, I. V., Lahtipera, M. and Paasivirta, J. (2007). "Sulfur organic compounds in bottom sediments of the eastern Gulf of Finland." *Environ Sci Pollut Res Int* **14**(6): 366-76.

Kloareg, B. and Quatrano, R. (1988). "Structure of the cell walls of marine algae and ecophysiological functions of the matrix polysaccharides." *Oceanogr. Mar. Biol. Ann. Rev.* **26**: 259-315.

Koshland, D. E. (1953). "Stereochemistry and the mechanism of enzymatic reactions." *Biol. Rev. Camb. Phylos. Soc.* **28**: 416-436.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." *J Mol Biol* **305**(3): 567-80.

Labes, A. and Schonheit, P. (2007). "Unusual starch degradation pathway via cyclodextrins in the hyperthermophilic sulfate-reducing archaeon Archaeoglobus fulgidus strain 7324." *J Bacteriol* **189**(24): 8901-13.

Labrude, P. and Becq, C. (2003). "[Pharmacist and chemist Henri Braconnot]." *Rev Hist Pharm (Paris)* **51**(337): 61-78.

Lahaye, M. and Robic, A. (2007). "Structure and functional properties of ulvan, a polysaccharide from green seaweeds." *Biomacromolecules* **8**(6): 1765-1774.

Lau J. M., M. N. M., Darvill A. G., Albersheim P (1987). "Treatment of rhamnogalacturonan I with lithium in ethylene diamine." *Carbohydrate Research* **168**(2): 245.

Lerouge, P., Roche, P., Faucher, C., Maillet, F., Truchet, G., Prome, J. C. and Denarie, J. (1990). "Symbiotic host-specificity of *Rhizobium meliloti* is determined by a sulphated and acylated glucosamine oligosaccharide signal." *Nature* **344**(6268): 781-784.

Leslie, A. G. W. (1992). "Recent changes to the MOSFLM package for processing film and image plate data." *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

Liesack, W., König, H., Schlesner, H. and Hirsch, P. (1986). "Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the Pirella Planctomyces group." *Arch Microbiol* **145**: 361-366.

Lietzke, S. E., Yoder, M. D., Keen, N. T. and Jurnak, F. (1994). "The Three-Dimensional Structure of Pectate Lyase E, a Plant Virulence Factor from *Erwinia chrysanthemi*." *Plant Physiology* **106**(3): 849-862.

Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S., Butler, M. K., Forde, R. J. and Fuerst, J. A. (2001). "Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell." *Arch Microbiol* **175**(6): 413-29.

Llobet-Brossa, E., Rossello-Mora, R. and Amann, R. (1998). "Microbial Community Composition of Wadden Sea Sediments as Revealed by Fluorescence In Situ Hybridization." *Appl Environ Microbiol* **64**(7): 2691-6.

Maenaka, K., Kawai, G., Watanabe, K., Sunada, F. and Kumagai, I. (1994). "Functional and structural role of a tryptophan generally observed in protein-carbohydrate interaction. TRP-62 of hen egg white lysozyme." *J Biol Chem* **269**(10): 7070-5.

Malet, C., Jimenez-Barbero, J., Bernabe, M., Brosa, C. and Planas, A. (1993). "Stereochemical course and structure of the products of the enzymic action of endo-1,3-1,4-beta-D-glucan 4-glucanohydrolase from Bacillus licheniformis." *Biochem J* **296 (Pt 3)**: 753-8.

Marsden, R. L. and Orengo, C. A. (2008). "Target selection for structural genomics: an overview." *Methods Mol Biol* **426**: 3-25.

Martinez-Fleites, C., Proctor, M., Roberts, S., Bolam, D. N., Gilbert, H. J. and Davies, G. J. (2006). "Insights into the synthesis of lipopolysaccharide and antibiotics through the structures of two retaining glycosyltransferases from family GT4." *Chem Biol* **13**(11): 1143-52.

Matthews, B. W. (1968). "Solvent content of protein crystals." *Journal of Molecular Biology* **33**(2): 491-497.

Maxam, A. M. and Gilbert, W. (1977). "A new method for sequencing DNA." *Proc Natl Acad Sci U S A* **74**(2): 560-4.

May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhöh, O., Weckwerth, W.*, et al.* (2008). "Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii." *Genetivs* **179**(1): 157-166.

Mayans, O., Scott, M., Connerton, I., Gravesen, T., Benen, J., Visser, J., Pickersgill, R. and Jenkins, J. (1997). "Two crystal structures of pectin lyase A from *Aspergillus* reveal a pH driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases." *Structure* **5**(5): 677-689.

McNeil, M., Darvill, A. G., Fry, S. C. and Albersheim, P. (1984). "Structure and function of the primary cell walls of plants." *Annu Rev Biochem* **53**: 625-63.

Médigue, C. and Moszer, I. (2007). "Annotation, comparison and databases for hundreds of bacterial genomes." *Research in Microbiology* **158**(10): 724-736.

Michel, G., Chantalat, L., Duee, E., Barbeyron, T., Henrissat, B., Kloareg, B. and Dideberg, O. (2001). "The kappa-carrageenase of *P. carrageenovora* features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases." *Structure* **9**(6): 513-525.

Mizoguchi, H., Mori, H. and Fujio, T. (2007). "Escherichia coli minimum genome factory." *Biotechnology and Applied Biochemistry* **46**: 157-167.
Mohnen, D. (1999). Biosynthesis of pectins and galactomannans. *Comprehensive Natural Products Chemistry*. K. Nakanishi. Amsterdam, Elsevier Science. **3:** 497–527.
Mohnen, D. (2008). "Pectin structure and biosynthesis." *Curr Opin Plant Biol* **11**(3): 266-77.
Moreno-Hagelsieb, G. and Latimer, K. (2008). "Choosing BLAST options for better detection of orthologs as reciprocal best hits." *Bioinformatics* **24**(3): 319-24.
Mulder, N. J. and Apweiler, R. (2008). "The InterPro database and tools for protein domain analysis." *Current Protocols in Bioinformatics* **Chapter 2**: Unit 2.7.
Myers, R. W., Lee, R. T., Lee, Y. C., Thomas, G. H., Reynolds, L. W. and Uchida, Y. (1980). "The synthesis of 4-methylumbelliferyl alpha-ketoside of N-acetylneuraminic acid and its use in a fluorometric assay for neuraminidase." *Anal Biochem* **101**(1): 166-74.
Navaza, J. (2001). "Implementation of molecular replacement in AMoRe." *Acta Crystallographica. Section D, Biological Crystallography* **57**(Pt 10): 1367-1372.
Neef, A., Amann, R., Schlesner, H. and Schleifer, K. H. (1998). "Monitoring a widespread bacterial group: in situ detection of *planctomycetes* with 16S rRNA-targeted probes." *Microbiology* **144**: 3257-3266.
Newstead, S. L., Potter, J. A., Wilson, J. C., Xu, G., Chien, C. H., Watts, A. G., Withers, S. G. and Taylor, G. L. (2008). "The structure of Clostridium perfringens NanI sialidase and its catalytic intermediates." *J Biol Chem* **283**(14): 9080-8.
Nicholls, H. (2007). "Sorcerer II: the search for microbial diversity roils the waters." *PLoS Biol* **5**(3): e74.
Nielsen, H., Brunak, S. and von Heijne, G. (1999). "Machine learning approaches for the prediction of signal peptides and other protein sorting signals." *Protein Engineering* **12**(1): 3-9.
Nothnagel, E. A., McNeil, M., Albersheim, P. and Dell, A. (1983). "Host-Pathogen Interactions : XXII. A Galacturonic Acid Oligosaccharide from Plant Cell Walls Elicits Phytoalexins." *Plant Physiol* **71**(4): 916-926.
Nyvall, P., Corre, E., Boisset, C., Barbeyron, T., Rousvoal, S., Scornet, D., Kloareg, B. and Boyen, C. (2003). "Characterization of mannuronan C-5-epimerase genes from the brown alga Laminaria digitata." *Plant Physiol* **133**(2): 726-35.
Ohno, S., Wolf, U. and Atkin, N. B. (1968). "Evolution from fish to mammals by gene duplication." *Hereditas* **59**: 169-187.
O'Neill, M. A., Ishii, T., Albersheim, P. and Darvill, A. G. (2004). "Rhamnogalacturonan II: structure and function of a borate cross-linked cell wall pectic polysaccharide." *Annu Rev Plant Biol* **55**: 109-39.
Osawa, T., Matsubara, Y., Muramatsu, T., Kimura, M. and Kakuta, Y. (2005). "Crystal structure of the alginate (poly alpha-l-guluronate) lyase from Corynebacterium sp. at 1.2 A resolution." *J Mol Biol* **345**(5): 1111-8.
O'Toole, N., Grabowski, M., Otwinowski, Z., Minor, W. and Cygler, M. (2004). "The Structural Genomics Experimental Pipeline: Insights From Global Target Lists." *Proteins - Structure Function and Bioinformatics* **56**: 201-210.
Pandey, A. and Mann, M. (2000). "Proteomics to study genes and genomes." *Nature* **405**: 837-846.
Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." *Methods in Enzymology* **183**: 63-98.
Pedros-Alio, C. (2006). "Genomics and marine microbial ecology." *Int Microbiol* **9**(3): 191-7.
Pernthaler, J. and Amann, R. (2005). "Fate of heterotrophic microbes in pelagic habitats: focus on populations." *Microbiol Mol Biol Rev* **69**(3): 440-61.
Piatigorsky, J. (2003). "Crystallin genes: specialization by changes in gene regulation may precede gene duplication." *J Struct Funct Genomics* **3**(1-4): 131-7.
Pickersgill, R., Jenkins, J., Harris, G., Nasser, W. and Robert-Baudouy, J. (1994). "The structure of Bacillus subtilis pectate lyase in complex with calcium." *Nat Struct Biol* **1**(10): 717-23.

Pitson SM, V. A., Beldman G (1996). "Stereochemical course of hydrolysis catalyzed by arabinofuranosyl hydrolases." *FEBS Letters* **398**(1): 7-11.
Potier, M., Mameli, L., Belisle, M., Dallaire, L. and Melancon, S. B. (1979). "Fluorometric assay of neuraminidase with a sodium (4-methylumbelliferyl-alpha-D-N-acetylneuraminate) substrate." *Anal Biochem* **94**(2): 287-96.
Rabus, R., Gade, D., Helbig, R., Bauer, M., Glockner, F. O., Kube, M., Schlesner, H., Reinhardt, R. and Amann, R. (2002). "Analysis of N-acetylglucosamine metabolism in the marine bacterium Pirellula sp. strain 1 by a proteomic approach." *Proteomics* **2**(6): 649-55.
Raetz, C. R. and Roderick, S. L. (1995). "A left-handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase." *Science* **270**(5238): 997-1000.
Read, R. J. (2001). "Pushing the boundaries of molecular replacement with maximum likelihood." *Acta Crystallogr D Biol Crystallogr* **57**(Pt 10): 1373-82.
Ridley, B. L., O'Neill, M. A. and Mohnen, D. (2001). "Pectins: structure, biosynthesis, and oligogalacturonide-related signaling." *Phytochemistry* **57**(6): 929-67.
Ried, J. L. and Collmer, A. (1986). "Comparison of pectic enzymes produced by Erwinia chrysanthemi, Erwinia carotovora subsp. carotovora, and Erwinia carotovora subsp. atroseptica." *Appl Environ Microbiol* **52**(2): 305-10.
Riemann, L., Leitet, C., Pommier, T., Simu, K., Holmfeldt, K., Larsson, U. and Hagstrom, A. (2008). "The native bacterioplankton community in the central baltic sea is influenced by freshwater bacterial species." *Appl Environ Microbiol* **74**(2): 503-15.
Robertsen, B. B. (1986). "Elicitors of the production of lignin-likecompounds in cucumberhypocotyls." *Physiological and molecular plant pathology* **28**(1): 137-148.
Rodríguez-Valera, F. (2004). "Environmental genomics, the big picture?" *FEMS Microbiol Lett* **231**(2): 153-158.
Roggentin, P., Rothe, B., Kaper, J. B., Galen, J., Lawrisuk, L., Vimr, E. R. and Schauer, R. (1989). "Conserved sequences in bacterial and viral sialidases." *Glycoconj J* **6**(3): 349-53.
Rossmann, M. G. and Blow, D. M. (1962). "Detection of Sub-Units within Crystallographic Asymmetric Unit." *Acta Crystallographica* **15**(Jan 10): 24-31.
Rost, B. (1999). "Twilight zone of protein sequence alignments." *Protein Eng* **12**(2): 85-94.
Ruane, K. M., Davies, G. J. and Martinez-Fleites, C. (2008). "Crystal structure of a family GT4 glycosyltransferase from Bacillus anthracis ORF BA1558." *Proteins* **73**(3): 784-787.
Russel, M. (1998). "Macromolecular assembly and secretion across the bacterial cell envelope: type II protein secretion systems." *J Mol Biol* **279**(3): 485-99.
Sanchez-Torres, P., Visser, J. and Benen, J. A. (2003). "Identification of amino acid residues critical for catalysis and stability in Aspergillus niger family 1 pectin lyase A." *Biochem J* **370**(Pt 1): 331-7.
Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., 3rd, Slocombe, P. M. and Smith, M. (1978). "The nucleotide sequence of bacteriophage phiX174." *J Mol Biol* **125**(2): 225-46.
Sanger, F., Nicklen, S. and Coulson, A. R. (1977). "DNA sequencing with chain-terminating inhibitors." *Proc Natl Acad Sci U S A* **74**(12): 5463-7.
Sanger, F., Thompson, E. O. and Kitai, R. (1955). "The amide groups of insulin." *Biochem J* **59**(3): 509-18.
Scavetta, R. D., Herron, S. R., Hotchkiss, A. T., Kita, N., Keen, N. T., Benen, J. A., Kester, H. C., Visser, J. and Jurnak, F. (1999). "Structure of a plant cell wall fragment complexed to pectate lyase C." *Plant Cell* **11**(6): 1081-1092.
Scheller, H., Jensen, J., Sorensen, S., Harolt, J. and Geshi, N. (2007). "Biosynthesis of pectin." *Physiologia Plantarum* **129**: 283–295.
Schinzel, R. and Nidetzky, B. (1999). "Bacterial alpha-glucan phosphorylases." *FEMS Microbiol Lett* **171**(2): 73-9.
Schlesner, H., Rensmann, C., Tindall, B. J., Gade, D., Rabus, R., Pfeiffer, S. and Hirsch, P. (2004). "Taxonomic heterogeneity within the Planctomycetales as derived by DNA-

DNA hybridization, description of Rhodopirellula baltica gen. nov., sp. nov., transfer of Pirellula marina to the genus Blastopirellula gen. nov. as Blastopirellula marina comb. nov. and emended description of the genus Pirellula." *Int J Syst Evol Microbiol* **54**(Pt 5): 1567-80.

Schloss, P. D. and Handelsman, J. (2005). "Links Metagenomics for studying unculturable microorganisms: cutting the Gordian knot." *Genome Biology* **6**(8): 229.

Schmidt, T. M., DeLong, E. F. and Pace, N. R. (1991). "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing." *Journal of Bacteriology* **173**(14): 4371-4378.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D. (2002). "ProDom: automated clustering of homologous domains." *Briefings in Bioinformatics* **3**: 246-251.

Sinnott, M. L. (1990). "Catalytic mechanisms of glycosyl transfer." *Chemical Reviews* **90**: 1171-1202.

Song, L., Kalyanaraman, C., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Brown, S., Imker, H. J., Babbitt, P. C., Almo, S. C., Jacobson, M. P*., et al.* (2007). "Prediction and assignment of function for a divergent N-succinyl amino acid racemase." *Nat Chem Biol* **3**(8): 486-91.

Sonnhammer, E. L., von Heijne, G. and Krogh, A. (1998). "A hidden Markov model for predicting transmembrane helices in protein sequences." *Proc Int Conf Intell Syst Mol Biol* **6**: 175-82.

Sprang, S. R. (1993). "On a (beta-) roll." *Trends Biochem Sci* **18**(9): 313-4.

Steinbacher, S., Seckler, R., Miller, S., Steipe, B., Huber, R. and Reinemer, P. (1994). "Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer." *Science* **265**(5170): 383-6.

Sterk, P., Kulikova, T., Kersey, P. and Apweiler, R. (2007). "The EMBL Nucleotide Sequence and Genome Reviews Databases." *Methods in Molecular Biology* **406**: 1-22.

Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." *Protein Expression and Purification* **41**(1): 207-234.

Styczynski, M. P., Moxley, J. F., Tong, L. V., Walther, J. L., Jensen, K. L. and Stephanopoulos, G. N. (2007). "Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery." *Analytical Chemistry* **79**(3): 966-973.

Sugawara, H., Ogasawara, O., Okubo, K., Gojobori, T. and Tateno, Y. (2008). "DDBJ with new system and face." *Nucleic Acids Research* **36**: D22-24.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997). "A genomic perspective on protein families." *Science* **278**(5338): 631-7.

Teeling, H., Lombardot, T., Bauer, M., Ludwig, W. and Glockner, F. O. (2004). "Evaluation of the phylogenetic position of the planctomycete 'Rhodopirellula baltica' SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees." *Int J Syst Evol Microbiol* **54**(Pt 3): 791-801.

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 Research* **22**: 4673-4680.

Tipton, K. and Boyce, S. (2000). "History of the enzyme nomenclature system." *Bioinformatics* **16**: 34-40.

Tjalsma, H., Antelmann, H., Jongbloed, J. D., Braun, P. G., Darmon, E., Dorenbos, R., Dubois, J. Y., Westers, H., Zanen, G., Quax, W. J*., et al.* (2004). "Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome." *Microbiological and Molecular Biology Review* **68**(2): 207-233.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA*, et al.* (2001). "The sequence of the human genome." *Science* **291**(5507): 1304-51.

Vetting, M. W., Frantom, P. A. and Blanchard, J. S. (2008). "Structural and enzymatic analysis of MshA from Corynebacterium glutamicum: substrate-assisted catalysis." *J Biol Chem* **283**(23): 15834-44.

Vincentelli, R., Canaan, S., Offant, J., Cambillau, C. and Bignon, C. (2005). "Automated expression and solubility screening of His-tagged proteins in 96-well format." *Anal Biochem* **346**(1): 77-84.

Vollmer, W., Blanot, D. and de Pedro, M. A. (2008). "Peptidoglycan structure and architecture." *FEMS Microbiol Rev* **32**(2): 149-67.

Watson, J. D. and Crick, F. H. (1953). "The structure of DNA." *Cold Spring Harb Symp Quant Biol* **18**: 123-31.

Willats, W. G., McCartney, L., Steele-King, C. G., Marcus, S. E., Mort, A., Huisman, M., van Alebeek, G. J., Schols, H. A., Voragen, A. G., Le Goff, A.*, et al.* (2004). "A xylogalacturonan epitope is specifically associated with plant cell detachment." *Planta* **218**(4): 673-81.

Williamson, R. (1969). "Purification of DNA by isopycnic banding in cesium chloride in a zonal rotor." *Anal Biochem* **32**(1): 158-63.

Woebken, D., Teeling, H., Wecker, P., Dumitriu, A., Kostadinov, I., Delong, E. F., Amann, R. and Glockner, F. O. (2007). "Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes." *Isme J* **1**(5): 419-35.

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. and Koonin, E. V. (2001). "Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context." *Genome Res* **11**(3): 356-72.

Wren, J. D. and Bateman, A. (2008). "Databases, data tombs and dust in the wind." *Bioinformatics* **24**(19): 2127-8.

Yamamoto, T., Maruta, K., Watanabe, H., Yamashita, H., Kubota, M., Fukuda, S. and Kurimoto, M. (2001). "Trehalose-producing operon treYZ from Arthrobacter ramosus S34." *Biosci Biotechnol Biochem* **65**(6): 1419-23.

Yeates, T. O. (1997). "Detecting and overcoming crystal twinning." *Methods Enzymol* **276**: 344-58.

Yoder, M. D. and Jurnak, F. (1995). "The Refined Three-Dimensional Structure of Pectate Lyase C from Erwinia chrysanthemi at 2.2 Angstrom Resolution (Implications for an Enzymatic Mechanism)." *Plant Physiol* **107**(2): 349-364.

Yoder, M. D., Keen, N. T. and Jurnak, F. (1993a). "New domain motif: the structure of pectate lyase C, a secreted plant virulence factor." *Science* **260**(5113): 1503-1507.

Yoder, M. D., Lietzke, S. E. and Jurnak, F. (1993b). "Unusual structural features in the parallel beta-helix in pectate lyases." *Structure* **1**(4): 241-51.

Zhou, W., Forouhar, F., Conover, K., Xiao, R., Acton, T. B., Montelione, G. T., Tong, L., Hunt, J. F. and (NESG), N. S. G. C. (2006). "Crystal Structure of the Putative Mannosyl Transferase (wbaZ-1) from Archaeoglobus fulgidus."

Zona, R., Chang-Pi-Hin, F., O'Donohue, M. J. and Janecek, S. (2004). "Bioinformatics of the glycoside hydrolase family 57 and identification of catalytic residues in amylopullulanase from Thermococcus hydrothermalis." *Eur J Biochem* **271**(14): 2863-72.

Résumé

La paroi des algues marines est caractérisée par une forte abondance de polysaccharides anioniques (sulfatés ou uroniques) qui n'ont pas d'équivalents chez les plantes terrestres. Ils constituent une source de carbone cruciale pour de nombreuses bactéries marines hétérotrophes qui jouent un rôle central dans le cycle global du carbone dans les océans. Parmi ces microorganismes, la planctomycète *Rhodopirellula baltica* apparaît comme un bon modèle pour identifier les enzymes spécifiques de ces polysaccharides algaux. Son génome contient de nombreuses polysaccharidases (29 GH, 4 PL, 17 CE et 59 GT, http://www.cazy.org/) ainsi que plus de 100 sulfatases ! Dans le but de comprendre les bases moléculaires du rôle écologique de *R. baltica*, j'ai procédé à une étude à moyen-débit des polysaccharidases de cette bactérie. Aidés de divers outils bioinformatiques, j'ai établi que les 120 polysaccharidases de *R. baltica* sont constitués de 165 modules structuralement indépendants. J'ai sélectionné 96 cibles parmi ces modules, en fonction de leur intérêt structural et/ou fonctionnel, et les ait inséré dans un vecteur avec étiquette poly-His. 92/96 gènes ont été clonés avec succès dans une souche *E. coli*. La surexpression des protéines recombinantes à été réalisée à basse température (20°C) dans un milieu de culture auto-inducteur. Leur expression et solubilité ont été testées par des analyses en SDS-PAGE et Dot-Blot. Au moins 30 protéines-cibles ont été exprimées sous forme soluble. Quatre polysaccharidases ont été surexprimées en plus grand volume, purifiées, caractérisées biochimiquement et cristallographiquement. J'ai pu mettre en évidence qu'au moins deux de ces enzymes avaient été annotées incorrectement (deux glycoside hydrolases des familles GH16 et GH57), et j'ai confirmé l'activité pectate lyase de l'une d'entre elles (famille PL1), dont plusieurs jeux de données ont été collectés. Parallèlement, j'ai réalisé une étude bioinformatique de l'ensemble des polysaccharidases de *R. baltica* pour réévaluer les annotations initiales de ces enzymes et affiner le rôle de dégradeur modèle que constitue cette bactérie. Cette analyse a permis de reconstruire les grandes voies métaboliques impliquant des polysaccharides complexes et, conjointement avec mes analyses expérimentales, de préparer le terrain à des analyses plus poussée de caractérisation de ces voies métaboliques.

Mots-clés : Génomique ; Polysaccharides ; Enzymes ; *Rhodopirellula baltica* ; Planctomycete ; Ecologie marine ; Biologie moléculaire ; Biochimie ; Surexpression ; Cristallographie ; Phylogénie.

Abstract

The cell wall of marine algae is characterized by the abundance of anionic polysaccharides (sulfated or uronic) which have no equivalent in land plants. These polysaccharides constitute a crucial carbon resource for numerous heterotrophic marine bacteria, which play a central role in the global carbon cycle in the sea. Among theses microorganisms, the planctomycete *Rhodopirellula baltica* appears as a good model to identify enzymes specific for algal polysaccharides. Its genome contains numerous carbohydrate active enzymes (29 GH, 4PL, 17 CE and 59 GT, http://www.cazy.org/) and more than 100 sulfatases! In order to understand the molecular basis of the ecological role of *R. baltica*, I have started a medium-throughput project. Using various bioinformatics tools, I have established that the 120 polysaccharidases from *R. baltica* are constituted by 165 independent structural modules. I have selected 96 targets, based on their structural and/or functional significance, cloned into a His-tag vector. 92/96 targets have been successfully transformed into *E. coli* strain. Recombinant proteins have been overexpressed at low temperature (20°C) in an auto-inducer culture medium. The expression and the solubility of the proteins have been tested by SDS-PAGE and Dot-Blot analyses. At least 30 target-proteins have been expressed under a soluble form. Four enzymes have been expressed on a wider scale, then purifed, characterized biochemically and studied by cristallographie. I could give evidence that at least two of those enzymes have been incorrectly annotated (two glycoside hydrolases of the families GH16 and GH57), and I confirmed the pectate lyase activity on one of them (family PL1), on which crystallographic study have been done. I've also performed some bioinformatic analyses of all of the polysaccharide-related enzymes of *R. baltica* in order to reevaluate the initial annotations of these enzymes, and to sharpen our understanding of the ecological role of this bacterium. These analyses lead me to reconstruct some of the main metabolisms implying polymers degradation. In correlation with my experimental data, I could settle the grounds for wider experiments in the study of those metabolisms.

Key-words: Genomics ; Polysaccharides ; Enzymes ; *Rhodopirellula baltica* ; Planctomycete ; Marine ecology ; Molecular biology ; Biochemistry ; Overexpression ; Crystallography ; Phylogeny.