Université de Montréal

Caractérisation *in silico* et purification des ligases à ARN de type RtcB de *Diplonema papillatum*

par

Alexandra Léveillé-Kunst

Département de biochimie et médecine moléculaire
Faculté de médecine

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en biochimie

Orientation générale

1er décembre 2020

Université de Montréal

Faculté de médecine

Ce mémoire intitulé

Caractérisation *in silico* et purification des ligases à ARN de type RtcB de *Diplonema papillatum*

présenté par

Alexandra Léveillé-Kunst

a été évalué par un jury composé des personnes suivantes :

Dre. Marlene Oeffinger
(président-rapporteur)

Dre. Gertraud Burger
(directeur de recherche)

Dre. Joelle Pelletier
(membre du jury)

Résumé

Les acides ribonucléiques (ARN) subissent plusieurs modifications post-transcriptionnelles avant de remplir leur rôle dans la cellule. Un des acteurs responsables de ces modifications sont les ligases à ARN. Il existe deux grandes familles de ligases à ARN soit les « ATP-grasp » et les « RtcB-like ». Malgré le fait que ces enzymes ont des rôles biologiques similaires dans la cellule, leurs mécanismes moléculaires sont très différents.

Notre équipe a découvert chez un eucaryote marin la présence de trois gènes codant pour des homologues de ligases à ARN de type RtcB. Ceci est pour le moins inhabituel considérant que la plupart des espèces eucaryotes n'ont qu'un seul gène codant pour des ligases de ce type. *Diplonema papillatum*, l'organisme en question, est un eucaryote dont l'étude a gagné en popularité au courant des dernières années dû au mode d'expression particulier de son matériel génétique mitochondrial. Celui-ci est fragmenté en plusieurs morceaux appelés modules qui, durant le processus de maturation de l'ARN, sont joints ensemble via épissage en *trans*.

Nous supposons que l'une des ligases de type RtcB présente chez *D. papillatum*, plus spécifiquement DpRTCB1, est un des acteurs principaux de ce phénomène d'épissage en *trans*. Nous pensons aussi que les deux autres RtcB présentes chez cet organisme, soit DpRTCB2 et DpRTCB3, ont chacune leur propre rôle dans la cellule. Nous avons donc modélisé la structure tertaire de ces protéines *in silico* donnant ainsi des indices quant à ce qui pourraient être requis comme cofacteurs par ces trois enzymes. Nous proposons aussi un système de classification des ligases de type RtcB en fonction de leurs rôles biologiques et de leurs variations au niveau des résidus composant le site actif de l'enzyme. Nous avons tenté de purifier des protéines de fusion DpRTCB pour de futurs essais enzymatiques afin de déterminer les rôles biologiques potentiels de ces enzymes. Toutefois, ces protéines formaient des corps d'inclusion rendant leur purification difficile. Ce faisant, nous démontrons les différentes techniques qui existent actuellement pour purifier des protéines à partir d'agrégats insolubles.

Ce mémoire prédit les potentiels cofacteurs et substrats nécessaires pour de futurs essais biochimiques des ligases DpRTCB. Nous établissons aussi une base robuste pour un système de classification des ligases de type RtcB. Ce document prodigue entre autres des solutions de base aux chercheurs désireux de purifier des protéines qui forment des corps d'inclusion avant de considérer passer à des méthodes de purification plus laborieuses et coûteuses.

Mots clés : Ligases à ARN de type RtcB, classification des ligases de type RtcB, *Diplonema papillatum*, modélisation *in silico*, purification de protéines

Abstract

Ribonucleic acids (RNA) undergo several post-transcriptional modifications before fulfilling their role in the cell. One of the actors responsible for these modifications are RNA ligases. There are two main families of RNA ligases, namely "ATP-grasp" and "RtcB-like". Despite the fact that these enzymes have similar biological roles in the cell, their molecular mechanisms are very different.

Our team has discovered in a marine eukaryote the presence of three genes encoding RtcB RNA ligase homologs. This is unusual considering that eukaryotes generally have only one gene encoding for an homolog of this ligase. *Diplonema papillatum*, the organism in question, is a eukaryote that has grown in popularity in recent years due to the particular mode of expression of its mitochondrial genetic material. Its genome is fragmented into several pieces called modules which, during the RNA maturation process, these modules undergo ligation via *trans*-splicing.

We posit that one of the RtcB-type ligases present in *D. papillatum*, more specifically DpRTCB1, is a major player in this *trans*-splicing phenomenon. We also believe that the other two RtcB ligases present in this organism, DpRTCB2 and DpRTCB3, each have its own role in the cell. We therefore established *in silico* models for these proteins which could hint at the cofactors required by these three enzymes. We also propose a classification system for RtcB-type ligases according to their biological roles and variations in known active site residues. We attempted to purify DpRTCB fusion proteins for future enzymatic assays in order to get a better understanding in the biological role of these enzymes. These proteins, however, formed inclusion bodies making their purification difficult. Thus, we demonstrate various techniques that currently exist to attempt to purify proteins from insoluble aggregates.

This Master's thesis attempts to predict the potential cofactors and substrates necessary for future biochemical assays of DpRTCB enzymes. We also establish a robust foundation for a classification system of RtcB-type ligases. Among other things, this document provides

basic solutions to researchers wishing to purify proteins that form inclusion bodies before considering switching to more laborious and expensive purification methods.

Keywords: RtcB-type RNA ligases, classification of RtcB-type RNA ligases,*Diplonema papillatum*, *in silico* modeling, protein purification

Table des matières

Résumé 5

Abstract 7

Liste des tableaux 13

Liste des figures 15

Liste des sigles et des abréviations 19

Remerciements 25

Chapitre 1. Introduction 27

1.1. Maturation de l'ARN 27
1.1.1. Épissage de l'ARN 27
1.1.2. Édition de l'ARN 28

1.2. Les ligases 28
1.2.1. Les ligases à ARN « ATP-grasp » 29
1.2.1.1. Découverte, classification and distribution des ligases à ARN ATP-grasp 29
1.2.1.2. Rôles biologiques des ligases à ARN ATP-grasp 30
1.2.1.3. Mécanisme moléculaire des ligases à ARN ATP-grasp 30
1.2.1.4. Structure et domaines protéiques des ligases à ARN ATP-grasp 30
1.2.2. Les ligases à ARN « RtcB-like » 32
1.2.2.1. Découverte, classification et distribution des ligases à ARN RtcB 32
1.2.2.2. Rôles biologiques des ligases à ARN RtcB 33
1.2.2.3. Mécanisme moléculaire des ligases à ARN RtcB 36
1.2.2.4. Structure et domaines protéiques des ligases à ARN RtcB 38
1.2.2.5. Stratégies de purification des ligases à ARN RtcB 40

1.3. *Diplonema papillatum* 40
1.3.1. Le génome nucléaire de *D. papillatum* 41
1.3.2. Le génome mitochondrial de *D. papillatum* 42

1.3.3. L'épissage en *trans* de l'ARN mitochondrial de *D. papillatum* 42
1.3.4. L'édition de l'ARN mitochondrial chez *D. papillatum* 44
1.4. Résultats préliminaires et hypothèses de recherche 44
1.5. Buts et objectifs 45
Chapitre 2. « *In silico* three-dimensional structure prediction of RtcB RNA ligases in *Diplonema papillatum* » 47
Préface au chapitre 2 47
Présentation de l'article 1 48
2.1. Abstract 49
2.2. Introduction 49
2.3. Materials and methods 51
2.3.1. Sub-cellular localization prediction 51
2.3.2. Three-dimensional structure prediction for DpRTCB ligases 51
2.3.3. Protein sequence collection 52
2.3.4. Phylogenetic analyses 53
2.4. Results and Discussion 53
2.4.1. Sequence analysis of DpRTCB proteins 53
2.4.2. Structure models of DpRTCB proteins 54
2.4.2.1. DpRtcB1 structure model 54
2.4.2.2. DpRtcB2 structure model 55
2.4.2.3. DpRtcB3 structure model 55
2.4.3. Phylogenetic relationships of RtcB-type sequences 56
2.4.4. Sequence variation in the active site of distinctive phylogenetic groups 57
2.4.5. Mapping the known biological roles on the phylogenetic tree of RtcB 59
2.4.6. Predicted roles for DpRTCB ligases 60
2.5. Conclusion 60
2.6. Figures 62
2.7. Tables 68
2.8. Supplementary data 69
Chapitre 3. Purification des ligases de type RtcB de *Diplonema papillatum* 77

3.1. Introduction 77

3.2. Matériel et méthodes 77
3.2.1. Isolation d'ARN polyadénylé 77
3.2.2. RT-PCR 78
3.2.3. Construction de plasmides pour la surexpression de protéines 78
3.2.4. Préparation de cellules compétentes *E. coli* 80
3.2.5. Transformation par choc thermique 82
3.2.6. Surexpression protéique 82
3.2.7. Lyse cellulaire des bactéries 82
3.2.8. Purification de protéines 83
3.2.9. Retrait du tag SUMO3 en N-terminus des protéines surexprimées 85
3.2.10. Immunobuvardage de type Western 85

3.3. Résultats 86
3.3.1. Construit des protéines recombinantes et choix de la souche bactérienne pour la surexpression 86
3.3.2. Surexpression des protéines recombinantes HS-DpRTCB pour la purification par affinité 86
3.3.3. Lyse des cellules exprimant HS-DpRTCB2 dans le but de purifier les protéines recombinantes 88
3.3.4. Essais de solubilisation de HS-DpRTCB2 92
3.3.5. Purification de HS-DpRTCB2 par affinité 95
3.3.6. Digestion par Ulp1 de la protéine purifiée HS-DpRTCB2 96
3.3.7. L'ajout d'un espaceur afin d'améliorer la solubilité de HS-DpRTCB1 98
3.3.8. Essais de solubilisation de HS-GS-DpRTCB1 à l'aide d'arginine 98

3.4. Conclusion 100

3.5. Remerciements 101

Chapitre 4. Discussion 103

4.1. Séquences protéiques des DpRTCB 103

4.2. Modélisation *in silico* des DpRTCB 104

4.3. Regroupement phylogénétique des ligases de type RtcB 104

4.4. L'expression hétérologue des protéines HS-DpRTCB 105

4.5. La faible solubilité des HS-DpRTCB est causée par des corps d'inclusion 106
4.6. Méthodes utilisées afin de replier les protéines suite à la dénaturation 107
4.7. Approches futures envisagées afin de purifier les protéines HS-DpRTCB 108
Chapitre 5. Conclusion 109
Références bibliographiques 111
Annexe A. Expression constitutive de la protéine recombinante HS-GS-DpRTCB1 119

Liste des tableaux

1.1 Comparaison des rôles biologiques entre les groupes de ligases à ARN ou ADN de type ATP-grasp et RtcB-like chez diverses espèces. 33
1.2 L'ajout de coiffe chez les ligases RtcB dépendant de leurs extrémités et de leurs acides nucléiques respectifs. 35

2.1 Predicted model accuracy and number of residues accurately modelled by Phyre2. 68
2.2 Select organisms with their respective RtcB UniProt reference ID. 74
2.3 Probability of mitochondrial targeting signal for diplonemid RTCB proteins. 75
2.4 Probability of other targeting signal for diplonemid RTCB proteins. 76
2.5 Accuracy of the predicted models obtained with SWISS-MODEL. 76

3.1 Amorces PCR utilisées pour les séquences DpRTCB1, DpRTCB2, DpRTCB3, DpRTCB1-GS et les construits de plasmides. 79
3.2 Potentiel d'agrégation de HS-DpRTCB1, HS-DpRTCB2, HS-DpRTCB3 et HS-GS-DpRTCB1. 90

Liste des figures

1.1 Schéma illustrant le mécanisme moléculaire de ligature des fragments d'ARN par les ligases à ARN ATP-grasp. 31

1.2 Schéma illustrant le mécanisme de liaison des fragments d'ARN par RtcB 37

1.3 Deux structures cristallographiques de la ligase RtcB de *Pyrococcus horikoshii* couplée avec ses cofacteurs respectifs 39

1.4 Arbre phylogénétique des eucaryotes 41

1.5 Traitement des modules d'ARN mitochondrial et épissage en *trans* chez *Diplonema papillatum* 43

2.1 Probability of mitochondrial targeting signal for diplonemid RTCB proteins. ... 62

2.2 Crystal structure of PhRtcB and models of other RtcB proteins predicted by Phyre2. 63

2.3 Two-dimensional topological representation of PhRtcB obtained from the database PDBsum (PDB ID: 4DWQ). 64

2.4 View of the active site to visualize interactions of amino-acid residues with cofactors. 65

2.5 Phylogenetic relationships among the RtcB-type family with color-shading indicating each RtcB clade. 66

2.6 Multiple sequence alignment of RtcB proteins with representatives of each clade along with *D. papillatum* RTCB sequences. 67

2.7 Multiple sequence alignment of known RTCB protein sequences in diplonemids and PhRtcB. 69

2.8 Prediction models obtained with SWISS-MODEL. 70

2.9 Three-dimensional structure overlay of residues within the active site of DpRTCB1 and PhRtcB. 71

2.10 Surface density representation of PhRtcB and DpRTCB2 with the groove present in the vicinity of the active site pocket. 71

2.11 Phylogenetic distribution of the RtcB-type family obtained with RAxML with the same alignment as in Figure 2.5. 72

2.12 Close-up on the phylogenetic distribution of Figure 2.5 and Supplementary Figure 2.11 73

3.1 Construits des plasmides pour HS-DpRTCB2 et HS-GS-DpRTCB1. 80

3.2 Schéma pour l'assemblage des plasmides utilisés pour la surexpression des protéines DpRTCB. 81

3.3 Séparation par Tris-glycine SDS-PAGE des lysats cellulaires des transformants Rosetta2-DpRTCB1, Rosetta2-DpRTCB2 et Rosetta2-DpRTCB3 avant et après induction. 87

3.4 Séparation sur gel Tris-glycine SDS-PAGE des lysats cellulaires des transformants pré-induits et induits à surexprimer la protéine HS-DpRTCB2 avec le surnageant et le culot résultant de la lyse 88

3.5 Profils d'agrégation de HS-DpRTCB1, HS-DpRTCB2, HS-DpRTCB3, HS-GS-DpRTCB1 et HS-EcRtcB obtenus à l'aide du logiciel Aggrescan. 91

3.6 Séparation sur gel Tris-tricine SDS-PAGE des échantillons suivant la surexpression et la lyse cellulaire des transformants Rosetta2-DpRTCB2. 92

3.7 Séparation sur gel Tris-glycine SDS-PAGE des cellules Rosetta2-DpRTCB2 induits lysés par « French press » après trois à cinq passages dans l'appareil. 93

3.8 Séparation sur gel Tris-tricine SDS-PAGE des échantillons suivant la surexpression et la lyse cellulaire des transformants Rosetta2-DpRTCB2 en présence de détergent ou d'urée. 94

3.9 Séparation sur gel Tris-tricine SDS-PAGE des échantillons obtenus durant la purification du lysat cellulaire des transformants ayant surexprimé la protéine HS-DpRTCB2. 95

3.10 Séparation des échantillons obtenus lors de l'échange de tampon sur gel Tris-tricine SDS-PAGE. 96

3.11 Séparation des échantillons par gel Tris-tricine SDS-PAGE de la protéine HS-DpRTCB2 suivant la digestion avec la protéase Ulp1. 97

3.12 Détection par immunobuvardage de type Western des protéines HS-DpRTCB1, HS-DpRTCB2 et HS-DpRTCB3 avec des anticorps anti-SUMO3. 98

3.13 Séparation des échantillons sur Tris-tricine SDS-PAGE obtenus lors de la lyse cellulaire des transformants Rosetta2-GS-DpRTCB1-GS ayant surexprimé leur protéine d'intérêt. 99

3.14 Détection par immunobuvardage de type Western des protéines HS-GS-DpRTCB1 suivant l'induction et la lyse des transformants avec des anticorps anti-SUMO3. 100

A.1 Détection par immunobuvardage de type Western des protéines HS-GS-DpRTCB1 suivant l'induction et la lyse des transformants avec des anticorps anti-SUMO3. 119

Liste des sigles et des abréviations

ADN	Acide désoxyribonucléique
ADNmt	ADN mitochondrial
AMP	Adénosine monophosphate
ARN	Acide ribonucléique
ARNi	ARN interference
ARNm	Acide ribonucléique messager
ARNt	Acide ribonucléique de transfert
AmRTCB1	*Artemidia motanka* RtcB1
AmRTCB2	*Artemidia motanka* RtcB2
ARNr	Acide ribonucléique ribosomal
ATP	Adénosine triphosphate

CGI-99	Facteur de transcription, de transport et de traduction de l'ARN
DDX1	Hélicase DEAD-box 1
DaRTCB1	*Diplonema ambulator* RtcB1
DaRTCB2	*Diplonema ambulator* RtcB2
DjRTCB1	*Diplonema japonicum* RtcB1
DjRTCB2	*Diplonema japonicum* RtcB2
DpRTCB1	*Diplonema papillatum* protéine RtcB1
DpRTCB1	*Diplonema papillatum* gène nucléaire RtcB1
DpRTCB2	*Diplonema papillatum* protéine RtcB2
DpRTCB2	*Diplonema papillatum* gène nucléaire RtcB2
DpRTCB3	*Diplonema papillatum* protéine RtcB3
DpRTCB3	*Diplonema papillatum* gène nucléaire RtcB3
EcRtcB	*Escherichia coli* RtcB

FnRTCB1	*Flectonema neradi* RtcB1
FnRTCB2	*Flectonema neradi* RtcB2
GMP	Guanosine monophosphate
GTP	Guanosine triphosphate
HAC1	Facteur de transcription HAC1
HsRTCB	*Homo sapiens* RtcB
LlRTCB1	*Lacrimina lanifica* RtcB1
LlRTCB2	*Lacrimina lanifica* RtcB2
LtRTCB	*Leishmania tarentolae* RtcB
Sen	Sous-unité de l'endonucléase d'épissage de l'ARNt
MazF	Endoribonucléase MazF
MkRtcB	*Methanopyrus kandleri* RtcB
MxRtcB1	*Myxococcus xanthus* RtcB1

MxRtcB2	*Myxococcus xanthus* RtcB2
MxRtcB3	*Myxococcus xanthus* RtcB3
MxRtcB4	*Myxococcus xanthus* RtcB4
MxRtcB5	*Myxococcus xanthus* RtcB5
MxRtcB6	*Myxococcus xanthus* RtcB6
NAD+	Nicotinamide adénine dinucléotide +
NgRTCB1	*Naegleria gruberi* RtcB1
NgRTCB2	*Naegleria gruberi* RtcB2
NkRTCB1	*Namystynia karyoxenos* RtcB1
NkRTCB2	*Namystynia karyoxenos* RtcB2
PhRtcB	*Pyrococcus horikoshii* RtcB
PNK	Polynucléotide kinase
ORF	Cadre de lecture ouvert

OXPHOS	Chaîne de phosphorylation oxydative
ReRTCB1	*Rhynchopus euleeides* RtcB1
ReRTCB2	*Rhynchopus euleeides* RtcB2
ReRTCB3	*Rhynchopus euleeides* RtcB3
RhRTCB1	*Rhynchopus humris* RtcB1
RhRTCB2	*Rhynchopus humbris* RtcB2
Rnl1	Ligase à ARN ATP-grasp 1
Rnl2	Ligase à ARN ATP-grasp 2
Rnl3	Ligase à ARN ATP-grasp 3
Rnl4	Ligase à ARN ATP-grasp 4
Rnl5	Ligase à ARN ATP-grasp 5
Rnl6	Ligase à ARN ATP-grasp 6
RtcA	Cyclase à ARN 3′-phosphate

SsRTCB1	*Sulcionema specki* RtcB1
SsRTCB2	*Sulcionema specki* RtcB2
SUMO3	Small ubiquitin related modifier 3
TcRTCB	*Trypanosoma cruzi* RtcB
TbRTCB	*Trypanosoma brucei* RtcB
TtRtcB	*Thermus thermophilus* RtcB
Tpt1	2′-phosphotransférase à ARNt
Trl1	Ligase à ARNt
Ulp1	Ubiquitin-like specific protease 1
UPR	Unfolded protein response
XBP1	X-box binding protein 1

Remerciements

J'aimerais tout d'abord remercier ma directrice de recherche Dre. Gertraud Burger de m'avoir accueillie dans son laboratoire d'abord comme stagiaire au baccalauréat et ensuite comme étudiante à la maîtrise. C'est en travaillant sur mon projet que je me suis découvert une passion pour la recherche et la biochimie. Cette expérience m'a permis de développer mon esprit scientifique et critique que j'applique aujourd'hui dans mon quotidien. C'est aussi grâce à ce projet que je me suis découvert un intérêt pour la bio-informatique que je n'aurais jamais pensé avoir.

J'aimerais aussi remercier les membres de mon laboratoire: le Dr. Matus Valach et Bhagya C. Thimmappa. D'abord, merci à Matus pour son support indéfectible et son aide inconditionnelle dans mon projet au courant des dernières années. Tes conseils m'ont été indispensables dans l'analyse de mes résultats et dans ma façon de les interpréter. Un second merci à Bhagya pour son support, nos discussions palpitantes et nos sessions de réflexion. J'ai apprécié les propos que nous avons échangés sur nos projets respectifs ou les thématiques scientifiques qui nous semblaient pertinentes.

Un gros merci à Matt Sarasin, Peniel Bustamente Villabolos et Savandara Besse pour leur support durant mon cheminement. Que ce soit en informatique ou au laboratoire, leur aide m'a toujours été d'une grande utilité. J'avais constamment cette impression de déranger avec mes questions, donc sachez que j'apprécie énormément le temps que vous avez pris pour me répondre malgré votre horaire chargé. J'aimerais aussi remercier sans ordre particulier les autres membres des laboratoires voisins et du département de biochimie qui m'ont soutenu durant mon cheminement: Pedro do Couto Bordignon, Louis Gendron, Nazli Kocatuğ, Musa Ozboyaci, Amruta Sahoo, Shamim Hasan, Lila Salhi, Philipe Lampron, Sebastien Truche, Shona Teijeiro et Elaine Meunier.

Finalement, j'aimerais remercier ma famille et mes meilleurs amis, Jean Bouchard et Andreea Seremet, pour leur support et leur patience au courant de mes années d'études. Sans leurs conseils et leur soutien durant mes moments plus difficiles je n'y serais jamais

parvenue. C'est grâce à eux que je suis maintenant la personne que je suis aujourd'hui et que je peux fièrement soumettre ce document qui est le fruit de mes 2 ans de travail acharné.

Chapitre 1

Introduction

Les acides ribonucléiques (ARN) sont des acteurs majeurs dans l'expression de l'information génique des cellules vivantes. Avant d'entamer leurs rôles respectifs, les molécules d'ARN doivent passer à travers plusieurs étapes de maturation, incluant le conditionnement de leurs extrémités, l'épissage et l'édition. Un groupe d'enzymes, les ligases à ARN, sont cruciales à cette maturation puisqu'elles sont impliquées dans l'épissage, l'édition et la réparation des ARN messagers (ARNm) et des ARN de transfert (ARNt). Vu l'ajout de nouveaux membres dans la famille des ligases à ARN au fil des années, leur classification et leur rôles biologiques ont été constamment mis à jour. Dans cette revue de littérature, il sera question des connaissances actuelles sur les ligases à ARN, plus particulièrement des ligases à ARN de la famille RtcB ainsi que de l'organisme que j'ai utilisé pour l'étude de ces enzymes dans le cadre de ma maîtrise, *Diplonema papillatum*.

1.1. Maturation de l'ARN

1.1.1. Épissage de l'ARN

Les ARN primaires contenant des introns subissent l'épissage qui peut opérer en *cis* ou en *trans*. Dans l'épissage prédominant procédant en *cis*, les introns sont retirés d'un transcrit précurseur contigu. Dans l'épissage en *trans*, les exons sont joints ensemble à partir de transcrits séparés. Les deux formes principales d'épissage se retrouvent au niveau des introns spliceosomaux, les introns du groupe I et du groupe II et les introns archéaux/ARNt [**1**].

Le retrait d'introns dit spliceosomaux requièrent la machinerie du spliceosome [**1**]. L'épissage en *cis* est prédominant pour les transcrits de gènes nucleaires. Toutefois, certaines espèces, tel le dinoflagellé *Karlodinium micrum*, le nématode *Caenorhabditis elegans* ou l'euglenozoaire *Trypasonoma brucei*, ajoutent par épissage en *trans*, un exon

non-codant 5-terminal, le « splice-leader ». L'exon contenu dans le « splice-leader » est ajouté à l'ARN rendant le transcrit complètement mature [**1, 2**]. Les introns du groupe I et du groupe II sont auto-catalytiques et diffèrent d'un groupe à l'autre de par leur structure secondaire et leur mécanisme moléculaire. Les introns du groupe I sont présents majoritairement au niveau des mitochondries et des plastides, mais se retrouvent aussi dans les gènes nucléaires de certains eucaryotes [**1, 3**]. Les introns du groupe II sont retrouvées dans les génomes d'organelles et aussi chez certaines espèces bactériennes et archées [**1, 4**]. Les introns archéaux/ARNt se retrouvent majoritairement dans les gènes des ARNt, mais chez les archées, ils sont présents au niveau des ARNr et des gènes protéiques et l'épissage de ces introns est effectué à l'aide d'endonucléases [**1**]. L'épissage de ces introns est effectué majoritairement par épissage classique en *cis*, mais l'épissage en *trans* de ces introns a pu être observé chez l'archée *Nanoarchaeum equitans* [**1, 5**] et chez certains eucaryotes [**6**].

1.1.2. Édition de l'ARN

Certains ARN requièrent des modifications post-transcriptionnelles supplémentaires, comme l'édition, avant d'être considérés comme des transcrits matures. Les types d'édition sont classifiés en trois différents types : l'insertion/délétion (indels) par l'ajout/retrait de nucléotides, la substitution de nucléotides et la substitution de nucléotides se produisant exclusivement au niveau des extrémités $5'$ ou $3'$ des boucles acceptrices de certains ARNt mitochondriaux [**7**]. Chez la mitochondrie, l'édition demeure un phénomène commun [**8**]. Dans cette organelle, les substitutions de type C-à-U (cytidine-à-uridine) représentent la forme la plus commune d'édition chez les plantes terrestres [**9**]. L'insertion et la délétion de nucléotides sont observées surtout dans les mitochondries des kinétoplastides [**10**]. Le groupement sœur aux kinétoplastides, les diplonémides et plus particulièrement *D. papillatum*, subissent aussi l'édition de leurs transcrits mitochondriaux comme décrit plus en détails ici-bas [**voir section 1.3.4**].

1.2. Les ligases

Les ligases sont des enzymes importantes dans la maturation de l'ARN puisqu'elles joignent des fragments d'ARN ensemble. Traditionnellement, les ligases d'acides nucléiques sont différentiées les unes des autres par leur usage soit d'ADN soit d'ARN comme substrat. Toutefois, comme les ligases à ADN ont probablement évolué à partir d'un domaine catalytique commun aux ligases à ARN, le système de classification actuel combine donc les deux types d'enzymes. Les ligases à ADN sont surtout impliquées dans la réparation des dommages subis par l'ADN, la réplication ainsi que la recombinaison génomique [**11, 12**]. Ces enzymes scellent les bris au niveau des ADN double-brins et utilisent en général

l'adénosine triphosphate (ATP) comme cofacteur. Or, il y a des exceptions : certaines ligases bactériennes utilisent plutôt la nicotinamide adénine dinucléotide$^+$ (NAD^+) comme cofacteur. Les enzymes coiffant les ARNm, malgré le fait qu'elles emploient le GTP, sont groupées dans la même catégorie que les ligases à ADN puisqu'elles possèdent le même domaine nucléotidyltransférase [**13, 14, 15**]. Les ligases à ARN, quant à elles, se divisent en deux catégories distinctes en fonction de leur cofacteurs respectifs étant soit le GTP ou l'ATP et sont connus sous le nom soit « ATP-grasp » (ou « T4-like ») ou « RtcB-like ».

1.2.1. Les ligases à ARN « ATP-grasp »

1.2.1.1. Découverte, classification and distribution des ligases à ARN ATP-grasp

Les ligases à ARN ont été découvertes pour la première fois chez des bactéries *Escherichia coli* infectées par des bactériophages T4 [**16**]. D'autres études menées chez les plantes et les champignons ont révélé que ces enzymes sont impliquées au niveau de la ligature des exons des ARNt [**17**]. À ce jour, la famille des ATP-grasp comprend six groupes différents classifiés selon le repliement distinct de leur domaine C-terminal, soit les ligases à ARN 1 à 6 (Rnl1-Rnl6) [**15, 18**]. Malgré le fait que toutes les ligases à ARN de type ATP-grasp possèdent un domaine commun, les ligases de type Rnl4 semblent avoir dévié récemment des ligases à ADN ATP-dépendantes [**11**]. Les deux groupes, Rnl1 et Rnl4, lient les bris simples brins au niveau de la boucle des tiges-boucles des ARNt, alors que les enzymes de type Rnl2 scellent les bris au niveau des ARN doubles brins formés par des duplexes ARN/ADN ou ARN/ARN [**15**]. Les ligases de type Rnl3 forment des ARN circulaires à partir d'ARN simples brins [**19, 20**] alors que les enzymes de type Rnl5 (qui ne possèdent pas de domaine C-terminal) réparent des bris dans des duplexes à ARN qui ont un domaine amino-terminal spécifique [**21, 22**]. Les enzymes de type Rnl6 sont pour le moment uniquement représentées par la ligase ATP-grasp appelée Trl1 retrouvée chez les champignons [**23**]. Cette ligase possède des substrats spécifiques qui seront abordés plus en détails dans la section sur les rôles biologiques de ligases ATP-grasp ici-bas [**voir section 1.2.1.2**].

Les ligases de type Rnl1 sont présentes chez les bactériophages, les baculovirus, les champignons, les plantes et plusieurs groupes de protistes [**12, 24**]. Les ligases de type Rnl2 se retrouvent dans les trois domaines du vivant et incluent les ligases éditrices des ARNm très bien étudiées chez *Trypanosoma* et *Leishmania* [**12, 25, 26**]. Les ligases de type Rnl3 et Rnl4 sont présentes dans les deux domaines procaryotiques et ont été étudiées en détails chez l'archée *Pyrococcus abyssi* et la bactérie *Clostridium thermocellum* [**15**]. Les ligases de type Rnl5 ont été retrouvées chez la bactérie *Deinococcus radiodurans* et l'euglezonaire *Naegleria gruberi* [**21, 22**]. Les ligases de type Rnl6 ont été uniquement observées chez les champignons

où leur domaine C-terminal a été étudié plus en détails chez *Chaetomium thermophilum* [**23**].

1.2.1.2. Rôles biologiques des ligases à ARN ATP-grasp

Chez les plantes et les champignons, les ligases ATP-grasp sont impliquées dans l'épissage d'introns des pré-ARNt en effectuant la ligature des exons. Ces ligases participent également à l'épissage non-conventionnel de l'ARNm du gène *HAC1* chez les mycètes, un gène codant pour un facteur de transcription impliqué dans la « unfolded protein response » (UPR). Chez *Saccharomyces cerevisiae* c'est la ligase à ARN Trl1p qui rejoint les exons; l'ARNm du gène *HAC1* est le seul dans la levure qui est épissé de façon non-spliceosomale [**17, 27**]. Finalement, les ligases à ARN ATP-grasp participent aussi à la réparation des dommages à l'ARN causés par des toxines telles que la zymocine et Shiga qui sont produites par certains champignons et bactéries, respectivement [**11**].

1.2.1.3. Mécanisme moléculaire des ligases à ARN ATP-grasp

Les ligases ATP-grasp agissent sur les extrémités 3′-OH et 5′-PO_4 des ARN, avec certaines de ces enzymes requérant Mg^{2+} comme cofacteur [**19**]. La ligature des ARN par les ligases ATP-grasp procède notamment en trois étapes majeures **(pour plus de détails, voir Figure 1.1)**. Ce processus requiert l'énergie formée par le lien phospho-anhydride de l'ATP qui est transféré à l'extrémité 5′ de l'ARN [**28**]. Le mécanisme catalytique des ligases à ARN ATP-grasp, décrit ici-bas, ressemble de très près à celui des ligases à ADN ATP-dépendantes, mais est assez différent de celui observé chez les ligases à ADN NAD^+-dépendantes [**11**].

Certains processus biochimiques dans la cellule produisent des extrémités d'ARN qui sont différentes de celles normalement requises par les ligases à ARN ATP-grasp. Par exemple, une fois l'intron retiré des ARNt chez les plantes, l'endonucléase laisse une extrémité 5′-OH. Avant la ligature, ces extrémités sont phosphorylées produisant ainsi des substrats adéquats pour la ligase. Cette phosphorylation est effectuée par les polynucléotides kinases (PNKs), partenaires importants des ligases à ARN ATP-grasp [**11**].

1.2.1.4. Structure et domaines protéiques des ligases à ARN ATP-grasp

Les ligases ATP-grasp comprennent un domaine protéine-kinase de type C-terminal et un domaine de repliement de type RAGNYA qui émanent fort possiblement d'une fusion ancestrale [**11**]. La ligature de l'ARN par ces enzymes a lieu au niveau d'un résidu lysine au sein du site actif de la protéine. Le site de liaison des nucléotides est composé d'une série de repliements α/β qui capturent la molécule d'ATP. Elles ont aussi des boucles

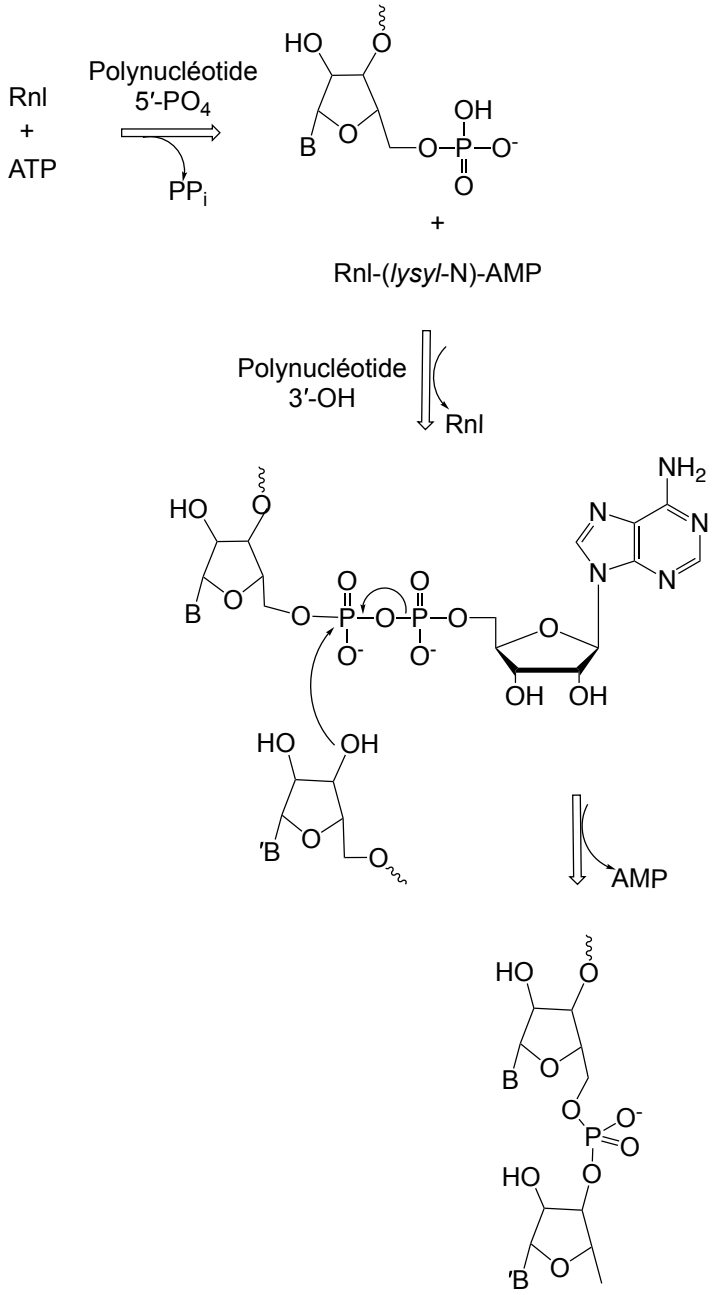

Fig. 1.1. Schéma illustrant le mécanisme moléculaire de ligature des fragments d'ARN par les ligases à ARN ATP-grasp. La ligation est initiée par (i) la ligase à ARN (Rnl) qui forme un intermédiaire covalent Rnl-(*lysyl*-N)-AMP en relâchant un pyrophosphate de l'ATP; suivi par (ii) l'adénylate formé qui est transféré à un polynucléotide 5′-PO_4 pour former un intermédiaire polynucléotide-(5′)pp-(3′)A ARN et finalement, (iii) Rnl catalyse l'attaque de l'extrémité 3′-OH sur l'intermédiaire N(5′)pp(3′)A ARN pour former la jonction entre les fragments d'ARN et relâcher un AMP.

inter-brin qui contiennent cinq motifs caractéristiques à cette famille [**12**]. Les ligases ATP-dépendantes montrent en général une extension C-terminale qui, comme mentionné dans la section précédente, forme la base du système de classification de ces enzymes en six différents sous-groupes. De plus, les ligases de type Rnl1 et Rnl3 possèdent toutes les deux un domaine N-terminal à quatre brins qui semble être critique pour leur interaction avec l'ARN [**11**].

1.2.2. Les ligases à ARN « RtcB-like »

Un deuxième groupe de ligases à ARN, qui est distinct de celui des ATP-grasp, est la superfamille des ligases à ARN GTP-dépendantes qui n'est constituée pour le moment que des membres du type RtcB. Il est à noter que, malgré que les ligases coiffantes de l'ARNm utilisent aussi le GTP comme cofacteur, elles sont plutôt classifiées avec les ligases ATP-grasp dû à leur structure protéique qui est similaire aux ligases à ADN ATP-dépendantes [**15**].

1.2.2.1. Découverte, classification et distribution des ligases à ARN RtcB

Les ligases à ARN RtcB ont été décelées pour la première fois il y a environ 40 ans lors d'études sur la maturation des ARNt dans des cellules HeLa [**29**]. À ce moment, l'enzyme responsable de cette activité demeurait inconnue. Les chercheurs ont finalement conclu que – contrairement aux plantes et aux champignons [**29**] – ce sont les ligases RtcB qui sont responsables de la ligature des exons lors de l'épissage des ARNt dans les cellules humaines et chez l'archée *Methanopyrus kandleri* [**30, 31**]. Chez les humains, cette ligase a été nommée HSPC117 (aussi C22orf28 ou FAAP) et classifiée comme membre de la famille protéique UPF0027 [**30**]. Les ligases du groupe RtcB sont présentes dans les trois domaines de la vie, soit chez les bactéries, les archées et les eucaryotes. Au sein des eucaryotes, les ligases RtcB ont été observées chez les métazoaires, et de nombreux protozoaires, mais elles sont absentes de la plupart des plantes et des champignons étudiés jusqu'à présent [**17, 30, 31, 32**].

Escherichia coli est la première espèce bactérienne chez laquelle la présence d'une ligase de la famille RtcB a été rapportée. Chez cet organisme, le gène *rtcb* forme un opéron *rtcBA* avec l'ARN 3′-phosphate cyclase (RtcA), régulé par RtcR, un activateur σ^{54}-spécifique [**32, 33**]. Toutefois, ce type de corégulation du gène *rtcB* avec *rtcA* est plutôt rare chez les bactéries. *Pelobacter propionicus* possède également un opéron contenant le gène *rtcB* et un de cofacteurs de cette enzyme appelé Archease. Cette dernière est requise chez

certains organismes (primordialement chez les métazoaires) pour accélérer le processus de ligature de l'ARN en permettant à RtcB de répéter son cycle catalytique [**33, 34, 35, 36, 37**].

Dû à la découverte récente des ligases de type RtcB, cette superfamille n'a pas encore été analysée avec autant de profondeur que les ligases ATP-grasp et, par conséquent, une sous-catégorisation des protéines RtcB n'a pu être établie. Il est probable que les enzymes RtcB seront éventuellement classifiées selon leur spécificité et leur préférence pour certains substrats [**38**].

1.2.2.2. Rôles biologiques des ligases à ARN RtcB

Chez divers organismes vivants, les ligases RtcB jouent différents rôles dans la cellule tels que l'épissage des ARNt, l'épissage non-conventionnel d'ARNm, la réparation d'ARN et l'ajout de coiffes sur divers acides nucléiques. Ces rôles sont similaires à ceux observés chez les ligases de type ATP-grasp **(voir Tableau 1.1)**.

Tableau 1.1. Comparaison des rôles biologiques entre les groupes de ligases à ARN ou ADN de type ATP-grasp et RtcB-like chez diverses espèces.

Rôles biologiques	ATP-grasp	RtcB-like
Épissage des exons des pré-ARNt dans le noyau[a]	Chez les plantes et les champignons	Chez les métazoaires et les archées
Épissage non-conventionnel d'ARNm durant l'UPR dans le cytosol à la surface de la membrane du réticulum endoplasmique[b]	L'ARNm *HAC1* chez les champignons	L'ARNm *XPB1* chez les métazoaires
Ajout de coiffe aux acides nucléiques dans le cytosol ou le noyau[c]	Coiffe en 5′ des ARNm chez les eucaryotes	Coiffe en 3′ ou en 5′ sur l'ADN et l'ARN chez les bactéries (*in vitro* seulement)
Réparation des dommages à l'ARN dans le cytosol[d]	Les ARNt chez les bactéries, les champignons et les plantes	L'ARNr 16S clivé par MazF chez *E. coli*
Édition de l'ARN[e]	Les ARN mitochondriaux chez les kinétoplastides	?

[a]Références: [**7, 15, 16**]
[b]Références: [**9, 28, 29**]
[c]Références: [**5, 19, 28**]
[d]Références: [**1, 28**]
[e]Références: [**9, 10**]

Épissage des ARNt médiés par les ligases à ARN RtcB chez les métazoaires. Chez les mammifères, RTCB/HSCP117 participe à l'épissage des ARNt. Ce phénomène a été démontré par l'inhibition de la maturation des ARNt en diminuant l'expression de cette

ligase à l'aide d'ARN-interférence (ARNi) [**30, 39**]. Le retrait des introns et la jonction des exons, qui apparaissent être deux processus distincts, a lieu dans le noyau [**17, 39, 40**]. De par sa fonction dans l'épissage des ARNt, il a été suggéré que RTCB/HSCP117 régulerait la synthèse protéique, puisque l'accumulation des pre-ARNt lors de stress cellulaire dans le cytoplasme a été démontrée d'inhiber la traduction [**41**]. Dans le cerveau de la souris, RTCB/HSCP117 se retrouve non seulement dans le noyau, mais aussi dans les granules de transport de l'ARNm des dendrites neuronales, indiquant que cette ligase est aussi impliquée dans d'autres voies de traitement de l'ARN, comme par exemple son transport. Les protéines retrouvées avec RTCB/HSCP117 dans ces granules de transport incluent des protéines DEAD-box, telle que l'hélicase DEAD-box 1 (DDX1) [**42**]. Cette hélicase interagit avec RTCB/HSCP117 et est impliquée dans la transcription, l'édition, l'épissage et les cycles catalytiques de l'ARN. RTCB/HSCP117 interagit aussi avec Archease, dont la présence est requise pour l'épissage des ARNt tel que mentionné ci-haut [**35, 36, 37**].

Épissage de l'ARNm *XPB1* médié par les ligases à ARN RtcB chez les métazoaires. Les ligases RtcB ont aussi un rôle dans l'épissage non-conventionnel de l'ARNm *XPB1* qui est l'homologue chez les métazoaires du gène *HAC1* des champignons [**41, 43, 44**]. L'épissage de l'ARNm *XBP1* ressemble à l'épissage non-conventionnel observé pour l'ARNm *HAC1* à la différence que la ligase Trl1p responsable de ce processus chez les champignons appartient au groupe des ATP-grasp [**voir section 1.1.2**]. Toutefois, dans les cellules animales dont le niveau de RTCB/HSCP117 est réduit, l'épissage de l'ARNm *XPB1* et des ARNt peut être restauré par la ligase ATP-grasp Trl1p mais uniquement en présence de la phosphatase à ARNt des champignons (Tpt1p) [**45, 46**]. Par contre, dans un mutant de *S. cerevisiae* dont *trl1* est non-fonctionnel, la ligase RtcB des animaux n'est pas en mesure de rétablir l'épissage de l'ARNm *HAC1* et des ARNt [**27, 44**], tandis que l'homologue d'*E. coli* en est capable [**17**].

Réparation de l'ARNr médiée par la ligase à ARN RtcB chez *E. coli*. *E. coli* possède un module toxine-antitoxine qui agit durant le processus de mort cellulaire programmée, appelé *mazEF* [**47**]. Activé lors d'un stress, ce module empêche la synthèse protéique en codant pour deux protéines : la toxine MazF et l'antitoxine MazE [**48**]. La nucléase MazF clive les ARN qui possèdent la séquence tri-nucléotidique ACA. L'ARN ribosomique 16S d'*E. coli* contient ce site à la position 1500-1502, 43 nt en amont de l'extrémité 3′. Le clivage de cet ARNr laisse une extrémité 2′, 3′-phosphate cyclique sur le fragment en amont et une extrémité 5′-OH sur le fragment en aval [**49, 50, 51, 52**]. Ces ribosomes tronqués seraient hypothétiquement des ribosomes « spécialisés », puisque la traduction de certains ARNm a toujours lieu dans la cellule malgré l'activation de MazF [**52, 53**]. Une fois le stress cellulaire dissipé, la ligase RtcB peut rattacher les fragments

d'ARNr 16S et ainsi réparer le ribosome [**49**].

L'ajout de coiffe sur les acides nucléiques médié par les ligases à ARN RtcB chez les bactéries. Il y a seulement cinq ans, on a découvert que les enzymes RtcB peuvent ajouter des coiffes aux extrémités 3′-PO_4 et 5′-PO_4 d'acides nucléiques (ARN et ADN) en utilisant le GTP. L'intermédiaire de cette réaction est un acide nucléique guanylé grâce à l'ajout d'un GMP à une extrémité phosphate [**38, 54, 55**]. L'ajout de coiffe est proposé de protéger les extrémités des acides nucléiques contre des exonucléases ou phosphodiestérases [**55**] et de rendre l'ADN prêt à la liaison avec un fragment approprié [**54**]. Toutefois, l'ajout de coiffe sur des acides nucléiques a seulement été observé *in vitro* pour des ligases présentes chez deux espèces bactériennes soit: la ligase RtcB de *E. coli* et deux des six homologues de ces ligases chez *Myxococcus xanthus*. La ligase RtcB d'*E. coli* (EcRtcB) ainsi qu'un des homologues de la bactérie *M. xanthus* (MxRtcB2) sont en mesure d'ajouter une coiffe à l'extrémité 3′-PO_4 de l'ADN [**38, 54**]. Un autre homologue de la ligase RtcB chez *M.xanthus* (MxRtcB3) est aussi capable d'ajouter une coiffe aux acides nucléiques. Toutefois, celle-ci peut le faire sur l'une ou l'autre des extrémités des acides nucléiques [**38**] **(voir Tableau 1.2)**. Conséquemment, on ne peut pas généraliser que les ligases RtcB aient une préférence pour l'ajout de coiffe sur l'une ou l'autre des extrémités des deux types d'acides nucléiques.

Tableau 1.2. L'ajout de coiffe chez les ligases RtcB dépendant de leurs extrémités et de leurs acides nucléiques respectifs.

	Acide nucléique		**Extrémité terminale**	
Homologue RtcB	ADN	ARN	3′-PO_4	5′-PO_4
E. coli RtcB	✓		✓	
M. xanthus RtcB2	✓		✓	
M. xanthus RtcB3	✓	✓	✓	✓

Il serait aussi possible que les ligases RtcB puissent réparer des ARN endommagés par des facteurs extracellulaires [**11**], et soit impliquées dans plusieurs autres processus, dont le transport d'ARN dans la cellule, comme suggéré par l'interaction des ligases RtcB avec CGI-99 dans les granules d'ARN [**37, 42**].

Questions ouvertes. Plusieurs questions demeurent quant aux ligases à ARN RtcB. Par exemple, chacune des six ligases RtcB présentes chez *Myxococcus* auraient-elles potentiellement leur propre rôle dans la cellule en interagissant différemment avec leurs substrats respectifs? Ou encore, ces ligases auraient-elles des tâches qui se chevauchent, comme si bien démontré chez les métazoaires où une seule ligase RtcB est responsable du processus

d'épissage des ARNt et de l'épissage non-conventionnel de l'ARNm *XPB1* durant l'UPR? Accomplissent-elles réellement des tâches différentes ou leurs rôles sont-ils semblables? Dans quelle mesure est-il pertinent pour certaines espèces de posséder plusieurs homologues d'une même enzyme?

1.2.2.3. Mécanisme moléculaire des ligases à ARN RtcB

Les ligases RtcB possèdent une activité phosphodiestérase, puisqu'elles sont en mesure de rompre les liens phosphate du GTP, et une activité ligase puisqu'elles scellent les extrémités 5′-OH avec les extrémités 3′-PO_4 ou 2′, 3′-phosphate cyclique de l'ARN [**32, 54, 56**]. Durant la ligature de l'ARN, l'enzyme procède en quatre étapes majeures, dont la première se déroule uniquement si l'ARN possède une extrémité 2′, 3′-phosphate cyclique **(pour plus de détails, voir Figure 1.2)** [**56, 57**]. L'activité de ligation de RtcB, qui nécessite normalement des ions Mn^{2+}, peut être inhibée (même en présence de Mn^{2+}) par des cations divalents tels que Zn^{2+} et Cu^{2+} et affectée, mais à un degré moins important, par des ions Ni^{2+} et Co^{2+}. Ces quatre cations inhibiteurs font compétition pour le site actif de l'enzyme [**32**]. Comme abordé plus en détails dans cette section, certaines ligases RtcB requièrent l'association de chaperonnes, telles que Archease [**35**], pour fonctionner convenablement.

Mécanisme d'épissage des ARNt catalysé par RtcB. L'épissage des ARNt chez les eucaryotes est initié lorsque les pré-ARNt sont coupés à la frontière intron-exon par le complexe ARNt-endonucléase formé de quatre sous-unités (Sen2, Sen34, Sen15, and Sen54 [**58**]) [**39**]. Le seul intron présent dans l'ARNt, d'une longueur de 14 à 60 nucléotides, est ensuite retiré du pré-ARNt [**58**]. L'exon en amont a une extrémité 2′, 3′-phosphate cyclique alors que l'exon en aval possède un groupement 5′-OH, tous deux étant des substrats pour RtcB dans la réaction décrite ci-dessus [**17, 30, 32**]. Une activité similaire pour l'épissage des ARNt a aussi été observée chez les archées [**31**]. Dans la voie d'épissage des ARNt, l'hélicase à ARN DDX1 hydrolyse l'ATP et maintient ainsi l'activité optimale de RtcB dans le complexe d'épissage [**37**]. Archease interagit également avec ce complexe et accélère l'activation et la guanylylation de l'extrémité 3′-PO_4. De plus, cette chaperonne permet à RtcB d'interagir avec des cofacteurs autre que le GTP tels que dGTP, ATP ou ITP [**36**].

Mécanisme d'épissage de l'ARNm *XPB1* catalysé par RtcB. Durant l'UPR chez les métazoaires, la kinase/endoribonucléase transmembranaire IRE1α du réticulum endoplasmique initie le retrait de l'intron contenu dans l'ARNm *XBP1*. Cette coupure crée une extrémité 2′, 3′-phosphate cyclique à l'extrémité 3′ de l'exon en 5′ et un groupement hydroxyle à l'extrémité 5′ de l'exon situé en 3′, qui peuvent être scellées directement par

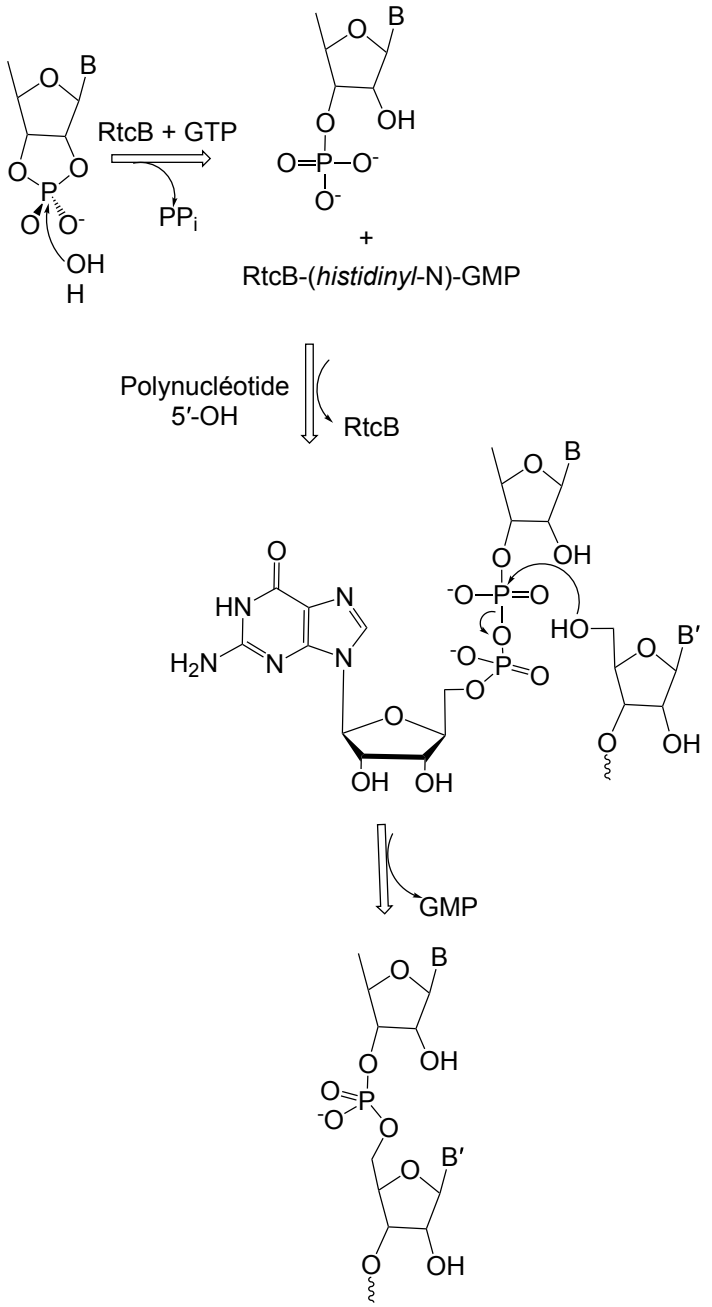

Fig. 1.2. Schéma illustrant le mécanisme de liaison des fragments d'ARN par RtcB; adapté de [**56**]. Le mécanisme débute avec (i) l'hydrolyse du 2′, 3′-phosphate cyclique de l'ARN pour obtenir un 3′-PO_4; (ii) RtcB forme un intermédiaire covalent RtcB-(*histidinyl*-N)-GMP en relâchant un pyrophosphate du GTP; (iii) le guanylate est transféré au 3′-PO_4 du polynucléotide pour former un intermédiaire ARN polynucléotide-(3′)pp(5′)G et finalement, (iv) RtcB catalyse l'attaque de l'extrémité 5′-OH sur l'intermédiaire ARN N(3′)pp(5′)G pour former la jonction entre les fragments d'ARN et relâcher un GMP.

RtcB selon le mécanisme décrit ci-dessus [**41, 43**].

Mécanisme de l'ajout de coiffe à l'ADN catalysé par RtcB. La ligase EcRtcB est en mesure d'ajouter *in vitro* un GMP à l'extrémité 3′-PO_4 de l'ADN produisant ainsi un ADN(3′)pp(5′)G. Les étapes de l'ajout de la coiffe se déroulent tel que mentionné ci-haut, mais suite à l'addition du GMP à l'extrémité 3′-PO_4 de l'ADN, RtcB ne joint pas l'ADN(3′)pp(5′)G à un autre acide nucléique. Donc, l'ADN guanylate résultant peut être vu comme un produit intermédiaire de la réaction de ligation [**38, 54**].

Les ligases à ARN RtcB dépendantes ou indépendantes des Archeases. Tel que mentionné plus haut, la ligase RtcB retrouvée chez les mammifères interagit avec le cofacteur Archease afin de demeurer catalytiquement active. Cette chaperonne accélère la ligature de l'ARN en permettant les renouvellements multiples de RtcB [**35, 36, 37**]. Par conséquent, la suppression d'Archease par ARNi *in vitro* rend déficiente la maturation des ARNt par la ligase RtcB humaine [**37**]. Les ligases RtcB de *Thermus thermophilus* (TtRtcB) et de *Pyrococcus horikoshii* (PhRtcB) interagissent également avec Archease afin de maintenir leur activité [**36**]. Or, certaines bactéries telle *Thermobifida fusca*, n'expriment pas d'Archease et l'activité de leur ligase RtcB n'est pas affectée par la présence d'une Archease hétérologue [**36**]. Ainsi, on classifie les enzymes RtcB en tant que ligases Archease-dépendantes ou Archease-indépendantes. Chez les ligases Archease-dépendantes, les sites de liaison Archease-RtcB sont hautement conservés, puisque l'Archease de *P. horikoshii* est capable d'activer la ligase RtcB de *T. thermophilus* et ce, malgré une identité de séquence relativement faible entre les Archeases de ces deux organismes [**36**].

1.2.2.4. Structure et domaines protéiques des ligases à ARN RtcB

Des structures cristallographiques ont été obtenues à partir d'enzymes RtcB provenant de bactéries et d'archées **(voir Figure 1.3)**. La ligase RtcB de l'archée *P. horikoshii* possède un repliement α/β caractéristique qui est lui-même encerclé de deux feuillets β. Des résidus histidines bien conservés, interagissant avec les ions Mn^{2+} et le GTP, sont présents au niveau du site actif de toutes les séquences homologues des enzymes RtcB étudiées [**11, 38, 59**]. La pochette active de la ligase RtcB est hydrophile en plus d'être profonde et large [**32, 56**]. Cinq résidus bien conservés, Asn95, Cys98 et trois His (His203, His234 et His329) coordonnent deux ions Mn^{2+}. Six autres acides aminés conservés, Phe204, Glu206, Gly 379, deux Ser (Ser 380 et Ser385) et Lys480, entrent en contact avec la nucléobase de la guanine, les résidus Ala406, Gly407 et Tyr451 orientent le ribose du GTP alors que His404 et deux Asn (Asn202 et Asn330) coordonnent les phosphates du GTP [**38, 59**]. Des résidus clés ont aussi été identifiés, par l'entremise d'ions sulfate, comme coordonnant les

phosphates de l'ARN (Gly69, Tyr70, Gly93, Tyr94, Asn95, Gly199, His203, Arg238, Gly239, Lys351 et Arg412) [**38, 59, 60**]. Chez la ligase RTCB/HSPC117 humaine, Cys122 est un résidu clé, puisque la mutation de cet acide aminé inhibe l'activité ligase de cette enzyme [**30**]. Il s'agit d'un des résidus qui coordonnent les ions Mn^{2+} [**61**], et est l'équivalent du Cys98 de PhRtcB mentionné ci-haut. L'analyse structurale de la ligase RtcB de *Pyrococcus* a révélé que le mécanisme de formation de l'intermédiaire de réaction par un système assisté d'ions métalliques ressemble à celui observé chez les ligases à ARN ATP-dépendantes [**62**].

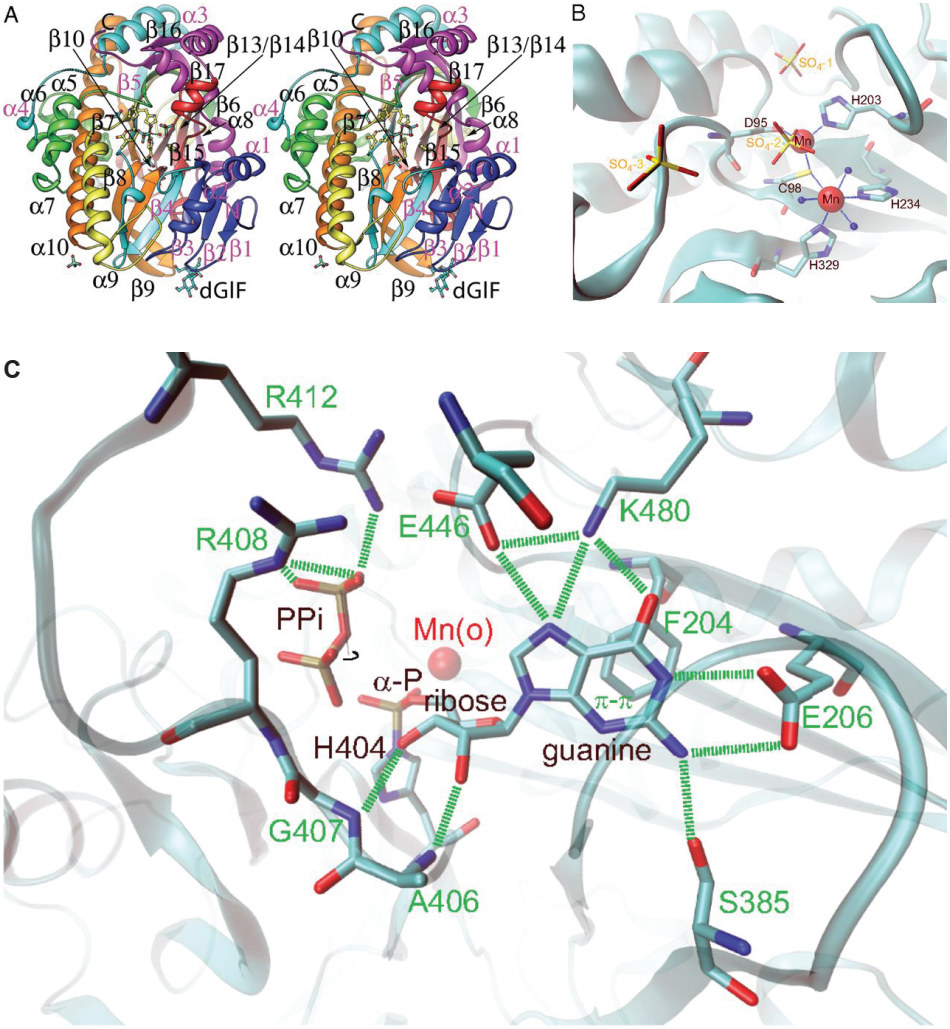

Fig. 1.3. Deux structures cristallographiques de la ligase RtcB de *Pyrococcus horikoshii* couplée avec ses cofacteurs respectifs; adapté de [**59**]. **A)** Première structure de la ligase RtcB de *P. horikoshii* couplée avec des ions Mn^{2+} et quatre ions sulfates. **B)** Une vue rapprochée du site actif de la structure présentée A) où sont identifiés les résidus interagissant avec les ions Mn^{2+}. **C)** Seconde structure de la même ligase RtcB couplée avec des ions Mn^{2+}, des ions sulfates et du GMP. Les résidus interagissant avec le ribose et la base guanylée du GMP ainsi qu'un pyrophosphate (PP_i) sont identifés en vert.

1.2.2.5. Stratégies de purification des ligases à ARN RtcB

La plupart des données sur les rôles biologiques [**voir section 1.2.2.2**], le mécanisme moléculaire [**voir section 1.2.2.3**] et la structure protéique [**voir section 1.2.2.4**] des ligases RtcB sont basées sur des essais enzymatiques ou des structures cristallographiques. Pour effectuer ces manipulations, l'enzyme devait être isolée et purifiée. Par exemple, la ligase RTCB/HSCP117 a été isolée à partir d'extraits de cellules HeLa qui démontraient une activité de ligature. Ces extraits ont été récoltés par chromatographie d'interactions hydrophobes, d'affinité, et échangeuse d'ions en plus d'utiliser des colonnes de dessalage [**30**]. De façon plus simple, les RtcB d'organismes procaryotes (par exemple *Pyrococcus* et *Myxococcus*) ont été exprimées dans, et isolées, à partir d'un organisme autre que celui dont elles proviennent originalement, notamment par surexpression dans des cellules d'*E. coli* [**31, 32, 38, 59**]. Les séquences codant pour les RtcBs possédaient toutes un tag His [**31, 32, 38, 59**] et certains de ces construits incorporaient un tag supplémentaire soit SUMO3 (« Small Ubiquitin-like Modifier 3 », aussi appelé Smt3 provenant de *S. cerevisiae*) [**32, 38**]. Le tag His sert à la purification des protéines par chromatographie d'affinité alors que le tag SUMO3 est utilisé afin de rendre la protéine soluble [**63**]. Suite à la chromatographie d'affinité, les protéines sont purifiées davantage par filtration sur gel ou par chromatographie à échange d'ions [**31, 32, 38, 59**].

1.3. *Diplonema papillatum*

Les organismes où l'on suspecte un rôle inouï des ligases RtcB sont les protistes appelés diplonémides. En rappel au système traditionnel de classification des eucaryotes, celui-ci se divisait en quatre grands règnes, soit : les animaux, les plantes, les champignons, et les protistes [**64**]. Le groupe des « protistes » incluait tous les eucaryotes qui n'appartenaient pas aux quatre autres règnes et est aujourd'hui reconnu d'être composé d'un vaste nombre des taxons extrêmement diversifiés desquels émergent les animaux, les champignons, et les plantes [**65**]. Cet ancien système taxonomique classifiait les protozoaires de façon négative, et donc des révisions ont été faites au cours des dernières décennies. Le groupe appelé diplonémides a récemment été découvert comme faisant partie des eucaryotes les plus abondants et diversifiés des océans [**66, 67, 68**]. Les diplonémides appartiennent au superphylum Discoba et au supergroupe Discicristata (qui comprend les heteroloboseans), et plus précisément au groupe des Euglenozoaires, qui inclut aussi les kinétoplastides et les euglénides **(voir Figure 1.4)** [**65**]. Depuis les dernières années, les diplonémides ont été étudiées non seulement par rapport à leur phylogénie et leur morphologie [**67, 69**], mais aussi au niveau génomique. Leurs génomes mitochondriaux se sont avérés particulièrement singuliers [**68, 70**]. La plupart de ces études ont été faites chez *Diplonema papillatum*,

l'espèce modèle de cette famille.

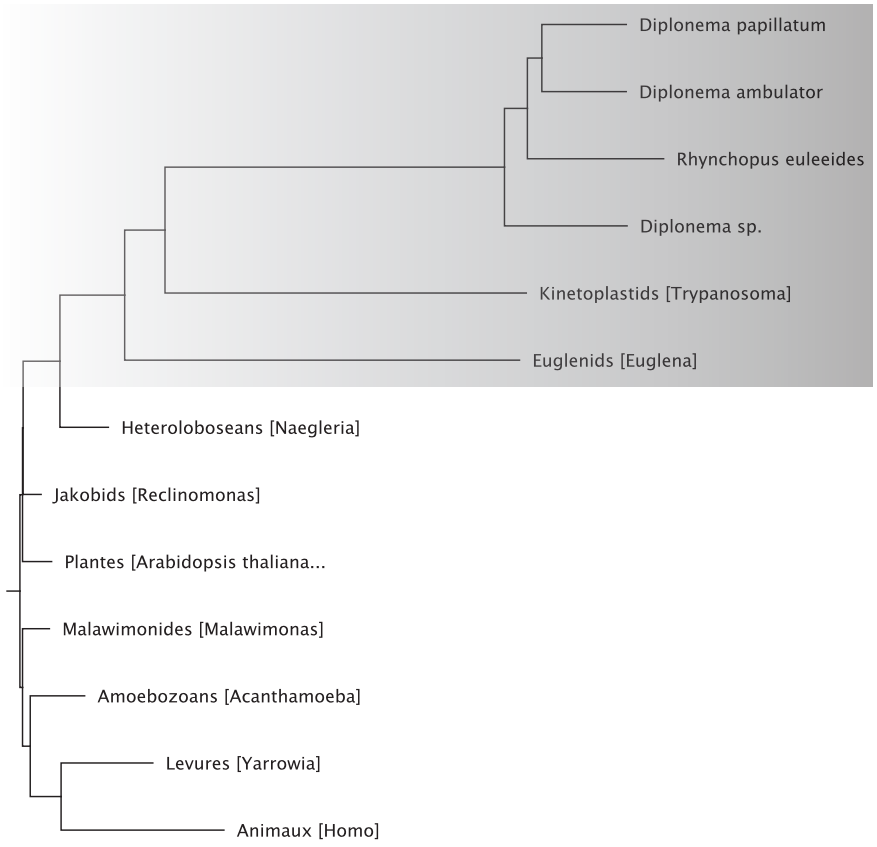

Fig. 1.4. Arbre phylogénétique des eucaryotes adapté de [**65, 71**]. Chaque superfamille mentionnée est représentée par une espèce spécifique indiquée entre parenthèses. Les Euglenozoaires (en gris) sont constitué de la famille des diplonémides (*Diplonema papillatum*, *Diplonema ambulator*, *Diplonema sp.* et *Rhynchopus euleeides*), des kinétoplastides (*Trypanosoma brucei*) et des euglénides (*Euglena gracilis*).

1.3.1. Le génome nucléaire de *D. papillatum*

Récemment, notre groupe a séquencé, assemblé et annoté le génome nucléaire de *D. papillatum* **(publication en préparation)**. La taille du génome est approximativement 180 Mpb, étant ∼7 fois plus grand que le génome de *Trypanosoma brucei* et ∼17 fois moins grand que le génome humain [**70, 72**]. Tous les gènes essentiels aux eucaryotes semblent être présents. Les données publiées à l'heure actuelle concernant le génome nucléaire des diplonémides sont limitées. Les données de séquençage complet du génome ont été obtenues (par séquençage de cellule unique) de la famille des Eupelagonemidae, qui sont phylogénétiquement très distants du groupe de diplonémides considérées comme « classiques » duquel fait partie *D. papillatum* [**73, 74**]. L'annotation et l'assemblage des

génomes nucléaires des membres de cette famille demeure toutefois fragmentaire. Le génome le plus complètement assemblé possède une lacune à plus de 90% de copies uniques de gènes orthologues pratiquement universellement conversés entre les espèces [**73**]. Toutefois, l'identité de certains gènes nucléaires chez les diplonémides ont aussi pu être obtenus à l'aide de méthodes de séquençage traditionnelles [**66, 75**].

1.3.2. Le génome mitochondrial de *D. papillatum*

Notre groupe a également séquencé le génome mitochondrial (ADNmt) de *D. papillatum* qui est composé de 81 chromosomes distincts de 6 kpb (classe A) ou 7 kpb (class B) de longueur, totalisant 600 kpb [**72, 76**]. Chacune de ses molécules d'ADN contient 95% de séquences répétitives. Les 5% restant correspondent à une séquence unique longue de 160-640 pb appelée « cassette ». Chaque cassette contient une partie d'un gène qui code pour une protéine ou un ARNr. Ces régions codantes appelées « modules » sont longues de 43-534 pb. Dans la cassette, les deux côtés des modules sont flanqués par approximativement 50 pb de séquences non-codantes [**77, 78**].

Les gènes mitochondriaux codent pour des protéines de la chaîne de phosphorylation oxydative (OXPHOS), soit les complexes I, III, IV, ATP synthase (*atp6*, *cob*, *cox1-3*, *nad1*, *nad4*, *nad5*, *nad7* et *nad8*) en plus de coder pour les ARNr de la petite et la grande sous-unité du mitoribosome (*rnl* et *rns*) [**78, 79**]. L'ADNmt de *D. papillatum* contient aussi six autres gènes (*y1-y6*), pour lesquels les produits protéiques demeuraient jusqu'alors inconnus. Nous avons découvert que ces protéines très divergentes font partie du Complèxe I de la chaîne respiratoire (*nad2* (*y3*), *nad3* (*y1*), *nad4L* (*y6*), *nad6* (*y5*) et *nad9* (*y2*)), Y4 étant possiblement une protéine accessoire de l'OXPHOS unique à *D. papillatum* [**71**].

1.3.3. L'épissage en *trans* de l'ARN mitochondrial de *D. papillatum*

D. papillatum utilise systématiquement l'épissage en *trans* pour générer les ARNm et ARNr mitochondriaux en joignant deux à onze courts transcrits (modules) **(voir Figure 1.5 pour les détails concernant la jonction des modules)** [**72, 79**]. Les transcrits mitochondriaux primaires *D. papillatum* (générés à partir de chromosomes différents) ne sont pas flanqués par des introns comme dans l'épissage en *trans* traditionnel [**1**]. Ce faisant, les structures secondaires ou motifs d'épissage caractéristiques d'introns connus (i.e., les introns spliceosomaux, du groupe I, du groupe II ou archéaux/ARNt) sont absents [**7, 80**]. Étonnamment, considérant la complexité de ce processus, les données expérimentales

indiquent qu'il n'y a virtuellement aucun désalignement ou jonction erronée de modules [**7, 81**]. Malgré que le processus de l'épissage en *trans* a été caractérisé en détail par notre équipe, le mécanisme moléculaire demeure toujours inconnu.

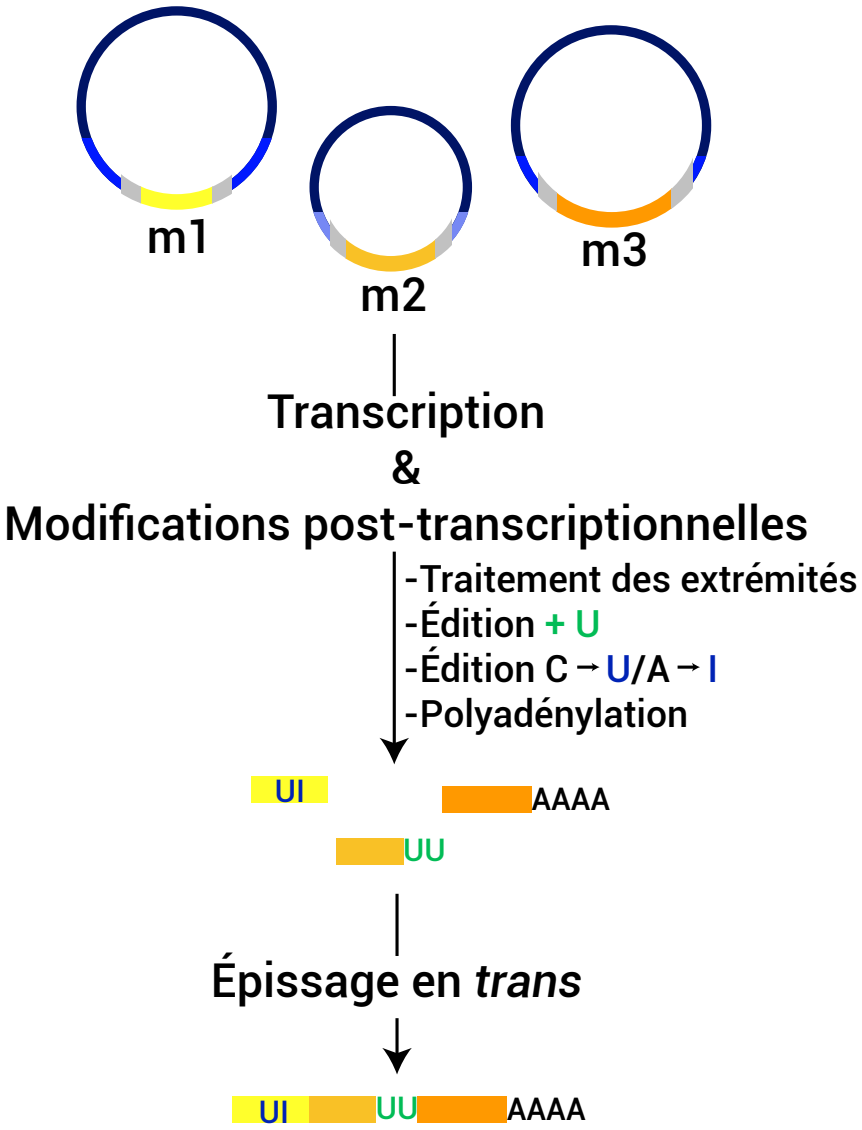

Fig. 1.5. Traitement des modules d'ARN mitochondrial et épissage en *trans* chez *Diplonema papillatum*; adapté de [**79**]. La maturation de l'ARNm mitochondrial débute avec la transcription de chaque cassette individuelle, chacune contenant un module. Ensuite, les nucléases retirent les portions non-codantes de l'ARN primaire afin de ne laisser que le module, qui code pour une partie de la protéine ou de l'ARNr. En poursuivant, la maturation implique l'édition de l'ARN, ce qui consiste en des substitutions de nucléotides en plus d'ajouts d'uridines aux modules. Finalement, la portion 3′ terminale de chaque ARNm est polyadénylée et les modules sont joints ensemble par épissage en *trans* afin de former l'ARNm complètement mature.

1.3.4. L'édition de l'ARN mitochondrial chez *D. papillatum*

Certains transcrits mitochondriaux chez *D. papillatum* sont fortement édités dans leurs séquences codantes. Ces changements varient de substitutions C-à-U et A-à-I (adénine-à-inosine) à l'addition d'une extension d'U (ajout de 1 à >50 nucléotides). Lorsque le génome mitochondrial entier est pris en considération, 15 des 18 gènes sont édités et ∼350 positions sont changées ou ajoutées [**7, 82**]. Ces modifications post-transcriptionnelles sont extrêmement précises et reconstituent la structure secondaire des ARNr, des cadres de lecture et des codons stop. En bref, l'édition rend les ARNm et ARNr non seulement fonctionnels, mais aussi considérablement plus similaires à leurs homologues chez d'autres organismes **(voir Figure 1.5 pour les détails concernant l'édition des modules)** [**7, 82**]. Malgré que le phénomène de l'édition de l'ARN a été étudié par notre groupe en détail, le mécanisme moléculaire sous-jacent reste énigmatique.

1.4. Résultats préliminaires et hypothèses de recherche

Tel que mentionné précédemment, la majorité des espèces encodent une seule ligase RtcB qui exerce apparemment plusieurs fonctions distinctes dans la cellule. Par contre, certains organismes possèdent plusieurs homologues de RtcB, ce qui nous amène à proposer que chacune de ces ligases exerce une fonction spécialisée. *Diplonema papillatum* est une des rares espèces dont le génome inclut plusieurs homologues de RtcB.

En effet, les résultats préliminaires de notre équipe de recherche indiquent que les trois ligases RtcB chez *D. papillatum* ont des fonctions distinctes. La première (DpRTCB1) possède un signal d'import mitochondrial prédit et est présente dans la mitochondrie selon nos données de spectrométrie de masse [**Burger *et al.*, non publié**]. De plus, la séquence génomique *DpRTCB1* est bien conservée chez les autres diplonémides recensées. La seconde ligase (DpRTCB2) est, au niveau de sa séquence protéique, plus similaire aux enzymes connues chez la famille des enzymes RtcB des métazoaires et des archées. La séquence de la troisième ligase (DpRTCB3) est similaire à celle de RtcB3 chez *Myxococcus*, qui est capable d'ajouter des coiffes aux acides nucléiques. DpRTCB3 se retrouve vraisemblablement dans le cytosol et/ou dans le noyau.

Nous postulons que DpRTCB1 est impliquée dans l'épissage en *trans* de la mitochondrie en joignant les différents modules constituant l'ARNm et l'ARNr. Étant donné que certaines ligases de type RtcB sont déjà impliqués dans la ligature des ARNm du cytosol, il n'est pas impossible que ce soit aussi le cas pour des transcrits mitochondriaux. De plus, des évidences en laboratoire ont démontré que les modules obtenus suivant la transcription

du génome mitochondrial possèdent des extrémités 3′-PO_4 et 5′-OH [**83**]. Étant donné sa localisation dans la mitochondrie, DpRTCB1 aurait des affinités pour des substrats et des facteurs distincts si on la compare aux ligases RtcB localisés dans le cytosol et le noyau. Ces différences seraient aussi reflétées dans sa structure protéique et son activité enzymatique. Nous postulons aussi que DpRTCB2 est responsable de l'épissage des ARNt ayant lieu dans le noyau et possiblement, de l'épissage de non-conventionnel d'ARNm au niveau du cytosol (comme cela a déjà été observé pour l'ARNm *XPB1* chez les métazoaires).

1.5. Buts et objectifs

Dans une perspective globale de découvrir de nouveaux mécanismes moléculaires, notre laboratoire vise à déceler le rôle biologique et l'enzymatique sous-jacente des enzymes RtcB chez *Diplonema*. La plus intéressante des trois ligases étant DpRTCB1, parce que, comme expliqué ci-haut, cette ligase est la candidate la plus probable d'être impliquée dans l'épissage en *trans* des transcrits mitochondriaux. Dans un premier temps, nous voulons comprendre les différences structurales entre DpRTCB1 et les ligases RtcB précédemment caractérisées chez des organismes modèles. Dans un second temps, il s'agira de produire DpRTCB1 en quantité suffisante en vue d'une future caractérisation fonctionnelle. Étant donné que DpRTCB2 est possiblement plus proche de la ligase PhRtcB que DpRTCB1, et que celle-ci est déjà bien caractérisée au niveau de son activité enzymatique, nous avons tout d'abord tenté d'optimiser les conditions de purification pour DpRTCB2 avant de les tester sur DpRTCB1.

Ainsi, les objectifs de mon projet de maitrise sont :

(1) Modéliser *in silico* la structure tridimensionnelle de DpRTCB1 avec une emphase sur ses résidus fonctionnels afin de pouvoir faire des prédictions sur les substrats et cofacteurs potentiellement requis par cette enzyme.

(2) Développer un protocole pour surexprimer et purifier DpRTCB1 en quantité suffisante pour des essais enzymatiques *in vitro*.

Chapitre 2

« *In silico* three-dimensional structure prediction of RtcB RNA ligases in *Diplonema papillatum* »

Préface au chapitre 2

Dans cet article, les structures tri-dimensionnelles des trois ligases de type RtcB présentes chez *D. papillatum* ont été prédites et étudiées afin d'envisager les rôles biologiques potentiels et les cofacteurs nécéssaires pour ces ligases. L'enzyme DpRTCB1 est présentée comme la protéine candidate responsable de la ligation des transcrits mitochondriaux lors de l'épissage en *trans*. Nous proposons ainsi un nouveau rôle biologique pour les ligases dépendantes du GTP. Il est aussi proposé que cette enzyme puisse utiliser des cofacteurs autre que le GTP. Nous présentons aussi un tout nouveau système de classification des ligases de type RtcB.

Présentation de l'article 1

Manuscript title: *In silico* three-dimensional structure prediction of RtcB RNA ligases in *Diplonema papillatum*"

Authors: Alexandra Léveillé-Kunst[1], Matus Valach[2] and Gertraud Burger[3]

Robert-Cedergren Center for Bioinformatics and Genomics and Department of Biochemistry and Molecular Medecine, Université de Montréal, Montréal, Québec, Canada.

[1]For further correspondence: alexandra.leveille-kunst@umontreal.ca
[2]For further correspondence: matus.valach@umontreal.ca
[3]For further correspondence: gertraud.burger@umontreal.ca

Liste et contributions des auteurs:

Alexandra Léveillé-Kunst: A fait l'analyse et la modélisation *in silico* des enzymes DpRTCB. A fait l'analyse et établi la distribution phylogénétique des ligases de type RtcB. A fait l'analyse des alignements multiples.

Matus Valach: A fourni les séquences des ligases de type RtcB prédites chez les diplonémides et les autres protozoaires mentionnés dans le manuscrit à partir des données transcriptomiques. A contribué à l'analyse de la distribution phylogénétique des ligases de type RtcB.

Manuscrit en préparation

Une validation expérimentale de la substitution du résidu serine 320 en cystéine chez la ligase de type RtcB de *E. coli* doit être completée.

2.1. Abstract

RtcB RNA ligases are crucial enzymes involved in biological processes such as tRNA splicing, unconventional mRNA splicing and rRNA repair observed across the tree of life. Here we investigate three inferred RtcB proteins encoded in the genome of the marine eukaryote *Diplonema papillatum* (DpRTCB), an unusual situation since eukaryotes carry almost always only a single representative of this protein family. To reveal distinctive features of diplonemid sequences, we analysed their phylogenetic affiliation and predicted their structures by *in silico* modeling. Combining this information, we infer that each protein plays a distinct biological role in *Diplonema*. Based on the phylogenetic distribution of RtcB-type ligases reported in the literature and described here, we propose a generalized classification system of this protein family predicting preferred substrates and cofactors of newly discovered RtcB-type proteins based on sequence information alone.

Keywords: RtcB, RNA ligase, classification, *in silico* structure prediction, diplonemids.

2.2. Introduction

RNA ligases are crucial enzymes acting in tRNA maturation and repair, unconventional mRNA splicing, rRNA repair and nucleic acid capping [**11**]. They are separated into two families with respect to their cofactors: members of the ATP-grasp family use ATP, whereas those of the RtcB-type family rely on GTP [**11**]. The ATP-grasp family has been extensively investigated over the past decades, which allowed to establish a robust classification system. In contrast, information about RtcB-type RNA ligases remains sparse, and no such framework exists for that protein family.

RtcB ligase-mediated tRNA maturation occurs in both metazoans and archaea in species such as humans (HsRTCB) or *Methanopyrus kandleri* (MkRtcB) [**30, 31**], while the unconventional splicing of the mRNA *XPB1* during the unfolded protein response (UPR) is confined to metazoans [**41, 43, 44, 46**]. The RtcB ligase in *Escherichia coli* (EcRtcB) has been reported to act in rRNA repair [**49**]. EcRtcB and two of the RtcB enzymes in *Myxococcus xanthus* (MxRtcB2 and MxRtcB3) are also able to cap nucleic acids *in vitro* [**38, 54**]. In general, RtcB ligases join RNA possessing 5′-OH ends to 3′-PO_4 or 2′, 3′-cyclic phosphate ends using GTP and Mn^{2+} as cofactors [**28, 32**]. To maintain full activity, certain RtcB ligases interact with proteins such as Archease or DEAD-box helicase 1 (DDX1) [**36, 37**]. Others like EcRtcB and MxRtcB1 do not require such factors [**27, 38**].

Two crystal structures are currently available for RtcB ligases: one from *Pyrococcus horikoshii* (PhRtcB, PDB ID: 4DWQ) [**60, 62**] and the other from *Thermus thermophilus* (TtRtcB, PDB ID: 2EPG). In both structures, RtcB enzymes display a specific fold composed of three β-sheets and at least 20 α-helices, one of which is considerably longer than the others (in PhRtcB, this helix possess 35 residues and measures a total of 50 Å) [**60, 62**]. On the surface of the protein, a characteristic groove composed of positively charged residues passes through the active site that is located in a deep hydrophilic pocket [**60, 62**]. The two conserved residues D95 and C98 (numbering relative to PhRtcB) coordinate the Mn^{2+} ions, and three histidines (H203, H234 and H329) and residues A406, G407 and Y451 coordinate the ribose of the GTP. The phosphates of the GTP are aligned by two asparagines (N202 and N330) together with H404, whereas the guanine base is aligned by six residues, F204, E206, G379, S380, S385, and K480 [**60, 62**]. In the structure, certain residues were also observed to interact with SO_4^{2-} ions, which were to mimic phosphate RNA groups. These amino acids include G69, Y70, G93, Y94, D95, G199, H203, R238, G239, K351 and R412, where D95 and H203 also interact with the Mn^{2+} ions and GTP, respectively [**38, 59, 60**].

Most organisms analysed to date have a single gene coding for an RtcB-type ligase. Only in certain bacteria have multiple RtcB-coding sequences been reported [**38**]. However, diplonemids appear to be an exception. Diplonemids are marine unicellular eukaryotes that belong to the Euglenozoan clade, which in turn, is part of the supergroup Discicristata (which also contains the heterolobosean clade) and the larger superphylum Discoba [**65**]. Diplonemids are among the most abundant eukaryotes in the ocean [**66, 68**] and are known for their unique *trans*-splicing mechanism of mitochondrial RNA transcripts [**77, 78**]. Nuclear genome analysis of the type species *Diplonema papillatum* **(Burger, Lukes, Williams *et al.*, unpublished)** indicates that this organism encodes three proteins of the RtcB family. This micro-eukaryote also possesses a gene coding for an ATP-grasp RNA ligase, which was previously characterized by bioinformatic means [**84**].

The mitochondrial genome of *D. papillatum* is composed of 81 chromosomes of 6 to 7 kbp. Only about 5% of each chromosome is coding and this coding sequence represents a single piece (referred to as module) of a gene [**76, 79**]. Each chromosome is transcribed separately. During transcript maturation, non-coding sequences are removed leaving only the coding sequence [**72, 79**]. Module transcripts also undergo end-processing and heavy editing with C-to-U and A-to-I substitutions and additions of U nucleobases. Subsequently, modules are ligated to a contiguous, conventional transcript via *trans*-splicing [**81, 82, 85**]. While the *trans*-splicing process has been characterized in detail, the actors involved remain unknown.

It has been proposed that module joining during diplonemid mitochondrial *trans*-splicing is catalyzed by an RNA ligase, which interacts not only with RNA substrates but also other cofactors (e.g., helicases) in a postulated larger complex called the "joinosome" [**70**]. We test here the hypothesis that the corresponding RNA ligase is a member of the RtcB family, and we provide support for RTCB1 being the joinosome ligase. We will pinpoint sequence and three-dimensional (3D)-structure differences between the three *Diplonema* DpRTCB, predict their cofactors, and determine their structural and phylogenetic relationship to other RTCB proteins whose biochemical function and biological role are known. We find that DpRTCB1 is distinct not only from the other DpRTCB but also from of the RtcB family members in other organisms, thus suggesting a novel biological role for RtcB ligases.

2.3. Materials and methods

2.3.1. Sub-cellular localization prediction

DpRTCB sequences were deduced from the genome and transcriptome data from *D. papillatum* **(Burger, Lukes, Williams *et al.*, unpublished)**. Inferred RTCB sequences from other diplonemids were retrieved from deposited transcriptomic data [**86**]. Mitochondrial targeting signals were predicted using TargetP v2.0 [**87**], MitoFates v1.1 [**88**], Predotar v10.4 [**89**], MITOPROT v1.101 [**90**], DeepLoc v1.0 [**91**], PredSL [**92**] and WoLF PSORT [**93**]. To determine the best suited parameters, we tested each software with a set of proteins from the TriTrypDB database for *Trypanosoma brucei* that are known to be imported in mitochondria. For TargetP and PredSL, we used parameters recommended for plant sequences. For Predotar, we used 'non-plant' parameters and for WoLFPSORT we used 'animal' parameters. MitoFates produced the same results with plant and animal parameters. For DeepLoc and MITOPROT, default parameters were used. The heatmap used to visualize the prediction signal was obtained with Heatmapper **(http://www.heatmapper.ca/expression/)**.

2.3.2. Three-dimensional structure prediction for DpRTCB ligases

Three-dimensional structure predictions for DpRTCB proteins and EcRtcB were obtained with the Protein Homology/analogY Recognition Engine v2.0 (Phyre2) [**94**] using intensive parameters. The predicted mitochondrial target signal was removed from the sequence of DpRTCB1 prior to modeling. The crystal structure of PhRtcB (PDB ID: 4DWQ) was obtained from the PDB database and used as a reference [**62**]. As a second, independent method of 3D-structure prediction, we used SWISS-MODEL [**95**]. The

obtained structures were visualized with either UCSF ChimeraX v1.0rc202005091901 [**96**] or UCSF Chimera v1.14 [**97**]. The models were superimposed with the MatchMaker tool of Chimera using default parameters. A two-dimensional topological model was obtained from the database PDBsum [**98**] for PhRtcB (PDB ID: 4DWQ).

2.3.3. Protein sequence collection

RtcB sequences from across the tree of life were downloaded from the RefSeq (NCBI Reference sequence database for proteins) for: bacteria (22 824), animal (761), archaea (843), protists (119), viruses (87), fungi (23) and plants (17). To expand the under-represented protists, we further included 186 sequences assembled and annotated [**99**] from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) [**100**]. Lastly, ten additional sequences from jakobids and malawimonads were identified in recent transcriptome data [**101**]. During sequence clean-up, sequences were verified through multiple sequence alignments using MUSCLE v3.8.1551 [**102**] if the sequence pool was <500 sequences, and using MAFFT v7.471 [**103**] if the sequence pool was larger. The alignments obtained were visualized with AliView v1.26 [**104**]. To reduce the complexity of the dataset, sequences were clustered with CD-HIT v4.8.1 [**105**] at a threshold of 80% (protists, animals, fungi, plants, viruses, and archaea) or 70% (bacteria) sequence identity. Inteins were manually removed from archaeal, bacterial and viral sequences based on the alignments. A number of dubious sequences were eliminated from the dataset, notably (i) <400 and >600 residue-long proteins (i.e., partial sequences and those with mis-annotated termini) were removed using SeqKit v0.13.2 [**106**]; (ii) all sequences that were annotated as "isoform", "partial", "multispecies", "low quality" or possessed unknown amino acids in the protein sequence (e.g., "X"); (iii), poorly aligning sequences and those containing gaps at residues in the active site; (iv), obvious contaminants (e.g., clearly bacterial proteins labeled as originating from a metazoan such as RefSeq sequence XP_018028639.1). The final dataset, including several curated reference RtcB sequences, as well as diplonemid RTCB **(Supplementary Table 2.2)**, sums up to a total of 1,225 sequences (906 sequences from bacteria, 135 from archaea, 117 from ∼100 protistan groups; 14 from animals; 16 from fungi; 10 from plants and 27 from viruses). Multiple sequence alignments of curated RtcB sequences were visualized with CLC Viewer v7.7.1 (Qiagen) or COBALT from the National Center for Biotechnology Information (NCBI) [**107**]. Sequence logos were generated with WebLogo3 [**108**]. Pairwise sequence identity between DpRTCB and other RtcB sequences was determined using blastp from the NCBI BLAST tools [**109**].

2.3.4. Phylogenetic analyses

Phylogenetic trees of RtcB ligases were constructed with IQ-TREE v2.0.3 [**110**] and RAxML v8.2.11 [**111**]. The multiple sequence alignment used to build the trees was obtained with MAFFT (see above) followed by sequences cleaning with trimAL v1.4.rev15 [**112**] to discard alignment columns creating more than 20% gaps. For the tree built by IQ-TREE, we used the best-fit model prediction (LG + I + G model) and ultrafast bootstrap with 1000 replicates. For the tree built by RaxML, we selected the LG model with PROTGAMMA parameters and empirical base frequencies and bootstrapping with 100 replicates. Trees were displayed with FigTree v1.4.4 **(http://tree.bio.ed.ac.uk/software/figtree/)**. The sequences used as clade references are listed in **Supplementary Table 2.2**. Sequences used for tree construction and Newick files can be obtained at the following links: **https://figshare.com/s/c0c1cabf69e870268afd** & **https://figshare.com/s/ed37fc5a34f89a15c33c**.

2.4. Results and Discussion

2.4.1. Sequence analysis of DpRTCB proteins

We analyzed in more detail the protein sequences of predicted diplonemid RTCB sequences by investigating the sub-cellular location to which these proteins are potentially targeted. RTCB1 sequences in diplonemids are predicted by most computational methods to be imported into mitochondria **(Figure 2.1)** while RTCB2 and RTCB3 counterparts appear to be strongly targeted to the cytoplasm **(Supplementary Table 2.3** & **Supplementary Table 2.4)**. Interestingly, *Diplonema papillatum* and *Rhynchopus euleeides* are the only eukaryotes currently known to carry three nuclear genes coding for RtcB-type proteins.

The active site in RtcB-type ligases has been extensively mapped in four proteins (from *P. horikoshii*, *E. coli*, *T. thermophilus*, and human) using both biochemical and structural data. A multiple sequence alignment of all predicted diplonemid homologs along with PhRtcB and EcRtcB revealed that essentially all residues which interact with the Mn^{2+} ions and GTP within the active site of the proteins are conserved **(Supplementary Figure 2.7)**. Most of the residues known to interact with SO_4^{2-} ions in the crystal structure of PhRtcB (which do not natively interact with the ligase but mimic the RNA phosphates) are also strongly conserved in diplonemid RTCB except for two tyrosine residues in PhRtcB (Y70 and Y94) **(Supplementary Figure 2.7)**. One of these tyrosine residues (Y70) is conserved in diplonemid RTCB2 homologs while it is substituted for a lysine in RTCB1 homologs and EcRtcB (K53) and an arginine in RTCB3 homologs. The other tyrosine

(Y94) is not conserved in any homolog as it is replaced by a valine in RTCB1 homologs and EcRtcB (V74), a phenylalanine in RTCB2 homologs and a glycine or an asparagine in RTCB3 homologs. Interestingly, RTCB1 in diplonemids have a cysteine (C330) residue connecting with the ribose of the GTP instead of a serine (S385 in PhRtcB), which is absolutely conserved across all other inferred diplonemid RTCB1 proteins **(Supplementary Figure 2.7)**.

Considering the localization prediction and the presence of all residues within the active site for all predicted diplonemid homologs, in the following, we use *Diplonema papillatum* ligases as representative sequences for diplonemid RTCB. This will allow a more detailed examination of the differences in critical residues belonging to the active site and structural differences compared to known RtcB.

2.4.2. Structure models of DpRTCB proteins

To get further insight into the structural impact of the detected sequence variations, we predicted the structures of DpRTCB1, DpRTCB2, and DpRTCB3 by homology modeling. We also modelled the biochemically well-characterized EcRtcB to represent Archease-independent ligases and as a representative of Archease-dependent homologs, we used the crystal structure of PhRtcB **(Figure 2.2 A)**. All modeled structures were predicted with high accuracy (>90%) for >80% of the protein length **(Table 2.1)**. To describe the differences between the predicted models and PhRtcB we refer hereafter to the two-dimensional topology representation available from the PDBsum database for PhRtcB **(Figure 2.3)**.

To understand the signatures of Archease-dependent and Archease-independent enzymes, we compared the predicted model of EcRtcB and the crystral structure of PhRtcB. The three main β-sheets are present in both structures; only the model of EcRtcB is missing a β-branch in one sheet **(Figure 2.2 B)**. These sheets are surrounded by several α-helices. The most notable difference of the EcRtcB structure is the absence of five α helices behind the active site of the enzyme **(Figure 2.3)**. These α-helices are likely the interaction platform of PhRtcB with certain cofactors, especially Archease, which EcRtcB does not to require [**32, 36**].

2.4.2.1. DpRtcB1 structure model

DpRTCB1 shares more identical residues with EcRtcB (∼60%) than with PhRtcB (∼30% identity). The three β-sheets are absolutely conserved in the DpRTCB1 model, and

the longest α-helix in this structure is shorter by two residues when compared to its homolog in PhRtcB. Similarly, to the predicted structure of EcRtcB, DpRTCB1 is also missing five α-helices behind the active site **(Figure 2.2 C)**. Based on these similarities in sequence and structure, we propose that DpRTCB1 is also an Archaease-independent enzyme.

As mentioned above, the highly conserved S385 residue (in PhRtcB) is substituted for a cysteine (C330) in DpRTCB1. To analyze how this change could influence the interaction with the nucleobase of the GTP, we overlapped the active site of the predicted DpRTCB1 model and the crystal structure of PhRtcB **(Figure 2.4** & **Supplementary Figure 2.9)**. The residue S385 of PhRtcB overlapped perfectly with the C330 of DpRTCB1. Despite other residues varying in their orientation (e.g., H347 in DpRTCB1 and H404 in PhRtcB), all critical residues within the active site are conserved.

2.4.2.2. DpRtcB2 structure model

DpRTCB2 has a higher sequence identity to PhRtcB (∼50%) compared to EcRtcB (∼30% identity). The longest α-helix within the structure is conserved between PhRtcB and DpRTCB2. There is a region of 103 residues containing α helices near the N-terminus where one of the conserved β-sheets is typically found **(Figure 2.2 D)**. Near that same area, the β-sheet of DpRTCB2 is shorter by four strands compared to its homolog in PhRtcB. Since DpRTCB2 has the highest sequence identity to PhRtcB out of the three protein sequences analysed, it is highly unlikely that this β-sheet should be truncated as depicted in the Phyre2 model. Considering that SWISS-MODEL was able to predict the β-sheet in DpRTCB2 with an equal number of strands as for PhRtcB, the truncated sheet may be the result of issues during the modelling by Phyre2 **(Supplementary Figure 2.8 C)**.

Additionally, the surfaces of DpRTCB2 and PhRtcB [**60**] display a groove from the top of the structure to the bottom going through the active site, which is absent from the structures of EcRtcB, DpRTCB1 and DpRTCB3. Most of the residues on the surface of the structures are positively charged and the residues within the active site are hydrophilic. This distinctive mark on the surface density of PhRtcB and DpRTCB2 **(Supplementary Figure 2.10)** suggests that interactions with the RNA substrates or a protein co-factor for both these enzymes are different in EcRtcB, DpRTCB1 and DpRTCB3.

2.4.2.3. DpRtcB3 structure model

DpRTCB3 shares relatively few identical residues with either PhRtcB (∼30%), or EcRtcB (∼30%); however, it is much more similar to MxRtcB3 (∼50% identity), a ligase

which was demonstrated *in vitro* to possess capping capability as opposed to RNA ligation [**38**]. All three major β-sheets are conserved in DpRTCB3 and the longest α-helix lacks five residues when compared to its homolog in PhRtcB **(Figure 2.2 E)**. However, out of all the structures, DpRTCB3 is the one which differs the most from PhRtcB. When compared to PhRtcB, DpRTCB3 has two extensions towards each end of the structure: an extension of 19 amino acids near the C-terminus and, near the N-terminus, an extension of 83 residues predicted to form three β-strands, folding into an additional β-sheet and two α-helices. Such extensions are not uncommon in RtcB-type ligases, however, this specific extension found in DpRTCB3 may indicate a different role for this enzyme compared to its other two *Diplonema* homologs.

2.4.3. Phylogenetic relationships of RtcB-type sequences

To understand the diversity of RtcB family members, we analyzed the phylogenetic distribution of RtcB-type proteins, along with diplonemid proteins, across the tree of life. We used two maximum likelihood (ML) approaches **(Figure 2.5** & **Supplementary Figure 2.11)**, which produced essentially identical results as to the phylogenetic affiliation of RtcB sequences. Differences between the two methods primarily reside in the rooting of certain clades and the bootstrapping values for each clade in the phylogenetic tree. Most nodes connecting the clades in the phylogenetic tree have high bootstrap values (>80) indicating strong statistical support. In the following, we use a nomenclature of RtcB proteins with a prefix referring to the species and a specific number for each distinct homolog **(for more details on the reference sequences used see Supplementary Table 2.2)**.

In the trees, RtcB proteins form several distinct phylogenetic clades **(Figure 2.5)**, which we numbered from I to IX. Some sequences diverged into clearly distinct, but closely related branches, so we additionally used letters to indicate such cases. The first clade, I.A, is formed primarily by bacterial sequences, such as MxRtcB1 and EcRtcB along with proteins from the Discicristata supergroup, which branch together with diplonemid RTCB1 homologs. The grouping of bacterial sequences with euglenozoan and heterolobosean proteins (i.e. NgRTCB1) may suggest a horizontal gene transfer (HGT) of the bacterial-type RtcB to a common ancestor of euglenozoans or discicristates. Curiously, the Clade I.A also contains a single sequence of metazoan origin from the early-diverging coral *Pocillopora damicornis*, also suggesting HGT, likely from bacteria.

Two other clades branch closely to I.A. Clade I.B contains sequences belonging mostly to algae, dinoflagellates, and diatoms, while Clade I.C contains only bacterial sequences. MxRtcB2, one of the previously biochemically characterized enzymes [**38**], has an unstable position based on bootstrap values and branches with similar probability as sister to Clade I.A or groups together with either I.B or I.C.

Clade II includes known Archease-dependent RtcB homologs from archaea and metazoans [**36, 37**]. It contains the vast majority of archaeal sequences, including MkRtcB and PhRtcB, the bacterial proteins MxRtcB5 and TtRtcB, as well as almost all metazoan homologs. Within this clade, a sub-clade is formed by sequences from metazoans and numerous sequences of protistan origin, including diplonemid RTCB2.

Clade III contains sequences from bacteria (e.g., MxRtcB3), fungi, plants as well as diplonemids, Naegleria (DpRTCB3, ReRTCB3 and NgRTCB2), and two other protists. Half of these sequences have extensions of ∼60 and ∼30 amino acids at their N- and C-termini, respectively.

Clade IV, divided into IV.A (e.g., MxRtcB4) and IV.B, contains only bacterial sequences. Clade V, which contains MxRtcB6, is mostly composed of bacterial, but also includes several viral sequences. Most proteins in these two clades (∼60%) carry an N-terminal extension of ∼20 residues.

2.4.4. Sequence variation in the active site of distinctive phylogenetic groups

The other clades found in this tree have sequences which display conspicuous variations in the known active site residues, more specifically in the residues interacting with the GTP and Mn^{2+} ions. The multiple sequence alignment of several RtcB-type representatives of each clade illustrates these variations at sites 2, 5, and 8 **(Figure 2.6)**.

In the clades mentioned below, distinctions between residues known to interact with SO_4^{2-} ions which mimic RNA phosphates in the crystal structure of PhRtcB, were not as striking as the ones observed for amino acids interacting with the GTP or Mn^{2+} ions. Overall, in all sequences, residues interacting with the SO_4^{2-} ions were either highly conserved at >90% (G69, G93, D95, G199, H203, R238, K351 and R412 in PhRtcB), moderately conserved at ∼50% or ∼70% (Y70 and G239 respectively) or poorly conserved

at ∼15% (Y94).

Clade VI sequences (from bacteria, archaea, viruses, and protist) possess significant variations at site 5 **(Figure 2.6)**. In this clade, the first residue of the sequence motif, a Gly in other clades, is replaced by a bulky hydrophobic residue such as Ile, Phe, Leu or Met (or an infrequent Val or Tyr). This residue is followed by a well conserved Asn instead of the otherwise typical Ser. The last amino acid, a Ser (with ∼70% prevalence), remains well conserved with variation in several subclades (Ile, Ala, Val, Thr, Cys, Leu and Gly).

Clade VII sequences are exclusively from bacteria and have a well-conserved site 5 composition, except for its last residue, which is more often a Gly (∼75%) instead of the otherwise typical Ser (∼5%), though ∼20% sequences have an Ala, a Thr or a Cys instead. Other variations are present at the site 8 **(Figure 2.6)** and involve a missing or misaligned Lys in this clade. Although lacking other significant active site residue variations, sequences of this clade notably feature specific ∼70 and ∼10 residues-long N- and C-terminal extensions, respectively.

Clade VIII sequences, mostly of bacterial origin, have variations at site 2 **(Figure 2.6)** with a substitution of the last Glu residue to a Phe or a Tyr (or rarely a Ser). However, the most notable changes are observed at site 5 in a similar fashion as seen in Clade VI. The first residue changes to hydrophobic residues such as Leu or Met instead of a Gly, while the second residue is a well conserved Asn rather than a Ser. Importantly, the usual terminal Ser of the motif is swapped to the hydrophobic Ile or Val in all clade members. Similarly the situation in Clade VII, the typical Lys is missing at site 8. Sequence-wise, members of this clade have the most divergent active sites.

Lastly, Clade IX sequences, exclusively from bacterial origin, display changes at site 2, with the last residue, typically a Glu, being an extremely well conserved Asp. Other changes occur at site 5 where the first two residues are mostly conserved, with a few substitutions of the first residue from a Gly to an Ala or Ser (∼10%) and the second residue is infrequently changed from a Ser to an Asn or Thr (∼10%). Most of the changes at this site reside in the last residue where nearly half of the sequences contain an Ala instead of a Ser.

2.4.5. Mapping the known biological roles on the phylogenetic tree of RtcB

We were also interested in determining how RtcB-type proteins group together with family members for which a biological role has been confirmed experimentally, namely HsRTCB, TtRtcB, EcRtcB, PhRtcB, MkRtcB and MxRtcB1 to 3.

The well-characterized RNA ligases EcRtcB and MxRtcB1 belong to the Clade I.A; both these enzymes have been demonstrated to not require Archease to display full activity [**28, 32, 38**]. We consider most likely that all members of this clade are Archease-independent proteins. Both MxRtcB1 and EcRtcB have also demonstrated capping activity *in vitro* on the 3′-PO_4 end of DNA [**38, 54**]. While MxRtcB2 does not clearly group with either Clade I.B or I.C, we can still expect that proteins from these two clades possess a biological role similar to MxRtcB2. MxRtcB2 has displayed better propensity to ligate DNA than RNA *in vitro* when compared to MxRtcB1 and also has a higher capping activity at 3′-PO_4 end of DNA [**38**]. Considering that MxRtcB2 as well, does not require Archease [**38**], we extrapolate that the entire Clade I is composed of Archease-independent proteins.

Clade II regroups sequences such as PhRtcB, MkRtcB and HsRTCB, which are known to participate in the maturation process of tRNA [**17, 31**]. Metazoan homologs not only contribute to the maturation of tRNA [**17, 30**], but also act in the unconventional splicing of the mRNA *XPB1* during UPR [**41, 43, 44, 46**]. The human enzyme specifically requires the helicase DDX1 along with Archease to carry out its role in tRNA splicing [**37**]. Despite TtRtcB also belonging to this clade, any contribution of RtcB ligases to the maturation process of tRNA in bacteria has yet to be documented. Additionally, since Clade I and II are most distant phylogenetically from one another **(Figure 2.5)**, the members of these two clades may have distinct biological roles or interacting proteins (e.g., Archease).

Lastly, MxRtcB3 within Clade III was demonstrated to display high *in vitro* capping activity [**38**], so it is conceivable that this applies in general to proteins of this clade. Despite missing information on the biological roles of ligases belonging to the Clades IV to IX, the variability of their active site residues and other sequence features such as N- or C-terminal extensions indicate that these proteins may display biological roles different from those already identified. Future studies focusing on representatives of these clades should lead to interesting new discoveries.

2.4.6. Predicted roles for DpRTCB ligases

Altogether, the phylogenetic distribution, the predicted cellular targeting, as well as the analysis of the individual sequences and their predicted structures give us insight into the potential biological roles carried out by RtcB-type ligases in *D. papillatum*. Considering the phylogenetic distribution and the structural similarity of DpRTCB2 to metazoan RtcB and PhRtcB, both Archease-dependent ligases, we posit that DpRTCB2 plays a role in tRNA maturation in the cell and could also possess a function in unconventional mRNA splicing in the cytosol. On the other hand, DpRTCB3 can be expected to cap nucleic acids *in vitro* due to the sequence similarity and phylogenetic affiliation to MxRtcB3. However, the exact biological role of this enzyme in the cell remains to be determined.

The most interesting RtcB member in *Diplonema* remains DpRTCB1. Based on its phylogenetic affiliation, active site residues and predicted cellular targeting signal, DpRTCB1 is strongly suspected to play a role in mitochondrial *trans*-splicing [**70**]. The most intriguing aspect of this enzyme is the presence of a cysteine residue instead of the strongly conserved serine residue in the active site of DpRTCB1. Although this substitution also rarely occurs in the Clades VI and VII in some bacteria (mostly Firmicutes), it seems absent from the well-characterized Clades I, II or III. Thus, diplonemid RTCB1 is clearly distinct from these other bacterial proteins having the Ser-to-Cys substitution, since members of the Clades VI and VII have additional deviations of active site residues (see above). We hypothesize that a swap in the Cys residue in diplonemid RTCB1 could impact on the interaction of this ligase with its cofactors and result in the use of ATP (instead of or in addition to GTP) or affect the speed of the ligation reaction. Cysteine residues can also be employed as "redox switches" in bioenergetic organelles such as mitochondria and chloroplasts. For example, oxidized cysteines can modulate protein function in response to the changes in the environment [**113, 114, 115**] such as the ATP synthase in spinach chloroplasts which are modulated by a cysteine redox switch to reduce ATP hydrolysis [**115**]. If RTCB1 indeed participates in the diplonemid mitochondrial *trans*-splicing, the Ser-to-Cys substitution could have evolved for temporarily slowing down the respiratory chain by turning off the mitochondrial transcript assembly when mitochondrial reactive oxygen species accumulate, and thus allow a return to a functional state of the organelle.

2.5. Conclusion

We have established here nine major classes of RtcB proteins, distinguished by sequence and structure signatures and predicted biological roles, preparing the stage for experimental validation by detailed biochemical analysis of class representatives. Phylogenetic inferences,

sequence comparison and structure prediction of RtcB-type ligases point to DpRTCB1 as the most interesting family member to be examined biochemically. In particular, we predict that the unique cysteine residue (C330) in the active site of this ligase, which is strictly conserved across diplonemid RTCB1 allows this enzyme to use cofactors other than GTP. This can be readily tested by replacing the cysteine (C330) of DpRTCB1 by a serine, and vice versa the serine (S315) in EcRtcB by cysteine. Further, it would be worthwhile by a similar approach, to test experimentally if the five helices in the structure of PhRtcB but absent from those predicted for DpRTCB1 and EcRtcB, indeed provide the platform for Archease-RtcB interactions. Thus, our bioinformatics analyses provide not only the basis for targeted biochemical characterization of the biochemical function and biological role of RtcB-family enzymes in *D. papillatum*; they also lay the groundwork for a robust RtcB-classification system.

2.6. Figures

Fig. 2.1. Probability of mitochondrial targeting signal for diplonemid RTCB proteins. Blue and red hues indicate high and low probability of a mitochondrial targeting signal being present, respectively.

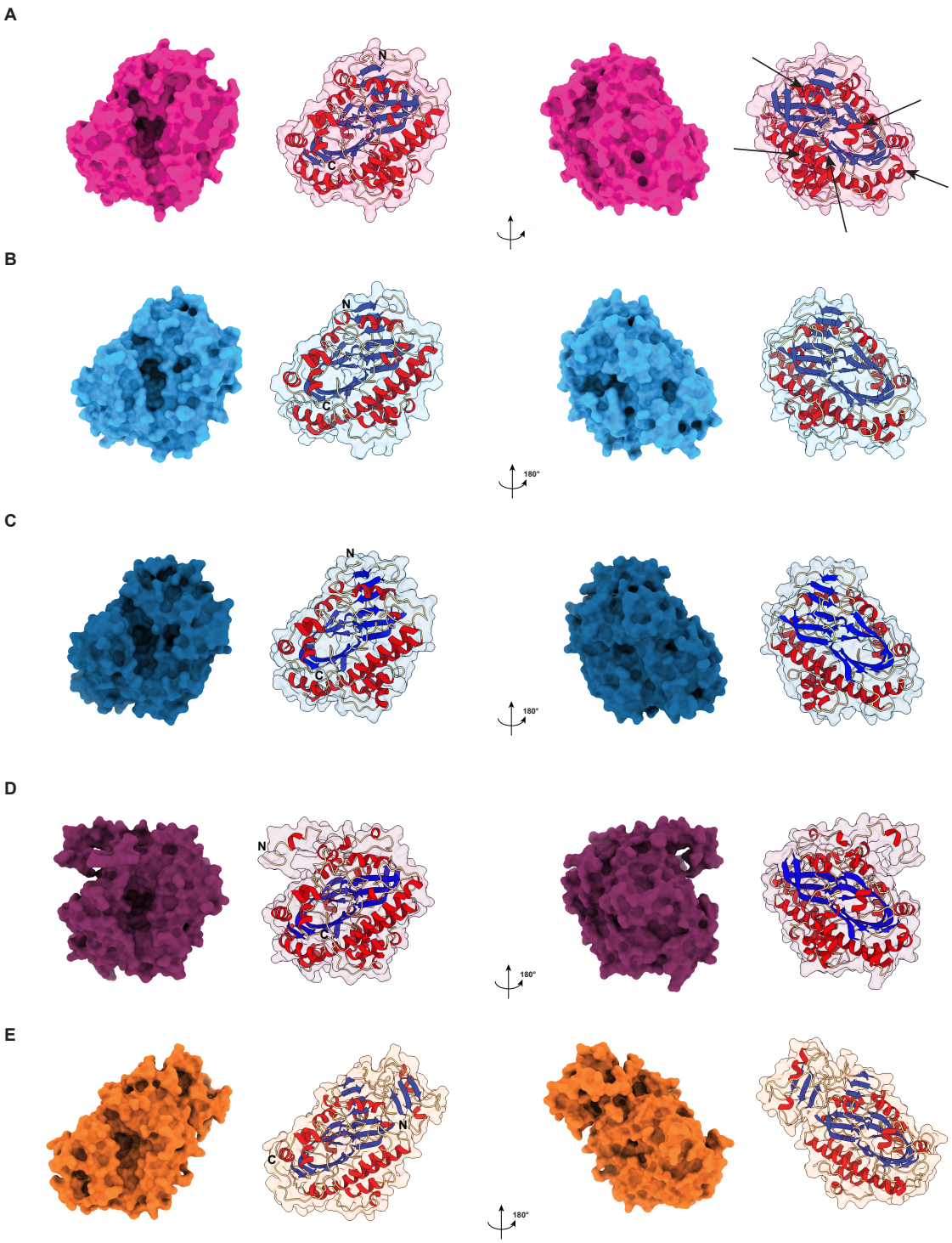

Fig. 2.2. Crystal structure of PhRtcB and models of other RtcB proteins predicted by Phyre2. **A)** Crystal structure obtained by X-ray diffraction of PhRtcB (PDB ID: 4DWQ). Helices mentioned to be lacking behind the active site in EcRtcB and DpRTCB1 are indicated with arrows. **B-E)** Predicted structures of: EcRtcB **(B)**, DpRTCB1 **(C)**, DpRTCB2 **(D)**, and DpRTCB3 **(E)**.

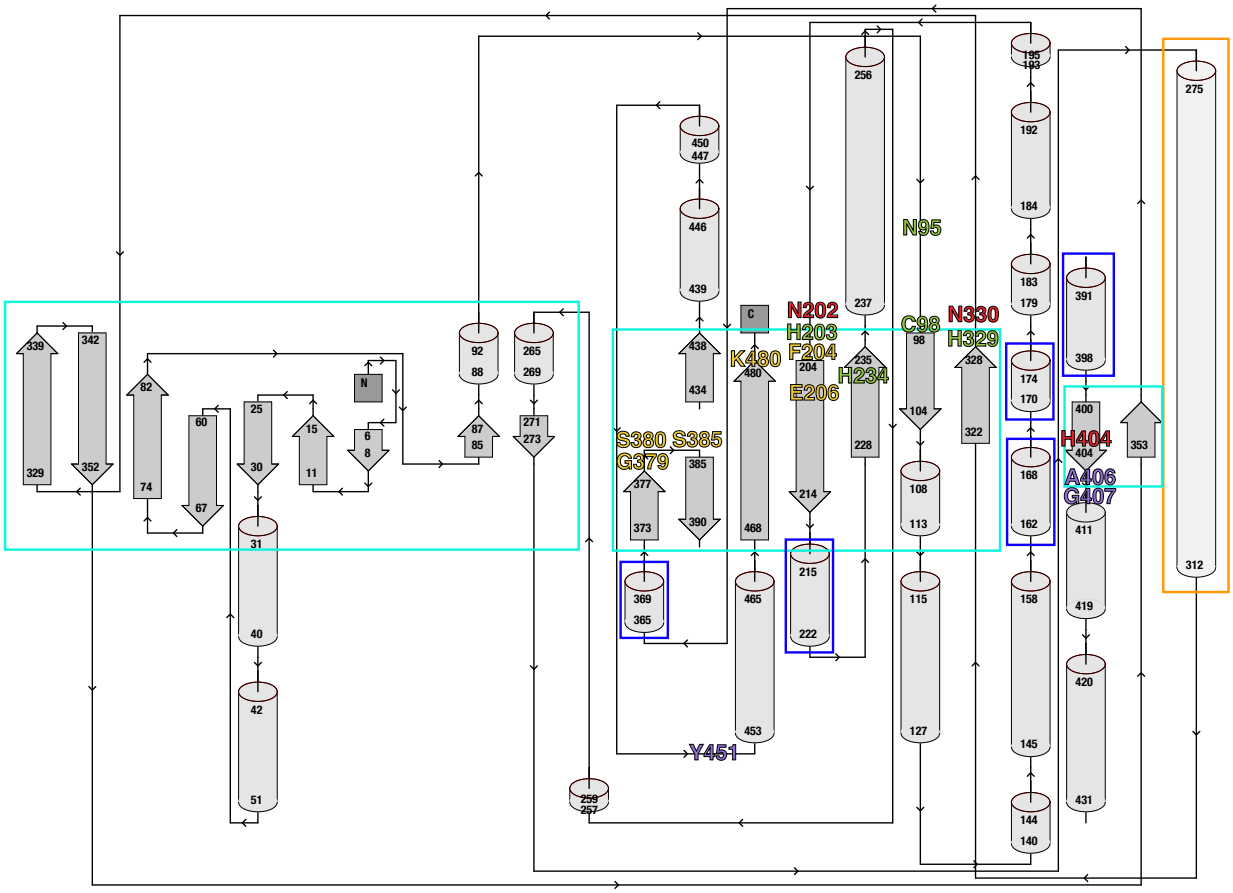

Fig. 2.3. Two-dimensional topological representation of PhRtcB obtained from the database PDBsum (PDB ID: 4DWQ). Delimitations in cyan represent the three conserved β-sheets. The α-helix contoured in orange is the longest helix in the structure. α-Helices in delimited in blue are present in PhRtcB but absent in EcRtcB and DpRTCB1. Residues critical to the active site of RtcB have been identified and are colored as those in **Supplementary Figure 2.7**.

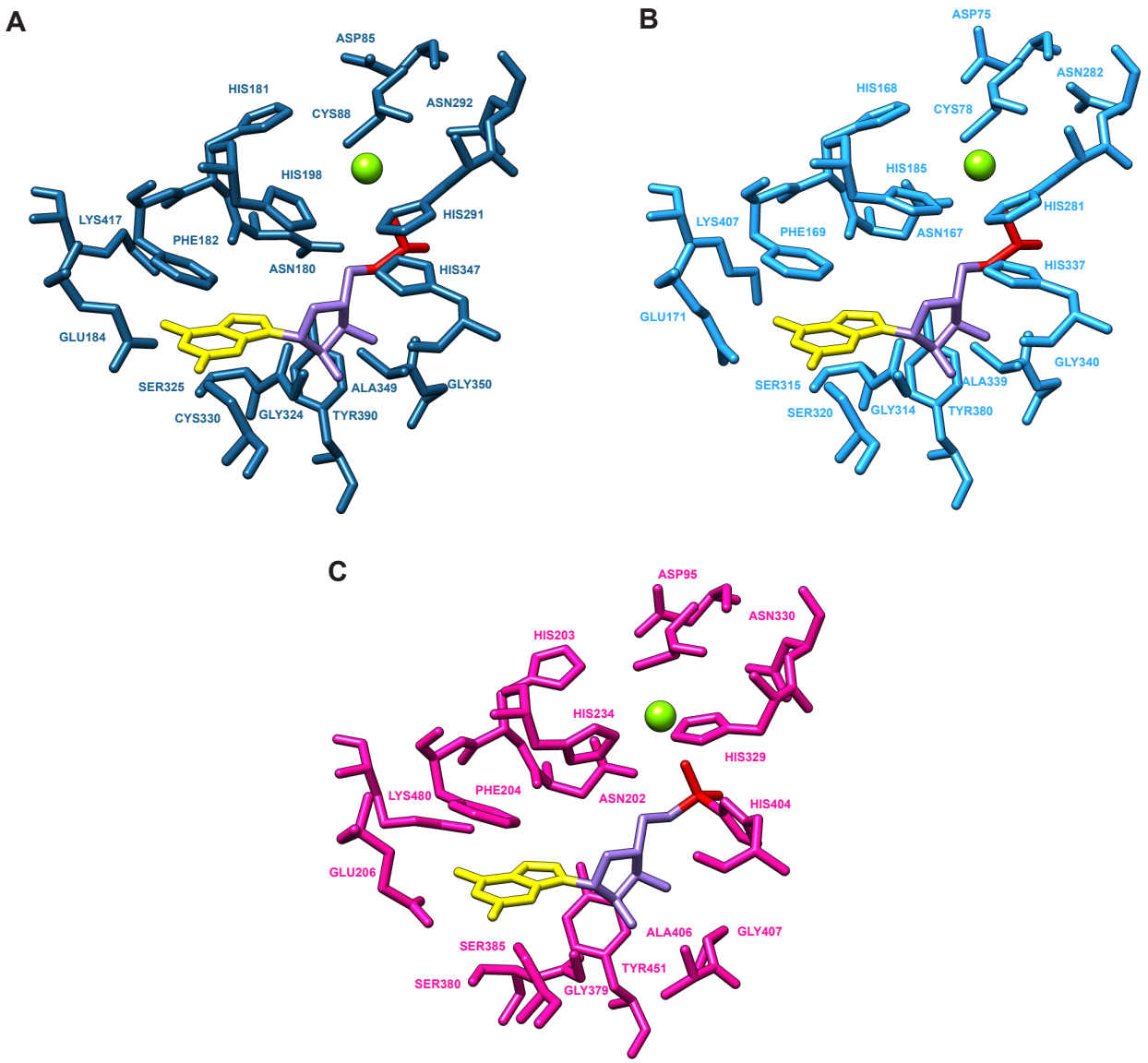

Fig. 2.4. View of the active site to visualize interactions of amino-acid residues with co-factors. Shown are the predicted models of **A)** DpRTCB1 and **B)** EcRtcB, as well as the crystal structure of **C)** PhRtcB. Cofactors are highlighted as in **Supplementary Figure 2.7** (GMP: guanine base in yellow, ribose in purple, phosphate in red, and manganese in green). Residues are numbered independently for each sequence.

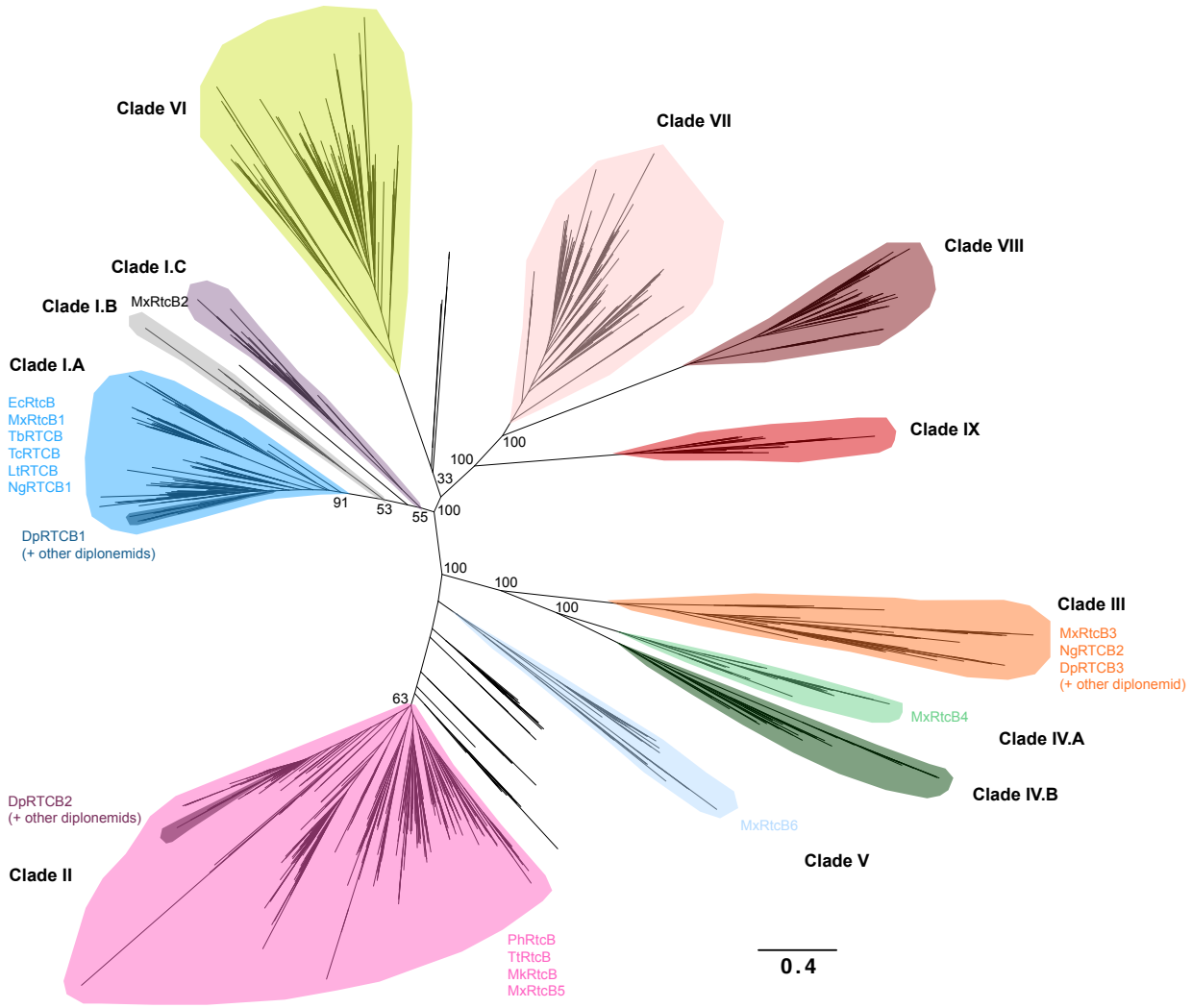

Fig. 2.5. Phylogenetic relationships among the RtcB-type family with color-shading indicating each RtcB clade. The tree was computed by IQ-TREE using 1225 sequences from bacteria, archaea, viruses, and diverse eukaryotes **(for details, see section 2.3 for Materials and methods and Supplementary Table 2.2)**.

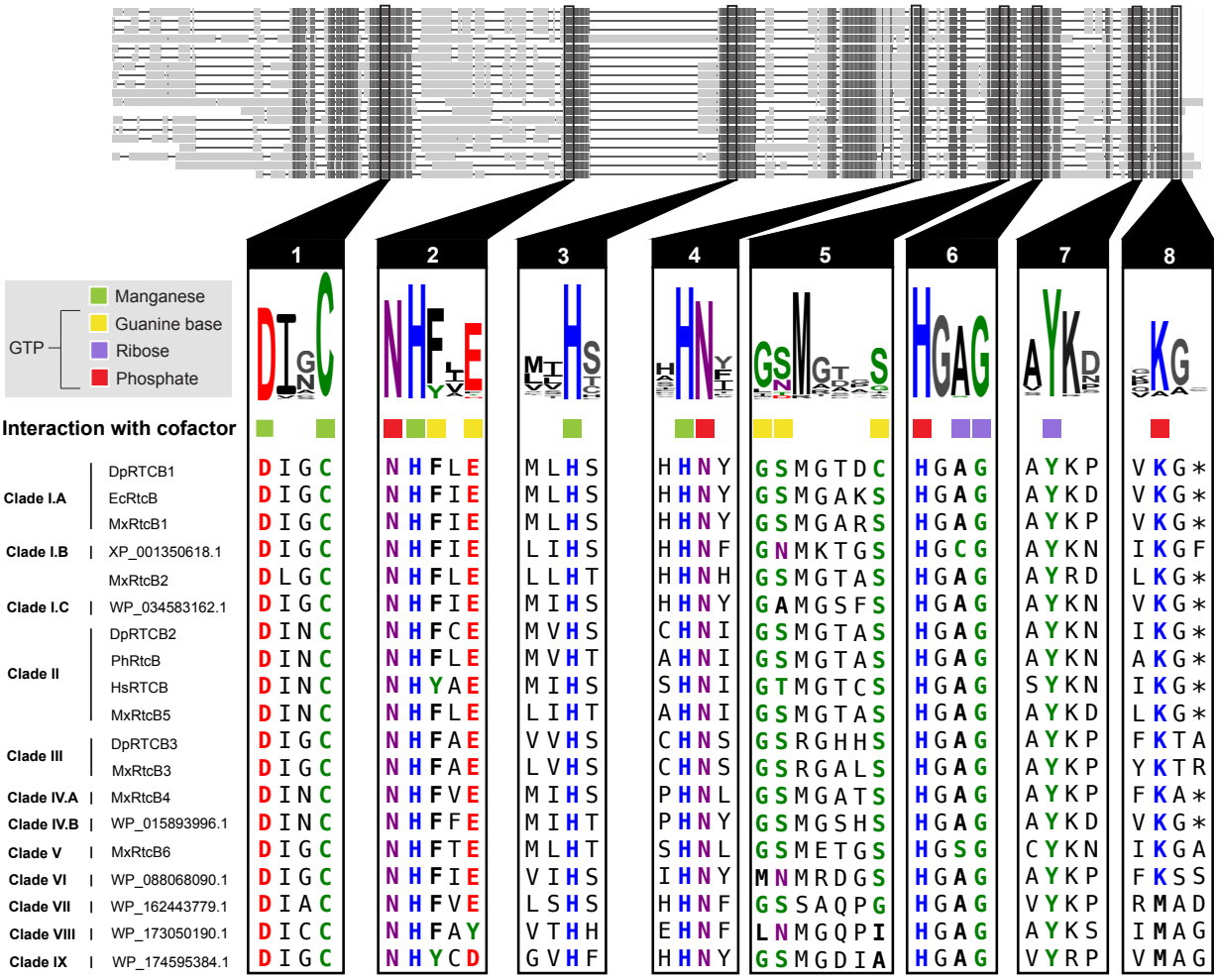

Fig. 2.6. Multiple sequence alignment of RtcB proteins with representatives of each clade **(see Figure 2.5)** along with *D. papillatum* RTCB sequences. Active site residues are identified according to their interaction with either the Mn^{2+} ions or GTP.

2.7. Tables

Table 2.1. Predicted model accuracy and number of residues accurately modelled by Phyre2.

Phyre2 structure	Total number of residues for the protein	Model prediction accuracy	Percentage of the sequence modelled with model prediction accuracy
EcRtcB	408	>90%	100%
DpRTCB1	418	>90%	100%
DpRTCB2	509	>90%	96%
DpRTCB3	517	>90%	80%

2.8. Supplementary data

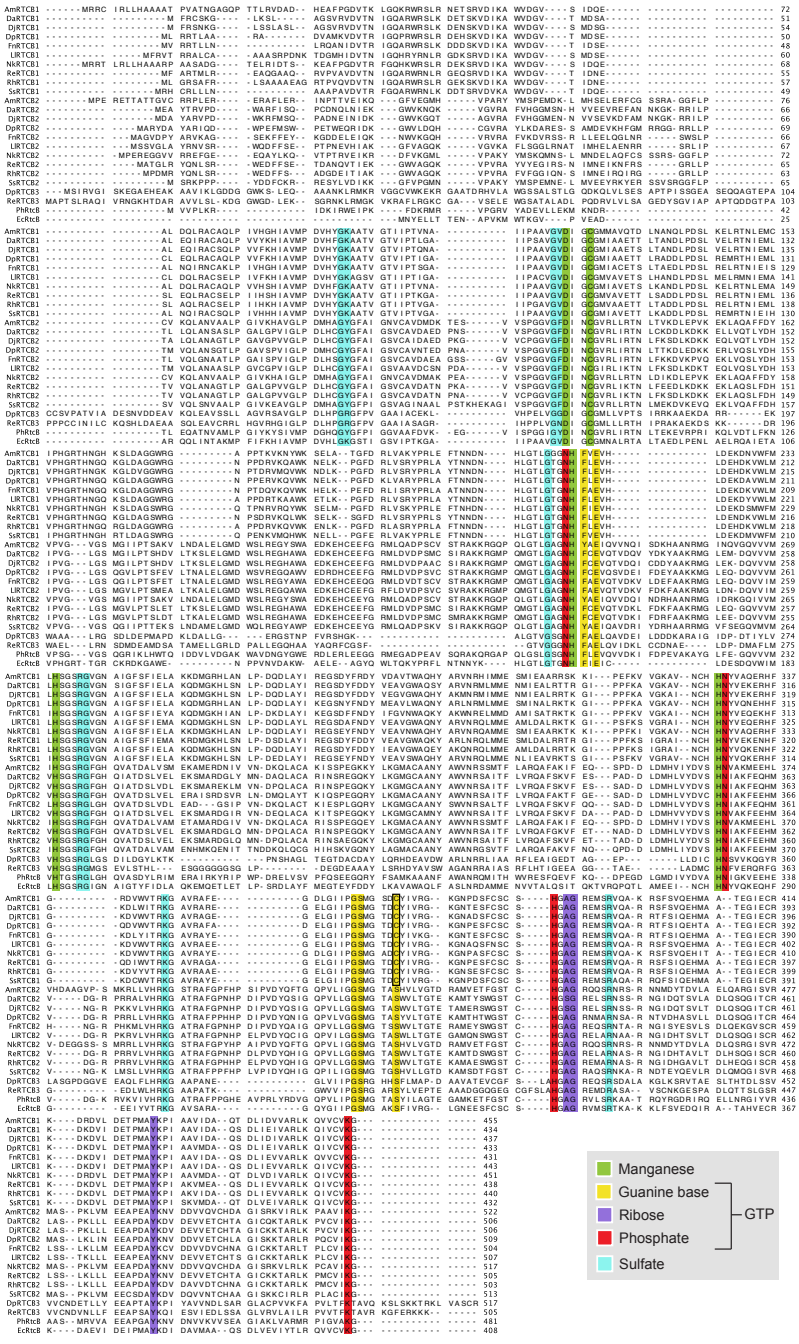

Fig. 2.7. Multiple sequence alignment of known RTCB protein sequences in diplonemids and PhRtcB. Conserved residues are identified according to their interactions with the Mn^{2+} ions (green), guanine base (yellow), ribose (purple), phosphate (red) and the crystallographic SO_4^{2-} ions which mimic phosphate RNA groups (light blue). The distinctive cysteine residue in diplonemid RTCB1 is framed in black. For more details on the sequences, see **Supplementary Table 2.2**.

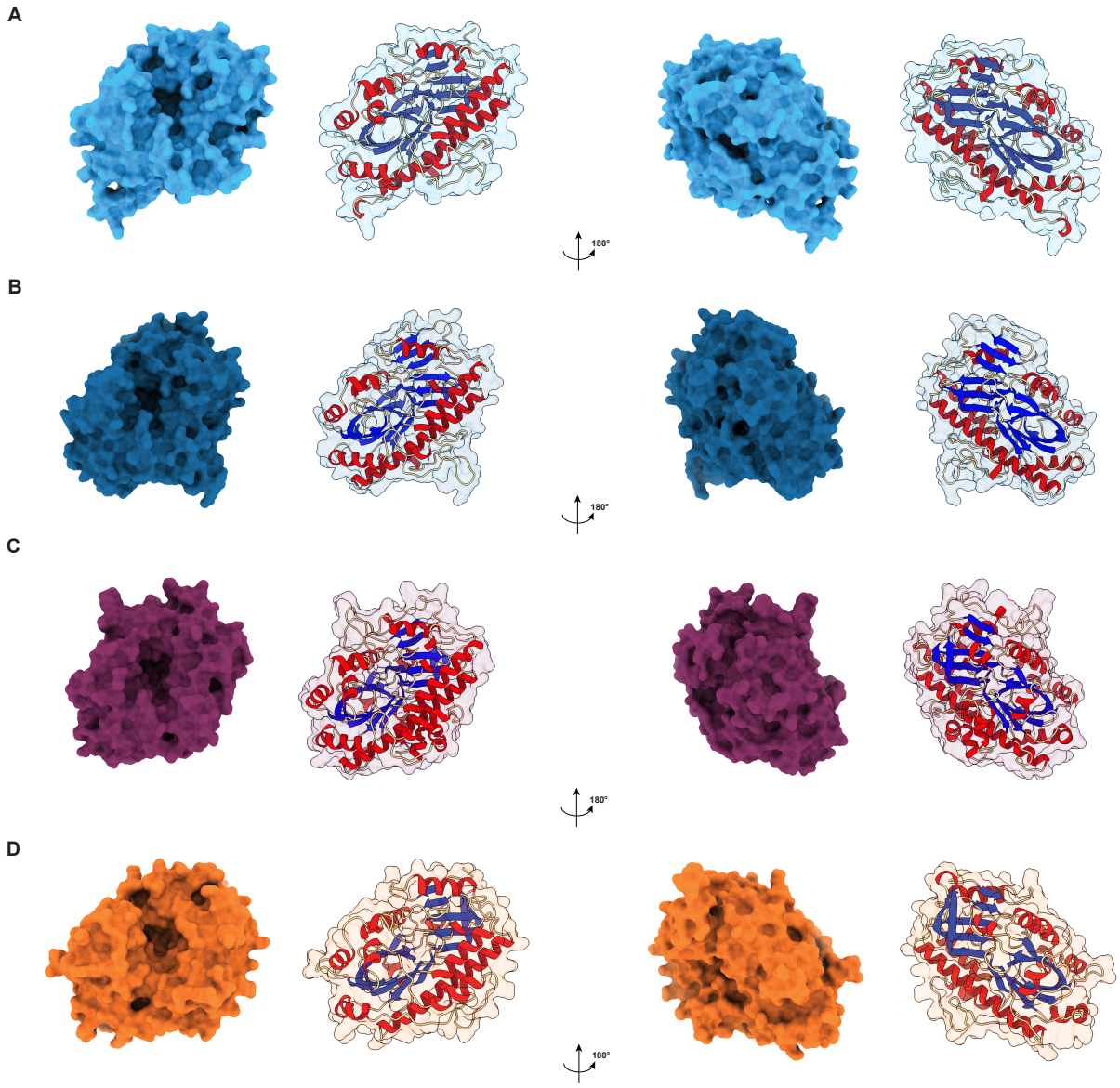

Fig. 2.8. Prediction models obtained with SWISS-MODEL. **A)** EcRtcB prediction model where in N-terminal 7 residues were not modelled. **B)** DpRTCB1 prediction model. **C)** DpRTCB2 prediction model where in N-terminal 9 residues were not modelled. **D)** DpRTCB3 prediction model where in N-terminal and in C-terminal, 105 and 17 residues respectively were not modelled.

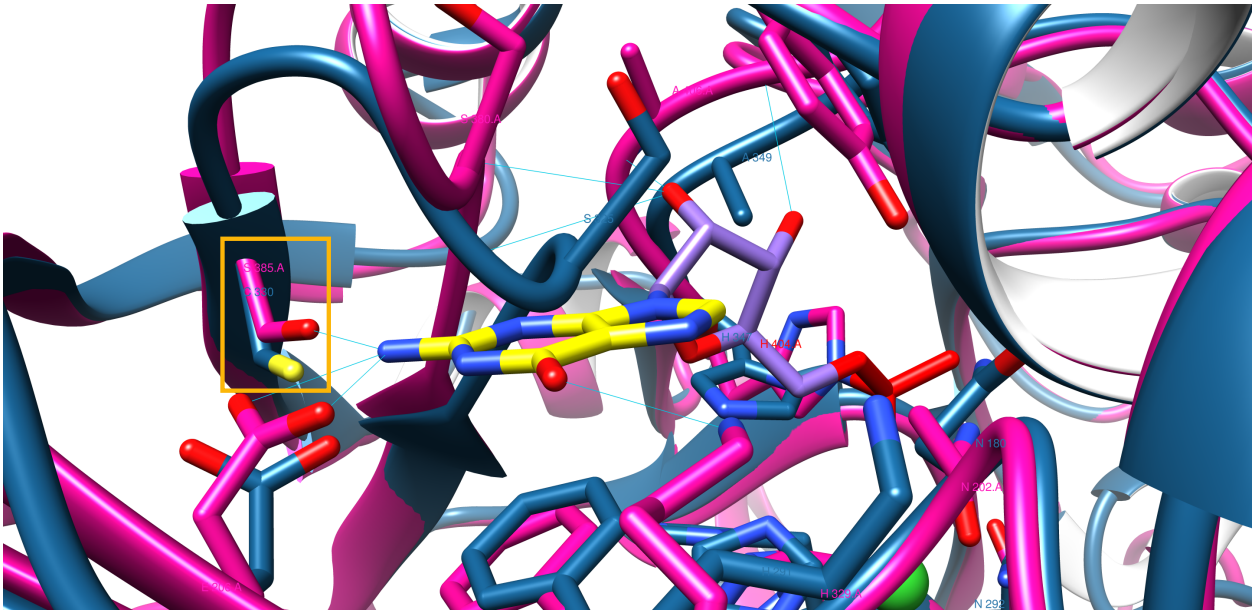

Fig. 2.9. Three-dimensional structure overlay of residues within the active site of DpRTCB1 (blue) and PhRtcB (pink). Structures are coupled with GMP (yellow for the guanine base, purple for the ribose and red for the phosphate) and Mn^{2+} (green). The cysteine residue specific to DpRTCB1 is delimited by the orange rectangle. The hydrogen bonds between the GMP and the residues in the crystallographic structure of PhRtcB are represented by cyan lines.

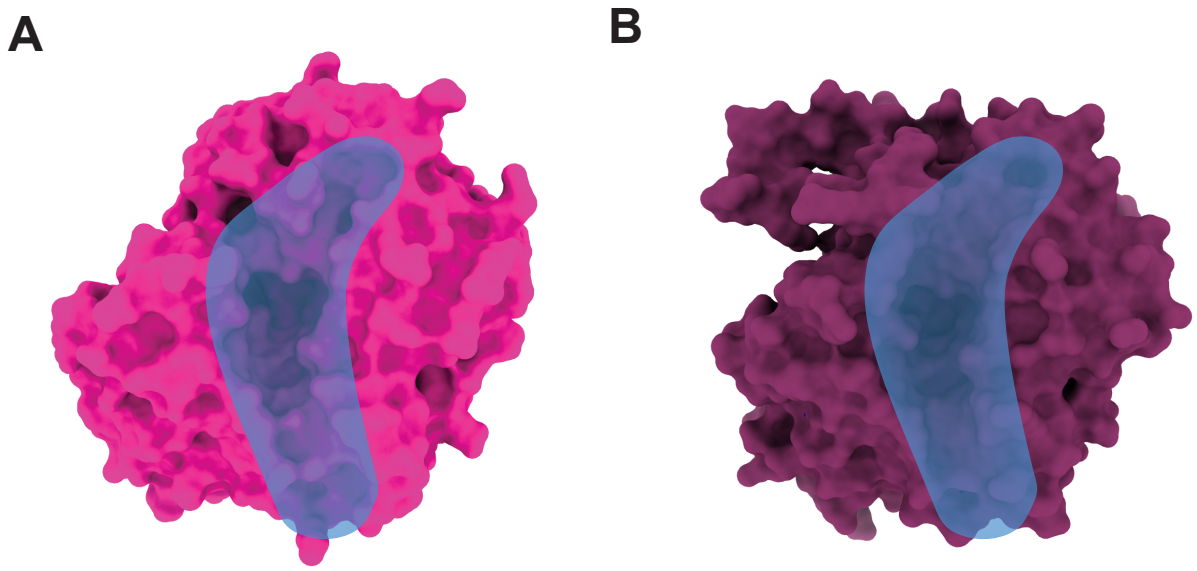

Fig. 2.10. Surface density representation of **A)** PhRtcB and **B)** DpRTCB2 with the groove present in the vicinity of the active site pocket highlighted in blue.

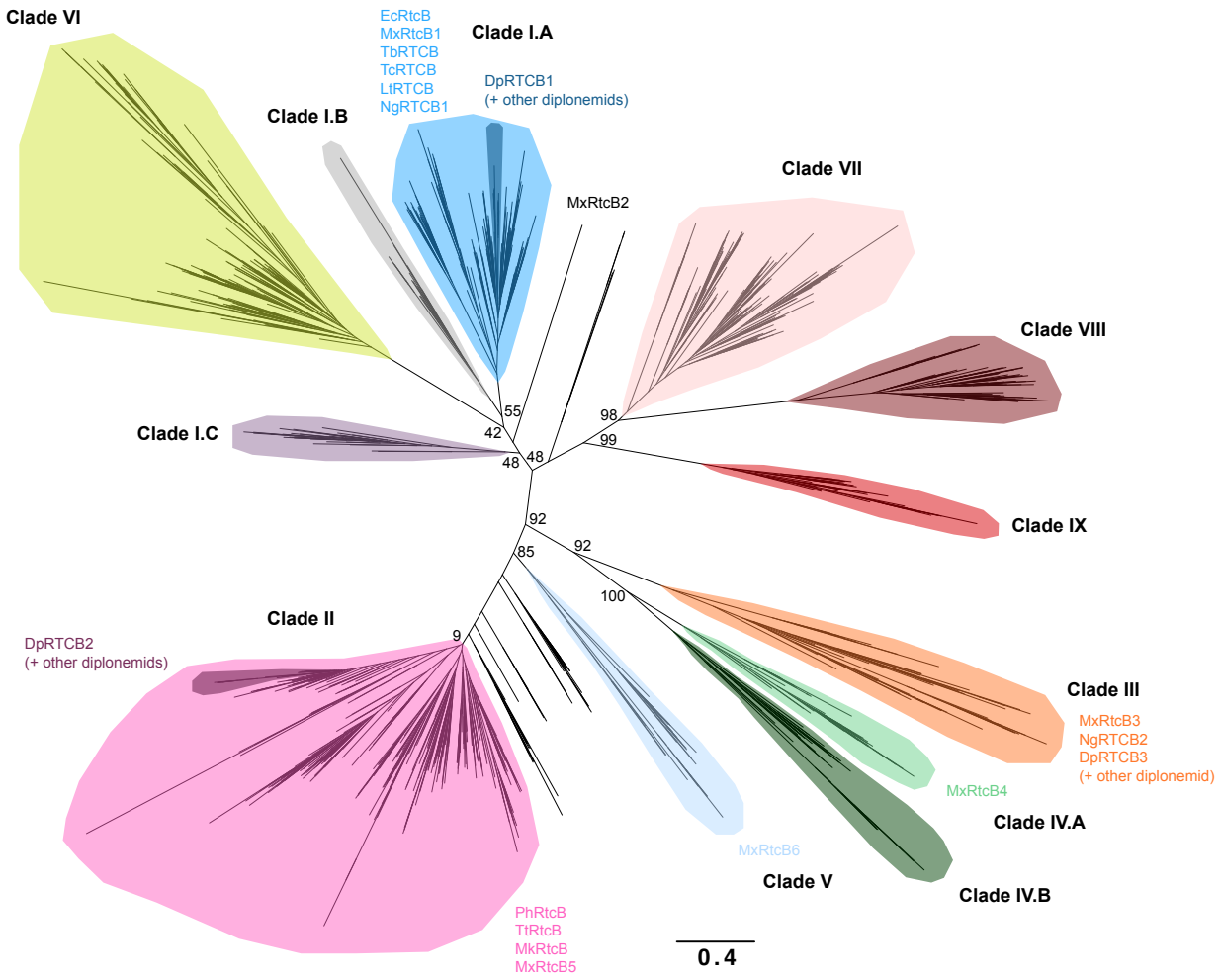

Fig. 2.11. Phylogenetic distribution of the RtcB-type family obtained with RAxML with the same alignment as in **Figure 2.5**. 1225 of sequences from bacteria, archaea, viruses, and diverse eukaryotes **(for details, see section 2.3 for Materials and methods and Supplementary Table 2.2)**.

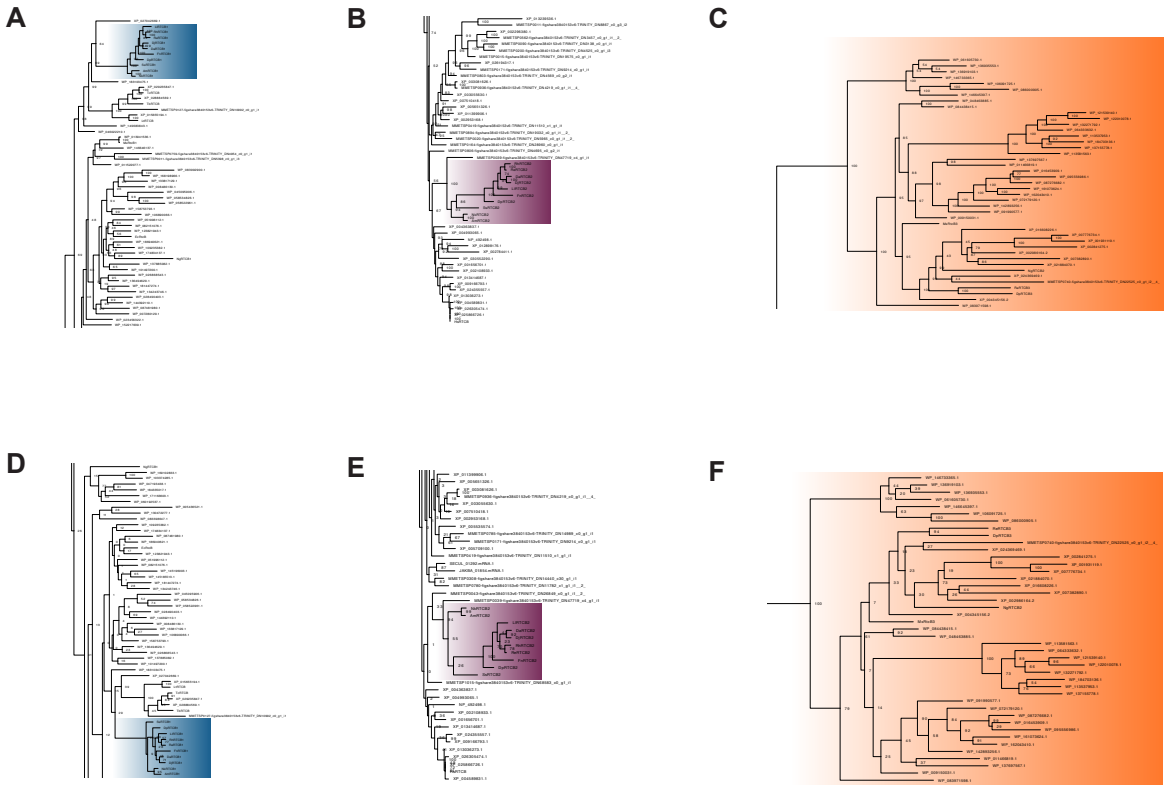

Fig. 2.12. Close-up on the phylogenetic distribution of **A-C) Figure 2.5** and **D-F) Supplementary Figure 2.11** for **A, D)** DpRTCB1, **B, E)** DpRTCB2, and **C, F)** DpRTCB3.

Table 2.2. Select organisms with their respective RtcB UniProt reference ID.

NCBI reference ID	Organism	RtcB nomenclature	Uniprot ID
NP_417879.1	*Escherichia coli*	EcRtcB	P46850
WP_011228915.1	*Thermus thermophilus*	TtRtcB	Q5SHE5
WP_011554960.1	*Myxococcus xanthus*	MxRtcB1	Q1D2I5
WP_020477908.1	*Myxococcus xanthus*	MxRtcB2	Q1DFL2
WP_011551357.1	*Myxococcus xanthus*	MxRtcB3	Q1DCX6
WP_011551632.1	*Myxococcus xanthus*	MxRtcB4	Q1DC50
WP_011555388.1	*Myxococcus xanthus*	MxRtcB5	Q1D1A0
WP_011550213.1	*Myxococcus xanthus*	MxRtcB6	Q1DG76
WP_148679988.1	*Methanopyrus kandleri*	MkRtcB	Q8TUS2
WP_010885677.1	*Pyrococcus horikoshii*	PhRtcB	O59245
XP_823328.1	*Trypasonoma brucei*	TbRTCB	Q389M0
XP_819860.1	*Trypanosoma cruzi*	TcRTCB	Q4DZR4
GET91792.1	*Leishmania tarentolae*	LtRTCB	A0A640KQK3
XP_002675141.1	*Naegleria gruberi*	NgRTCB1	D2VBK9
XP_002678559.1	*Naegleria gruberi*	NgRTCB2	D2VLJ2
NP_055121.1	*Homo sapiens*	HsRTCB	Q9Y3I0
-	*Diplonema papillatum*	DpRTCB1	
-	*Diplonema papillatum*	DpRTCB2	
-	*Diplonema papillatum*	DpRTCB3	
-	*Diplonema ambulator*	DaRTCB1	
-	*Diplonema ambulator*	DaRTCB2	
-	*Diplonema japonicum*	DjRTCB1	
-	*Diplonema japonicum*	DjRTCB2	
-	*Lacrimina lanifica*	LlRTCB1	
-	*Lacrimina lanifica*	LlRTCB2	
-	*Rhynchopus euleeides*	ReRTCB1	
-	*Rhynchopus euleeides*	ReRTCB2	
-	*Rhynchopus euleeides*	ReRTCB3	
-	*Rhynchopus humris*	RhRTCB1	
-	*Rhynchopus humris*	RhRTCB2	
-	*Flectonema neradi*	FnRTCB1	
-	*Flectonema neradi*	FnRTCB2	
-	*Sulcionema specki*	SsRTCB1	
-	*Sulcionema specki*	SsRTCB2	
-	*Artemidia motanka*	AmRTCB1	
-	*Artemidia motanka*	AmRTCB2	
-	*Namystynia karyoxenos*	NkRTCB1	
-	*Namystynia karyoxenos*	NkRTCB2	

Table 2.3. Probability of mitochondrial targeting signal for diplonemid RTCB proteins[a].

Protein	TargetP	MitoFates	PredSL	Predotar	MITOPROT	WoLFPSORT	DeepLoc
AmRTCB1	**0.717838**	0.383	**0.999389**	**0.44**	**0.9795**	**28.5**	**0.6074**
DaRTCB1	**0.732252**	**0.731**	**0.999468**	**0.43**	**0.8276**	**31.5**	**0.4964**
DjRTCB1	**0.790537**	0.109	**0.999391**	**0.38**	**0.5529**	**28.5**	**0.3430**
DpRTCB1	**0.558573**	**0.814**	**0.997766**	**0.57**	**0.6086**	**24**	**0.8542**
FnRTCB1	**0.596351**	0.164	**0.998619**	**0.76**	**0.9089**	**18**	**0.5152**
LlRTCB1	**0.530869**	0.045	**0.999794**	**0.64**	**0.8841**	**24**	**0.9514**
NkRTCB1	**0.630932**	0.769	0.999806	0.58	0.9538	28.5	0.794
ReRTCB1	0.400083	**0.609**	**0.999294**	**0.9**	**0.7726**	**16.5**	**0.8702**
RhRTCB1	**0.556126**	0.147	**0.999805**	**0.69**	**0.8249**	**16.5**	**0.7586**
SsRTCB1	**0.602674**	0.206	**0.999745**	**0.02**	**0.6520**	**25.5**	**0.9577**
AmRTCB2	0.000062	0.000	0.012489	0	**0.6805**	0	0.0124
DaRTCB2	0.000352	0.026	0.035648	0	0.2071	0	0.1022
DjRTCB2	0.000273	0.029	0.034165	0.02	0.1646	1	0.1298
DpRTCB2	0.000017	0.021	0.030079	0	0.2605	1	0.1495
FnRTCB2	0.000018	0.000	0.004321	0.01	0.0798	2	0.0716
LlRTCB2	0.063228	0.145	0.588346	0	0.4034	0	0.2249
NkRTCB2	0.00006	0.012	0.000275	0	0.1085	1	0.0443
ReRTCB2	0.004555	0.015	0.086936	0.06	0.3099	1	0.1425
RhRTCB2	0.000099	0.003	0.139442	0	0.0507	2	0.0492
SsRTCB2	0.000021	0.000	0.004232	0	0.2471	8	0.033
DpRTCB3	0.000614	0.000	0.019943	0.01	0.1194	0	0.0516
ReRTCB3	0.007691	0.088	**0.995102**	0.09	0.1122	0	**0.4239**

[a]In bold are the values predicted by each software of the presence of a targeting signal for a given sequence.

Table 2.4. Probability of other targeting signal for diplonemid RTCB proteins[a].

Protein	TargetP	MitoFates	PredSL	Predotar	MITOPROT	WoLFPSORT cytoplasm/nucleus	DeepLoc cytoplasm/nucleus
AmRTCB1	0.267089	0.617	0.000611	0.55	0.0205	2.5/0	0.2029/0.0039
DaRTCB1	0.266212	0.269	0.000532	0.56	0.1724	0/0	0.2571/0.0055
DjRTCB1	0.199868	0.891	0.000609	0.61	0.4471	0/0	0.2981/0.0037
DpRTCB1	0.439434	0.186	0.002234	0.43	0.3914	5/1	0.0642/0.0046
FnRTCB1	0.402903	0.836	0.001381	0.24	0.0911	14/0	0.2026/0.0048
LlRTCB1	0.467928	0.955	0.000206	0.36	0.1159	2/1	0.0165/0.0002
NkRTCB1	0.367519	0.231	0.000194	0.42	0.0462	0/0	0.0834/0.0029
ReRTCB1	0.598968	0.391	0.000706	0.10	0.2274	3/1	0.0299/0.0006
RhRTCB1	0.440742	0.853	0.000195	0.30	0.1751	7.5/0	0.0575/0.0032
SsRTCB1	0.391388	0.794	0.000255	0.94	0.3480	0/1	0.0155/0.0001
AmRTCB2	0.997101	1.000	0.987511	0.99	0.3195	**18.5/6.5**	**0.638/0.0026**
DaRTCB2	0.989558	0.974	0.964352	0.99	0.7929	**20/10**	**0.4642/0.057**
DjRTCB2	0.998621	0.971	0.965835	0.98	0.8354	**22/7**	**0.4574/0.0765**
DpRTCB2	0.999874	0.979	0.969921	0.99	0.7395	**16.5/9.5**	**0.468/0.0288**
FnRTCB2	0.999382	1.000	0.995679	0.99	0.9202	**19.5/7.5**	**0.4775/0.0085**
LlRTCB2	0.784701	0.855	0.411654	0.99	0.5966	**17.5/10.5**	**0.4304/0.0193**
NkRTCB2	0.998911	0.988	0.999725	0.99	0.8915	**18.5/6.5**	**0.5482/0.0154**
ReRTCB2	0.994655	0.985	0.913064	0.93	0.6901	**19.5/8.5**	**0.4994/0.0082**
RhRTCB2	0.997824	0.997	0.860558	0.99	0.9493	**21/6**	**0.6205/0.0278**
SsRTCB2	0.999616	1.000	0.995768	0.99	0.7529	**13.5/6.5**	**0.5435/0.0315**
DpRTCB3	0.998599	1.000	0.980057	0.99	0.8806	**20/8**	**0.3342/0.0022**
ReRTCB3	0.991637	0.912	0.004898	0.91	0.8878	**14/6**	0.1656/0.0065

[a]In bold are the values predicted by each software of the presence of a targeting signal for a given sequence.

Table 2.5. Accuracy of the predicted models obtained with SWISS-MODEL.

SWISS-MODEL structure	PDB ID reference	Sequence identity (%) to reference structure	Residues not modelled starting from N-terminus	Residues not modelled starting from C-terminus
EcRtcB	4.dwq.2.A	30.46%	7	-
DpRTCB1	4.dwq.2.A	30.34%	-	-
DpRTCB2	4.dwq.1.A	49.15%	9	-
DpRTCB3	4.dwq.1.A	31.12%	105	17

Chapitre 3

Purification des ligases de type RtcB de *Diplonema papillatum*

3.1. Introduction

Tel que mentionné dans l'introduction ainsi que dans le chapitre 2 de ce mémoire, nous pensons que DpRTCB1 est impliquée dans l'épissage en *trans* des ARN mitochondriaux chez *Diplonema*. Nous avons aussi des indications quant aux rôles biologiques des deux autres DpRTCB : DpRTCB2 agirait dans l'épissage des ARNt et DpRTCB3 serait liée à l'ajout de coiffe aux acides nucléiques chez *Diplonema*. Toutefois, ces hypothèses doivent être validées expérimentalement. Le second objectif de ce projet était donc de purifier les protéines DpRTCB en quantité et en qualité suffisantes pour effectuer des essais enzymatiques. Une connaissance sur la préférence des substrats de ces enzymes pourrait nous renseigner sur la fonction de ces ligases. Ce chapitre décrit les différentes méthodes de purifications qui ont été employées afin de purifier ces protéines.

3.2. Matériel et méthodes

3.2.1. Isolation d'ARN polyadénylé

L'ARN a été extrait des cellules de *D. papillatum* en utilisant une solution de « Home-made Trizol substitute » [**116, 117**]. Les ARN poly-A ont été enrichis par purification sur une cellulose oligo(dT) (ThermoFisher Scientific/Ambion Cat. Nr. AM10020). Le mélange de poudre de cellulose a été suspendu dans un tampon d'élution (10 mM Tris-HCl, 0.1% SDS et 10 mM EDTA pH 7.5) et le surnageant a été retiré par centrifugation. La cellulose a été mélangée avec 100 mM NaOH et le surnageant a été retiré par centrifugation. Le tampon d'adhésion (10 mM Tris-HCl, 0.5 M NaCl, 0.1% SDS et 1 mM EDTA pH 7.5) a été ajouté à la cellulose et celle-ci a été incubée pour 3 minutes tout en étant mélangée

occasionnellement. Le surnageant a été retiré par centrifugation et l'étape précédente a été répétée cinq fois. Le tampon d'application (100 mM Tris-HCl, 1% SDS et 10 mM EDTA pH 7.5) a été ajouté à l'échantillon d'ARN, qui a été incubé pendant 5 minutes à 70°C, puis a été transféré sur la glace pour 2 minutes. Du NaCl a été ajouté à l'échantillon d'ARN jusqu'à une concentration finale de 0.5 M et celui-ci a ensuite été ajouté à la cellulose pour suivre avec une incubation de 15 minutes à température pièce. Le surnageant a été retiré par centrifugation et le tampon de lavage (10 mM Tris-HCl, 0.25 M NaCl et 1 mM EDTA pH 7.5) a été ajouté à la cellulose. Le surnageant a été retiré par centrifugation et l'étape précédente a été répétée trois fois. Le tampon d'élution a été ajouté à la cellulose, mélangé et l'ARN a été élué deux fois par incubation à 50°C pour 2 minutes. Le surnageant a été collecté par centrifugation et la cellulose a été nettoyée une fois avec de l'eau DEPC et lavée trois avec le tampon d'adhésion. L'échantillon a été ajouté de nouveau sur la cellulose et les étapes de purifications précédemment mentionnées ont été répétées en commençant par l'étape de dénaturation de l'ARN. L'éluat final a été centrifugé (14000×g, 15 minutes, température pièce) et l'ARN a été précipité durant la nuit à -20°C avec 20 mg/mL de glycogène, 0.12 volume de 3M d'acétate de sodium pH 5.3 et 1.1 volume d'isopropanol. L'échantillon a été centrifugé (18000×g, 45 minutes, 4 °C) et le culot a été incubé pour 10 minutes dans de l'éthanol 70%. L'échantillon a été centrifugé (18000×g, 10 minutes, 4°C) pour retirer l'éthanol et une fois le culot sec, l'ARN a été solubilisé avec de l'eau DEPC.

3.2.2. RT-PCR

La transcription inverse (RT) de l'ARN purifié a été effectuée à l'aide de la « SuperScript IV Reverse Transcriptase » (ThermoFisher Scientific) et des hexamères aléatoires. L'amplification de l'ADNc par l'amplification en chaîne par polymérase (PCR) a été effectuée avec la « Q5® High-Fidelity DNA polymerase » (New England Biolabs) avec des amorces spécifiques aux séquences codantes *DpRTCB1*, *DpRTCB2* et *DpRTCB3* **(voir Tableau 3.1)**. Les produits d'amplification ont été séparés sur un gel d'agarose et extraits du gel à l'aide de la trousse « QIAquick Gel Extraction Kit » (Qiagen). *DpRTCB1-GS* a été obtenu après avoir rajouté dans le mélange réactionnel un oligo codant pour une répétition de penta-(glycine-sérine).

3.2.3. Construction de plasmides pour la surexpression de protéines

Les séquences codantes des gènes *DpRTCB1*, *DpRTCB2*, *DpRTCB3* et *DpRTCB1-GS*) ont été insérées dans des plasmides pET28b-His_6SMT3 (pour *DpRTCB1*; un don du Dr.

Tableau 3.1. Amorces PCR utilisées pour les séquences *DpRTCB1*, *DpRTCB2*, *DpRTCB3*, *DpRTCB1-GS* et les construits de plasmides.

Nom	**Séquence (5′ à 3′)**	**Utilisation**
dp320	ATGCTGCGGCGAACGCTC	*DpRTCB1* ORF
dp321	CCCCTTGACGCACACAATTTGC	*DpRTCB1* ORF
dp322	GATTGGTGGATCCGTGGACGTGACGAGA	*DpRTCB1* ORF-NEBuilder
dp323	GGTGCTCGAGTTACCCCTTGACGCACA	*DpRTCB1* ORF-NEBuilder
dp324	GCGTCAAGGGGTAACTCGAGCACCACC	pET-28-SMT3-*DpRTCB1*-NEBuilder
dp325	TCGTCACGTCCACGGATCCACCAATCTG	pET-28-SMT3-*DpRTCB1*-NEBuilder
dp326	TTCTGCCCGTTGTGCGTG	*DpRTCB1* vérification de séquençage Sanger
dp327	GTTCAAGGCGGTCGGCAA	*DpRTCB1* vérification de séquençage Sanger
dp360	GGATCCGGTAGCGGTTCTGGGAGCGGATCGGATATC	BamHI-5xGS-EcoRV oligo
dp363	CAGAGAACAGATTGGTGGATCCGGTAGC	SUMO3-GS_linker for NEBuilder – utilisation sur dp360 (dp363+dp364)
dp364	CTTCATCGCGACGTCGGCCGATCCGCTCCC	GS_linker-*DpRTCB1* pour NEBuilder – utilisation sur dp360 (dp363+dp364)
dp365	TCCACCAATCTGTTCTCTGTGAGCCTC	SUMO3 pour NEBuilder – utilisation sur dp324
dp328	ATGGCGCGGTATGATGCATAC	*DpRTCB2* ORF
dp329	TTATCCTTTGATGACACACTGCGG	*DpRTCB2* ORF
dp332	ATTGGTGGATCCATGGCGCGGTATGAT	*DpRTCB2* ORF-NEBuilder
dp344	TCGAGTGCGGCTTATCCTTTGATGACACACT	*DpRTCB2* ORF-NEBuilder
dp334	CATCAAAGGATAAGCCGCACTCGAGCAC	pET-28-SMT3-*DpRTCB2*-NEBuilder
dp335	CATACCGCGCCATGGATCCACCAATC	pET-28-SMT3-*DpRTCB2*-NEBuilder
dp340	GCACCAGCCGCTCCTTT	*DpRTCB2* vérification de séquençage Sanger
dp341	GGCTTCTCGTGCATAGGA	*DpRTCB2* vérification de séquençage Sanger
dp345	GCTCGCCAATTCGGGAAC	*DpRTCB2* vérification de séquençage Sanger
dp330	ATGAGCATCCGCGTGGGGA	*DpRTCB3* ORF
dp331	TTACCGGCACGACGCGACT	*DpRTCB3* ORF
dp336	TTGGTGGATCCATGAGCATCCGCGTG	*DpRTCB3* NEBuilder
dp337	GTGCTCGAGTTACCGGCACGACGC	*DpRTCB3* NEBuilder
dp338	CGTCGTGCCGGTAACTCGAGCACCAC	pET-28-SMT3-*DpRTCB3*-NEBuilder
dp339	ACGCGGATGCTCATGGATCCACCAATC	pET-28-SMT3-*DpRTCB3*-NEBuilder
dp342	GGACCAGCTTTTCGCAA	*DpRTCB3* vérification de séquençage Sanger
dp343	CTTTGCTGGACATCTGCC	*DpRTCB3* vérification de séquençage Sanger

John Pascal, Université de Montréal) et pET28b-His$_{10}$SMT3 (pour *DpRTCB2*, *DpRTCB3* et *DpRTCB1-GS*; un don du Dr. Stewart Shuman, Sloan Kettering Institute, New York, U.S.A.) **(voir Figure 3.1)**. Le plasmide codant pour la protéine « ubiquitin-like specific protease 1 » (Ulp1) (aussi un don du Dr. Shuman) a été utilisé afin de surexprimer et purifier la protéine Ulp1 pour cliver le tag SUMO3 des protéines DpRTCB. Les plasmides ont été amplifiés par PCR en utilisant la « Platinum™ SuperFi™ DNA Polymerase » (ThermoFisher Scientific). Les produits obtenus ont été séparés par électrophorèse sur un gel d'agarose et extraits du gel à l'aide de la trousse « Monarch® DNA Gel Extraction

Kit » (New England Biolabs). Les séquences *DpRTCB* et les plasmides amplifiés ont été assemblés à l'aide de la trousse « NEBuilder® HiFi DNA Assembly Cloning Kit » (New England Biolab) **(voir Figure 3.2)**. Afin de vérifier l'assemblage des plasmides, des cellules Rosetta2 (DE3) *E. coli* (Novagen) ont été transformées par choc thermique [**pour plus de détails, voir la section 3.2.5**] avec le plasmide construit. Les cellules ont été cultivées au courant de la nuit à 37°C avec agitation constante dans un milieu LB auquel a été ajouté 50 µg/mL de kanamycine et 35 µg/mLde chloramphénicol. Les plasmides ont été extraits des cellules à l'aide de la trousse « Monarch® Plasmid Miniprep Kit » (New England Biolab) et envoyés pour séquençage à l'Institut de recherche en immunologie et cancérologie de l'Université de Montréal (IRIC) **(voir Tableau 3.1 pour les amorces utilisées)**.

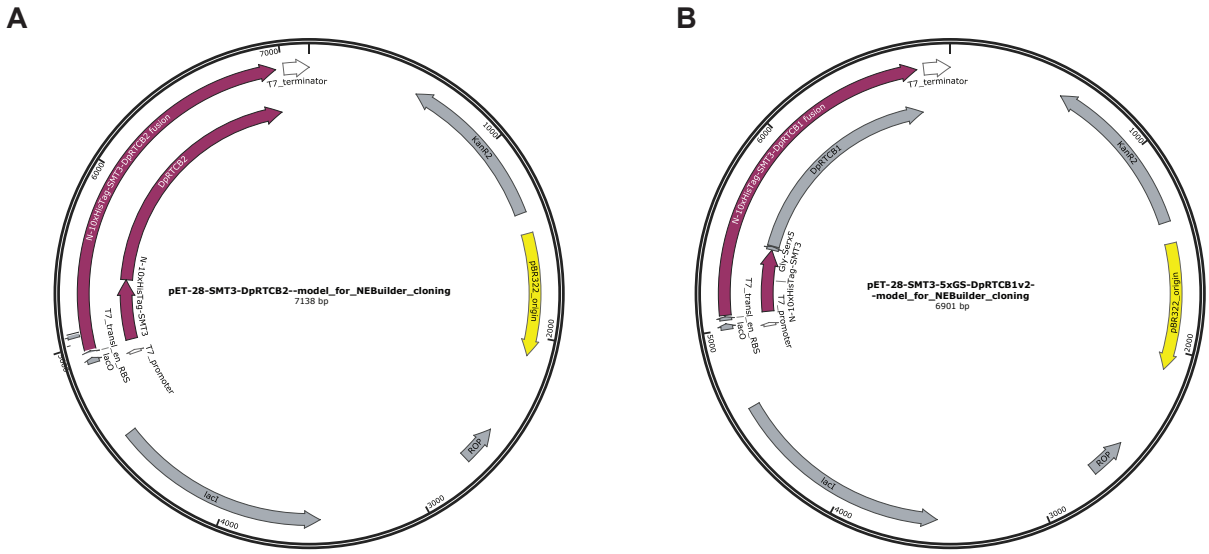

Fig. 3.1. Construits des plasmides **A)** pour HS-DpRTCB2 et **B)** HS-GS-DpRTCB1 utilisés pour la transformation avec leur gène de résistance aux antibiotiques (KanR2), le tag SUMO3 et histidine (N-10xHisTag-SMT3) et l'insertion de l'espaceur (Gly-Ser×5) avec l'aide du logiciel SnapGene.

3.2.4. Préparation de cellules compétentes *E. coli*

Des échantillons congelés de la souche Rosetta2 ont été décongelés et les cellules ont ensuite été cultivées pendant la nuit à 37 °C sur un milieu LB agar auquel a été ajouté 35 µg/mL de chloramphénicol. La préparation des cellules compétentes a été effectuée selon le protocole établi par Inoue *et al.* [**118**]. Des colonies individuelles ont été sélectionnées et propagées dans un milieu SOB (2% Tryptone, 0.5% extrait de levure, 8.6 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$ et 10 mM $MgSO_4$ pH 7.0) avec une agitation constante jusqu'à atteindre une $DO_{600\ nm}$ de 0.4-0.6. La culture a été placée sur glace (10 minutes) et les cellules ont

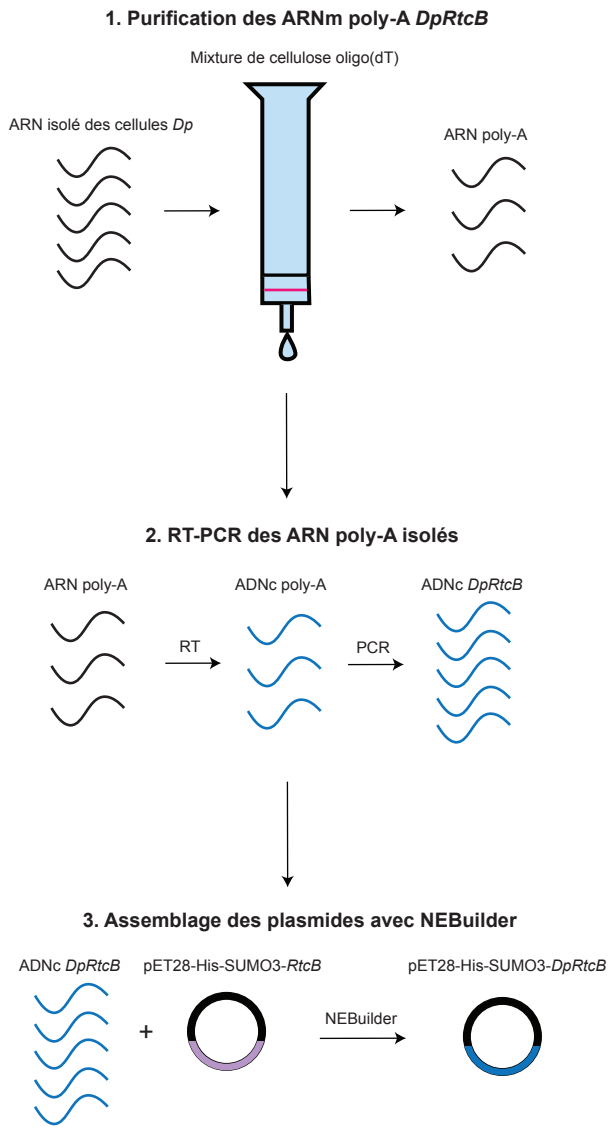

Fig. 3.2. Schéma pour l'assemblage des plasmides utilisés pour la surexpression des protéines DpRTCB. **1)** Les ARN poly-A ont d'abord été obtenus par purification d'ARN isolé à partir de cellules complètes à l'aide de cellulose oligo(dT). **2)** Une fois les ARN poly-A obtenus, ils ont été soumis à la transcription inverse (RT) afin d'obtenir l'ADN codant (ADNc) et les séquences DpRTCB spécifiquement ont été amplifiés à l'aide de l'amplification en chaîne par polymérase (PCR). **3)** Les séquences codantes pour les DpRTCB ont été assemblées avec le plasmide pET28-His-SUMO3-RtcB en utilisant la trousse d'assemblage NEBuilder®.

été récoltées par centrifugation (1800×g, 10 minutes, 4°C). Le culot a été suspendu dans un milieu TB (10 mM PIPES, 250 mM KCl, 15 mM $CaCl_2$ et 5.5 mM $MnCl_2$ pH 6.7) et mis sur glace (10 minutes). La culture a été centrifugée (1800×g, 10 minutes, 4°C) et suspendue dans du TB auquel a été ajouté 7% de diméthyle sulfoxide (DMSO). Les cellules ont été aliquotées, congelées dans de l'azote liquide et entreposées à -80°C pour une transformation

future par choc thermique.

3.2.5. Transformation par choc thermique

La transformation par choc thermique de cellules Rosetta2 a été effectuée suivant le protocole établi par Inoue *et al.* [**118**]. Les échantillons de cellules, auquel 5 ng de plasmide a été ajouté, ont été incubés sur la glace pendant 30 minutes. Les cellules ont été soumises à un choc thermique (45 secondes, 42°C) et placées sur la glace pour 2 minutes avant d'y ajouter du SOC (SOB auquel a été ajouté 20 mM de glucose). Les cellules ont été incubées pour 1 heure à 37°C avec une agitation constante avant d'être placées sur un milieu LB agar auquel a été ajouté 50 µg/mL de kanamycine et 35 µg/mL de chloramphénicol. Les plaques inoculées ont été incubées pendant la nuit à 37°C.

3.2.6. Surexpression protéique

Le protocole suivant a été employé pour une surexpression routinière des protéines chez les bactéries. Une seule colonie a été sélectionnée à partir du milieu d'agar puis propagée à 37°C pendant la nuit avec une agitation constante dans un milieu LB liquide auquel a été ajouté 50 µg/mL kanamycine et 35 µg/mL chloramphénicol. La culture résultante a été transférée dans un milieu LB liquide d'un volume 50 fois supérieur à la culture initiale, auquel a été ajoutée la même concentration de kanamycine et chloramphénicol. Les cellules ont été propagées à 30°C avec agitation constante jusqu'à atteindre une DO_{600nm} située entre 0.4-0.5. Une fois la DO atteinte, un échantillon contrôle de 1 mL a été prélevé de la culture et solubilisé dans le tampon de charge 1× Tris-Tricine SDS-PAGE (3% w/v SDS, 1.5% v/v β-mercaptoéthanol, 7.5% w/v glycérol, 0.01% bleu de Coomassie G-250 et 37.5 Tris-HCl pH 7.0) ou 1× Tris-glycine SDS-PAGE (2% SDS; 10% glycérol; 0.002% bromophénol bleu et 0.06 M Tris-HCl pH 6.8) dépendant du gel utilisé pour séparer les protéines. La surexpression de protéine dans la culture a été induite avec de l'isopropyl-β-D-thiogalactopyranoside (IPTG) à une concentration finale de 0.1 mM. Les protéines ont été surexprimées pour ∼18h à 18°C avec agitation constante. La surexpression en milieu contenant 2 mM de Mn^{2+} a été entreprise de manière identique, mais à 11°C.

3.2.7. Lyse cellulaire des bactéries

Suivant la surexpression, un échantillon de 1 mL a été prélevé de la culture cellulaire ayant surexprimé la protéine d'intérêt puis l'échantillon a été solubilisé dans un tampon de charge 1× Tris-tricine SDS-PAGE ou 1× Tris-glycine SDS-PAGE. Les cellules ont été

récoltées par centrifugation (4000×g, 4°C, 20 minutes) et suspendues dans le tampon de lyse (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 1 mM DTT) auquel a été ajouté 1× « cOmplete™, EDTA-free Protease Inhibitor Cocktail » (Roche). Le pH du tampon de lyse variait en fonction du point isoélectrique de la protéine d'intérêt (DpRTCB1 and DpRTCB1-GS: pH 7.4, DpRTCB2: pH 8.0, DpRTCB3: pH 8.0 et Ulp1: pH 8.0). Dans les conditions où il y avait une plus haute concentration de sel, le tampon de lyse contenait 1 M de NaCl. Pour les conditions de lyse avec l'arginine, 0.5 M d'arginine a été ajouté au tampon de lyse et le pH de la solution a été ajusté à la valeur correspondante pour la protéine à purifier avec du HCl.

Avant la lyse cellulaire, du lysozyme a été ajouté à la solution pour atteindre une concentration finale de 0.2 mg/mL et les cellules ont été incubées sur la glace pour 1 heure. Pour les échantillons qui contenaient de la « Benzonase® Nuclease » (Millipore Sigma), la nucléase a été ajoutée à une concentration de 50 U/mL suivant la lyse cellulaire et le tampon de lyse a été ajusté pour contenir 150 mM de NaCl et 2 mM de Mg^{2+}. L'échantillon a été incubé à température pièce pour 30 minutes. La sonication (2 minutes totales, avec des cycles de 10 secondes d'activité and 45 secondes éteint) ou le « French press » (à une pression de 1000 psi) ont été utilisés pour la rupture des cellules. Pour les échantillons dans des conditions dénaturantes, 4 M d'urée a été ajouté au lysat cellulaire suivant la lyse et avant la centrifugation (15000×g, 4°C, 20 minutes). Suivant la centrifugation, le surnageant a été récupéré et le culot a été suspendu dans le même volume de tampon de lyse que celui du surnageant. Suivant la séparation entre le surnageant et le culot, certaines conditions impliquaient l'incubation du culot avec 0.1% de sodium lauroyl sarcosinate (sarkosyl) ou 4 M d'urée pour 30 minutes et 24 heures ou 15 et 30 minutes respectivement. Les échantillons qui ont été récupérés pour l'analyse de la lyse cellulaire étaient de 100 µL et ont été séparés sur des gels de 12% Tris-glycine SDS-PAGE ou 10% Tris-tricine SDS-PAGE [**119**]. La quantité correspondante de tampon de charge a été ajoutée aux échantillons et ceux-ci ont été séparés sur gel pour 2 à 3 heures à 80V à température pièce.

3.2.8. Purification de protéines

La résine « HisPur™ Cobalt Superflow Agarose » (ThermoFisher Scientific) a été utilisée pour la purification des protéines DpRTCB et Ulp1. Le pH des tampons utilisés pour la purification variait en fonction de la protéine à purifier tel que mentionné dans la section sur la lyse cellulaire bactérienne [**voir section 3.2.7**]. Pour la purification en conditions dénaturantes, la composition des tampons différait des tampons natifs de par l'ajout d'urée (4 M en concentration finale). Le processus de purification pour toutes les

protéines a été effectué en fonction des recommandations données par le manufacturier avec quelques modifications [**120**]. Le volume de résine à utiliser a été déterminé en fonction de la quantité de protéine attendue à être purifiée qui a été calculée à l'étape précédente (la quantité de protéine attendue pour tous les échantillons était ∼1.5 à 3 mg). Toutes les périodes d'incubation mentionnées dans le protocole ont été effectuées à 4°C et toutes les étapes de centrifugations ont été faites à 700×g pour 2 minutes à 4°C. Pour chacune des étapes de purification, un échantillon de 100 µL a été collecté pour une analyse en aval de la purification.

La résine a été équilibrée avec du tampon d'équilibration (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 5 mM imidazole) pour l'équivalent deux fois le volume utilisé de résine avant d'être centrifugé. Le surnageant a été ajouté à la colonne et incubé pour une période de 30 minutes sur un mélangeur rotatif bout à bout avant d'être centrifugé. La résine a été nettoyée avec un tampon de lavage (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 20 mM imidazole) pour l'équivalent de deux fois le volume utilisé de résine puis a été centrifugée. Cette étape a été répétée pour une seconde fois. Les protéines ont été éluées avec le tampon d'élution (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 150 mM imidazole) pour dix fois le volume utilisé de résine et incubé sur un mélangeur rotatif bout à bout pour 10 minutes avant la centrifugation. Cette étape a été répétée trois fois. Les protéines purifiées ont été nettoyées de l'imidazole et de l'urée par centrifugation de la fraction éluée dans un AMICON (3000×g, 1 min, 4°C) puis en ajoutant un tampon de suspension (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 1 mM DTT) au volume manquant après la centrifugation. Les protéines ont été concentrées dans le plus petit volume possible (∼250 à 500 µL). Du glycérol 100% a été ajouté à l'échantillon concentré jusqu'à atteindre une concentration finale de 50% de glycérol. Les échantillons ont été congelés à l'aide de l'azote liquide puis entreposés à -80°C.

L'analyse en aval a été effectuée sur un gel de 12% Tris-glycine SDS-PAGE ou 10% Tris-tricine SDS-PAGE [**voir section 3.2.7**]. Les gels ont été colorés pendant 1h à l'aide d'une solution de bleu de Coomassie (0.1% bleu de Coomassie R-250, 50% méthanol et 10% acide acétique) puis décolorés à l'aide d'une solution de décoloration (40% méthanol et 10% acide acétique). Les protéines ont été quantifiées directement à partir des résultats obtenus sur gel à l'aide du logiciel Image Lab v6.0.1 (Bio-Rad) et en séparant des quantités fixes d'albumine de sérum bovin (BSA). Pour une confirmation de la lyse cellulaire et de l'efficacité de purification, un immunobuvardage de type Western a été fait sur les échantillons obtenus durant la purification de façon similaire à ce qui est mentionné dans la section sur des immunobuvardages de type Western ici-bas [**voir section 3.2.10**].

3.2.9. Retrait du tag SUMO3 en N-terminus des protéines surexprimées

La protéase Ulp1 a été utilisée pour cliver DpRTCB2 au site Smt3 la nuit durant dans un tampon de digestion (50 mM Tris-HCl, 150 mM NaCl, 10% glycérol et 1 mM DTT pH 8.0) tel que mentionné dans le protocole établi par Tanaka *et al.* [**28**]. Comme contrôle de clivage par Ulp1, une protéine contrôle (∼140 kDa) offerte par Dr. John Pascal ayant un tag SUMO3 a été utilisée pour les tests enzymatiques de Ulp1 et comme contrôle pour les anticorps anti-SUMO3 utilisés dans les immunobuvardages de type Western. La digestion protéique a été effectuée la nuit durant dans le tampon de digestion.

3.2.10. Immunobuvardage de type Western

Les quantités de protéines des échantillons ont été déterminées par la méthode de Bradford et les échantillons ont été séparés sur gel SDS-PAGE selon la méthode utilisée dans la section concernant la lyse cellulaire bactérienne [**voir section 3.2.7**]. Les échelles « PageRuler Prestained Protein Ladder » (ThermoFisher Scientific) et « Biotinylated Protein Ladder » (Biotinylated Protein Ladder Detection Pack #7727 Cell Signaling) ont aussi été chargées sur gel comme marqueurs moléculaires de taille. Les procédures d'immunobuvardage de type Western ont été effectuées suivant les instructions données par Cell Signaling [**121**]. La membrane « Amersham™ Hybond™ PVDF membrane 0.45µm » (GE Healthcare) a été préincubée 10 secondes dans du méthanol, puis 10 secondes dans de l'eau distillée et ensuite 5 minutes dans le tampon de transfert (25 mM Tris-HCl, 192 mM glycine et 20% v/v méthanol pH 8.3) avant le transfert de protéines la nuit durant à 10V et à 4°C. Suivant le transfert, la membrane a été incubée sur une plateforme rotative pour 1 heure à 4°C dans un tampon de blocage TBST (137 mM NaCl, 20 mM Tris-HCl et 0.1% v/v Tween 20 pH 7.6) auquel a été ajouté 5% m/v de lait écrémé en poudre. La membrane a été nettoyée trois fois pour 5 minutes avec le TBST avant l'incubation la nuit durant à 4°C sur la plateforme rotative avec le tampon de blocage et les anticorps primaires.

Les anticorps primaires anti-SUMO3 (obtenus par le Dr. Julius Lukes, Czech Academy of Science, République Tchèque) ont été dilués suivant le ratio 1:1000 pour anti-SUMO3 (anticorps primaires produits chez le lapin). L'échelle moléculaire biotinylée a été visualisée à l'aide de l'anticorps primaire suivant une dilution de 1:1000 (« Biotinylated Protein Ladder Detection Pack #7727 », Cell Signaling). Suivant l'incubation avec l'anticorps primaire, la membrane a été lavée avec du TBST trois fois pour 5 minutes. L'anticorps secondaire anti-lapin (« anti-rabbit IgG HRP-linked Antibody #7074 », Cell Signaling) à une dilution de 1:5000 a été incorporé au tampon de blocage avant d'être ajouté à la

membrane qui a été incubée pour 1 heure à 4°C sur la plateforme rotative. Suivant l'incubation, la membrane a été lavée trois fois pour 5 minutes avec du TBST et a été incubée pour 1 minute dans le réactif « Signal Fire Reagent ECL » (Cell Signaling) pour la détection.

3.3. Résultats

3.3.1. Construit des protéines recombinantes et choix de la souche bactérienne pour la surexpression

Pour exprimer les gènes RtcB de *Diplonema* chez *E. coli* nous avons utilisé le plasmide de base pET28b contenant le gène de résistance à l'antibiotique kanamycine [**28**]. Le promoteur T7 du plasmide a permis l'expression des protéines recombinantes DpRTCB ayant un tag His-SUMO3 en position N-terminale. DpRTCB1 a été étiquetée avec six résidus histidines, alors que DpRTCB2 et DpRTCB3 avec dix. Dans les résultats, nous appellerons les protéines recombinantes RTCB de *Diplonema* HS-DpRTCB.

Les codons tels qu'AGA, AGG, AUA, CUA, GGA, CCC et CGG sont très fréquents chez *Diplonema*, une large quantité des ARNt correspondants sont donc nécessaires lors de la surexpression des protéines HS-DpRTCB. Les cellules d'*E. coli* sont communément utilisées pour la surexpression, toutefois, les niveaux d'ARNt permettant la lecture des codons mentionnés ci-haut sont bas chez cette bactérie. Nous avons donc utilisé la souche Rosetta2 d'*E. coli* qui possède des copies supplémentaires de ces gènes ARNt par l'entremise d'un plasmide. Ce plasmide est maintenu dans la bactérie grâce à un gène de résistance à l'antibiotique chloramphénicol. Par conséquent, les cellules transformées ont été sélectionnées sur un milieu auquel a été ajouté les antibiotiques kanamycine et chloramphénicol. Seuls les clones formant des colonies individuelles ont été isolés, et pour chaque expérience de transformation, les clones contenant le plasmide, soit les colonies Rosetta2-DpRTCB, ont été examinés plus en détails.

3.3.2. Surexpression des protéines recombinantes HS-DpRTCB pour la purification par affinité

Afin d'évaluer la surexpression protéique des clones Rosetta2 recombinants, les lysats cellulaires des transformants ont été séparés par SDS-PAGE **(Figure 3.3)**. Suite à l'induction avec l'IPTG, les échantillons prélevés des transformants démontrent qu'une bande est plus intense que les autres, une des caractéristiques de la surexpression protéique. L'augmentation de la quantité de protéines suivant la surexpression a été calculée à partir

des valeurs d'intensité du signal des protéines sur gel. Ces valeurs, obtenues à l'aide du logiciel Image Lab, se basent sur les nuances de gris causées par une variation en quantité de protéines sur l'image du gel. Ces valeurs d'intensité ont été comparées entre les échantillons pré-induits et induits pour les bandes ayant des tailles moléculaires similaires. L'augmentation en protéines d'intérêt lors de la surexpression a donc été évaluée à ∼30x pour HS-DpRTCB3 et à ∼35x plus élevée pour HS-DpRTCB1 et HS-DpRTCB2. Les tailles apparentes de HS-DpRTCB1 et HS-DpRTCB2 correspondent aux tailles théoriques attendues soit ∼60.1 kDa et ∼70.4 kDa, respectivement. Toutefois, la bande représentant HS-DpRTCB3, malgré que cette protéine ait une taille théorique de ∼68.7 kDa apparaît, pour une raison inconnue, ∼10 kDa plus large. Nous suspectons que ceci est causé par un phénomène appelé « gel shifting » qui a déjà été observé pour une grande variété de protéines. Celui-ci peut se produire en partie à cause de la structure tertiaire d'une protéine [**122**] ou à cause du type de gel utilisé [**123**]. La phosphorylation des protéines [**124**] ou la quantité de résidus polaires [**125**] peuvent aussi affecter la migration des protéines.

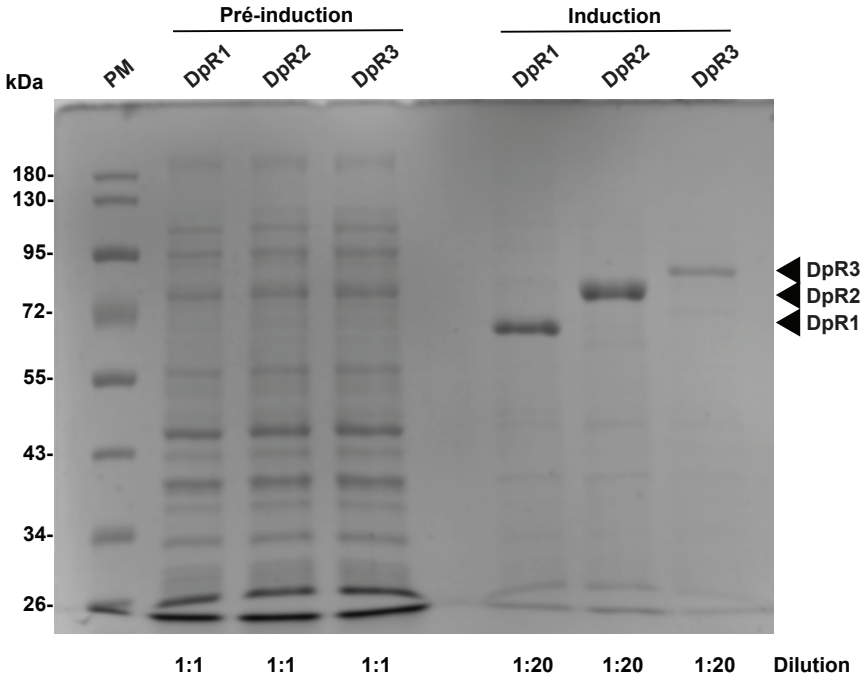

Fig. 3.3. Séparation par Tris-glycine SDS-PAGE des lysats cellulaires des transformants Rosetta2-DpRTCB1 (DpR1), Rosetta2-DpRTCB2 (DpR2) et Rosetta2-DpRTCB3 (DpR3) avant et après induction. Les échantillons induits ont été dilués 20 fois comparé aux échantillons préinduits.

3.3.3. Lyse des cellules exprimant HS-DpRTCB2 dans le but de purifier les protéines recombinantes

Afin de purifier les protéines recombinantes, nous avons lysé les cellules dans un tampon à base de Tris-HCl pour lequel le pH a été ajusté en fonction du point isoélectrique des protéines HS-DpRTCB [**voir section 3.2.7**]. Nous avons commencé avec la purification de HS-DpRTCB2, puisqu'elle ressemble au niveau de sa séquence à la RtcB de *Pyrococcus horikoshii* (∼50%) qui a été précédemment surexprimée et purifiée avec succès à partir de cellules d'*E. coli*. Une fois les conditions optimisées pour cette protéine, nous souhaitions appliquer ces paramètres aux autres HS-DpRTCB.

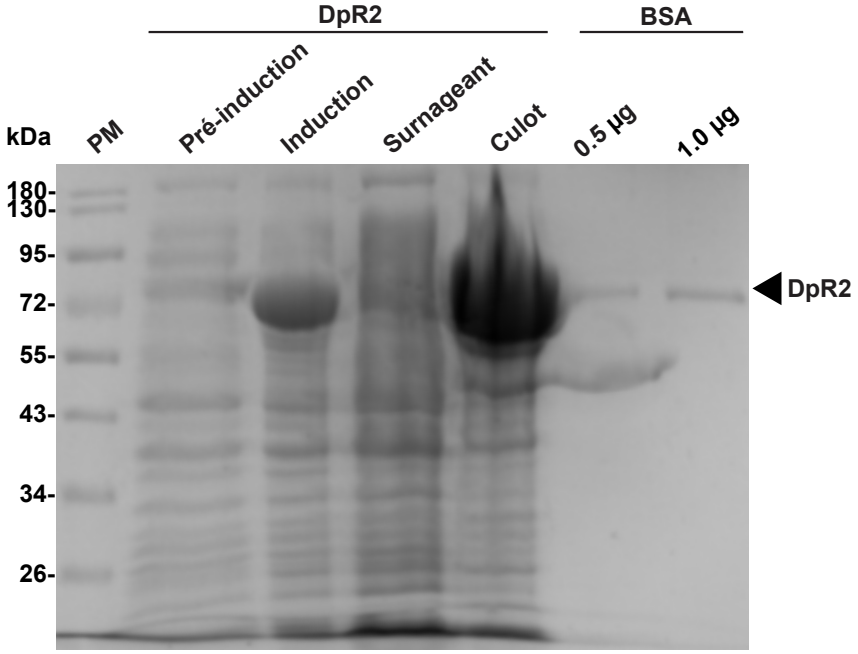

Fig. 3.4. Séparation sur gel Tris-glycine SDS-PAGE des lysats cellulaires des transformants pré-induits et induits à surexprimer la protéine HS-DpRTCB2 (DpR2) avec le surnageant et le culot résultant de la lyse. L'albumine de sérum bovin (BSA) a été ajoutée sur le gel en même temps que les échantillons.

La fraction soluble du lysat cellulaire (surnageant) a été séparée par centrifugation de la fraction insoluble (culot). La migration sur SDS-PAGE démontre que la plupart (∼90%) de la protéine HS-DpRTCB2 est insoluble **(Figure 3.4)**. Nous avons observé un patron similaire pour les protéines HS-DpRTCB1 et HS-DpRTCB3 [**données non montrées**]. Ceci était prévisible étant donné que toutes les protéines recombinantes devaient être hydrophobes basées sur leur score GRAVY respectif de: -0.42 (HS-DpRTCB1), -0.46 (HS-DpRTCB2) et -0.34 (HS-DpRTCB3) **(http://www.gravy-calculator.de)**. Ce qui

est toutefois plus surprenant, est que leur score GRAVY est très similaire à celui de la protéine recombinante His_{10}-SUMO3-EcRtcB (HS-EcRtcB) de -0.41 qui elle est soluble [**32**]. Cette comparaison a été faite avec cette protéine spécifiquement (et non pas PhRtcB), puisque le construit de HS-EcRtcB est identique à celui des protéines HS-DpRTCB. Le potentiel d'agrégation est aussi grand pour les trois protéines diplonémides et se comparent bien aux protéines amyloïdes qui ont tendance à agréger. Toutefois, ces valeurs sont aussi similaires pour HS-EcRtcB. Selon le logiciel Aggrescan [**126**], les valeurs a^4v de la somme de la séquence normalisée pour 100 résidus (Na^4vSS) sont de -12.9 pour HS-DpRTCB1, -14.3 pour HS-DpRTCB2 et -16.1 pour HS-DpRTCB3 **(Tableau 3.2** et **Figure 3.5 A, B, C et E)**. Ces résultats indiquent donc que l'insolubilité des protéines HS-DpRTCB résiderait potentiellement dans la structure secondaire ou tertiaire de ces protéines plutôt qu'au niveau de leur séquence primaire, puisque contrairement aux HS-DpRTCB, la protéine recombinante HS-EcRtcB, qui possède un construit similaire, est soluble.

Tableau 3.2. Potentiel d'agrégation de HS-DpRTCB1, HS-DpRTCB2, HS-DpRTCB3 et HS-GS-DpRTCB1. Les valeurs obtenues pour la protéine recombinante HS-EcRtcB et des données concernant des protéines amyloïdes [**126**] ont été obtenues aux fins de comparaison.

Nom de la séquence	HS-DpRTCB1	HS-DpRTCB2	HS-DpRTCB3	HS-GS-DpRTCB1	HS-EcRtcB	Protéines amyloïdes
a3v Moyenne de séquence (a3vSA)	-0.124	-0.139	-0.157	-0.139	-0.107	-0.120
Nombre de zones chaudes (nHS)	16.00	15.00	15.00	16.00	15.00	5.86
nHS normalisé pour 100 résidus (NnHS)	2.963	2.373	2.344	2.883	2.825	2.890
Zone du profile au-dessus du seuil (AAT)	41.484	40.512	42.48	41.635	43.079	24.510
Total de surface de la zone chaude (THSA)	35.497	32.939	34.25	35.497	34.617	21.260
Surface totale (TA)	-58.544	-77.389	-89.62	-68.004	-47.941	-26.420
AAT par résidu (AATr)	0.077	0.064	0.066	0.075	0.081	0.130
THSA par résidu (THSAr)	0.066	0.052	0.054	0.064	0.065	0.110
a4v de la somme de la séquence normalisée pour 100 résidus (Na4vSS)	-12.90	-14.30	-16.10	-14.30	-11.10	-12.96

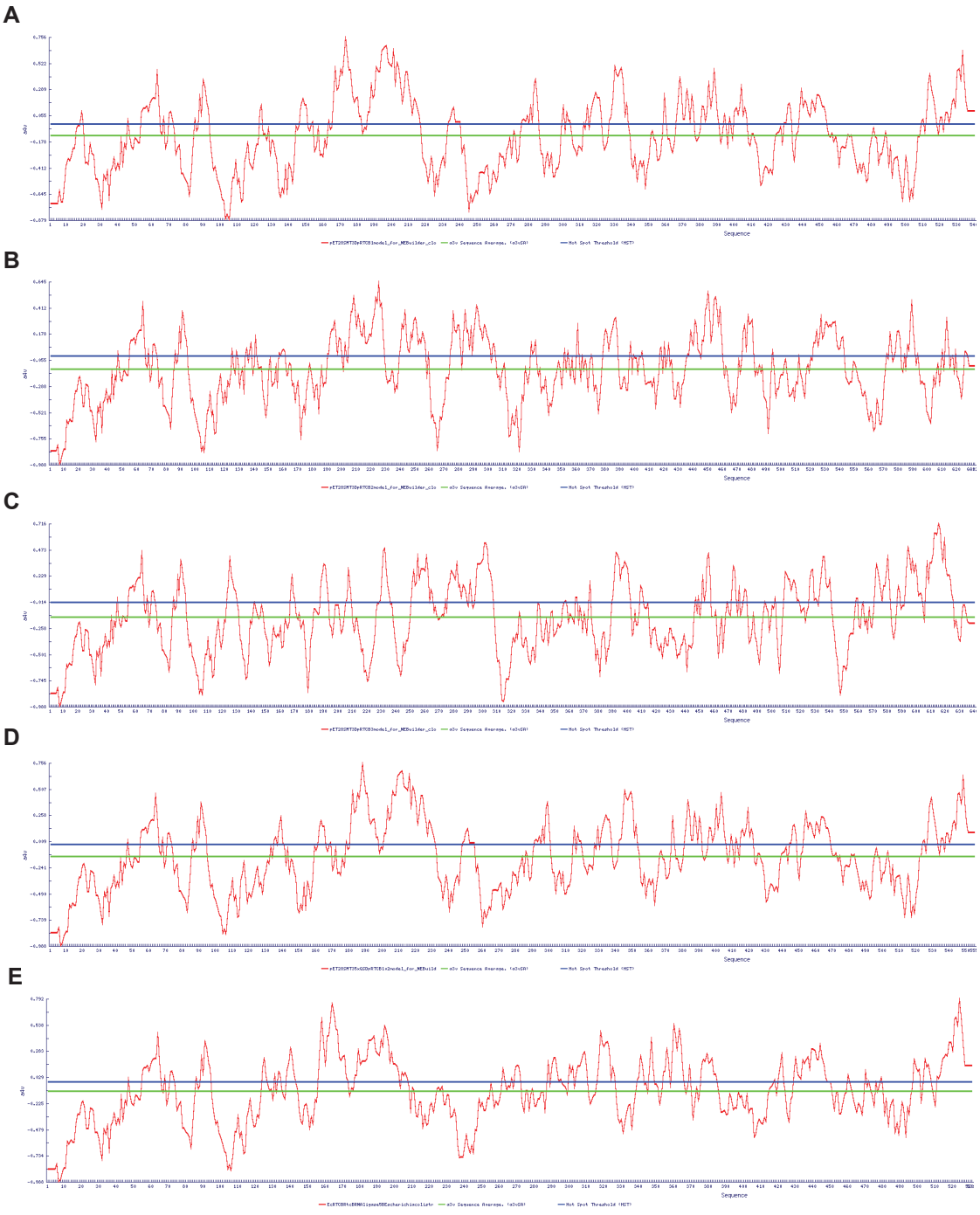

Fig. 3.5. Profiles d'agrégation de **A)** HS-DpRTCB1, **B)** HS-DpRTCB2, **C)** HS-DpRTCB3, **D)** HS-GS-DpRTCB1 et **E)** HS-EcRtcB obtenus à l'aide du logiciel Aggrescan. La ligne en bleu représente le seuil d'agrégation (fixé à 0.02) et la ligne en vert la moyenne a^3v (la prédiction de zones chaudes d'agrégations [**127**]) sur la séquence entière d'acides aminés (a^3vSA).

3.3.4. Essais de solubilisation de HS-DpRTCB2

Vu la faible solubilité de HS-DpRTCB2, nous avons testé diverses techniques afin de rendre cette protéine soluble. Tout d'abord, nous avons testé si HS-DpRTCB2 serait plus soluble dans un milieu plus concentré en sels en ajoutant une plus haute concentration de NaCl dans le tampon de lyse (300 mM à 1 M). De plus, par un traitement avec la nucléase Benzonase® après la lyse cellulaire, nous avons testé si les protéines étaient liées à des acides nucléiques. La séparation du surnageant et du culot sur gel SDS-PAGE montre que la majeure partie (∼80%) des protéines demeurait toujours dans le culot **(Figure 3.6)**.

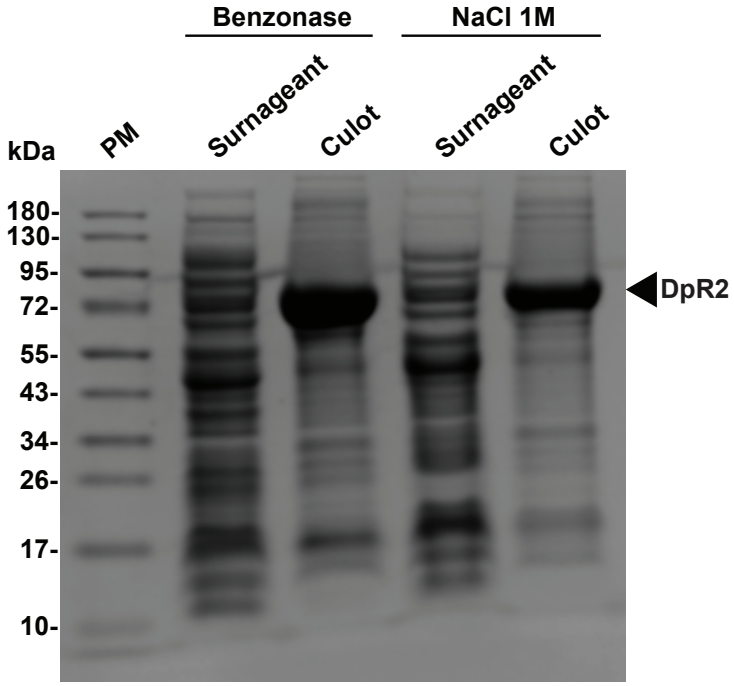

Fig. 3.6. Séparation sur gel Tris-tricine SDS-PAGE des échantillons suivant la surexpression et la lyse cellulaire des transformants Rosetta2-DpRTCB2 (DpR2). Les lysats ont été soit incubés avec la nucléase Benzonase, soit traités avec une quantité de NaCl accrue dans le tampon de lyse (1 M).

Suite à ces résultats, nous supposions que la faible fraction de protéine soluble était due à une piètre lyse des cellules qui avait été effectuée par sonication. C'est pourquoi nous avons changé la méthode de lyse des cellules en utilisant le « French press », mais la lyse par ces deux méthodes a donné des résultats similaires en termes de solubilité de HS-DpRTCB2 **(Figure 3.7 A)**.

Les résultats ci-haut suggéraient que les protéines étaient prisonnières dans des corps d'inclusion. Les corps d'inclusion sont typiquement des agrégats insolubles de protéines

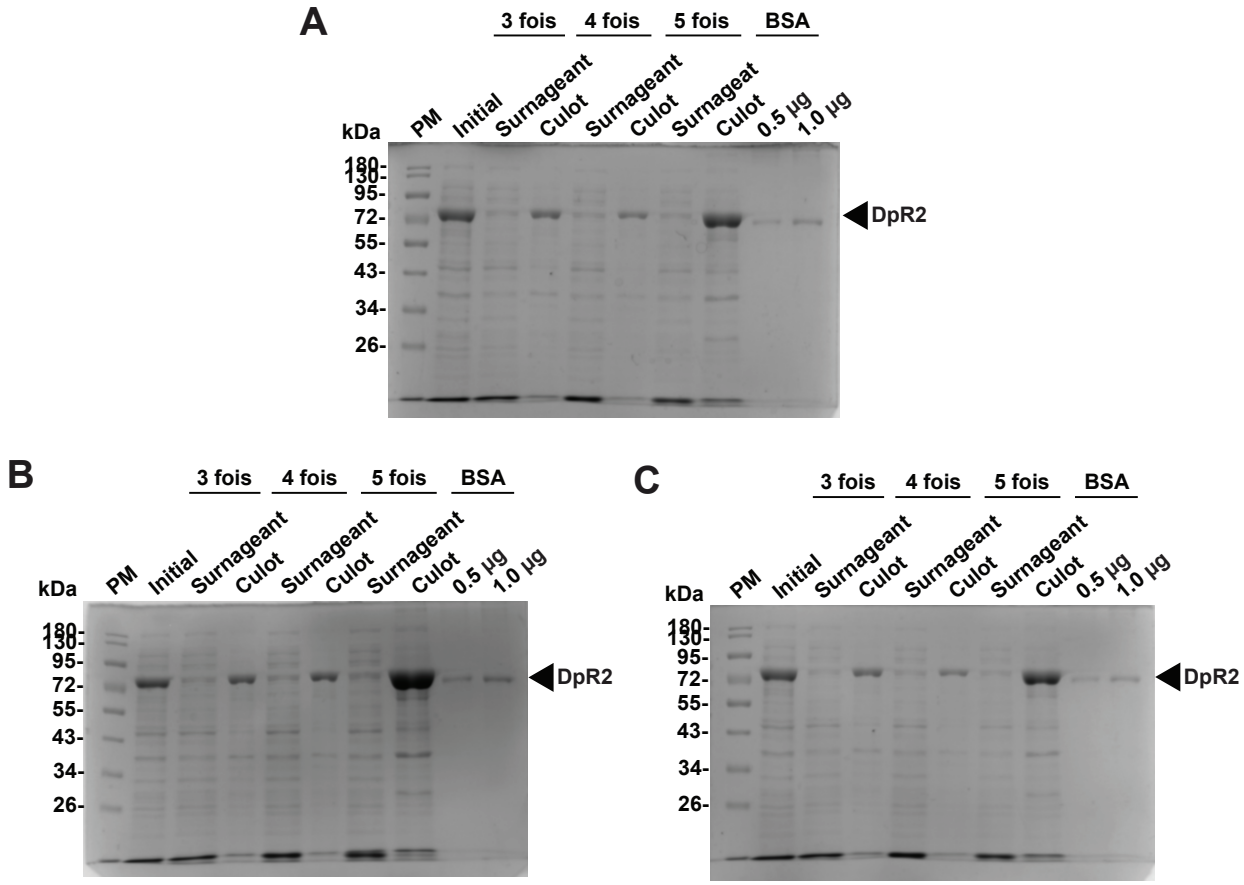

Fig. 3.7. Séparation sur gel Tris-glycine SDS-PAGE des cellules Rosetta2-DpRTCB2 (DpR2) induits lysés par « French press » après trois à cinq passages dans l'appareil. Le surnageant et le culot suivant chaque passage ainsi que l'albumine de sérum bovin (BSA) ont été chargés sur gel. Pour chaque condition, la surexpression a été effectuée à **A)** 18°C sans manganèse, **B)** 11°C dans un milieu auquel a été ajouté 2 mM de manganèse et **C)** à 11°C sans manganèse.

non-repliées ou mal repliées [**128, 129**], mais une certaine fraction de ces protéines peuvent aussi être bien repliées et entièrement fonctionnelles [**130**]. Nous avons tout d'abord testé si soit l'ajout du cofacteur (Mn^{2+}) de RtcB ou la culture des cellules à basse température aiderait au repliement des protéines les rendant ainsi plus solubles. Or, suivant la lyse, les protéines HS-DpRTCB2 démontraient le même comportement qu'en l'absence du cofacteur et à température normalement utilisée pour la surexpression **(Figure 3.7 B et C)**.

Puisque nos essais n'aidaient pas à solubiliser les protéines durant la lyse, nous avons ensuite tenté de déterminer s'il était possible d'extraire les protéines directement du culot suivant la lyse. Nous avons essayé de libérer avec du détergent et de l'urée la protéine HS-DpRTCB2 des corps d'inclusion dans le culot. Ceci impliquait une incubation du culot avec 0.1% de sarkosyl ou 4 M d'urée pour deux périodes d'incubation différentes, puis de

récolter le surnageant suivant cette incubation **(Figure 3.8 A)**. Le sarkosyl n'a pas libéré les protéines des corps d'inclusion, mais l'urée a réussi à solubiliser ∼25% des protéines HS-DpRTCB2 du culot.

Étant donné que l'urée était le seul composé qui libérait les protéines du culot, ceci nous a fourni une piste afin d'effectuer la rupture des cellules en utilisant un tampon de lyse contenant de l'urée. Nous avons testé ces conditions non seulement sur les clones Rosetta2-DpRTCB2, mais aussi sur les transformants Rosetta2-DpRTCB1 et Rosetta2-DpRTCB3 ayant tous surexprimé leur protéine respective **(Figure 3.8 B)**. La dénaturation de HS-DpRTCB2 à l'aide de l'urée a permis de libérer jusqu'à ∼50% de la protéine dans le surnageant **(Figure 3.8 B)** vis-à-vis de ∼30% en l'absence d'urée **(Figure 3.8 C)**.

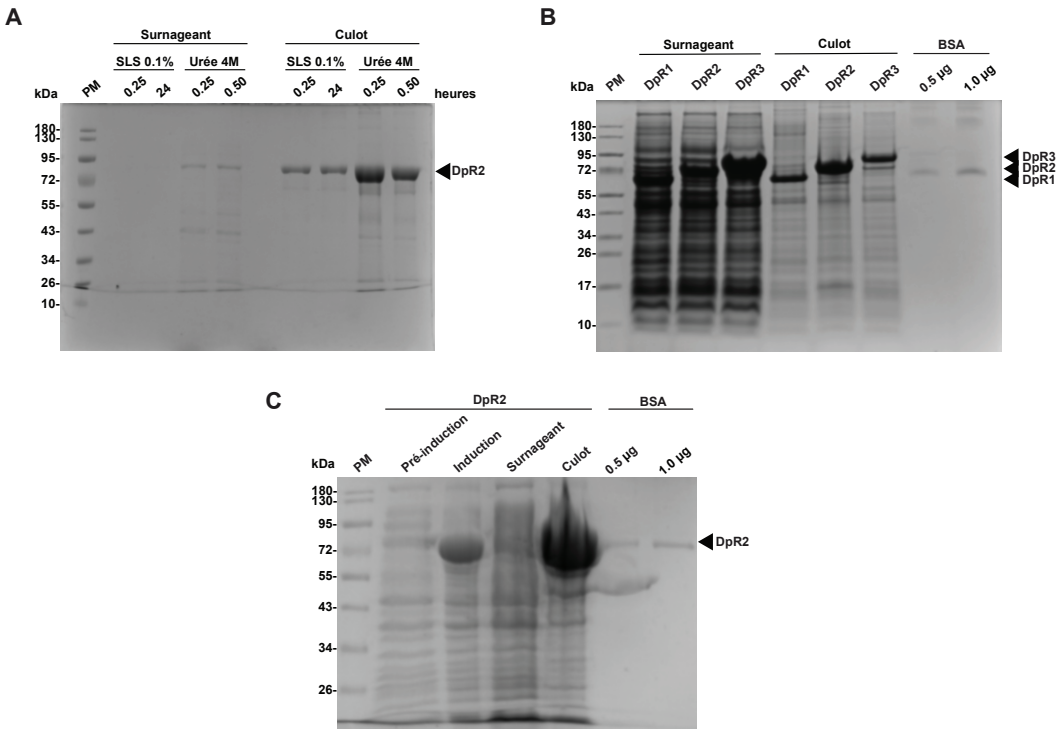

Fig. 3.8. Séparation sur gel Tris-tricine SDS-PAGE des échantillons suivant la surexpression et la lyse cellulaire des transformants Rosetta2-DpRTCB2 (DpR2) en présence de détergent ou d'urée. **A)** Incubation du culot obtenu durant la lyse cellulaire des transformants Rosetta2-DpRTCB2 ayant surexprimé la protéine d'intérêt et récolte du surnageant et du culot suivant cette incubation. Les culots ont été incubés avec 4 M d'urée ou 0.1 % de sodium lauryol sarcosinate (SLS) respectivement. **B)** Échantillons du surnageant et du culot des transformants Rosetta2-DpRTCB1 (DpR1), Rosetta2-DpRTCB2 (DpR2) et Rosetta2-DpRTCB3 (DpR3) suite à la lyse cellulaire dans des conditions dénaturantes avec 4 M d'urée. **C)** Cette figure est identique à la **Figure 3.4**.

3.3.5. Purification de HS-DpRTCB2 par affinité

Nous avons tenté de purifier la portion soluble de la protéine HS-DpRTCB2 obtenue grâce à la lyse cellulaire avec l'urée. Nous avons utilisé des tampons durant la purification permettant à la protéine de demeurer dans des conditions natives (sans urée) ou des conditions dénaturantes (avec urée). La purification a été effectuée par un passage du lysat à travers une colonne de cobalt car ces cations bivalents lient moins fortement l'étiquette His qu'une colonne de nickel, mais assurerait une meilleure pureté de la protéine résultante.

Des échantillons des différentes étapes de purification de la protéine ont été séparés sur gel SDS-PAGE **(Figure 3.9 A et B)**. HS-DpRTCB2 purifiée dans des conditions dénaturantes possède un rendement de ∼5% alors que celle purifiée dans des conditions natives de ∼2%. Le bas rendement obtenu durant la purification est dû à une faible liaison de la protéine à la colonne. Ce comportement ne changeait pas si l'on compare l'utilisation d'un tampon à base de phosphate ($Na \bullet HPO_4$) comparé à celui que nous utilisions à base de Tris **(Figure 3.9 C et D)**.

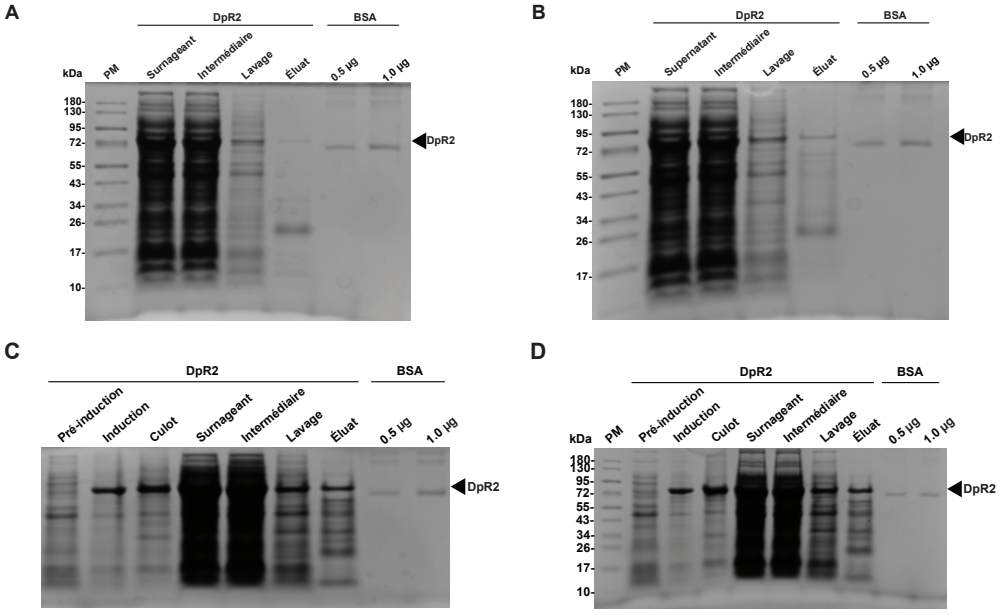

Fig. 3.9. Séparation sur gel Tris-tricine SDS-PAGE des échantillons obtenus durant la purification du lysat cellulaire des transformants ayant surexprimé la protéine HS-DpRTCB2 (DpR2). Un volume équivalent pour chaque étape de purification a été chargé. La lyse cellulaire a été effectuée en conditions dénaturantes avec 4 M d'urée. L'albumine de sérum bovin (BSA) a aussi été chargée avec les échantillons. Le surnageant a été purifié initialement soit **A)** en conditions natives ou **B)** en conditions dénaturantes. Par la suite, la même protéine a été purifiée en conditions dénaturantes avec des tampons contenant soit **C)** Tris soit **D)** phosphate ($Na \bullet HPO_4$).

Avant de procéder au clivage du tag de la protéine HS-DpRTCB2, il faut retirer de l'échantillon l'urée ainsi que l'imidazole, ce dernier étant requis pour relâcher la protéine de la colonne d'affinité, puisque la protéase n'est pas active en présence de ces produits. Nous avons donc testé l'échange de tampon pour la protéine HS-DpRTCB2 en utilisant l'ultrafiltration. Des échantillons ont été prélevés aux différentes étapes de l'expérience et séparés sur gel SDS-PAGE **(Figure 3.10)**. Apparemment, la majeure partie de la protéine recombinante demeure attachée à la membrane de filtration causant davantage de perte de la protéine puisque le pourcentage de protéine récupérée par cette procédure est ∼50%. Finalement, nous avons utilisé la dialyse comme méthode d'échange de tampon, mais la protéine a précipité durant la procédure [**données non affichées**]. Il semble que la protéine requiert de l'urée pour demeurer soluble.

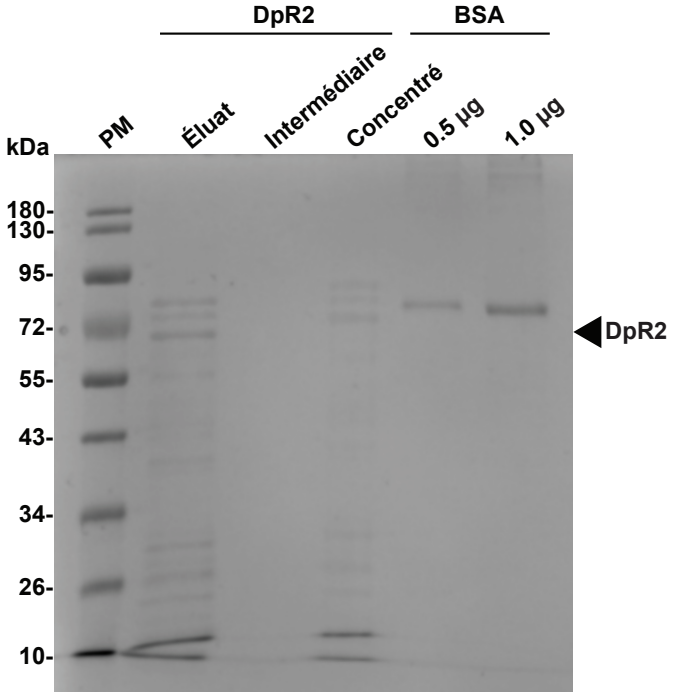

Fig. 3.10. Séparation des échantillons obtenus lors de l'échange de tampon sur gel Tris-tricine SDS-PAGE. La concentration de la protéine purifiée HS-DpRTCB2 dans des conditions dénaturantes avec 4 M d'urée par ultrafiltration. L'albumine de sérum bovin (BSA) a aussi été chargée sur gel.

3.3.6. Digestion par Ulp1 de la protéine purifiée HS-DpRTCB2

Suivant l'échange de tampon, nous avons procédé au retrait du tag His_{10}-SUMO3 de la protéine recombinante HS-DpRTCB2 avec la protéase Ulp1 que nous avons préparée et

purifiée en laboratoire [**voir Méthodes, section 3.2**]. Cette protéase, qui clive spécifiquement au niveau N-terminal entre la séquence protéique et le tag SUMO3, a été utilisée à différentes concentrations et différents temps d'incubation pour digérer HS-DpRTCB2. La séparation sur SDS-PAGE démontre que HS-DpRTCB2 conserve toujours son tag peu importe la concentration d'Ulp1 utilisée ou le temps de digestion puisque la taille de la protéine avec tag est ∼70.4 kDa alors que celle de la protéine sans tag devrait être ∼56.3 kDa **(Figure 3.11 A)**.

Suspectant qu'il y ait eu des problèmes lors de la préparation de la protéase Ulp1 maison, nous avons effectué deux expériences contrôles : (i) le clivage de HS-DpRTCB2 avec une Ulp1 préparée par un autre laboratoire et démontrée active et (ii) le traitement d'une protéine portant un tag SUMO3 laquelle a été démontrée préalablement d'être clivable. La protéase et la protéine contrôles ont été offertes par le Dr. John Pascal de notre département. Tout d'abord, le clivage de la protéine HS-DpRTCB2 avec la protéine contrôle Ulp1 a donné des résultats similaires à ce qui a déjà été observé avec la Ulp1 maison : le tag His_{10}-SUMO3 demeure joint à la protéine de fusion **(Figure 3.11 B)**. Par la suite, l'immunobuvardage de type Western avec des anticorps anti-SUMO3 démontre que la Ulp1 maison est en mesure de cliver la protéine contrôle possédant le tag SUMO3 **(Figure 3.12)**. Nous en avons conclu que le repliement protéique de HS-DpRTCB2 rendait son tag indisponible pour la digestion.

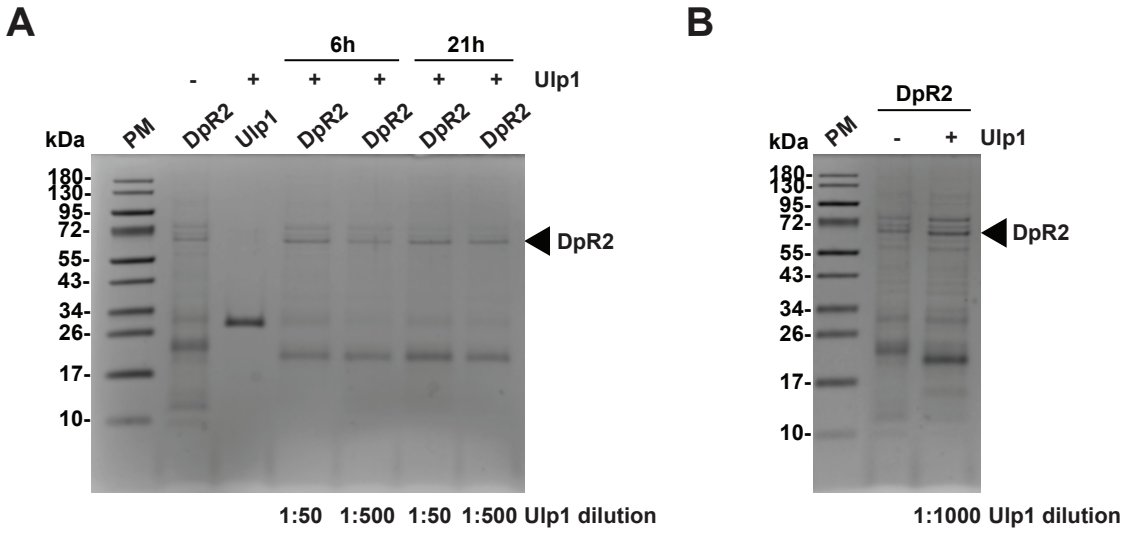

Fig. 3.11. Séparation des échantillons par gel Tris-tricine SDS-PAGE de la protéine HS-DpRTCB2 (DpR2) suivant la digestion avec la protéase Ulp1. La protéine recombinante a été **A)** digérée avec la protéase Ulp1 purifiée en laboratoire à différentes concentrations et pour différentes périodes de temps ou **B)** digérée à l'aide d'une protéine Ulp1 contrôle.

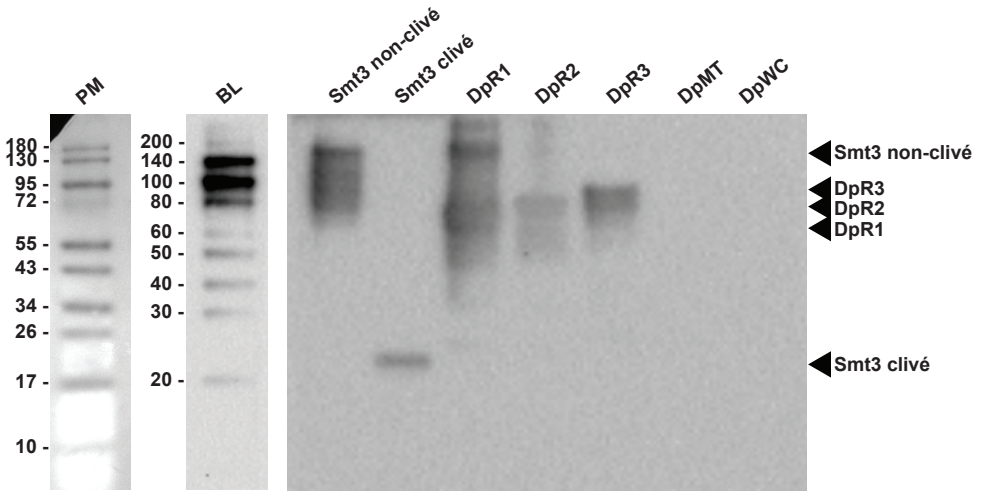

Fig. 3.12. Détection par immunobuvardage de type Western des protéines HS-DpRTCB1 (DpR1), HS-DpRTCB2 (DpR2) et HS-DpRTCB3 (DpR3) avec des anticorps anti-SUMO3. Une protéine contrôle possédant le tag SUMO3 (∼140 kDa) non clivé et clivé par la protéase Ulp1 faite maison ont été chargés sur gel. L'échelle moléculaire biotinylée (BL) a été chargée comme contrôle de taille et détectée à l'aide d'anticorps anti-biotine. Des lysats cellulaires (DpWC) et mitochondriaux (DpMT) de *Diplonema* ont été chargés sur gel comme contrôles.

3.3.7. L'ajout d'un espaceur afin d'améliorer la solubilité de HS-DpRTCB1

Afin de mieux exposer le tag des protéines recombinantes DpRTCB, nous avons refait le construit du plasmide et introduit une répétition flexible penta-(glycine-sérine) avant le tag His_{10}-SUMO3. Ces expériences ont été effectuées avec DpRTCB1. Toutefois, avec un score GRAVY de -0.45 et des valeurs Na^4vSS de -14.30, la nouvelle protéine recombinante HS-GS-DpRTCB1 (∼61.7 kDa) est prédite de présenter les mêmes problèmes de solubilité que les autres HS-DpRTCB **(Tableau 3.2 et Figure 3.5 D)**. Effectivement, suivant la surexpression et la lyse cellulaire, HS-GS-DpRTCB1 démontre un même comportement d'agrégation que HS-DpRTCB2 qui ne possède pas l'espaceur **(Figure 3.13 A)**. Donc, l'extension glycine-sérine n'a pas rendu le tag plus accessible au solvant.

3.3.8. Essais de solubilisation de HS-GS-DpRTCB1 à l'aide d'arginine

Il a été préalablement démontré que l'arginine facilite la solubilisation des protéines et favorise leur repliement [**131**]. Nous avons testé ces conditions avec HS-GS-DpRTCB1. La première condition que nous avons testée était l'ajout de 0.5 M d'arginine au milieu de culture avant la croissance bactérienne. Or, malgré l'ajustement du pH à une valeur

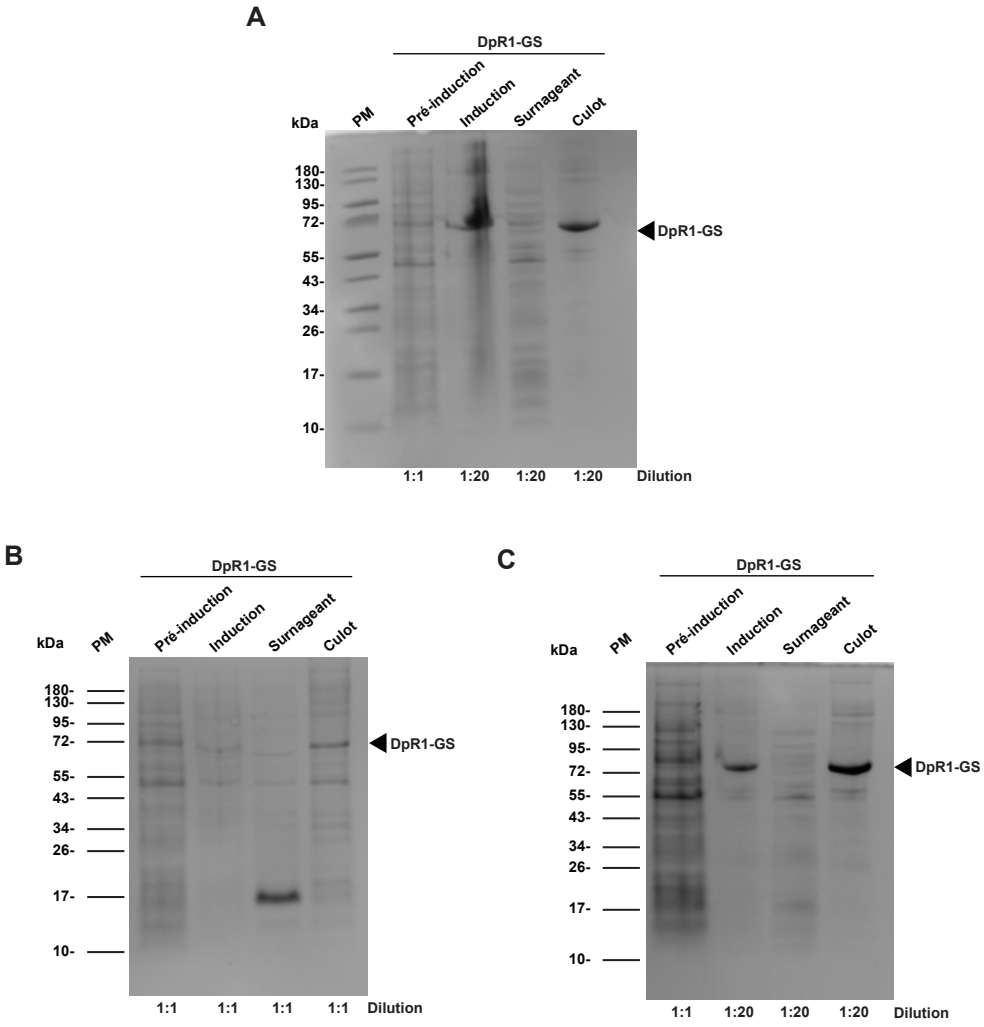

Fig. 3.13. Séparation des échantillons sur Tris-tricine SDS-PAGE obtenus lors de la lyse cellulaire des transformants Rosetta2-GS-DpRTCB1-GS (DpR1-GS) ayant surexprimé leur protéine d'intérêt. Des échantillons des cellules avant l'induction, suivant l'induction et les surnageant et culots obtenus lors de lyse cellulaire ont été chargé sur gel. Les transformants ont été **A)** lysés dans des conditions natives **B)** ont exprimé leur protéine suite à l'ajout de 0.5 M d'arginine dans le milieu au moment de l'induction pour être par la suite lysé et **C)** ont été induits dans les conditions mentionnées en A) puis lysés dans un tampon auquel a été ajouté 0.5 M d'arginine. Le pH du tampon a été ajusté avec du HCl afin de respecter le pH du tampon préalablement utilisé pour la lyse des transformants Rosetta2-DpRTCB1.

physiologique de ∼7.5, les cellules étaient incapables de croître [**données non affichées**]. Par la suite, au moment d'induction, nous avons ajouté 0.5 M d'arginine au milieu. Suivant la surexpression, la quantité de cellules récoltée était ∼50% inférieure à n'importe quelle autre condition que nous avions préalablement testée [**données non affichées**]. Par conséquent, les quantités de protéines obtenues étaient considérablement plus basses lors de la surexpression et de la lyse **(Figure 3.13 B)**. Comme dernière tentative, nous avons

ajouté de l'arginine au tampon de lyse. Suite à ce traitement ∼20% de HS-DpRTCB1-GS était soluble comparée à ∼30% dans l'absence d'arginine **(Figure 3.13 A et C et Figure 3.14)**. En conclusion, l'arginine n'a pas rendu HS-GS-DpRTCB1 plus soluble, au contraire, la protéine est devenue moins soluble.

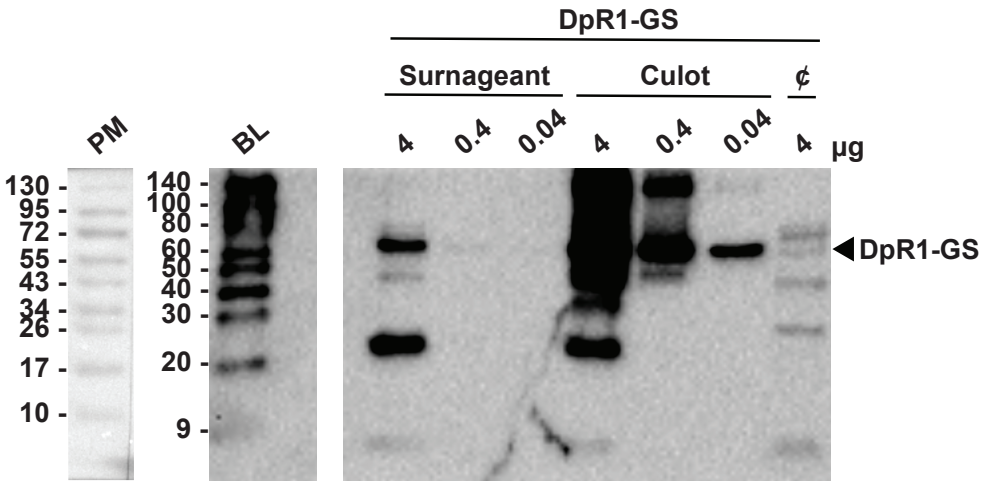

Fig. 3.14. Détection par immunobuvardage de type Western des protéines HS-GS-DpRTCB1 (DpR1-GS) suivant l'induction et la lyse des transformants avec des anticorps anti-SUMO3. L'échelle biotinylée (BL) a été utilisée comme marqueur de taille et a été détectée à l'aide d'un anticorps anti-biotine. La lyse des cellules s'est effectuée dans un tampon contenant 0.5 M d'arginine.

3.4. Conclusion

Nos résultats démontrent que les protéines recombinantes HS-DpRTCB sont très peu solubles et que seul des composés dénaturants tel que l'urée aide à solubiliser ces protéines. À noter que l'extension glycine-sérine et l'ajout d'arginine n'ont pas été appliqués à HS-DpRTCB2 et 3, puisque les chances de surmonter le problème de solubilité n'étaient pas très élevées au vu des potentiels d'agrégation de ces deux protéines qui sont plus hauts que HS-DpRTCB1. Les travaux futurs comprennent la surexpression protéique à l'aide de nouvelles souches bactériennes (tel que Lumo21 (DE3) ou Rosetta2 pLyS [**128**]) qui permettent de réguler l'expression protéique réduisant ainsi les chances d'obtenir des corps d'inclusion. Une autre option impliquerait l'immunoprécipitation des protéines DpRTCB discutée plus en détail dans la **section 4.6**.

3.5. Remerciements

Nous aimerions remercier le Dr. Stewart Shuman (Sloan Kettering Institute, États-Unis) pour le don du plasmide dérivé pET28b permettant l'expression de la protéine recombinante EcRtcB-SUMO3. Nous aimerions aussi remercier le Dr. John Pascal (Université de Montréal, Canada) pour le don des contrôles de la protéase Ulp1 et de la protéine recombinante contrôle SUMO3. Nous aimerions aussi reconnaître la contribution de Stéphanie Dufresne (Université de Montréal, Canada) pour la construction des plasmides exprimant les protéines recombinantes HS-DpRTCB2 et HS-DpRTCB3.

Chapitre 4

Discussion

Tel que mentionné précédemment, *D. papillatum* est un organisme particulier reconnu non seulement pour l'assemblage de ses transcrits mitochondriaux par épissage massif en *trans*, mais aussi pour être un des rares eucaryotes qui possède trois gènes codant pour des ligases à ARN de type RtcB. Le but de notre étude était de caractériser ces ligases par des approches bio-informatiques et biochimiques. Notamment, nous avons modélisé *in silico* les structures tridimensionnelles et établi l'arbre phylogénétique des ligases DpRTCB. De plus, nous avons surexprimé les ligases DpRTCB afin d'obtenir des protéines pures pour des essais enzymatiques. Toutefois, la purification de ces enzymes n'a pas été un succès avec les méthodes standards utilisées.

4.1. Séquences protéiques des DpRTCB

L'analyse des séquences protéiques des DpRTCB a révélé que DpRTCB1 possède un signal de localisation mitochondrial [**voir section 2.4.1**]. Il est donc fort possible que cette ligase soit impliquée dans l'épissage en *trans* unique observé dans la mitochondrie des diplonémides [**70**]. En contraste, la localisation prédite de DpRTCB2 et DpRTCB3 est majoritairement cytosolique.

Toutes les DpRTCB ont un haut niveau de conservation des résidus critiques nécessaires à l'activité enzymatique des ligases de type RtcB [**voir section 2.4.1**]. Toutefois, DpRTCB1 possède une cystéine dans son site actif, qui est hautement conservé chez les RTCB1 des diplonémides [**voir section 2.4.1**] et demeure rare chez d'autres séquences de type RtcB, plutôt qu'une sérine qui est plus hautement conservée en cette même position. Nous postulons que ce résidu permet aux ligases RTCB1 d'interagir avec des cofacteurs autres que le GTP par exemple l'ATP. D'autres hypothèses qui peuvent expliquer cette variation incluent un changement au niveau de la vitesse de réaction de cette ligase ou

encore la capacité d'agir en tant qu'interrupteur redox pouvant activer la ligase dépendant des conditions mitochondriales. Ces interrupteurs ont déjà été étudiés chez la mitochondrie [**113, 114**] et chez les chloroplastes comme par exemple lors de la réduction d'hydrolyse de l'ATP par oxydation des cystéines au niveau de l'ATP synthase de l'épinard [**115**]. L'effet de la substitution de la sérine en cystéine pourrait être examiné par des essais enzymatiques *in vitro*.

4.2. Modélisation *in silico* des DpRTCB

Les prédictions de structure tridimensionnelle des DpRTCB indiquent que presque tous les résidus participant au site actif de DpRTCB1 sont conservés. Malgré que leur orientation ne soit pas exactement la même que celle de la structure de référence, leurs positions au sein des modèles sont identiques. L'analyse structurale de DpRTCB1 démontre que la cystéine présente au niveau du site actif pourrait interagir avec la base guanylée de la même façon que la sérine à la position correspondante chez PhRtcB [**60, 62**] et celle prédite chez EcRtcB [**voir section 2.4.2**].

La modélisation structurale a apporté également des informations sur les régions qui interagissent possiblement avec l'un des cofacteurs des ligases de type RtcB [**voir section 2.4.2**]. La structure de PhRtcB possède cinq hélices α qui se retrouvent en proximité du site actif de l'enzyme [**60, 62**] et qui, en contraste, sont absentes dans des structures prédites de EcRtcB et DpRTCB1. Puisque PhRtcB est une ligase Archease-dépendante et que EcRtcB ne l'est pas, nous suspectons que le cofacteur Archease interagit avec les hélices présentes chez PhRtcB. Ceci ajoute davantage à l'hypothèse que DpRTCB1 est possiblement Archease-indépendante. L'obtention d'une structure cristallographique d'une RtcB Archease-indépendante permettrait de confirmer la présence ou l'absence de ces hélices.

4.3. Regroupement phylogénétique des ligases de type RtcB

En analysant la phylogénie de ligases RtcB présentes à travers les trois domaines de la vie (ainsi que chez les virus), il a été possible de classifier celles-ci en au moins neuf clades majeur distincts [**voir section 2.4.3 et section 2.4.4**]. Les ligases de type RtcB dont le rôle biologique est connu se regroupent dans des clades cohérents (Clades I à III) alors que d'autres clades semblent se former grâce à différentes caractéristiques structurales (Clades IV et V) ou par une divergence significative des acides aminés interagissant avec

les ions Mn^{2+} et le GTP (Clades VI à IX). Ainsi, nous pouvons prédire les potentiels rôles cellulaires exercés par les ligases de type RtcB retrouvées chez *Diplonema*. Plus précisément, DpRTCB2 s'associe avec le Clade II formée par les ligases des archées et des métazoaires impliquées dans l'épissage des ARNt, où les ligases des métazoaires peuvent aussi effectuer l'épissage non-conventionnel de l'ARNm *XPB1* durant l'UPR [**41, 43, 44, 46**]. Cela indique l'implication de DpRTCB2 dans la maturation des ARNt (encodés par le génome nucléaire) et possiblement dans une sorte d'épissage non-conventionnel de certains ARNm cytosoliques. Considérant que toutes les enzymes connues appartenant au Clade II sont Archease-dépendantes [**36, 37**], nous proposons que DpRTCB2 dépende aussi de ce cofacteur qui est en effet encodé par le génome nucléaire de *Diplonema*. Dans le cas de DpRTCB3, celle-ci s'associe avec le Clade III où se retrouve la ligase MxRtcB3 qui a été démontrée *in vitro* d'ajouter des coiffes aux acides nucléiques sans discerner si le substrat est de l'ADN ou de l'ARN, ou si l'extrémité coiffée est le $3'$-PO_4 ou le $5'$-PO_4 [**38**]. DpRTCB3 pourrait donc être en mesure d'ajouter des coiffes guanylées aux acides nucléiques de manière similaire à MxRtcB3. Finalement, DpRTCB1 se place dans le Clade I ensemble avec les séquences EcRtcB, MxRtcB1, les RTCB des kinétoplastides et les RTCB1 des autres diplonémides. Les enzymes connues retrouvées dans ce clade sont Archease-indépendantes, suggérant que ce soit aussi le cas pour DpRTCB1. De plus, les ligases RTCB1 des diplonémides sont bien isolées dans ce clade, suggérant un rôle biologique unique pour ces enzymes.

4.4. L'expression hétérologue des protéines HS-DpRTCB

Suivant les analyses *in silico* nous avons procédé avec l'expression des protéines recombinantes DpRTCB dans *E. coli* afin de purifier ces ligases pour de futurs essais enzymatiques. Lors de nos expériences en laboratoire nous avons observé que HS-GS-DpRTCB1 est exprimée de façon constitutive, donc sans induction avec l'IPTG **(voir Figure A.1)**. Nous suspectons que cela ait aussi été le cas pour les autres protéines HS-DpRTCB. D'autres chercheurs ont déjà observé ce phénomène appelé « leaky expression » chez certaines souches d'*E. coli* (par exemple BL21) exprimant des protéines hétérologues sous le control du promoteur *lac* [**128, 132, 133**]. Ce phénomène peut se produire à cause du milieu utilisé pour la croissance cellulaire qui contient de la tryptone, un peptone obtenu à partir de la digestion de la caséine par la trypsine [**134**]. Puisque l'IPTG imite un métabolite du lactose appelé allolactose, la tryptone impure pourrait activer le promoteur. Le glucose a aussi été démontré d'avoir un effet sur la production de la polymérase T7 [**135, 136**] en ralentissant son expression par la répression du promoteur [**137**]. Toutefois, même en ajoutant la plus petite quantité possible d'inducteur et en incubant les transformants pour de longues

périodes à basse température, cette « leaky expression » des protéines HS-GS-DpRTCB1 n'a pas pu être supprimée. La « leaky expression » a peut-être causé un environnement plus favorable à l'agrégation protéique.

4.5. La faible solubilité des HS-DpRTCB est causée par des corps d'inclusion

Nous avons initialement pensé que les protéines HS-DpRTCB étaient prises avec les composantes insolubles bactériennes tel que l'ADN génomique, ou alternativement, qu'elles étaient insolubles à cause d'une lyse cellulaire partielle. Puisque nous n'avons pas été en mesure de relâcher les protéines suite à un traitement à la nucléase ou par sonication (ce qui détruit les acides nucléiques par force de cisaillement [**138**]) [**voir section 3.3.3**], nous avons conclu que les protéines HS-DpRTCB étaient emprisonnées dans des corps d'inclusion. Ce phénomène est commun chez les bactéries et est souvent observé lorsqu'*E. coli* est utilisé pour l'expression de protéines hétérologues [**128, 129**]. Les corps d'inclusion, contrairement à la pensée commune, ne sont pas toujours des agrégats de protéines mal repliées. Ils peuvent être composés d'un mélange de protéines bien repliées et non repliées, que l'on appelle des corps d'inclusions relâchés, à partir desquels on peut récupérer des protéines entièrement fonctionnelles [**130**]. Un rapport suggère que la sonication peut transformer des corps d'inclusion relâchés en agrégats compacts [**139**]. Toutefois, il est peu probable que les corps d'inclusion insolubles que nous observons soient le résultat de la sonication car nous avons obtenu des résultats similaires avec la sonication et le « French press ».

Plusieurs stratégies ont été développées afin d'éviter la formation des corps d'inclusion, d'aider à solubiliser ces agrégats ou de récupérer des protéines correctement repliées à partir de corps d'inclusions partiellement solubles. Certaines de ces stratégies impliquent l'extraction de protéines à l'aide de conditions non-dénaturantes comme par exemple de : (i) réduire les quantités d'inducteur, (ii) faire croître les cellules à des basses températures [**140**], (iii) ajouter les cofacteurs qui favorisent le repliement des protéines [**141, 142, 143**], (iv) ajouter un tag qui aide à la solubilisation tel que SUMO3 ou la « maltose-binding protein » (MBP) [**144**] ou une extension flexible d'acide aminés qui rend le tag accessible au solvant pour la purification d'affinité [**145**], (v) utiliser un plasmide qui encode une chaperonne aidant au repliement des protéines et qui est exprimée en même temps que la protéine d'intérêt [**129**], (vi) varier la concentration de sels, la composition des tampons et leur pH [**146**], (vii) ajouter du détergent [**147**] ou (viii) ajouter de l'arginine dans le milieu de croissance des clones ou du tampon de lyse qui peut aider à replier et solubiliser la protéine [**131**]. Finalement, des conditions plus rudes, telles que la dénaturation de

la protéine, par exemple en utilisant des agents chaotropiques comme l'urée [**148**] ou le chlorure de guanidium [**149**] peuvent aussi permettre d'extraire les protéines des corps d'inclusion. Toutefois, l'utilisation de conditions dénaturantes impliquent que par la suite, les protéines doivent être repliées sous leur forme native [**voir section 4.6**].

Nous avons testé la plupart des astuces qui ont été démontrées comme étant efficaces dans d'autres systèmes. Premièrement, les protéines recombinantes HS-DpRTCB possèdent une extension SUMO3 en plus du tag histidine afin d'améliorer leur solubilité. Malgré que le tag His-SUMO3 ait été utilisé avec succès dans la purification d'autres protéines RtcB, pour les protéines DpRTCB, ce tag demeurait apparemment inaccessible pour purification par affinité ou pour digestion par Ulp1 [**voir section 3.3.5**] **et** [**section 3.3.6**]. Ceci indique que la structure même des protéines HS-DpRTCB rend le tag inaccessible au solvant. Même l'insertion d'un penta-glycine-sérine entre la séquence RTCB1 et le tag n'a pas résolu ce problème [**voir section 3.3.7**]. Il se peut que l'extension n'ait pas été assez longue pour exposer le tag. Nous avons aussi essayé de solubiliser les protéines contenues dans le culot à l'aide de détergent ou encore de tenter de rendre ces protéines solubles durant la rupture cellulaire en changeant la composition du tampon de lyse par l'ajout d'arginine ou en modifiant sa composition de base. Or, ces conditions n'ont pas libéré les protéines des corps d'inclusion. Seules des conditions plus rudes telles que l'urée ont permis de solubiliser les HS-DpRTCB. Ceci suggère que les protéines à l'intérieur des corps d'inclusion forment des agrégats compacts et ne peuvent être relâchées à l'aide de méthodes plus douces. Le chlorure de guanidium aurait aussi pu être utilisé comme agent chaotropique. Par contre, ce composé précipite en présence de SDS [**150**] et aurait rendu difficile le suivi des étapes de purification des protéines sur SDS-PAGE. Une alternative potentielle que nous n'avons pas testée dû à un conflit de marqueur de sélection est l'utilisation d'un plasmide permettant la co-expression d'une chaperonne aidant au repliement protéique.

4.6. Méthodes utilisées afin de replier les protéines suite à la dénaturation

Bien que faible, le meilleur rendement obtenu durant la purification de la protéine HS-DpRTCB2 était sous des conditions dénaturantes. Toutefois, ceci requérait que la protéine dénaturée soit ensuite repliée sous sa forme native. Ceci a pu être obtenu à l'aide de la dialyse afin de retirer l'urée, mais malgré tout, la protéine a précipité [**voir section 3.3.5**]. Lorsque nous tentions de retirer l'urée par ultrafiltration, au moins la moitié de la protéine restait adsorbée de façon irréversible au filtre [**voir section 3.3.5**]. Des méthodes alternatives pour retirer les composants dénaturants de l'échantillon protéique

purifié impliquant une dilution ou l'utilisation de puces microfluidiques [**151**] n'ont pas été essayées ici puisqu'elles très sont coûteuses ou demandent beaucoup de temps.

4.7. Approches futures envisagées afin de purifier les protéines HS-DpRTCB

En somme, nos tentatives de solubiliser les protéines HS-DpRTCB et de les purifier avec un rendement acceptable n'ont pas réussi. Dans des situations comme la nôtre, plusieurs options sont possibles afin de faire face à ces problèmes. L'une d'entre elles est d'exprimer la protéine dans un organisme autre qu'*E. coli*. Notre collègue Marek Sebesta de CEITEC Brno en République Tchèque, a essayé de surexprimer les protéines recombinantes DpRTCB dans des cellules d'insectes, où elles ont été fusionnées avec une « maltose-binding protein » (MBP) [**données non affichées**]. Le tag MBP est communément utilisé pour la purification afin d'augmenter la solubilité de la protéine [**152**]. Or, ces protéines de fusion DpRTCB ont démontré le même patron d'agrégation que dans les cellules d'*E. coli*.

De futures expériences envisagées prévoient l'utilisation d'une autre souche d'*E. coli*, soit Lemo21 (DE3), qui permet d'ajuster minutieusement le taux de protéines surexprimées, ou encore l'expression homologue (i.e. dans *Diplonema*) et purifier les DpRTCB par immunoprécipitation en utilisant des anticorps qui reconnaissent un tag précis (par exemple la protéine A) [**153**]. L'alternative la plus prometteuse serait d'exprimer les protéines DpRTCB sans tag de fusion à l'aide d'un système « cell-free » qui utilise la machinerie d'expression protéique d'*E. coli* (par exemple PureExpress de NEB) et dont les enzymes possèdent un tag permettant leur retrait du mélange suite à la réaction. Puisque l'inaccessibilité du tag des HS-DpRTCB est considéré comme l'un des problèmes majeurs dans la purification de ces protéines, leur surexpression sous une forme non-étiquetée est l'une des approches futures priorisées.

Chapitre 5

Conclusion

Les ligases à ARN de type RtcB sont impliquées dans plusieurs processus différents dans la cellule tel que l'épissage, la maturation et la réparation de l'ARN. Pour l'étude de ces ligases, *Diplonema papillatum* est l'organisme de choix, puisqu'il est l'un des rares eucaryotes possédant plusieurs ligases de ce type. De plus, DpRTCB1 est d'un intérêt particulier grâce à son implication potentielle dans le processus d'épissage en *trans* dans la mitochondrie. Toutes nos tentatives de purifier les ligases RTCB de *Diplonema* avec des méthodes biochimiques classiques ont échoué, mais une approche nouvelle, soit l'expression dans un système « cell-free », promet de résoudre le tenace problème d'insolubilité que nous avons rencontré.

L'obtention de protéines DpRTCB servirait non seulement à examiner leur activité enzymatique, mais aussi à produire des anticorps spécifiques pour une future immunolocalisation. Il sera aussi possible d'effectuer une co-immunoprécipitation pour détecter les protéines qui interagissent avec ces enzymes ainsi que les substrats précis de ces ligases. Ceci fournirait davantage d'information sur les ligases de type RtcB qui, lorsque comparées à leurs cousines les ligases ATP-grasp, représentent une famille d'enzymes qui a été très peu explorée.

Références bibliographiques

[1] Moreira S, Breton S, Burger G. Unscrambling genetic information at the RNA level. Wiley Interdiscip Rev RNA. 2012;3(2):213–28.

[2] Pettitt J, Harrison N, Stansfield I, Connolly B, Muller B. The evolution of spliced leader *trans*-splicing in nematodes. Biochem Soc Trans. 2010;38(4):1125–30.

[3] Lasda EL, Blumenthal T. *Trans*-splicing. Wiley Interdiscip Rev RNA. 2011;2(3):417–34.

[4] McNeil BA, Semper C, Zimmerly S. Group II introns: versatile ribozymes and retroelements. Wiley Interdiscip Rev RNA. 2016;7(3):341–55.

[5] Randau L, Calvin K, Hall M, Yuan J, Podar M, Li H, et al. The heteromeric *Nanoarchaeum equitans* splicing endonuclease cleaves noncanonical bulge-helix-bulge motifs of joined tRNA halves. Proc Natl Acad Sci U S A. 2005;102(50):17934–9.

[6] Soma A, Sugahara J, Onodera A, Yachie N, Kanai A, Watanabe S, et al. Identification of highly-disrupted tRNA genes in nuclear genome of the red alga, *Cyanidioschyzon merolae* 10D. Sci Rep. 2013;3(1):2321.

[7] Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G. Novel modes of RNA editing in mitochondria. Nucleic Acids Res. 2016;44(10):4907–19.

[8] Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986;46(6):819–26.

[9] Rudinger M, Fritz-Laylin L, Polsakiewicz M, Knoop V. Plant-type mitochondrial RNA editing in the protist *Naegleria gruberi*. RNA. 2011;17(12):2058–62.

[10] Simpson L, Thiemann OH, Savill NJ, Alfonzo JD, Maslov DA. Evolution of RNA editing in trypanosome mitochondria. Proc Natl Acad Sci U S A. 2000;97(13):6986–93.

[11] Burroughs AM, Aravind L. RNA damage in biological conflicts and the diversity of responding RNA repair systems. Nucleic Acids Res. 2016;44(18):8525–8555.

[12] Shuman S, Lima CD. The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. Curr Opin Struct Biol. 2004;14(6):757–64.

[13] Martin IV, MacNeill SA. ATP-dependent DNA ligases. Genome Biol. 2002;3(4):REVIEWS3005.

[14] Wilkinson A, Day J, Bowater R. Bacterial DNA ligases. Mol Microbiol. 2001;40(6):1241–8.

[15] Unciuleac MC, Goldgur Y, Shuman S. Structure and two-metal mechanism of a eukaryal nick-sealing RNA ligase. Proc Natl Acad Sci U S A. 2015;112(45):13868–73.

[16] Silber R, Malathi VG, Hurwitz J. Purification and properties of bacteriophage T4-induced RNA ligase. Proc Natl Acad Sci U S A. 1972;69(10):3009–13.

[17] Popow J, Schleiffer A, Martinez J. Diversity and roles of (t)RNA ligases. Cell Mol Life Sci. 2012;69(16):2657–70.

[18] Becker HF, L'Hermitte-Stead C, Myllykallio H. Diversity of circular RNAs and RNA ligases in archaeal cells. Biochimie. 2019;164:37–44.

[19] Brooks MA, Meslet-Cladiere L, Graille M, Kuhn J, Blondeau K, Myllykallio H, et al. The structure of an archaeal homodimeric ligase which has RNA circularization activity. Protein Sci. 2008;17(8):1336–45.

[20] Torchia C, Takagi Y, Ho CK. Archaeal RNA ligase is a homodimeric protein that catalyzes intramolecular ligation of single-stranded RNA and DNA. Nucleic Acids Res. 2008;36(19):6218–27.

[21] Unciuleac MC, Shuman S. Characterization of a novel eukaryal nick-sealing RNA ligase from *Naegleria gruberi*. RNA. 2015;21(5):824–32.

[22] Raymond A, Shuman S. *Deinococcus radiodurans* RNA ligase exemplifies a novel ligase clade with a distinctive N-terminal module that is important for 5′-PO_4 nick sealing and ligase adenylylation but dispensable for phosphodiester formation at an adenylylated nick. Nucleic Acids Res. 2007;35(3):839–49.

[23] Banerjee A, Ghosh S, Goldgur Y, Shuman S. Structure and two-metal mechanism of fungal tRNA ligase. Nucleic Acids Res. 2019;47(3):1428–1439.

[24] Lopes RR, Silveira Gde O, Eitler R, Vidal RS, Kessler A, Hinger S, et al. The essential function of the *Trypanosoma brucei* Trl1 homolog in procyclic cells is maturation of the intron-containing tRNATyr. RNA. 2016;22(8):1190–9.

[25] Blanc V, Alfonzo JD, Aphasizhev R, Simpson L. The mitochondrial RNA ligase from *Leishmania tarentolae* can join RNA molecules bridged by a complementary RNA. J Biol Chem. 1999;274(34):24289–96.

[26] Palazzo SS, Panigrahi AK, Igo RP, Salavati R, Stuart K. Kinetoplastid RNA editing ligases: complex association, characterization, and substrate requirements. Mol Biochem Parasitol. 2003;127(2):161–7.

[27] Tanaka N, Meineke B, Shuman S. RtcB, a novel RNA ligase, can catalyze tRNA splicing and *HAC1* mRNA splicing in vivo. J Biol Chem. 2011;286(35):30253–7.

[28] Tanaka N, Chakravarty AK, Maughan B, Shuman S. Novel mechanism of RNA repair by RtcB via sequential 2′,3′-cyclic phosphodiesterase and 3′-phosphate/5′-hydroxyl ligation reactions. J Biol Chem. 2011;286(50):43134–43.

[29] Laski FA, Fire AZ, RajBhandary UL, Sharp PA. Characterization of tRNA precursor splicing in mammalian extracts. J Biol Chem. 1983;258(19):11974–80.

[30] Popow J, Englert M, Weitzer S, Schleiffer A, Mierzwa B, Mechtler K, et al. HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science. 2011;331(6018):760–4.

[31] Englert M, Sheppard K, Aslanian A, Yates r J R, Soll D. Archaeal 3′-phosphate RNA splicing ligase characterization identifies the missing component in tRNA maturation. Proc Natl Acad Sci U S A. 2011;108(4):1290–5.

[32] Tanaka N, Shuman S. RtcB is the RNA ligase component of an *Escherichia coli* RNA repair operon. J Biol Chem. 2011;286(10):7727–31.

[33] Das U, Shuman S. 2′-phosphate cyclase activity of RtcA: a potential rationale for the operon organization of RtcA with an RNA repair ligase RtcB in *Escherichia coli* and other bacterial taxa. RNA. 2013;19(10):1355–62.

[34] Genschik P, Drabikowski K, Filipowicz W. Characterization of the *Escherichia coli* RNA 3′-terminal phosphate cyclase and its σ54-regulated operon. J Biol Chem. 1998;273(39):25516–26.

[35] Desai KK, Cheng CL, Bingman CA, Phillips J G N, Raines RT. A tRNA splicing operon: Archease endows RtcB with dual GTP/ATP cofactor specificity and accelerates RNA ligation. Nucleic Acids Res. 2014;42(6):3931–42.

[36] Desai KK, Beltrame AL, Raines RT. Coevolution of RtcB and Archease created a multiple-turnover RNA ligase. RNA. 2015;21(11):1866–72.

[37] Popow J, Jurkin J, Schleiffer A, Martinez J. Analysis of orthologous groups reveals Archease and DDX1 as tRNA splicing factors. Nature. 2014;511(7507):104–7.

[38] Maughan WP, Shuman S. Characterization of 3′-phosphate RNA ligase paralogs RtcB1, RtcB2, and RtcB3 from *Myxococcus xanthus* highlights DNA and RNA 5′-phosphate capping activity of RtcB3. J Bacteriol. 2015;197(22):3616–24.

[39] Paushkin SV, Patel M, Furia BS, Peltz SW, Trotta CR. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3′ end formation. Cell. 2004;117(3):311–21.

[40] Schmidt CA, Giusto JD, Bao A, Hopper AK, Matera AG. Molecular determinants of metazoan tricRNA biogenesis. Nucleic Acids Res. 2019;47(12):6452–6465.

[41] Kosmaczewski SG, Edwards TJ, Han SM, Eckwahl MJ, Meyer BI, Peach S, et al. The RtcB RNA ligase is an essential component of the metazoan unfolded protein response. EMBO Rep. 2014;15(12):1278–85.

[42] Kanai Y, Dohmae N, Hirokawa N. Kinesin transports RNA: isolation and characterization of an RNA-transporting granule. Neuron. 2004;43(4):513–25.

[43] Jurkin J, Henkel T, Nielsen AF, Minnich M, Popow J, Kaufmann T, et al. The mammalian tRNA ligase complex mediates splicing of *XBP1* mRNA and controls antibody secretion in plasma cells. EMBO J. 2014;33(24):2922–36.

[44] Lu Y, Liang FX, Wang X. A synthetic biology approach identifies the mammalian UPR RNA ligase RtcB. Mol Cell. 2014;55(5):758–70.

[45] Unlu I, Lu Y, Wang X. The cyclic phosphodiesterase CNP and RNA cyclase RtcA fine-tune noncanonical *XBP1* splicing during ER stress. J Biol Chem. 2018;293(50):19365–19376.

[46] Ray A, Zhang S, Rentas C, Caldwell KA, Caldwell GA. RTCB-1 mediates neuroprotection via *XBP-1* mRNA splicing in the unfolded protein response pathway. J Neurosci. 2014;34(48):16076–85.

[47] Manwar MR, Shao C, Shi X, Wang J, Lin Q, Tong Y, et al. The bacterial RNA ligase RtcB accelerates the repair process of fragmented rRNA upon releasing the antibiotic stress. Sci China Life Sci. 2020;63(2):251–258.

[48] Nariya H, Inouye M. MazF, an mRNA interferase, mediates programmed cell death during multicellular *Myxococcus* development. Cell. 2008;132(1):55–66.

[49] Temmel H, Muller C, Sauert M, Vesper O, Reiss A, Popow J, et al. The RNA ligase RtcB reverses MazF-induced ribosome heterogeneity in *Escherichia coli*. Nucleic Acids Res. 2017;45(8):4708–4721.

[50] Aizenman E, Engelberg-Kulka H, Glaser G. An *Escherichia coli* chromosomal "addiction module" regulated by guanosine 3′, 5′-bispyrophosphate: a model for programmed bacterial cell death. Proc Natl Acad Sci U S A. 1996;93(12):6059–63.

[51] Culviner PH, Laub MT. Global analysis of the *E. coli* toxin MazF reveals widespread cleavage of mRNA and the inhibition of rRNA maturation and ribosome biogenesis. Mol Cell. 2018;70(5):868–880 e10.

[52] Sauert M, Wolfinger MT, Vesper O, Muller C, Byrgazov K, Moll I. The MazF-regulon: a toolbox for the post-transcriptional stress response in *Escherichia coli*. Nucleic Acids Res. 2016;44(14):6660–75.

[53] Vesper O, Amitai S, Belitsky M, Byrgazov K, Kaberdina AC, Engelberg-Kulka H, et al. Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in *Escherichia coli*. Cell. 2011;147(1):147–57.

[54] Das U, Chakravarty AK, Remus BS, Shuman S. Rewriting the rules for end joining via enzymatic splicing of DNA 3′-PO4 and 5′-OH ends. Proc Natl Acad Sci U S A. 2013;110(51):20437–42.

[55] Das U, Chauleau M, Ordonez H, Shuman S. Impact of DNA3′pp5′G capping on repair reactions at DNA 3′ ends. Proc Natl Acad Sci U S A. 2014;111(31):11317–22.

[56] Chakravarty AK, Shuman S. The sequential 2′, 3′-cyclic phosphodiesterase and 3′-phosphate/5′-OH ligation steps of the RtcB RNA splicing pathway are GTP-dependent. Nucleic Acids Res. 2012;40(17):8558–67.

[57] Chakravarty AK, Subbotin R, Chait BT, Shuman S. RNA ligase RtcB splices 3′-phosphate and 5′-OH ends via covalent RtcB-(histidinyl)-GMP and polynucleotide-(3′)pp(5′)G intermediates. Proc Natl Acad Sci U S A. 2012;109(16):6072–7.

[58] Abelson J, Trotta CR, Li H. tRNA splicing. J Biol Chem. 1998;273(21):12685–8.

[59] Englert M, Xia S, Okada C, Nakamura A, Tanavde V, Yao M, et al. Structural and mechanistic insights into guanylylation of RNA-splicing ligase RtcB joining RNA between 3′-terminal phosphate and 5′-OH. Proc Natl Acad Sci U S A. 2012;109(38):15235–40.

[60] Okada C, Maegawa Y, Yao M, Tanaka I. Crystal structure of an RtcB homolog protein (PH1602-extein protein) from *Pyrococcus horikoshii* reveals a novel fold. Proteins. 2006;63(4):1119–22.

[61] Nandy A, Saenz-Mendez P, Gorman AM, Samali A, Eriksson LA. Homology model of the human tRNA splicing ligase RtcB. Proteins. 2017;85(11):1983–1993.

[62] Desai KK, Bingman CA, Phillips J G N, Raines RT. Structures of the noncanonical RNA ligase RtcB reveal the mechanism of histidine guanylylation. Biochemistry. 2013;52(15):2518–25.

[63] Malakhov MP, Mattern MR, Malakhova OA, Drinker M, Weeks SD, Butt TR. SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. J Struct Funct Genomics. 2004;5(1-2):75–86.

[64] Whittaker RH. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. Science. 1969;163(3863):150–60.

[65] Adl SM, Bass D, Lane CE, Lukes J, Schoch CL, Smirnov A, et al. Revisions to the classification, nomenclature, and diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66(1):4–119.

[66] Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, et al. Extreme diversity of diplonemid eukaryotes in the ocean. Curr Biol. 2016;26(22):3060–3065.

[67] de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348(6237):1261605.

[68] Lukes J, Flegontova O, Horak A. Diplonemids. Curr Biol. 2015;25(16):R702–4.

[69] David V, Archibald JM. Evolution: plumbing the depths of diplonemid diversity. Curr Biol. 2016;26(24):R1290–R1292.

[70] Faktorová D, Valach M, Kaur B, Burger G, Lukeš J. Chapter 6. In: Cruz-Reyes J, Gray MW, editors. Mitochondrial RNA editing and processing in diplonemid protists. Nucleic Acids and Molecular Biology. Cham: Springer International Publishing; 2018. p. 145–176.

[71] Valach M, Leveille-Kunst A, Gray MW, Burger G. Respiratory chain Complex I of unparalleled divergence in diplonemids. J Biol Chem. 2018;293(41):16043–16056.

[72] Burger G, Valach M. Perfection of eccentricity: mitochondrial genomes of diplonemids. IUBMB Life. 2018;70(12):1197–1206.

[73] Gawryluk RMR, Del Campo J, Okamoto N, Strassert JFH, Lukes J, Richards TA, et al. Morphological identification and single-cell genomics of marine diplonemids. Curr Biol. 2016;26(22):3053–3059.

[74] Okamoto N, Gawryluk RMR, Del Campo J, Strassert JFH, Lukes J, Richards TA, et al. A revised taxonomy of diplonemids including the Eupelagonemidae n. fam. and a type species, *Eupelagonema oceanica* n. gen. & sp. J Eukaryot Microbiol. 2019;66(3):519–524.

[75] Qian Q, Keeling PJ. Diplonemid glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and prokaryote-to-eukaryote lateral gene transfer. Protist. 2001;152(3):193–201.

[76] Marande W, Lukes J, Burger G. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell. 2005;4(6):1137–46.

[77] Marande W, Burger G. Mitochondrial DNA as a genomic jigsaw puzzle. Science. 2007;318(5849):415.

[78] Vlcek C, Marande W, Teijeiro S, Lukes J, Burger G. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res. 2011;39(3):979–88.

[79] Burger G, Moreira S, Valach M. Genes in hiding. Trends Genet. 2016;32(9):553–565.

[80] Kiethega GN, Turcotte M, Burger G. Evolutionarily conserved *cox1 trans*-splicing without *cis*-motifs. Mol Biol Evol. 2011;28(9):2425–8.

[81] Valach M, Moreira S, Hoffmann S, Stadler PF, Burger G. Keeping it complicated: mitochondrial genome plasticity across diplonemids. Sci Rep. 2017;7(1):14166.

[82] Valach M, Moreira S, Faktorova D, Lukes J, Burger G. Post-transcriptional mending of gene sequences: looking under the hood of mitochondrial gene expression in diplonemids. RNA Biol. 2016;13(12):1204–1211.

[83] Kiethega GN, Yan Y, Turcotte M, Burger G. RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. RNA Biol. 2013;10(2):301–13.

[84] Moreira S, Noutahi E, Lamoureux G, Burger G. Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2. BMC Struct Biol. 2015;15:20.

[85] Valach M, Moreira S, Kiethega GN, Burger G. *Trans*-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res. 2014;42(4):2660–72.

[86] Kaur B, Zahonova K, Valach M, Faktorova D, Prokopchuk G, Burger G, et al. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res. 2020;48(5):2694–2708.

[87] Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. 2019;2(5):e201900429.

[88] Fukasawa Y, Tsuji J, Fu SC, Tomii K, Horton P, Imai K. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics. 2015;14(4):1113–26.

[89] Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics. 2004;4(6):1581–90.

[90] Claros MG. MitoProt, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci. 1995;11(4):441–7.

[91] Almagro Armenteros JJ, Sonderby CK, Sonderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33(24):4049.

[92] Petsalaki EI, Bagos PG, Litou ZI, Hamodrakas SJ. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics Bioinformatics. 2006;4(1):48–55.

[93] Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585–7.

[94] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–58.

[95] Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296–W303.

[96] Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. 2018;27(1):14–25.

[97] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–12.

[98] Laskowski RA, Jablonska J, Pravda L, Varekova RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci. 2018;27(1):129–134.

[99] Johnson LK, Alexander H, Brown CT. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. GigaScience. 2019;8(4).

[100] Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12(6):e1001889.

[101] Valach M, Alcazar JAG, Sarrasin M, Lang BF, Gray MW, Burger G. An unexpectedly complex mitoribosome in *Andalucia godoyi*, a protist with the most bacteria-like mitochondrial genome. Mol Biol Evol. 2020.

[102] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

[103] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

[104] Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276–8.

[105] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.

[106] Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016;11(10):e0163962.

[107] Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics. 2007;23(9):1073–9.

[108] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.

[109] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

[110] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530–1534.

[111] Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35(21):4453–4455.

[112] Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.

[113] Mailloux RJ. Cysteine Switches and the Regulation of Mitochondrial Bioenergetics and ROS Production. Adv Exp Med Biol. 2019;1158:197–216.

[114] Mailloux RJ, Jin X, Willmore WG. Redox regulation of mitochondrial function with emphasis on cysteine oxidation reactions. Redox Biol. 2014;2:123–39.

[115] Yang JH, Williams D, Kandiah E, Fromme P, Chiu PL. Structural basis of redox modulation on chloroplast ATP synthase. Commun Biol. 2020;3(1):482.

[116] Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G, Lang BF. Construction of cDNA libraries: focus on protists and fungi. Methods Mol Biol. 2009;533:33–47.

[117] Valach M. RNA extraction using the 'home-made' Trizol substitute; 2016. Accessed: 2020-09-30. https://www.protocols.io/view/RNA-extraction-using-the-home-made-Trizol-substitu-eiebcbe.

[118] Inoue H, Nojima H, Okayama H. High efficiency transformation of *Escherichia coli* with plasmids. Gene. 1990;96(1):23–8.

[119] Schagger H. Tricine-SDS-PAGE. Nat Protoc. 2006;1(1):16–22.

[120] Scientific T. Instructions: HisPur^{TM} Cobalt Superflow Agarose; 2012. Accessed: 2020-09-30. https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FMAN0011805_HisPurTM_Cobalt_SupfLw_Agarose_UG.pdf&title=VXNlciBHdWlkZTogIEhpc1B1clRNIENvYmFsdCBTdXBlcmZsb3cgQWdhcm9zZQ==.

[121] Technology© CS. Western Blotting Protocol; 2005. Accessed: 2020-09-30. https://www.cellsignal.com/contents/resources-protocols/western-blotting-protocol/western.

[122] Rath A, Glibowicka M, Nadeau VG, Chen G, Deber CM. Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. Proc Natl Acad Sci U S A. 2009;106(6):1760–5.

[123] Murphy RM, Lamb GD. Important considerations for protein analyses using antibody based techniques: down-sizing Western blotting up-sizes outcomes. J Physiol. 2013;591(23):5823–31.

[124] Kinoshita E, Kinoshita-Kikuta E, Koike T. Separation and detection of large phosphoproteins using Phos-tag SDS-PAGE. Nat Protoc. 2009;4(10):1513–21.

[125] Walkenhorst WF, Merzlyakov M, Hristova K, Wimley WC. Polar residues in transmembrane helices can decrease electrophoretic mobility in polyacrylamide gels without causing helix dimerization. Biochim Biophys Acta. 2009;1788(6):1321–31.

[126] Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics. 2007;8(1):65.

[127] Sanchez de Groot N, Pallares I, Aviles FX, Vendrell J, Ventura S. Prediction of "hot spots" of aggregation in disease-linked polypeptides. BMC Struct Biol. 2005;5(1):18.

[128] Baneyx F. Recombinant protein expression in *Escherichia coli*. Curr Opin Biotechnol. 1999;10(5):411–21.

[129] Rinas U, Garcia-Fruitos E, Corchero JL, Vazquez E, Seras-Franzoso J, Villaverde A. Bacterial inclusion bodies: discovering their better half. Trends Biochem Sci. 2017;42(9):726–737.

[130] Ventura S, Villaverde A. Protein quality in bacterial inclusion bodies. Trends Biotechnol. 2006;24(4):179–85.

[131] Tsumoto K, Umetsu M, Kumagai I, Ejima D, Philo JS, Arakawa T. Role of arginine in protein refolding, solubilization, and purification. Biotechnol Prog. 2004;20(5):1301–8.

[132] Studier FW. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. J Mol Biol. 1991;219(1):37–44.

[133] Dubendorff JW, Studier FW. Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with *lac* repressor. J Mol Biol. 1991;219(1):45–59.

[134] Bertani G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. J Bacteriol. 1951;62(3):293–300.

[135] Pan SH, Malcolm BA. Reduced background expression and improved plasmid stability with pET vectors in BL21 (DE3). Biotechniques. 2000;29(6):1234–8.

[136] Grossman TH, Kawasaki ES, Punreddy SR, Osburne MS. Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. Gene. 1998;209(1-2):95–103.

[137] Sorensen HP, Mortensen KK. Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. J Biotechnol. 2005;115(2):113–28.

[138] Wittwer CT, Makrigiorgos GM. In: Rifai N, Horvath AR, Wittwer CT, editors. Nucleic Acid Techniques. Elsevier; 2018. p. 47–86.

[139] Peternel S, Komel R. Active protein aggregates produced in *Escherichia coli*. Int J Mol Sci. 2011;12(11):8275–87.

[140] Singh A, Upadhyay V, Upadhyay AK, Singh SM, Panda AK. Protein recovery from inclusion bodies of *Escherichia coli* using mild solubilization process. Microb Cell Fact. 2015;14:41.

[141] Bushmarina NA, Blanchet CE, Vernier G, Forge V. Cofactor effects on the protein folding reaction: acceleration of alpha-lactalbumin refolding by metal ions. Protein Sci. 2006;15(4):659–71.

[142] Goedken ER, Keck JL, Berger JM, Marqusee S. Divalent metal cofactor binding in the kinetic folding trajectory of *Escherichia coli* ribonuclease HI. Protein Sci. 2000;9(10):1914–21.

[143] Wittung-Stafshede P. Role of cofactors in protein folding. Acc Chem Res. 2002;35(4):201–8.

[144] Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: advances and challenges. Front Microbiol. 2014;5:172.

[145] Klein JS, Jiang S, Galimidi RP, Keeffe JR, Bjorkman PJ. Design and characterization of structured protein linkers with differing flexibilities. Protein Eng Des Sel. 2014;27(10):325–30.

[146] Zayas JF. Chapter 2. In: Solubility of Proteins. Berlin, Heidelberg: Springer Berlin Heidelberg; 1997. p. 6–75.

[147] Peternel S, Grdadolnik J, Gaberc-Porekar V, Komel R. Engineering inclusion bodies for non denaturing extraction of functional proteins. Microb Cell Fact. 2008;7:34.

[148] Santos CA, Beloti LL, Toledo MA, Crucello A, Favaro MT, Mendes JS, et al. A novel protein refolding protocol for the solubilization and purification of recombinant peptidoglycan-associated lipoprotein from *Xylella fastidiosa* overexpressed in *Escherichia coli*. Protein Expr Purif. 2012;82(2):284–9.

[149] Palmer I, Wingfield PT. Preparation and extraction of insoluble (inclusion-body) proteins from *Escherichia coli*. Curr Protoc Protein Sci. 2012;Chapter 6:Unit6 3.

[150] Bornhorst JA, Falke JJ. Purification of proteins using polyhistidine affinity tags. Methods Enzymol. 2000;326:245–54.

[151] Yamaguchi H, Miyazaki M. Refolding techniques for recovering biologically active recombinant proteins from inclusion bodies. Biomolecules. 2014;4(1):235–51.

[152] Sorensen HP, Mortensen KK. Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. Microb Cell Fact. 2005;4(1):1.

[153] Trahan C, Aguilar LC, Oeffinger M. Single-step affinity purification (ssAP) and mass spectrometry of macromolecular complexes in the yeast *S. cerevisiae*. Methods Mol Biol. 2016;1361:265–87.

Annexe A

Expression constitutive de la protéine recombinante HS-GS-DpRTCB1

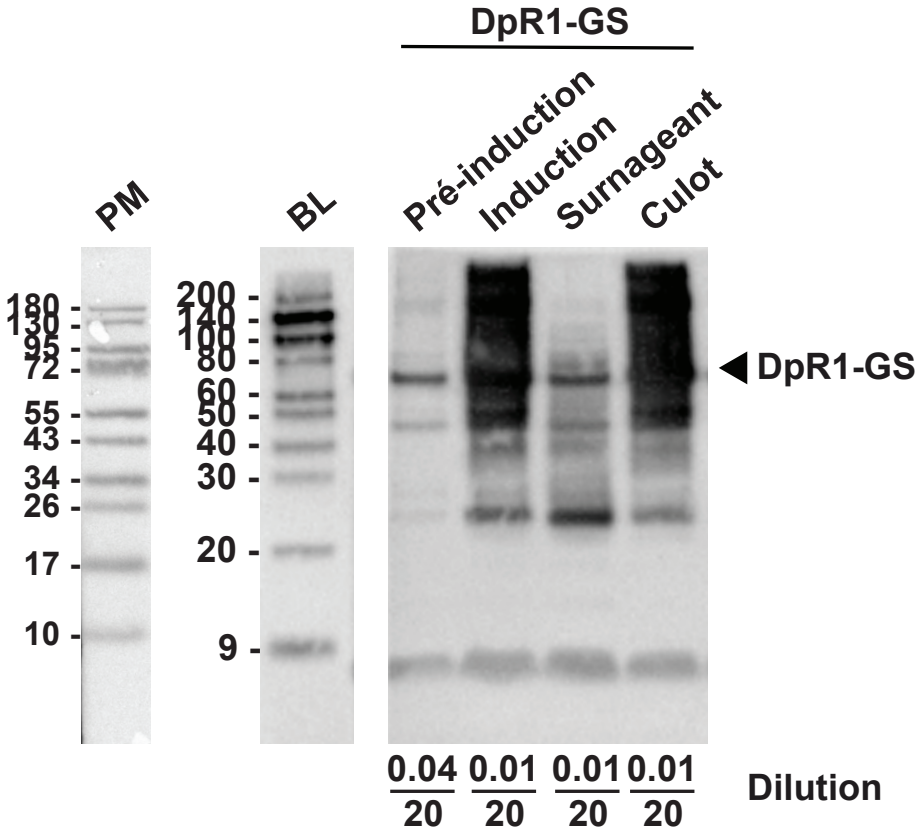

Fig. A.1. Détection par immunobuvardage de type Western des protéines HS-GS-DpRTCB1 (DpR1-GS) suivant l'induction et la lyse des transformants avec des anticorps anti-SUMO3. L'échelle biotinylée (BL) a été utilisée comme marqueur de taille et a été détectée à l'aide d'un anticorps anti-biotine. La lyse des cellules s'est effectuée dans un tampon contenant 0.5 M d'arginine.