

**An investigation into the structural  
determinants of the positional specificity  
of hydrolysis of *myo*-inositol  
hexakisphosphate by HP phytases**

Isabella Maria Acquistapace

A thesis submitted for the degree of  
Doctor of Philosophy in Biomolecular Science

University of East Anglia  
School of Biological Sciences

November, 2018

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK copyright law. In addition, any quotation or extract must include full attribution.

# Abstract

Phytases are phosphatases which catalyse the stepwise hydrolysis of the phosphomonoester bonds of *myo*-inositol hexakisphosphate (IP<sub>6</sub>). IP<sub>6</sub> is the major storage form of phosphorous in grains, seeds and beans found in animal feeds, but also can act as an antinutrient. Feed conversion ratios for monogastric animals are improved by addition of phytases to feed preparations in a market with an annual worldwide turnover of more than \$500m. New phytases are continually sought to satisfy this demand. An alternative approach is to reengineer known phytases, tuning their positional stereospecificity with the aim of generating more efficient IP<sub>6</sub> hydrolysis.

To help shed light on the structural basis for differences in positional stereospecificity between phytases of clade 2 of the family of histidine phosphatase superfamily (HP2), the role of the catalytic proton donor in the *Escherichia coli* phytase, AppA, was investigated in comparison with that observed for multiple inositol polyphosphate phosphatases (MINPPs). Four AppA active site mutants were generated by site-directed mutagenesis of the wild type “HDT” motif to replace it with the MINPPs-like “HAE”. Shifts in the ratio of IP<sub>5</sub> intermediates were detected. In the process, the 6-phytase AppA, was converted to a 1/3- or 1/3/4/6-phytase. High resolution X-ray structures of the mutants revealed the primary roles played by size and polarity of specificity pockets in determining positional stereospecificity traits.

The first structural evidence for a large  $\alpha$ -domain closure motion and its involvement in catalysis in a HP2 phytase is also reported, the MINPP of *Bifidobacterium longum* susp. *infantis*. Ligand binding-driven conformational changes correlate with the presence of a unique active site loop insertion. This loop significantly reduces the size of the active site and contributes to the coordination of the substrate.

Many MINPPs lack the positional stereospecificity displayed by other HP2 phytases. In principle, this property could be utilized for the more efficient dephosphorylation of IP<sub>6</sub>. To identify sequence determinants of this characteristic, 15 MINPPs were selected by genome mining. The resulting recombinant enzymes were tested and classified as either positionally or non-positionally stereospecific. X-ray crystal structures of an enzyme from each group revealed conserved active site residues which were mutated *in silico*, and the IP<sub>6</sub> binding poses for each mutant predicted by molecular docking. This process revealed hotspot residues which can potentially direct positional stereospecificity in this family of phytases.

The work reported in this thesis has led to the identification of potential residue determinants of positional stereospecificity in HP2 phytases and has prepared the ground for future engineering studies possibly leading to more efficient animal feed enzymes.

# Acknowledgements

Many people deserve a thankyou here and I am sure someone will be forgotten.

First, I thank my supervisors Andrew Hemmings and Charles Brearley. I thank them to have given me this life-changing opportunity and to have supported me along the way when help was needed. Andrew Hemmings has been an essential, wise guide in the area of structural biology and in the organization of the project. However, thanks need be given for his constant push towards the simplification of research problems, which led me to significantly improve my abilities in processing and delivering scientific concepts. Charles Brearley has been a very good mentor in all matters concerning enzymology. I thank him for the conversations on phytases but also for his example of methodical and never-ending scientific curiosity. I also thank ABVista for providing that important funding and for opportunities for information exchange.

I thank all the people in the research group: Arthur Li and Hayley Whitfield for their essential help in the first stages of my PhD, Melissa Salmon in particular to have given me her very efficient protocol for Gateway cloning and for sharing “cat stories”, Caroline Laurendon for her smile, support and nice conversations, Raquel Faba-Rodriguez for the nice time spent together and Hans Pfalzgraf for sharing his passion for science. Thanks also go to all those people in BIO 2.30, past and present, for the collaborative environment which allowed me to work efficiently. A special thanks goes to Marcus Edwards for his support and help in tackling everyday lab problems.

A huge thankyou goes to my family, to my grandparents and to my grandmother Ines, because without her love, example and strength, I would not be here now. To my parents, to all the members of my large family and to my “third grandmother” Maria, for their constant presence and uniqueness and because they have always been my shelter. I need to thank many friends in Italy and the UK. Special thanks go to Cristina, Laura, Lina, Athina, Martina, Anna, Chantal, Mauro and Stefano. However, many more people contributed to this endeavour than I can include by name and to these individuals I send my warm thanks.

# Abbreviations

<b>Abs</b>	absorbance
<b>Amp</b>	ampicillin
<b><i>EcAppA</i> HAT</b>	<i>Escherichia coli</i> AppA D326A
<b><i>EcAppA</i> HET</b>	<i>Escherichia coli</i> AppA D326E
<b><i>EcAppA</i> HDE</b>	<i>Escherichia coli</i> AppA T327E
<b><i>EcAppA</i> HAE</b>	<i>Escherichia coli</i> AppA D326A, T327E
<b>ASU</b>	asymmetric unit
<b>BLAST</b>	Basic Local Alignment Search Tool [1, 2]
<b><i>Bt</i>MINPP</b>	<i>Bifidobacterium longum</i> susp. <i>infantis</i> ATCC 15697 MINPP
<b><i>Bt</i>MINPP</b>	<i>Bacteroides thetaiotaomicron</i> MINPP
<b>Cam</b>	chloramphenicol
<b>CV</b>	column volume
<b>FT</b>	flow through
<b>FTU</b>	phytase units (quantity of enzyme that release 1 $\mu$ M of inorganic phosphate from sodium phytate per minute)
<b>Gent</b>	gentamycin
<b>HP</b>	histidine phosphatases
<b>HP1</b>	histidine phosphatases – clade 1
<b>HP2</b>	histidine phosphatases – clade 2
<b>HEPES</b>	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
<b>HPLC</b>	high performance liquid chromatography
<b>IHS</b>	inositol hexasulphate
<b>IMAC</b>	Immobilized Metal Affinity Chromatography
<b>IPs</b>	inositol polyphosphate
<b>IP<sub>1-6</sub></b>	inositol mono/bis/tris/tetra/penta/hexaphosphate
<b>IPTG</b>	isopropyl- $\beta$ -D-1-thiogalattopiranoside
<b>LB</b>	lysogeny broth
<b>LIC</b>	Ligation Independent Cloning
<b>MINPP</b>	multiple inositol phosphate phosphatase
<b>MM equation</b>	Michaelis Menten equation
<b>MW</b>	molecular weight

<b>MWCO</b>	molecular weight cut-off
<b>NPS</b>	non-position stereospecific enzymes
<b>OD<sub>600</sub></b>	optical density at 600nm
<b>o/n</b>	over-night
<b>o/w</b>	over-weekend
<b>p1</b>	<i>Flammeovirga pacifica</i> MINPP
<b>p2</b>	<i>Arsukibacterium</i> sp. MJ3 MINPP
<b>p3</b>	<i>Prevotella brevis</i> MINPP
<b>p4</b>	<i>Bacteroides</i> sp. CAG:927 MINPP
<b>p5</b>	<i>Bacteroides intestinalis</i> MINPP
<b>p6</b>	<i>Bacteroides</i> sp. CAG:545 MINPP
<b>p7</b>	<i>Bacteroides</i> sp. CAG:770 MINPP
<b>p8</b>	<i>Bacteroides</i> sp. CAG:617 MINPP
<b>p9</b>	<i>Bacteroides bacterium</i> CF MINPP
<b>p10</b>	<i>Fibrobacter succinogenes</i> MINPP
<b>p11</b>	<i>Oxalobacteraceae bacterium</i> AB MINPP
<b>p12</b>	<i>Aeromicrobium</i> sp. Root 236 MINPP
<b>p13</b>	<i>Aeromicrobium</i> sp. Root 495 MINPP
<b>p14</b>	<i>Streptacidiphilus jeojiense</i> MINPP
<b>p15</b>	<i>Amycolatopsis jejuensis</i> MINPP
<b>p16</b>	<i>Streptomyces</i> sp. AW19M42 MINPP
<b>Pi</b>	inorganic phosphate
<b>pNPP</b>	<i>para</i> -nitrophenyl phosphate
<b>POI</b>	Protein of Interest
<b>PS</b>	positional stereospecific enzymes
<b>PAGE</b>	polyacrylamide gel electrophoresis
<b>PCR</b>	polymerase chain reaction
<b>PDB</b>	Protein Data Bank
<b>PEG</b>	polyethylene glycol
<b>R.m.s.d.</b>	root mean square deviation
<b>SDS</b>	sodium dodecyl sulphate
<b>SIRAS</b>	single isomorphous replacement with anomalous scattering

<b>Spec</b>	spectinomycin
<b>Str</b>	streptomycin
<b>Tet</b>	tetracycline
<b>Tris</b>	2-Amino-2-hydroxymethyl-propane-1,3-diol

# List of contents

Abstract .....	1
Acknowledgements .....	3
Abbreviations .....	4
List of contents .....	7
List of Figures .....	11
List of Tables .....	13
CHAPTER 1 .....	14
1. Introduction.....	14
1.1. Phytases in the animal feed industry- a short history.....	14
1.2. Inositol and their functions.....	15
1.3. Phytases and their classes .....	16
1.3.1. Beta propeller phytases (BPPs).....	17
1.3.2. Purple acid phosphatases (PAPs) .....	17
1.3.3. Protein tyrosine phosphatase-like phytases (PTPLPs).....	17
1.3.4. Histidine phosphatases (HP2) phytases.....	18
1.4. <i>Escherichia coli</i> AppA phytase ( <i>EcAppA</i> ).....	19
1.5. Bacterial MINPPs – A timeline .....	20
1.6. Aims of my project.....	25
CHAPTER 2 .....	26
2. Experimental protocols .....	26
2.1. Plasmids .....	26
2.2. Growth media and bacterial strains.....	30
2.3. Transformations.....	31
2.3.1. Preparation of competent cells .....	31
2.3.2. Transformation .....	31
2.4. Cloning .....	32
2.4.1. In-fusion cloning.....	32
2.4.2. Gateway cloning.....	33
2.5. Cloning quality checks .....	36
2.5.1. Cell lysis and plasmid length detection on agarose gel.....	36
2.5.2. Colony PCR .....	37
2.5.3. DNA digestion.....	37
2.5.4. Sequencing .....	38
2.6. Methods of enzymes purification .....	38
2.6.1. Purification of His-tagged proteins.....	39

2.6.2.	Purification of un-tagged periplasmic <i>EcAppA</i> .....	40
2.7.	Biochemical and biophysical techniques of phytases characterisation .....	42
2.7.1.	Phosphatase activity assay - monitoring pNPP hydrolysis .....	42
2.7.2.	Phytase activity assay – determination of released phosphate by molybdenum blue method.....	42
2.7.3.	HPLC separation of inositol polyphosphates .....	43
2.7.4.	Differential scanning calorimetry.....	44
2.8.	X-ray crystallography.....	45
2.8.1.	Protein crystals and their growth.....	45
2.8.2.	X-ray diffraction experiments.....	48
2.8.3.	X-ray data processing.....	51
2.8.4.	Methods for the solution of protein structures and refinement.....	52
2.9.	Molecular docking techniques.....	59
CHAPTER 3	.....	60
3.	Probing the catalytic flexibility of the <i>Escherichia coli</i> 6-phytase AppA.....	60
3.1.	Gene constructs.....	61
3.2.	Expression trials, solubility and purification.....	63
3.3.	Site-directed mutagenesis of active site residues.....	68
3.4.	pH profiles.....	70
3.5.	Analysis of the product of IP <sub>6</sub> hydrolysis by HPLC.....	72
3.6.	Enzyme kinetics studies .....	74
3.7.	Inhibition by inositol hexasulphate (IHS).....	77
3.8.	Crystallisation with inositol hexasulphate .....	78
3.9.	X-ray crystal structure determination.....	81
3.10.	Structural analysis .....	84
3.11.	Discussion .....	94
CHAPTER 4	.....	97
4.	<i>Bifidobacterium longum</i> MINPP and its domain motion during the catalytic cycle .....	97
4.1.	Expression.....	97
4.2.	Purification.....	98
4.3.	Characterisation.....	99
4.4.	Crystallisation of the complex <i>B</i> /MINPP:IHS and <i>B</i> /MINPP E401Q .....	101
4.5.	X-ray crystal structure determination.....	104
4.6.	Structural analysis.....	110
4.7.	Conservation of domain flexibility.....	117
4.8.	Discussion.....	124
CHAPTER 5	.....	125

5.	Genome mining of positionally non-stereoselective MINPPs .....	125
5.1.	Collection of standard sequences.....	126
5.2.	Grouping of standard sequences.....	128
5.2.1.	Substrate specificity.....	128
5.2.2.	Proton donor.....	129
5.2.3.	Phylogenesis.....	131
5.3.	Analysis of standards.....	133
5.3.1.	X-ray crystal structures .....	133
5.3.2.	Identification of conserved motifs and their function .....	138
5.4.	Collection of novel MINPPs sequences and clustering .....	144
5.5.	Selection of MINPPs of interest.....	146
5.5.1.	Definition of active site pockets.....	146
5.5.2.	Conserved motifs in the active site of <i>Actinobacteria</i> .....	148
5.5.3.	Conserved motifs in the active site of <i>Bacteria</i> (excluding <i>Actinobacteria</i> ) .....	150
5.5.4.	Phylogenesis of bacterial MINPPs sequences .....	152
5.6.	Analysis of the 16 selected sequences prior synthesis .....	154
5.6.1.	Biological sources .....	154
5.6.2.	Analysis of the conservation of HAP2 and MINPP motifs.....	156
5.6.3.	Phylogenesis.....	158
5.6.4.	Active-site prediction .....	160
5.6.5.	Genes optimization and synthesis.....	161
5.7.	Discussion.....	162
CHAPTER 6 .....		165
6.	Characterization of 15 novel MINPPs .....	165
6.1.	Cloning .....	165
6.2.	Expression trials .....	170
6.3.	Purification.....	171
6.4.	pH profiles and phosphatase activity .....	175
6.5.	Analysis of the product of IP <sub>6</sub> hydrolysis by HPLC.....	177
6.6.	Biochemical and biophysical characterisation of p8 and p15 .....	183
6.6.1.	pH profile.....	183
6.6.2.	Enzyme kinetics .....	184
6.6.3.	Inhibition by inositol hexasulfate (IHS) .....	186
6.6.4.	Differential scanning calorimetry.....	187
6.7.	X-ray crystal structures determination.....	189
6.7.1.	Crystallization and X-ray diffraction.....	190
6.7.2.	X-ray crystal structures solution of p8 and p15.....	192

6.8.	Analysis of the crystal structures .....	195
6.9.	Docking studies .....	198
6.10.	Discussion .....	202
CHAPTER 7 .....		205
7.	Final discussion.....	205
CHAPTER 8 - APPENDIX.....		210
8.	Supplemental information – Chapter 2.....	210
8.1.	List of protein sequences.....	210
8.1.1.	<i>Bifidobacterium longum</i> MINPP .....	210
8.1.2.	<i>Escherichia coli</i> AppA .....	211
8.1.3.	15 novel MINPPs phytases .....	212
8.2.	List of primers .....	217
8.2.1.	In-fusion cloning.....	217
8.2.2.	Gateway cloning.....	217
8.2.3.	Site-directed mutagenesis of <i>EcAppA</i> .....	219
8.3.	Supplemental information – Chapter 3.....	221
8.3.1.	Summary table of <i>EcAppA</i> expression trials.....	221
8.4.	Supplemental information – Chapter 5.....	222
8.4.1.	Standards multiple sequence alignments .....	222
8.4.2.	Representative sequences multiple sequence alignments .....	225
8.4.3.	3D modelling of selected 16 MINPPs.....	228
8.4.4.	Disulphide bridges prediction.....	229

# List of Figures

Figure 1.2.1. <i>myo</i> -inositol hexakisphosphate. ....	16
Figure 1.3.4.1. catalytic mechanism of the histidine phosphatase superfamily. ....	18
Figure 1.5.1. alignment of the active site of four MINPPs whose functionality was tested by site-directed mutagenesis. ....	23
Figure 2.1.1. pOPINA and pOPINB. ....	28
Figure 2.1.2. pDONR <sup>TM</sup> 207, pDEST <sup>TM</sup> 17 and pET-28a(+). ....	29
Figure 2.8.1.1. graph representing the crystallization process. ....	47
Figure 2.8.2.1. configuration of beamlines I03 and I04 at Diamond Light Source synchrotron (Didcot, UK). ....	49
Figure 2.8.2.2. Graphic description of Bragg's law. ....	50
Figure 2.8.4.1. Example of Patterson calculation in 2D space. ....	55
Figure 3.1.1. <i>EcAppA</i> constructs. ....	61
Figure 3.1.2. pOPINA <i>EcAppA</i> cloning. Panel of figures. ....	62
Figure 3.2.1. <i>EcAppA</i> expression trials and solubility test. Panel of figures. ....	64
Figure 3.2.2. Purification of <i>EcAppA</i> -H. ....	66
Figure 3.2.3. Purification of <i>EcAppA</i> . ....	67
Figure 3.2.4. Purification of H- <i>EcAppA</i> . ....	68
Figure 3.3.1. Site-directed mutagenesis of <i>EcAppA</i> ....	69
Figure 3.4.1. pH profile of <i>EcAppA</i> wild-type and mutants. ....	70
Figure 3.5.1. Sequential hydrolysis of IP <sub>6</sub> by <i>EcAppA</i> HAE. ....	72
Figure 3.5.2. HPLC chromatogram of reactions stopped when total IP <sub>5</sub> peaks area is equal to 34-40 % of total IPs. ....	73
Figure 3.5.3. HPLC chromatogram of reactions stopped when total Pi/IP <sub>1</sub> peaks area is equal to 28 % of total IPs. ....	74
Figure 3.7.1. Inhibition assay by IHS. ....	77
Figure 3.8.1. <i>EcAppA</i> crystals. ....	80
Figure 3.9.1. <i>EcAppA</i> :IHS crystal on LithoLoop after diffraction. ....	83
Figure 3.9.2. <i>EcAppA</i> PDB id: 4TSR, <i>EcAppA</i> HET:IHS complex and omit single-difference Fourier map of the substrate analogue IHS in <i>EcAppA</i> HAE mutant. ....	83
Figure 3.10.1. Structural overview of <i>EcAppA</i> . ....	84
Figure 3.10.2. Structural overview of conformational changes in <i>EcAppA</i> . ....	85
Figure 3.10.3. Active site pockets of <i>EcAppA</i> . ....	88
Figure 3.10.4. Active site representation of the E:IHS complex for wild-type and mutants. ....	90
Figure 3.10.5. Shifts in binding of IHS to the wild-type and mutant enzymes. ....	92
Figure 3.10.6. Possible enzyme-substrate intermediate of reaction ....	93
Figure 4.1.1. Overexpression trials of <i>B</i> /MINPP in Rosetta (DE3) pLysS. ....	98
Figure 4.2.1. Purification of <i>B</i> /MINPP. ....	99
Figure 4.3.1. pH profile of <i>B</i> /MINPP. ....	100
Figure 4.3.2. Test of <i>B</i> /MINPP recovery after heating. ....	100
Figure 4.4.1. Crystals used in diffraction experiments. ....	102
Figure 4.4.2. Time of flight mass spectrometry positive electrospray ionisation of a purified <i>B</i> /MINPP sample. ....	103
Figure 4.5.1. Ramachandran plot for all non-Pro/Gly residues Phi and Psi angles. ....	105
Figure 4.5.2. Omit single-difference Fourier map of the substrate analogue IHS. ....	105
Figure 4.5.3. Ramachandran plot for all non-Pro/Gly residues Phi and Psi angles. ....	108
Figure 4.5.4. Omit single-difference Fourier map of the phosphate portion on NEP45. ....	108
Figure 4.6.1. Structural overview of <i>B</i> /MINPP. ....	111
Figure 4.6.2. Active site configuration of <i>B</i> /MINPP in the IHS bound form. ....	112
Figure 4.6.3. Snapshots of the catalytic cycle of <i>B</i> /MINPP. ....	113
Figure 4.6.4. <i>B</i> /MINPP conformational changes during the catalytic cycle. ....	114
Figure 4.6.5. Conformational changes of <i>B</i> /MINPP. ....	116
Figure 4.6.6. Domain movement topology in <i>B</i> /MINPP:IHS. ....	116

Figure 4.7.1. Mechanical hinge residues of <i>B</i> /MINPP coloured by conservation. ....	118
Figure 4.7.2. R48 rotation and downstream loop motion in IHS binding. ....	118
Figure 4.7.3. Molecular Phylogenetic analysis by Maximum Likelihood method inferred for bacterial MINPPs. ....	122
Figure 5.2.1.1. Sets representation of HP, branch 2. ....	129
Figure 5.2.2.1. Sets representation of PD motifs distribution in HP, branch 2. ....	130
Figure 5.2.3.1. Molecular Phylogenetic analysis by Maximum Likelihood method. ....	132
Figure 5.3.1.1. Crystal structures of standard sequences of class A and B. page 139 ....	134
Figure 5.3.1.2. Crystal structures of standard sequences of class C and D. page 139 ....	135
Figure 5.3.1.3. Structural similarity dendrogram of HP2. ....	137
Figure 5.3.2.1. Multi-sequence alignment of all standard sequences. ....	140
Figure 5.3.2.2. Structural representation of HP2 conserved motifs in <i>B</i> /MINPP. ....	141
Figure 5.3.2.3. Multi-sequence alignment of MINPPs standard sequences. ....	142
Figure 5.3.2.4. Structural representation of MINPPs conserved motifs in <i>B</i> /MINPP. ....	143
Figure 5.5.1.1. Active site pockets in <i>Bifidobacterium longum</i> subsp. <i>infantis</i> MINPP ( <i>B</i> /MINPP). ....	147
Figure 5.5.2.1. Multiple sequence alignment of the active site of 18 <i>Actinobacteria</i> representative sequences. ....	149
Figure 5.5.3.1. Active site multiple sequence alignment of 72 <i>Bacteria</i> representative sequences ....	152
Figure 5.5.4.1. Molecular Phylogenetic analysis by Maximum Likelihood method. ....	153
Figure 5.6.2.1. Multi-sequence alignment of selected MINPPs. ....	156
Figure 5.6.4.1. Multi-sequence alignment of the active site residues of selected MINPPs. ....	160
Figure 6.1.1. Agarose gels (1%) of the PCR for the amplification of the entry clones. ....	167
Figure 6.1.2. Transformation plates of the BP reactions between the donor vector (pDONR207) and the gene of interest (Gol). ....	168
Figure 6.1.3. Quality check of the cloning – screening for positive inserts. ....	169
Figure 6.3.1. 12% acrylamide SDS-PAGE of the purifications of proteins p8, p10, p15. ....	173
Figure 6.3.2. 12% acrylamide SDS-PAGE of the IMAC purifications of proteins p4, p16 and p2. ....	173
Figure 6.3.3. SDS-PAGE of the final 15 MINPPs samples stocked -80°C. ....	174
Figure 6.4.1. Phosphatase activity against pNPP of the proteins p1-p16. ....	176
Figure 6.5.1. A) Reactions conditions of IP <sub>6</sub> hydrolysis by 15 MINPPs. ....	178
Figure 6.5.2. HPLC chromatograms of the products of IP <sub>6</sub> hydrolysis by the 15 selected MINPPs. ....	180
Figure 6.5.3. Active site alignments of non-positional stereospecific MINPPs (NPS, panel A) and positional stereospecific MINPPs (PS, panel B). ....	182
Figure 6.6.1.1. pH profile of p8 and p15. ....	184
Figure 6.6.3.1. Inhibition of the enzymes A) p15 and B) p8 by IHS. ....	187
Figure 6.6.4.1. DSC experiment on p8 and p15. ....	188
Figure 6.7.1.1. p8 crystal mounted on LithoLoops. ....	191
Figure 6.7.1.2. p15 protein crystals. ....	191
Figure 6.7.2.1. Omit single-difference Fourier map of the substrate analogue IHS. ....	194
Figure 6.8.1. X-ray crystal structures of p8 and p15. ....	196
Figure 6.9.1. Docking studies. ....	200
Figure 6.9.2. Docking studies. ....	201
Figure 8.3.1.1. Summary of the expression trials carried out for <i>EcAppA</i> . ....	221
Figure 8.4.1.1. : Multi-sequence alignment of group A – HP that are no phytases. ....	222
Figure 8.4.1.2. Multi-sequence alignment of group C – MINPPs. ....	222
Figure 8.4.1.3. Multi-sequence alignment of group B – bacterial phytases. ....	223
Figure 8.4.1.4. Multi-sequence alignment of group D – fungal phytases. ....	224
Figure 8.4.2.1. Pairwise sequence alignment of <i>Bifidobacterium longum</i> subsp. <i>infantis</i> MINPP ( <i>B</i> /MINPP) vs <i>Bifidobacterium pseudocatenulatum</i> MINPP ( <i>Bp</i> MINPP). ....	225
Figure 8.4.2.2. Multiple sequence alignment of 18 <i>Actinobacteria</i> representative sequences. ....	226
Figure 8.4.2.3. Multiple sequence alignment of 72 <i>Bacteria</i> representative sequences. ....	227
Figure 8.4.3.1. 3D modelling of 16 selected MINPPs. ....	228
Figure 8.4.4.1. S-S bridges prediction in the 16 selected MINPPs. ....	229

# List of Tables

Table 2.1.1. Plasmids used for protein overexpression in bacterial hosts.....	27
Table 2.2.1. Bacterial strains employed in protein overexpression or plasmids amplification.....	30
Table 2.6.1. Compositions of buffers used in IMAC. ....	39
Table 2.6.1.1. Compositions of buffers used in gel filtration. ....	40
Table 3.6.1. Kinetic parameters for the hydrolysis of inositol hexasulphate.....	76
Table 3.8.1. <i>EcAppA</i> macromolecular structures deposited in the PDB. ....	78
Table 3.9.1. Summary of the macromolecular structures collected for <i>EcAppA</i> wild-type and mutants. ....	82
Table 3.9.2. Refinement statistics of selected <i>EcAppA</i> wild-type and mutant structures.....	82
Table 4.5.1. Solution of <i>B/MINPP</i> macromolecular structures. ....	104
Table 4.5.2. Data collection and refinement statistics for the <i>B/MINPP:IHS</i> complex. ....	106
Table 4.5.3. Data collection and refinement statistics for the structure of <i>B/MINPP E401Q</i> . ....	109
Table 4.7.1. Summary of the results of the DynDom analysis.....	120
Table 4.7.2. Activity of bifidobacterial phytases on different phosphorylated substrates. [21]. ....	123
Table 5.1.1. List of HP-clade2 sequences whose activity has been confirmed experimentally. ....	126
Table 5.3.1.1. Matrix of the root mean square deviations calculated from the pairwise alignment of eight standard HP2 proteins. ....	138
Table 5.4.1. Number of clusters and proton donor distribution in <i>Genus</i> for which multiple sequences were mined. ....	145
Table 5.6.1. Bacterial sources and ecological niches.....	155
Table 5.6.5.1. % GC content of the 16 MINPPs selected for gene synthesis. ....	161
Table 6.2.1. Expression trials of 15 MINPPs.....	170
Table 6.2.2. Best conditions of expression identified for each protein .....	171
Table 6.5.1. Summary table describing the distributions of individual IP <sub>5</sub> % produced by 15 MINPPs and some of their sequence features.....	181
Table 6.6.2.1. Kinetic parameters for the hydrolysis of inositol hexasulphate at 37°C, 5min reactions.....	185
Table 6.7.2.1. Data collection and refinement statistics. ....	193

# CHAPTER 1

## 1. Introduction

### 1.1. Phytases in the animal feed industry- a short history

Phytases are phosphatases that can catalyse stepwise hydrolysis of *myo*-inositol hexakisphosphate (IP<sub>6</sub>) phosphomonoester bonds. The discovery of the first phytase was reported at the beginning of the 20th century [3], but it is only in the last few decades, after the introduction of phytases from *Aspergillus niger* and *Escherichia coli* into the animal-feed additive market – a business with a world-wide turnover of more than \$500 million US per year – that these enzymes have become a hot-spot in enzyme research.

*Aspergillus (ficuum) niger* NRRL 3135 (ATCC 66876) phytase was first isolated by the International Minerals and Chemical Corporation in the 1960s during a screen of phytase producing organisms. Their aim was to find enzymes able to hydrolyse phytate in soybean and other plant meals which were suitable for commercialisation [4]. An engineered version of *A. niger* phytase was released into the global market only thirty years later (1991) under the name of Natuphos®. The phytase market grew slowly for 15 years, as the supplementation of animal feed with inorganic phosphate was more cost-effective. After 2008, the increase in cost of inorganic phosphate and the introduction of environmentally-friendly regulations in several states in the USA led to renewed efforts in phytase research resulting in a new generation of commercial enzymes based on engineered versions of the *E. coli* phytase, AppA, and a huge expansion of the phytase global market [5].

Multiple phytases are currently on the market. In addition to the engineered *E. coli* phytase (Quantum and Quantum blue – ABVista, PhyzymeXP - Danisco), enzymes from *Citrobacter braakii* (Ronozyme Hiphos – Novozyme), *Buttaxiella* sp. (AextraPHY – Danisco), *Aspergillus niger* (Natuphos, BASF) and *Peniophora lycii* (Ronozyme NP – Novozyme) can be found. Despite the large number of products on the market, enzymes able to efficiently hydrolyse IP<sub>6</sub> to completion are still needed. Phytate is often partially degraded by currently available enzymes, with the release of inorganic phosphate, minerals and lower phosphorylated inositols that contribute to the diet of animals. However, total hydrolysis would increase inositol availability, a molecule with antioxidant properties which is involved in multiple metabolic

processes. A new study for example correlates levels of inositol to the maintenance of a healthy immune system in the gut of carp [6]. To achieve this goal, the current strategy is the addition of a higher dose of enzymes to animal feeds (super-dosing, circa 1500-2000 FTU/kg of QuantumBlue - ABVista), however an alternative option could be to develop more efficient enzymes or enzyme mixtures.

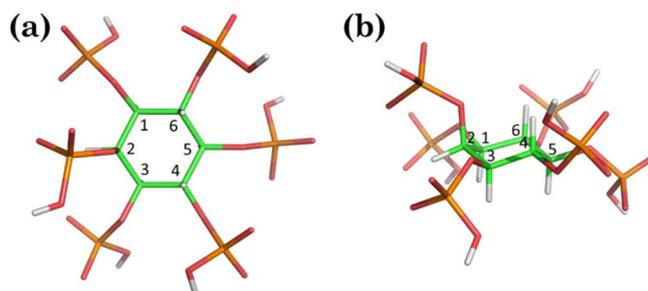
## 1.2. Inositol and their functions

The prime substrates of phytases are inositol phosphates (IP<sub>x</sub>), organic phosphorylated compounds ubiquitous in living cells. *Myo*-inositol hexakisphosphate (or IP<sub>6</sub>) is the predominant form of storage of phosphorus in plant seeds comprising 50-80% of total phosphate [7]. The salt form of IP<sub>6</sub> is commonly known as phytate and can chelate various inorganic ions such as Ca<sup>2+</sup>, Fe<sup>2+</sup>, Zn<sup>2+</sup>, Cu<sup>2+</sup>, Mg<sup>2+</sup>, Mn<sup>2+</sup>. These accumulate in cereals, nuts and beans which are released together with phosphate by phytases during plants germination to become anabolism building blocks [8].

Phytate is considered the primary source of phosphate in animal diet. The gut flora of bi-gastric animals produces phytases which are very efficient in the release of phosphate and minerals. However, the microbiome of mono-gastric animals such as pigs, horses, fishes and humans, is largely unable to degrade phytates to completion leading to the accumulation of these compounds in faeces. As consequence of intensive farming, high concentrations of phosphate can be locally discharged into soil and aquatic ecosystems disturbing the equilibrium of the phosphate cycle in these environments. To avoid environmental hazard and to save money by reducing the amount of inorganic phosphate supplemented into the animal diet, agricultural and animal feed industries started an intensive research program on phytases more than 20 years-long. Currently, phytases are common nutritional supplements present in most animal feeds on the market [5, 9, 10].

Inositols are polyols of cyclohexane which are present in nature in 9 different stereoisomeric orientations. The most common form is *myo*-inositol (cis-1,2,3,5-trans-4,6-cyclohexanehexol – **Figure 1.2.1**). This is also the conformation adopted by IP<sub>6</sub>, an optically inactive meso compound (an achiral molecule, though containing 4 chiral centres), which present a plane of symmetry with either 5-equatorial/1-axial or 5-axial/1-equatorial phosphate groups. In fact, the hydroxyl groups of *myo*-inositol can be replaced by phosphate groups generating molecules with different phosphorylation levels. Depending on the number of substituents, inositol phosphates can be labelled as IP<sub>1</sub>, IP<sub>2</sub>, IP<sub>3</sub>, IP<sub>4</sub>, IP<sub>5</sub> or IP<sub>6</sub>, but also the inositol pyrophosphates IP<sub>7</sub>, IP<sub>8</sub> and IP<sub>9</sub>,

which contain one or more diphosphate groups [11]. IP<sub>6</sub>, present phosphate groups covalently linked through phosphomonoester bonds to each of the 6 carbons of the inositol ring [8, 11].



**Figure 1.2.1. *myo*-inositol hekakisphosphate.**

The boat conformer of *myo*-inositol hexakisphosphate is presented in a 5-equatorial/1-axial conformation. Carbon is coloured in green, oxygen in red, phosphorus in orange, hydrogen in white. (a) top view; (b) side view.

### 1.3. Phytases and their classes

Phytases (or *myo*-inositol hexaphosphate phosphohydrolases) are enzymes widely spread along the evolutionary tree of life. Numerous enzymes able to hydrolyse the phosphomonoester bonds of inositol hexaphosphate have been identified from plants, animals and bacteria. They differ from each other depending on their molecular weight, substrate specificity, mechanism of catalysis, metal ion cofactors, pH and temperature optima. Considering their catalytic process and structures, phytases can be divided in four classes:  $\beta$ -propeller phytases (BPPs), purple acid phytases (PAPs), protein tyrosine phosphatase-like phytases (PTPLPs) and histidine phosphatases phytases (HP2 phytases) [5].

The International Union of Biochemistry and Molecular Biology (IUBMB) has grouped phytases into three classes according to the phosphate position they hydrolyse preferentially: 3-phytases (Enzyme Commission nomenclature: EC 3.1.3.8), 5-phytases (Enzyme Commission nomenclature: EC 3.1.3.72) and 6-phytases (Enzyme Commission nomenclature: EC 3.1.3.26) [12]. 3-/5-/6-phytases follow a standard sequential order of phosphate hydrolysis, on the other hand multiple inositol polyphosphate phosphatases (MINPPs - Enzyme Commission nomenclature: EC 3.1.3.62) can produce mixed IP<sub>5</sub> isomers [12]. Phytases are also able to hydrolyse sequentially IP<sub>5</sub> and lower phosphorylated intermediates releasing, eventually, all the phosphates on IP<sub>6</sub>. In fact, the axial 2-phosphate of IP<sub>6</sub> can be hydrolysed by only two known phytases that are found in the fungi *Debaryomyces castellii* CBS 2923 and *Schwanniomyces occidentalis* HP [13].

### 1.3.1. Beta propeller phytases (BPPs)

BPPs are secreted by bacteria, such as *Bacillus*, in the soil and aquatic environment to utilize phytate as a source of phosphate [14]. It is perhaps due to the harsh environmental conditions that evolution has made these enzymes able to catalyse the complete dephosphorylation of IP<sub>6</sub> at a broad range of pHs (pH optimum around 7) and temperatures. Their structure has a characteristic six bladed  $\beta$ -propeller fold and two binding sites for phosphate, the “cleavage site” in which the hydrolysis reaction starts and an allosteric “affinity site” which enhances substrate affinity [15]. These enzymes hydrolyse phosphate bonds through a direct attack of a metal-bridging water molecule on the IP<sub>6</sub> phosphomonoester bonds involving four Ca<sup>2+</sup> ions which stabilise the transition state [15].

### 1.3.2. Purple acid phosphatases (PAPs)

PAPs are pink/purple coloured metallohydrolases which catalyse the breakdown of IP<sub>6</sub> involving two cofactors: one is Fe(III) and the second can be Zn(II), Mn(II) or Fe(II). These enzymes are also known as tartrate-resistant acid phosphatases (TRAPs) as they are insensitive to inhibition by tartrate. Various PAPs phytases have been purified and characterized during years (the majority comes from plants), but a structure is not available in the literature yet. The first gene found was discovered in Soybean and was described in 2001. PAP phytases have mostly been identified in plants and recently in bacteria. Most family members share seven amino acids involved in metal cofactor chelation and show four conserved amino acid sequence motifs [16].

### 1.3.3. Protein tyrosine phosphatase-like phytases (PTPLPs)

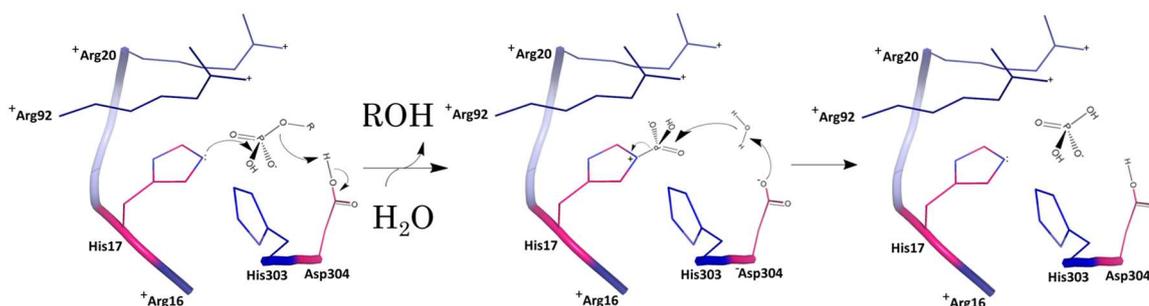
PTPLPs, also known as cysteine phytases, are homologs of the well-characterized eukaryotic protein tyrosine phosphatases (PTP). PTP-like phytases have a peculiar fold, they also share the conserved active site motif "HCEAGVGR" and a loop (named "P loop") which undertakes a substrate-induced conformational change during phosphate hydrolysis [17]. They breakdown IP<sub>6</sub> following standard Michaelis-Menten kinetics and a two-step general acid-general base catalysis with the formation of a Cys-phosphate intermediate. Strangely, this class of phytases can release the product only when the hydrolysis of a second substrate is concluded, in the meantime the product waits in a specific "standby" pocket [17-20]

### 1.3.4. Histidine phosphatases (HP2) phytases

HP2 phytases belong to the branch two of histidine phosphatases (HP2). They are the class of phytases most widely studied (around 50 phytases structures have been deposited in the PDB to date). Histidine phosphatases were named in the past “histidine acid phosphatases” because preliminary works on HPs focused on proteins with low pH optimum. Currently, many enzymes that are active at neutral/basic pH are known (e.g. *B. longum* MINPP [21], *H. sapiens* MINPP1 [22], *L. longiflorum* MINPP [23]). They are called “histidine phosphatases” because of the characteristic histidine acting as a nucleophile in the first stage of the catalysed reaction of hydrolysis of phosphomonoester bonds [24-26].

They share an  $\alpha/\beta$ -domain (structurally conserved in all histidine phosphatases) and the active site motifs involved in acid-base catalysis. These motifs are the “RHGxRxP”, positioned at the N-terminal, and the “HDx/HAE”, which includes the proton donor. The catalytic histidine in HPs is enclosed in motif “RHGxRxP” [25, 27-29], of which the “RHG” triplet is conserved in all HPs while the region “xRxP” is only present in high molecular weight phosphatases (HPs-clade 2 or HP2s) [26, 30]. Also, the “xP” portion is not present in all HP2s [21]. Asp/(Glu) of the “HDx/(HAE)” motif is the most probable proton donor [27, 31].

These enzymes hydrolyse the substrate in a two-step reaction. First, an active site catalytic histidine residue acts as nucleophile, attacking one of the phosphates of IP<sub>6</sub>. This leads to the formation of a histidine-phosphate enzyme intermediate and to the production of a lower phosphorylated inositol polyphosphate. Then, the phosphate is released, and the enzyme returns to its native state (**Figure 1.3.4.1**). Active site residues assist the process establishing hydrogen bonds with the phosphate groups.



**Figure 1.3.4.1. catalytic mechanism of the histidine phosphatase superfamily.**

Residues numbered as in *E. coli* AppA. 1) Nucleophilic attack. Phosphorylation of the active-site histidine and release of IP<sub>5</sub>. 2) Proton donation by the proton donor (PD) and breakdown of the phosphomonoester scissile bond on the phospho-histidine intermediate. 3) Release of an inorganic phosphate.

## 1.4. *Escherichia coli* AppA phytase (*EcAppA*)

The *E. coli* phytase AppA was discovered in 1987 [32] and its catalytic mechanism was first described in 1992 [25]. In this article, the roles of the catalytic core residues “RHGxR” (aa 37-41), “R113” and “H325” were tested for the first time by site-directed mutagenesis. The *E. coli* phytase shares with a wide range of HPs an “HDx” proton donor (aa 325-326), however, other variants of the motif have been identified in the family. The proton donor motif “HAE”, for example, seems to be characteristic of MINPPs. This apparently small difference could influence the enzyme positional stereospecificity of IP<sub>6</sub> hydrolysis (**Chapter 3**).

In HPs, a carboxylic group has been considered essential for proton donation since the 70s, from early works by Saini and Etten (1979) on human prostatic acid phosphatase [33]. A subsequent confirmation of a role in proton donation came from site-directed mutagenesis of the His and Asp of the “HDx” proton donor motif by Ostanin and Etten (1993). The authors studied the effect of mutations on kinetic parameters in *E. coli* AppA [31]. When the aspartic acid Asp326 was mutated to Ala, at pH 2.5,  $V_{\max}$  was drastically reduced without registering a change in  $K_M$ . However,  $K_M$  seemed to be pH dependent with a decrease over the pH range 3.5-6.0. Also, the decrease in  $V_{\max}$ , despite the binding of the substrate not being affected, appears to be nevertheless dependent on the substrate pKa. In the mutant,  $V_{\max}$  is higher, for substrates with better leaving group, e.g. higher  $V_{\max}$  for pNPP than for fructose-1,6-diphosphate, suggesting an involvement of Asp326 in proton donation. Results from the mutation of the adjacent residue His325 to Ala showed a stronger loss in activity, although not substrate dependent. The authors suggested that Asp326 may be involved in proton donation for the hydrolysed substrate to be released, while His325 could contribute to the maintenance of the protonation state of the Pi-intermediate facilitating the donation of the Pi group to water or an acceptor, e.g. ethylene glycol [31].

The first crystal structure of *EcAppA* was solved in 2000 [24]. *EcAppA* has a classical HP2 phytase fold consisting of an  $\alpha$ -domain and an  $\alpha/\beta$ -domain, at the interface of whose, the active site is positioned. Active site dimensions have been correlated to substrate specificity. Liu *et al.* compared the *E. coli* phytase AppA and *A. fumigatus* phytase, reporting that a helix (aa 209-228) present in the *E. coli* enzyme and absent in the *A. fumigatus* HP, decreases the active site volume of the bacterial phytase and restricts substrate specificity to phytate only. Instead, the *A. fumigatus* phytase is able to degrade a broader range of phosphorylated compounds [34].

*EcAppA* phytase hydrolyses phosphates on IP<sub>6</sub> in a sequential fixed order, which depends on the position of the orthophosphate on the inositol ring. Typically, sequential cleavage of IP<sub>6</sub> by this enzyme leads to the generation of one predominant IP<sub>5</sub> species I(1,2,3,4,5)P<sub>5</sub> (also referred as 6-OH IP<sub>5</sub>). Further attack yields to I(2,3,4,5)P<sub>4</sub>, I(2,4,5)P<sub>3</sub> or I(1,2,4)P<sub>3</sub>, I(1,2)P<sub>2</sub> or I(2,5)P<sub>2</sub> or I(4,5)P<sub>2</sub> and finally I(2)P<sub>1</sub> or I(5)P<sub>1</sub> [35].

## 1.5. Bacterial MINPPs – A timeline

In 1997, when the first MINPP enzyme was characterized, it appeared clear that MINPPs are members of the HP2s family (Pfam ID: PF00328, His\_Phos\_2).

The first MINPP identified was a rat (*Rattus norvegicus*) hepatic phytase and its sequence has aligned with 61% and 55% identity to *A. niger* and *Myceliophthora* HPs phytases in the region between the amino acids 48 and 65, centred on the RHGxRxP motif [36]. It was named multiple inositol phosphate phosphatase because of its broader substrate specificity in comparison with e.g. *A. niger* HP. *Rattus norvegicus* MINPP could attack IP<sub>6</sub> with no preference between equatorial phosphates, producing three IP<sub>5</sub> peaks, two IP<sub>4</sub> peaks and one IP<sub>3</sub> peak on an HPLC chromatogram of the products of hydrolysis (data not shown). In contrast, it was highly selective on I(1,3,4,5)P<sub>4</sub> and I(1,3,4,5,6)P<sub>5</sub> generating, respectively, I(1,4,5)P<sub>3</sub> and I(1,4,5,6)P<sub>4</sub>. Its activity was also assayed against diphosphoinositol polyphosphates [37]. The enzyme showed its ability to hydrolyse [5-β-<sup>32</sup>P]PP-IP<sub>5</sub> liberating [<sup>32</sup>P]P with no formation of [5-β-<sup>32</sup>P]PP-IP<sub>4</sub>. It attacked (PP)2-IP<sub>4</sub>, however, no hydrolysis was detected on smaller substrates like I(1,4)P<sub>2</sub>, I(1,4,5)P<sub>3</sub>, I(1,3,4)P<sub>3</sub>, pNPP, PP, Glu-6P, Fru-1,6BP or ATP [36].

No other study has been carried out since on such a wide range of substrates, despite this, available data show that MINPPs do not have a uniform substrate specificity. Notably, it has been demonstrated that *R. norvegicus* MINPP, as well as MINPP from *H. sapiens*, *G. gallus* and *D. discoideum* are able to remove, like bisphosphoglycerate mutase, the phosphate from the carbon in position 3 of 2,3-bisphosphoglycerate expanding the regulating capacity of the Rapoport–Luebering glycolytic shunt [38]. A substrate apparently shared among all MINPPs is IP<sub>6</sub>, however, when moving to smaller phosphomonoesters, specificity becomes mostly unpredictable. A major reason is because a preferential substrate/function of MINPPs *in vivo* is still uncertain despite evidence of a role of these enzymes in a variety of cellular processes and organisms is established. Also, rarely *in vivo* studies are

coupled with *in vitro* kinetics which could establish catalytic parameters on cellular substrates.

Accounts on specificity were reported for *Mus musculus* MINPP1, which can hydrolyse [H3]I(1,3,4,5)P<sub>4</sub>, but, unlike *R. norvegicus* MINPP, is also able to attack pNPP [22] like yeast and *E. coli*-like phytases. The chick HiPER1 can hydrolyse pNPP, ATP and ribose-1-P even if phytate appears to be a better substrate [39]. Rat hepatic MINPP and other member of the family are known for their unspecific IP<sub>6</sub> hydrolysis [22], contrarily, *Lilium longiflorum* MINPP prefers the cleavage of Pi at the C5 position of the inositol ring [23]. Another point of discordance in the family is the extent of IP<sub>6</sub> hydrolyses. Some enzymes, like *L. longiflorum* MINPP [23], seem not to degrade IP<sub>6</sub> to completion, to I(2)P<sub>1</sub>, but stop at IP<sub>3</sub>. I(2)P<sub>1</sub>, the inositol ring with only an axial Pi group left, is generally considered to be the final product of a HP-catalysed hydrolysis reaction. Until now, only two HP-clade2 are known to generate a Pi-free inositol: *Debaryomyces castellii* CBS 2923 and *Schwanniomyces occidentalis* HP [13]. Bacterial MINPPs like *B. longum* MINPP, *B. pseudocatenulatum* MINPP and *B. thetaiotaomicron* MINPP attack IP<sub>6</sub> without a strong positional specificity proceeding to generate Pi and presumably I(2)P<sub>1</sub>. *B. longum* and *pseudocatenulatum* also retain some activity on pNPP, acetyl-phosphate and fructose 1,6-bisphosphate, even if this is <10% in comparison with phytate hydrolysis. *Bt*MINPP has a 20-36-fold higher V<sub>max</sub> for I(1,2,3,5)P<sub>4</sub> than for IP<sub>6</sub> [21, 40-42].

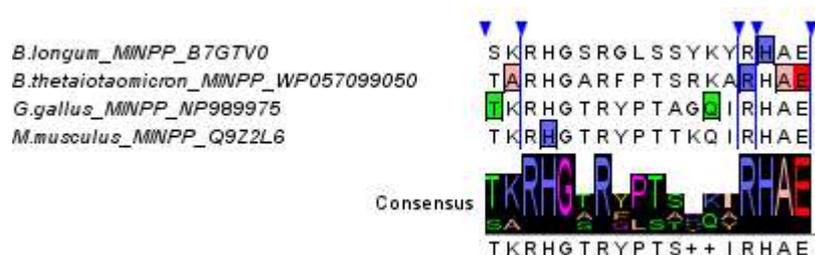
The second protein to be characterized was, in 1998, the chick HiPER1 (Histidine Phosphatase of the Endoplasmic Reticulum-1) [43]. Its sequence showed that avian MINPPs retain a XDEL endoplasmic reticulum retention signal. This destination is in common with other MINPPs from higher organism e.g. the mammalian *R. norvegicus* [36], *M. musculus* [22], *H. sapiens* [22, 44], or the plant *T. aestivum* [45], *H. vulgare* [45], *A. thaliana* [45], *Z. mays* [45]. Describing HiPER1, Romano et al. (1998) pointed out that, in addition to the "RHGxRxP" motif, in this protein (and more generally in MINPPs), the residues R131 and H332 are also conserved [43]. Both residues are present in the whole HP family (clade1 and 2). The first residue, R131, is important in the enzyme-substrate binding interactions. The second residue, H332, is positioned in the proton donor region and is a candidate, with aspartic and glutamic acid, alone or in concert with the latter, for the role of proton donor in the final stage of IP<sub>6</sub> hydrolyses. Mutations of either residues in *EcAppA* did not inactivate the enzyme but reduced its activity more than 200-fold [24-26, 29].

In 1999, the DNA fragments coding for *Homo sapiens*, *Mus musculus* and two *Drosophila melanogaster* MINPPs were sequenced for the first time [22]. Also, the MINPP sequence of *Arabidopsis thaliana* was available within GenBank [46] along with the sequences of the MINPPs of *Rattus norvegicus* and *Gallus HiPER1*. Their translated protein sequences could be analysed for gene structures, conservation and the first phylogram was inferred for the HP family including MINPPs, yeast phytases, lysosomal acid phosphatases from various organisms, the human prostatic acid phosphatase and *E. coli* AppA. The genomic organisation for *H. sapiens*, *M. musculus* and *G. gallus* MINPPs was recognised to be identical, unlike the more distant *D. melanogaster* and *A. thaliana* MINPPs. The yeast enzymes appeared to be closer to MINPPs than any other HP. Chi et al. (1999) hypothesized that MINPPs diverged from yeast enzymes just after they diverged from lysosomal and prostatic acid phosphatase [22]. They suggested that MINPPs enzymes could group together as a separate branch within the histidine phosphatase family. Another important information provided, even though not highlighted by Chi et al. (1999) is the different proton donor motif of *D. melanogaster* MINPPs [22]. The two proteins display a “HST” or a “HSG” triplet, where the only residue able to act as a proton donor is the conserved histidine [22].

In MINPPs, the classical role of aspartic acid was presumed to be taken by the glutamic acid of the conserved “HAE” triplet. Exceptions to this rule were reported for *D. melanogaster* and other species by Cheng and Andrew (2015) [47], while the work on MINPPs described in this thesis was on going. MINPPs containing proton donor motif outliers were: “HGE” - *Danio rerio* MIPPB, “HSE” - *Dictyostelium discoideum* MIPP2, “HSG” - *Anopheles gambiae* and *Drosophila melanogaster* MIPP2, “HST” - *Drosophila melanogaster* MIPP1, “HEV” - *Cryptococcus neoformans* Phytase 11 but also “HDT” - *Anopheles gambiae* MIPP1. The article, however, was mainly focused on another subject, on *in vivo* studies that suggests an advantageous role for extracellularly expressed *D. melanogaster* MIPP1 in epithelial cells during collective migration in embryonic trachea [47]. Considering the proposed importance of an Asp or Glu residue in the active site of histidine phosphatases [27, 29, 31, 33], it could be argued that the enzymes described above may be characterised by a reduced activity or even by differenced in substrate specificities in comparison with classical “HAE” MINPPs. The lack of a candidate proton donor could suggest that they may be pseudoenzymes, in other words, enzymes predicted not to maintain the basic enzyme function shared by the family [48]. However, these proteins, even though characterised by an unusual proton donor, seem to have a clear biological role *in vivo*.

The expression of MIPP1 and MIPP2 in *D. melanogaster* is very finely tuned and it supports this evidence: MIPP2 (proton donor: HGS) is expressed ubiquitously, MINPP1 (proton donor: HST) - is dynamically expressed and deletion or protein over-expression leads respectively to decrease or increase in filopodia number [47]. Also, it cannot be ruled out that another active site residue, which may be positioned elsewhere in the amino acid sequence, could counterbalance the loss of the classical aspartic acid or glutamic acid of the HDx/HAE triplet by acting as proton donor in their place. Another uncommon proton donor triplet is found in a maize MINPPs [45] that carries a LAE motif, this time the histidine is replaced by a leucine, no bacterial MINPPs with this substitution has been found yet.

In 2006 a second article was published on HiPER1 in support of the genetical modification of poultry to constitutively secrete the MINPP into their digestive tract [39]. In this article mutations in the chicken MINPP were introduced to test the involvement in catalysis of two further amino acids in addition to the RHGxR, R, Hxx motifs. The first residue to be mutated in a MINPP was the catalytic histidine (H89A) in the mouse enzyme by Chi et al. (1999) which resulted in a complete inactivation of the protein towards pNPP [22]. Instead in the chicken MINPP, the residues T27(G) and Q78(A) were at test. Both mutations resulted in a dramatic decrease of  $K_M$  for IP<sub>6</sub> hydrolysis, though this change was coupled with a 96% decrease in  $V_{max}$  for the T27G mutant and a 13% decrease for the Q78A mutant. The overall catalytic efficiency ( $V_{max}/K_M$ ) for the Q78A mutant was doubled. Also, substrate specificity was slightly changed for this mutant: it is 2.4-fold less active towards pNPP and 1-fold more active on ribose 1-phosphate, ATP rate of hydrolysis is unchanged.



**Figure 1.5.1. alignment of the active site of four MINPPs whose functionality was tested by site-directed mutagenesis.**

The mutated amino acids are highlighted in boxes and coloured depending on their chemical properties. *B. longum* MINPP was made inactive by an H401N mutation of the histidine in the proton donor triplet [42]. The substitutions A31Y, H59A, R183D, A324D, E325N, A324D/E325N were introduced in *B. thetaiotaomicron* MINPP [40]. Residues T27G and Q78A were mutated in the chicken HiPER1 [39]. *Mus musculus* MINPP was made inactive by an H89A mutation of the catalytic histidine [22].

In recent years, the catalytic relevance of other residues was studied by site-directed mutagenesis of *Bt*MINPP [40, 41]. Mutations A31Y, R183D, A324D, E325N and A324D/E325N were introduced in this enzyme. For the A31Y substitution, a 2.5-fold decrease in activity was registered. The R183D mutant instead, 10-fold less active, caused a loss in the enzyme ability to hydrolyse the phosphate groups at positions 4 and 6 of the inositol ring of IP<sub>6</sub> [41], the predominant IP<sub>5</sub> species produced being I(1/3-OH)P<sub>5</sub>. Mutants A324D, E325N and A324D/E325N were more than 100-fold less active and generated respectively predominant peaks of I(4/6-OH)P<sub>5</sub>, I(1/3-OH)P<sub>5</sub>, I(1/3-OH)P<sub>5</sub> [40]. Mutations in MINPPs are described in **Figure 1.5.1**.

In 2006, two MINPPs isoforms were characterized from lily pollen [23]. These MINPPs distinguished themselves for their preferential IP<sub>6</sub> hydrolysis of the Pi group on C5 of the inositol ring (the final product is *cis*-I(1,2,3)P<sub>3</sub>), but also for the requirement of Ca<sup>2+</sup> ions for activity and the consequent inhibitory effect of EDTA. In 2007, four wheat and three barley isoforms were added to the list of known phytases [45]. In line with the behaviour of most of HP, they showed to be inhibited by metal ions. As expected from their conserved motifs, they cluster together with other plant MINPPs in a phylogenetic tree of HPs and PAPs. In 2012, the first bacterial MINPP homologues in the Gram-positive *Bifidobacterium pseudocatenulatum* ATCC 27919 (*Bp*MINPP) and *Bifidobacterium longum* subsp. *infantis* ATCC 15697 (*B*MINPP) were identified [21]. Both enzymes showed to have a preference for the hydrolysis of phytate above many other substrates (activity is decreased more than 90% in pNPP, acetyl-phosphate, fructose 1,6-bisphosphate; no activity is detected in phosphoenolpyruvate, AMP, ADP, ATP, glyceraldehyde-3-phosphate, glucose-1-phosphate, glucose 6-phosphate, fructose 1-phosphate), a pH optimum of 5.5, activation by Ca<sup>2+</sup> at low concentrations (<2mM) in particular at neutral pH, inhibition by Ca<sup>2+</sup> at high concentrations (>4mM), weak inhibition by EDTA [21].

In 2014, the characterisation of a MINPP from the Gram negative *Bacteroides thetaiotaomicron* (*Bt*MINPP) was reported. It stimulates Ca<sup>2+</sup> mobilization in human colonic epithelial cells after its delivery is mediated by outer membrane vesicle fusion [40, 41]. The article also described the first crystal structures solved for an enzyme of the MINPPs family in complex with Pi or the substrate analogue IHS (a representation can be found in Chapter 5, PDB id: 4fdu). This MINPP has the highest V<sub>max</sub> ever reported in this family of enzymes, corresponding to 178 μmol/mg/min. It can hydrolyse indiscriminately IP<sub>6</sub> on each Pi position except for the axial C2-Pi and with a V<sub>max</sub> 20-36-fold higher for I(1,2,3,5)P<sub>4</sub> than with IP<sub>6</sub>. It has three activity

optima at 2.5, 4.0 and 7.5, a pH profile that overlap with the pH of the human gastrointestinal tract [41].

Bacterial homologues of *Bt*MINPP were found in *Bacteroidetes*, *Fusobacteria*, *Spirochaetae*, *Actinobacteria*,  $\alpha$ -/ $\beta$ -/ $\gamma$ -*Proteobacteria*. No representatives were found in *Firmicutes* a common human microbiota of the GI tract. Inferring a phylogenetic tree of the MINPPs family, bacterial MINPPs appear to group in a separate branch from the eukaryotic counterparts, but all are related to their higher organism homologs [41].

## 1.6. Aims of my project

The aim of this work was to explore positional stereo-specificity in HP phytases in order to discover or engineer enzymes able to process *myo*-inositol hexakisphosphate to completion.

To achieve this goal, the active site of *E. coli* AppA was engineered. The characteristic proton donor motif “HDT” was replaced by a MINPP-like proton donor “HAE”. Four mutations were introduced and characterized by biochemical, biophysical and structural methods.

Also, the X-ray macromolecular crystal structure of a MINPP from the Gram+ bacterium of the human gut flora *Bifidobacterium longum* susp. *infantis* was solved in complex with the substrate analogue inositol hexasulphate to identify the structural features determining its unique IP<sub>5</sub> product profile. When compared to the structure of the apo-enzyme [40], the complex revealed a large  $\alpha$ -domain movement. The sequence determinants of the motion were studied, as well as their conservation in the family of MINPPs.

Finally, a genome mining analysis was carried out to search for novel MINPPs. This work lead to the identification of potential sequence determinants of positional stereo-specificity of inositol hexakisphosphate hydrolysis.

# CHAPTER 2

## 2. Experimental protocols

This chapter aims to provide a general description of routine methods adopted in the production, chemical-physical characterization and engineering of heterologous proteins.

### 2.1. Plasmids

Plasmids are the vectors used in this work to carry genes encoding for proteins of interest. They are extrachromosomal genetic elements that replicate autonomously, can vary in size (<2 to >1000 kbp), in geometry (linear or circular double strand DNA), in copy number (from 1 to hundreds of copies per chromosome) and can be found in *Bacteria*, *Archaea* and lower *Eukaryotes*. Their beneficial role in these organisms is diverse, for example plasmids can confer resistance to antibiotics, but also can allow strains to be able to survive in harsh environments (e.g. giving them the ability to use alternative carbon sources/electron donors) or to produce virulence factors [49].

Vectors used in heterologous protein expression are designed accordingly to the users' need, though they maintain the backbone of an original bacterial plasmid. In fact, to be amplified by a bacterial host, their origin of replication (*ori*) needs to be recognized. In addition, *ori* controls their copy number: commonly used *ori* in *E. coli* are the pColE1, for low copy number plasmids of 15-20 elements per chromosome (e.g. pBR322 [50], pET28a-c, pOPINA/B, pDEST17), or a pColE1 derivative: pUC *ori*, for high copy number plasmids of 500-700 elements per chromosome (e.g. pTriEx2, pOPINF, pDONR2017).

Commercial plasmids typically carry a selectable marker, allowing users to clone genes of interest downstream of a strong promoter which can be easily induced and tightly regulated by a repression system. They also contain primer binding sites upstream and downstream the insert to allow its amplification by PCR and sequencing.

Commercial vectors for bacterial expression grant users the possibility to finely tune protein expression; for example, they can assure cytoplasmic localisation or periplasmic translocation, they can provide fusion tags able to enhance solubility (e.g. glutathione S-transferase: GST-tag, thioredoxin: Trx-tag, N-utilisation

substance: Nus-tag) or catalyse disulphide bonds formation (e.g. Trx-tag, DsbA-tag, DsbC-tag).

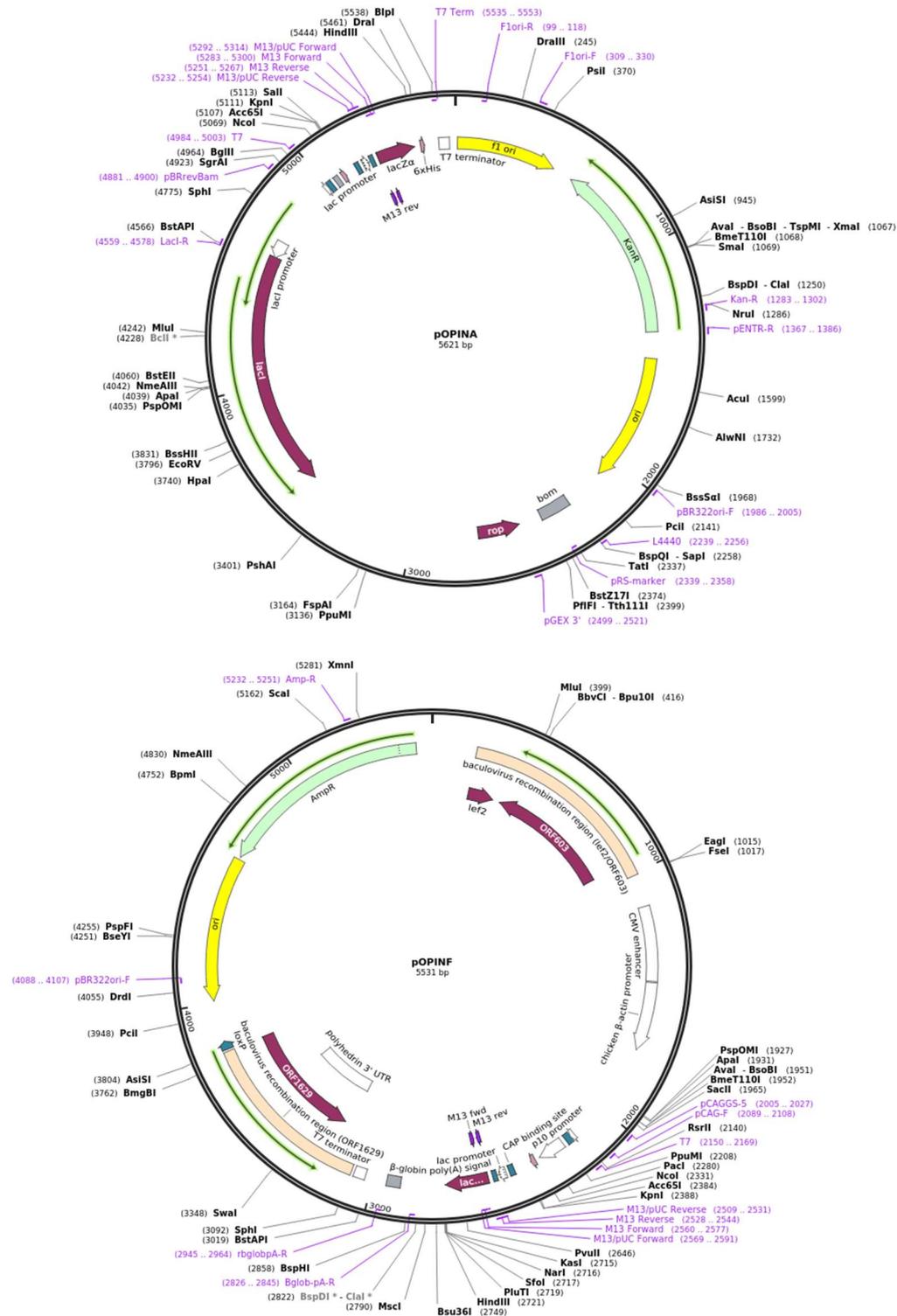
Plasmids used in this work for protein expression in bacterial hosts are: pOPINA (**Figure 2.1.1**), pOPINB (**Figure 2.1.1**), pOPINF (**Figure 2.1.1**), pET28a (**Figure 2.1.2**) and pDEST17 (**Figure 2.1.2**). Some of their features are summarized in the following table:

**Table 2.1.1. Plasmids used for protein overexpression in bacterial hosts.**

POI: protein of interest; HIS6: six histidine tag; 3C:His-tagged 3C-protease.

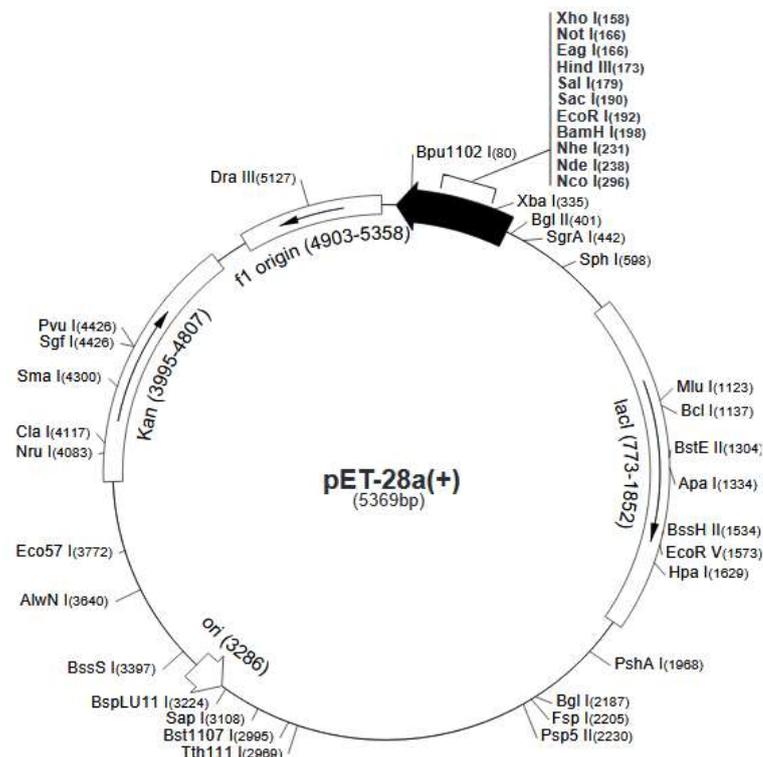
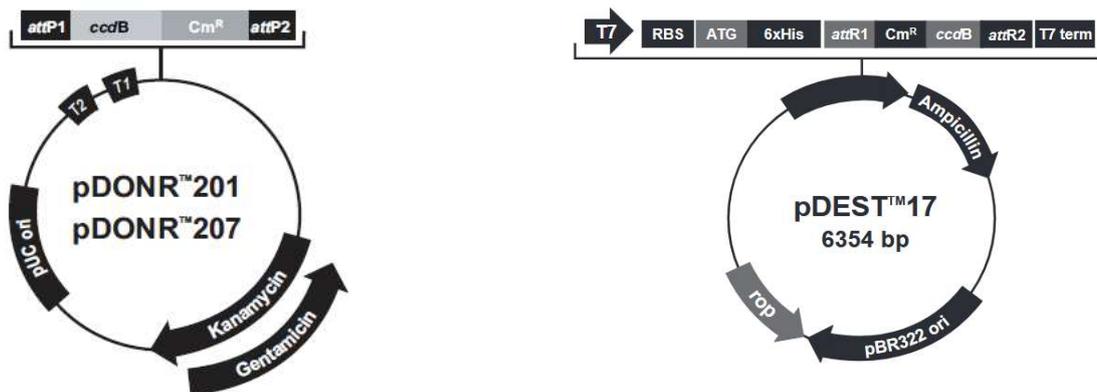
Plasmid	Resistance	Construct	Ori /copy nr.	Cloning	Other features
pOPINA	Kan (50 µg/mL)	POI-KHIS6	pColE1 (15-20)	In-Fusion	<i>lacI</i> – <i>lac</i> Operator
pOPINB	Kan (50 µg/mL)	HIS6-3C-POI	pColE1 (15-20)	In-Fusion	<i>lacI</i> – <i>lac</i> Operator
pOPINF	Amp (100 µg/mL)	HIS6-3C-POI	pUC (500-700)	In-Fusion	<i>Baculovirus</i> recombination region
pET28a	Kan (50 µg/mL)	HIS6-3C-POI	pColE1 (15-20)	Restriction Digest/Ligation	<i>lacI</i> – <i>lac</i> Operator
pDONR207	Gent (20 µg/mL)	HIS6-3C-POI	pUC (500-700)	Gateway	<i>ccdB</i>
pDEST17	Amp (100 µg/mL)	HIS6-3C-POI	pColE1 (15-20)	Gateway	<i>ccdB</i>

The plasmids in **Table 2.1.1**, are compatible with bacteriophage T7 expression systems and are IPTG-inducible. Gene of interests cloned into these vectors are under the control of a strong T7 promoter recognized selectively by an IPTG-inducible T7 RNA polymerase (whose gene is integrated in the genome of  $\lambda$ DE3 lysogen bacterial cells under the control of a *lac* promoter). The transcription of T7 RNA polymerase and of the insert is inhibited by the binding of the protein LacI to an operator region, downstream of their promoters. IPTG can inhibit the binding of LacI to the operator region allowing the *E. coli* RNA polymerase to start the transcription of the T7 RNA polymerase which will then bind to the now free T7 promoter and start transcription of the gene of interest.



**Figure 2.1.1. pOPINA and pOPINF.**

**Top plasmid: pOPINA.** T7 expression system, allows selection by kanamycin and  $\alpha$ -complementation of the  $\beta$ -galactosidase gene. It carries a copy of *lacI* under the control of lac promoter. **Bottom plasmid: pOPINF.** Contains a promoter for T7 bacterial expression. It is a p10 *Baculovirus* expression system and a CMV+chicken  $\beta$ -actin promoter is present upstream the cloning site. It can be used either for bacterial expression or for expression in insects and mammalian cells. It is selected by ampicillin resistance.



**Figure 2.1.2. pDONR™207, pDEST™17 and pET-28a(+).**

**Top left plasmid: pDONR207.** Entry vector for Gateway cloning (see characteristic *attP1* and *attP2* recombination sites). Vectors with insert can be selected by gentamycin resistance. Also, empty vectors own a *camR* gene and *ccdB* gene in the cloning site. This allow the selection using CcdB protein: a DNA gyrase inhibitor. Only *E. coli* strains containing a mutant DNA gyrase resistant to CcdB or a CcdA antitoxin can survive when the empty vector is present. This allow during cloning the auto-elimination of common containment hosts transformed with empty vectors. Instead in resistant strain, empty plasmids containing cells can be selected by chloramphenicol resistance. **Top right plasmid: pDEST17.** Destination vector for Gateway cloning. Vectors with insert are selected by ampicillin/carbenicillin resistance. Empty vectors contain, as pDONR207 plasmids, a *camR* gene and *ccdB* gene. It is an expression vector; thus, it contains a ribosome binding site, Met triplet and an N-term His-tag upstream the cloning site and a T7 terminator sequence downstream. **Bottom plasmid: pET-28a(+).** A traditional plasmid for restriction/ligation cloning. It carries a gene for kanamycin resistance and a gene encoding for the LacI repressor. It is a T7 expression vector and contains upstream the multiple cloning site a ribosome binding site, Met triplet. It allows the addition of N- or C-term His-tag to the insert.

## 2.2. Growth media and bacterial strains

*Growth media.* Miller's LB-broth (10 g tryptone, 5 g yeast extract, 10 g NaCl) was the nutrient-rich media used in all bacterial cultures. LB solutions were autoclave-sterilized before use and selective pressure was applied during cell growth by the addition of antibiotics.

*Bacterial strains.* All hosts for protein expression were IPTG-inducible T7 expression system strains ( $\lambda$ DE3 lysogens) derived from *Escherichia coli* B or K-12 (Table 2.2.1). In "pLysS" strains, a plasmid encoding T7 lysozyme, a natural inhibitor of T7 RNA polymerase, is maintained to annul basal expression of genes under the control of the T7 promoter.

**Table 2.2.1. Bacterial strains employed in protein overexpression or plasmids amplification.**

Strains	Antibiotic resistance	Key Features
BL21(DE3)	-	Protein expression host. <i>E. Coli</i> B derived. <i>lon</i> and <i>ompT</i> protease deficient.
BL21 (DE3) pLysS	Cam (30 µg/mL)	Protein expression host. <i>E. Coli</i> B derived. <i>lon</i> and <i>ompT</i> protease deficient.
Rosetta 2 (DE3) pLysS	Cam (30 µg/mL)	Protein expression host. <i>E. Coli</i> B derived. <i>lon</i> and <i>ompT</i> protease deficient. It expresses the tRNA for rare codons AUA, AGG, AGA, CUA, CCC, GGA and CGG.
Rosetta-gami 2 (DE3) pLysS	Tet (10 µg/mL) Cam (30 µg/mL) Str - chromosomic	Protein expression host. <i>E. Coli</i> K-12 derived. Express the tRNA for rare codons AUA, AGG, AGA, CUA, CCC, GGA and CGG. <i>trx</i> and <i>gor</i> mutations for enhanced formation of S-S bridges in the cytoplasm, Kan sensitive. Leu auxotroph.
SHuffle T7 Express	Spec (50 µg/mL) Str - low levels	Protein expression host. <i>E. Coli</i> B derived. Expresses constitutively the disulphide bond isomerase and chaperone DsbC. <i>lon</i> and <i>ompT</i> protease deficient. <i>endA1</i> deficient. Resistant to phage T1 ( <i>fhuA2</i> ).
SHuffle T7 Express lysY	Spec (50 µg/mL) Str - low levels	Protein expression host. <i>E. Coli</i> B derived. Expresses constitutively the disulphide bond isomerase and chaperone DsbC. <i>lon</i> and <i>ompT</i> protease deficient. <i>endA1</i> deficient. Resistant to phage T1 ( <i>fhuA2</i> ). Control of T7 RNA Polymerase by lysozyme
Stellar	-	Cloning host. <i>endA1</i> , <i>recA</i> and <i>hsdR</i> deficient. Allow blue-white colour screening.
XL-1 blue	Tet (10 µg/mL)	Cloning host. <i>endA1</i> , <i>recA</i> and <i>hsdR</i> deficient. Allow blue-white colour screening.

## 2.3. Transformations

In *Bacteria*, plasmids can be easily transferred by conjugation, spreading quickly, in a colony, the traits they encode for. However, most *E. coli* commercial vectors are engineered for low mobilization (plasmids specifically designed for conjugation are an exception to the rule). *Bacteria* can donate low mobilisation plasmids only through vertical transfer during duplication. However, plasmids need to be placed into the host cell in the first place. Two methods are most frequently employed for the insertion of vectors in bacterial hosts: chemical transformation and electroporation. In this thesis, chemical transformation is the preferred approach.

In chemical transformation, cells previously made competent for the uptake of a foreign plasmid are incubated with it in an ice-cold  $\text{CaCl}_2$  solution, which could shield the negative charges of Lipid A in order to allow the passage of the DNA, in turn negatively charged. A short heat shock generates convective waves throughout the membranes to facilitate the entry of the plasmid into the cell.

In-house competent cells were prepared for routine experiments, but commercial aliquots were used in cloning (Stellar<sup>TM</sup>, Clontech) and mutagenesis (XL10-Gold<sup>®</sup> Ultracompetent Cells, Stratagene).

### 2.3.1. Preparation of competent cells

Two 100 mL flasks of LB were inoculated with 2 mL o/n culture to which suitable antibiotics were added. Flasks were incubated at 37 °C, with shaking (180 rpm), until they reached OD 0.3-0.4. Cells were aliquoted in 4 falcon tubes of 50 mL and were incubated on ice, for 15 min. They were harvested at 2000 xg, 4 °C, for 5 min, and the media was discarded. Pellets were resuspended in a total of 30 ml solution 0.1 M  $\text{CaCl}_2$  and left on ice, for 30 min. Cells were again harvested at 2000 xg, for 5 min, at 4 °C, and the media was discarded. Pellets were resuspended in a total of 4 ml solution 0.1 M  $\text{CaCl}_2$ , 30% glycerol. Finally, they were aliquoted, snap frozen and stored at -80 °C. All media and solutions were sterilized before use.

### 2.3.2. Transformation

Plasmids were mixed with a stock aliquot of competent cells to reach a final DNA concentration of 1-2 ng/ $\mu\text{L}$ . Cells were incubated on ice for 30 min, and heat-shocked for 30-45 sec, at 42 °C. They were left on ice for 1-2 min, to recover, and

supplemented by preheated SOC media (42 °C) of volume 3x larger the volume of the cell solution. Cells were incubated at 37 °C shaking, for 45 min, plated in sterile conditions on antibiotic selective LB-agar, and left at 37 °C o/n for colony development.

## 2.4. Cloning

Cloning consists in the insertion of a DNA fragment into a vector. The latter can then be transformed into an organism to allow storage and/or transcription of the fragment. A wide range of methods have been developed to accomplish this objective. Classical cloning procedures consist in the ligation of a plasmid and a gene of interest both previously cleaved by a class II endonuclease.

In this thesis, Gateway and In-Fusion cloning were used. Gateway cloning is a straightforward efficient LIC method, in use since the late '90s. Insertion of a gene of interest is obtained through recombination by Invitrogen proprietary enzymes: LR and BP clonase, engineered to selectively recognised *att*-sites. In-Fusion cloning is an alternative LIC method developed by the Oxford Protein Production Facility (UK). Genes of interest are inserted in pOPIN Suite vectors and integration is catalysed by the proprietary In-Fusion enzyme. This method not involve recombination.

### 2.4.1. In-fusion cloning

In the In-Fusion technology, two linear DNA fragments, sharing the same extensions, are fused together by the Clontech's proprietary enzyme, In-Fusion. This one-step reaction of ligation independent cloning promises >90% cloning efficiency in high-throughput settings and it is amenable to automation. The techniques have proved to be versatile: a large number of inserts from different sources have been tested and a many pOPIN vector have been developed to suit user needs [51]. The procedure includes steps of: pOPIN vectors digestion, gene amplification, In-Fusion Reaction, transformation of the in-fused plasmid.

*pOPIN vector digestion.* pOPIN vectors were linearized by digestion at positions adjacent to the two sequences recognized by the In-Fusion enzyme. pOPINA and pOPINB were the vectors used. They were amplified in DH5 $\alpha$ , purified, and linearized by KpnI and HindIII. In a typical digestion, 1 $\mu$ g of plasmid was cut by 10

U of the restriction enzymes KpnI and HindIII (New England Biolabs) in 1x CutSmart buffer, for 1 h, at 37 °C, before enzyme inactivation at 80 °C, for 20 min.

*Gene amplifications.* Genes were amplified by PCR to build a stock at concentration suitable for cloning. Two extensions, recognized by the In-Fusion enzyme, were fused at 5' and 3' of the insert, by PCR. DNA electrophoresis (1% agarose gel) was set up to visualize PCR results and to purify the gene from gel before cloning (NucleoSpin Gel and PCR clean-up kit (Macherey-Nagel)). Sequences and primers are attached in **Supplemental information 8.1 and 8.2**. Polymerase chain reactions consisted of an initial denaturation of the double helix at 98 °C, for 3 min, followed by 30 cycles of: denaturation (98 °C, for 15 sec), primer annealing (at the lowest melting temperature between the two primers + 3 °C, for 30 sec), and extension (72 °C, for 45 sec). A final extension step at 72 °C for 10 min was included before cooling to 4 °C. Reactions were set up to contain 1x HF buffer (which provides 1.5 mM of MgCl<sub>2</sub>), 0.2 mM dNTPs, 0.5 μM primers mixture, 70 pg/μL plasmid or 3 ng/μL bacterial genome and 0.02 U/μL Phusion DNA polymerase.

*In-Fusion reaction.* In-fusion reactions occur between the digested plasmid and the gene which, as a result, will be ligated in a new circular plasmid vector. To a mix of 4 μL of digested plasmid and gene in equal concentrations, 1 μL of In-Fusion HD Enzyme Premix was added. In-Fusion reactions were incubated at 50 °C between 20 to 30 min.

*Transformation of 'In-fused' plasmid.* In-fusion reactions were transformed into 50 μL Stellar competent cells for plasmid amplification and blue-white screening. Cells were plated on agar supplemented by suitable antibiotics, 1 mM IPTG, 40 μg/mL XGal; cells containing a plasmid with an insert are complementation deficient and won't be able to produce a functional β-galactosidase able to hydrolyse X-gal, and therefore appear as white colonies. White colonies were screened for the presence of the correct insert (**Chapter 2.5**).

#### **2.4.2. Gateway cloning**

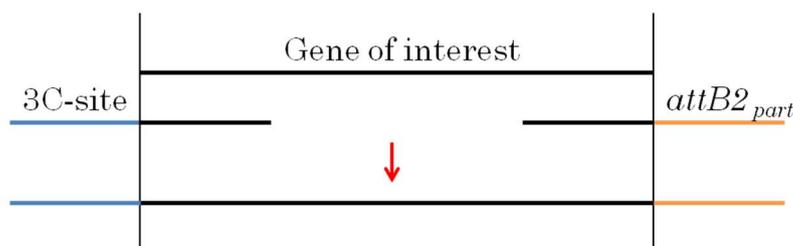
In Gateway cloning, the insertion of DNA fragments into plasmids follows the mechanism adopted by the bacteriophage λ to integrate its DNA into the chromosome of *E. coli*. This consist in a highly specific recombination in correspondence *att* sites,

whose length can vary from 25 bp to 242 bp. Two enzymes catalyse reactions, BP clonase, which recombines *attB*-sites with *attP*-sites and LR clonase which catalyses the reverse reaction. Recombination is very precise because each site is specifically recognized by clonases: *att*-sites are distinguished by the presence or absence of "arms" upstream or downstream of a 25 bp recognition region containing a 7 bp asymmetric overlap, in which the DNA is cleaved. Moreover, the recognition regions were mutated to generate new subtypes of *att*-sites able to recombine only with each other (for example *attB1* will recombine only with *attP1*, *attB2* with *attP2*, and so on) [52]. In this work, to ensure directionality of the cloning, 5' *attB1*-extensions have been used to recombine with *attP1*-sites on pDONR207 and 3' *attB2*-extensions to recombine with *attP2*-sites on pDONR207, which, upon recombination, will produce respectively *attL1/attR1*-sites and *attL2/attR2*-sites.

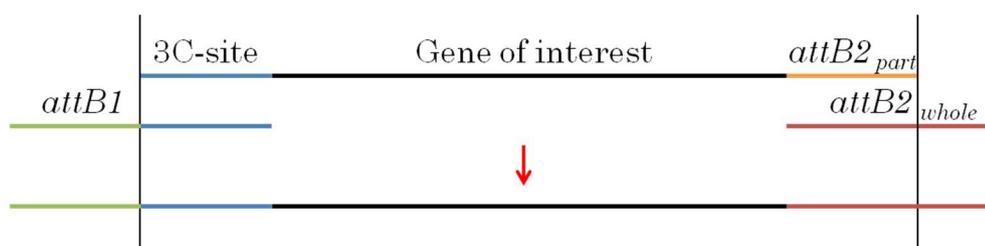
Gateway cloning was used to incorporate 15 genes, close relatives of eukaryote MINPPs, into the expression vector pDEST17. A 3C-protease cleavage site was added between the *attB1*-site and the genes of interest, to enable the removal of the 5' region containing N-term His-tag + *att*-site. The procedure includes in three main steps: generation of an "Entry clone", integration of the Entry clone into a donor vector (pDONR207) and transfer of the insert into a destination vector (pDEST17).

*Generation of an Entry clone.* The genes of interest were amplified with the two flanking extensions, *attB1* and *attB2*, respectively at 5' and 3', and a 3C-protease cleavage site was also included. The amplified genes will be incorporated into pDONR207 and then transferred to pDEST17 which will also add an N-terminal His-tag to the constructs. Because of the additional 3C-protease cleavage site, 5' extensions were too long to be added in one step, and therefore forward and reverse primers were designed for a two-step PCR:

- The first forward primer contains a region complementary to the gene and the 3C-protease cleavage site, the first reverse primer a region complementary to the gene and part of the *attB2* extension.



- The second forward primer contains the 3C-protease cleavage site and the whole *attB1* extension, the reverse primer the whole *attB2* site.



Sequences of the primers are reported in **Supplemental information 8.2**.

The genes to be cloned were codon optimized for *E. coli* to simplify cloning (averaged GC content) and expression. Fifteen polymerase chain reactions run in parallel with positive and negative controls for the first round of amplification. Reactions started with an initial denaturation at 98 °C, for 3 min, followed by 30 cycles of: denaturation (98 °C for 15 sec), primers annealing (50 °C, for 30 sec), extension (72 °C, for 45 sec). A final extension step at 72 °C for 10 min was included before cooling to 4°C. Reactions were set up in master-mix to contain 1x HF buffer (which provides 1.5 mM of MgCl<sub>2</sub>), 0.2 mM dNTPs, 0.5 μM primers mixture, 200 pg/μL plasmids and 0.02 U/μL Phusion DNA polymerase. The two rounds of amplification were set up using the same reagent mixture but with different primers and templates. The templates for the first PCR were the plasmids containing the codon optimized genes. For the second PCR, 1 μL from the previous round of amplification was used. The temperature program of the second round of PCR also differed from the previous: an initial denaturation (98 °C, for 3 min) was followed by 5 cycles of denaturation (98 °C, for 15 sec), annealing (45 °C, for 30 sec), extension (72 °C, for 1 min) and 15 cycles of denaturation (98 °C, for 15 sec), annealing (55 °C, for 30 sec), extension (72 °C, for 1 min) before a final extension step (72 °C, for 10 min) and cooling to 4 °C.

*Integration of the fragment into a "Donor vector".* The fragment was inserted into pDONR207 (which confers gentamycin resistance) through recombination catalysed by BP clonase in correspondence with the *attB1* x *attP1* and *attB2* x *attP1* sites. During the process *attB* sites were modified to *attL1* and *attL2*. A typical BP reaction contains 1 μL of donor plasmid pDONR207 (100 ng/μL), 1 μL of the PCR for the amplification of the fragment of interest with *attB* extensions, 1μL of BP clonase, 2 μL of TE buffer. Reactions were set up, vortexed and incubated o/n at 25 °C before inactivation of BP clonase by the addition of 0.5 μL Proteinase K, at 37 °C for 10 min. 2 μL of the resulting mixture were transformed into 20 μL Stellar competent cells, plated on selective agar (20 μg/μL Gentamycin), and incubated o/n at 37 °C. Positive

insertion were screened by colony PCR (see **Chapter 5.2**) and confirmed by sequencing.

*Transfer of the fragment into a "Destination vector"*. Recombination between "Donor vector" and "Destination Vector" catalysed by LR clonase allows the integration of the fragment of interest into the expression vector (pDEST17), in correspondence with *attL* x *attR* sites. LR clonase catalyses the reverse reaction of BP clonase, for this reason after LR recombination *attL* sites return to their original sequence (we can call them again *attR*). A typical LR reaction contains 1  $\mu$ L of destination plasmid pDEST17 (100 ng/ $\mu$ L), 1  $\mu$ L of donor plasmid (100 ng/ $\mu$ L), 0.5  $\mu$ L of LR clonase, 2  $\mu$ L of TE buffer. Reactions were set up, vortexed and incubated o/n, at 25°C, before inactivation of LR clonase by the addition of 0.5  $\mu$ L Proteinase K, 37 °C, for 10 min. 2  $\mu$ L of reaction were transformed into 25  $\mu$ L Stellar competent cells, plated on selective agar (20  $\mu$ g/ $\mu$ L Gentamycin) and incubated o/n at 37 °C. Positive insertion was confirmed by NruI digestions (**Chapter 5.3**).

## 2.5. Cloning quality checks

After cloning, regardless of the technique used, there is the need to confirm the correct insertions of a fragment of interest. Many approaches can be used, though final confirmation usually comes with DNA sequencing. This section reports the quality check methods used in this project to search for of the correct integration of genes of interest in expression vectors.

### 2.5.1. Cell lysis and plasmid length detection on agarose gel

This method allows the investigator to pick a single colony, to lyse it and to separate its DNA content on agarose gel (0.8 % (w/v)) by size. A colony was resuspended in 10  $\mu$ L LB media. 2  $\mu$ L were inoculated on a master plate and the rest was added to 30  $\mu$ L lyses buffer (10 % sucrose w/v, 100 mM NaOH, 60 mM KCl, 5 mM EDTA, 0.25 % SDS, 0.05 % Bromophenol Blue) warmed to 37 °C and incubated at 42 °C for 5 min. The solution was placed on ice for 5 min, spun for 10 min at maximum speed, and the supernatant was loaded into the agarose gel. Gels run in TAE buffer, at an electric potential difference of 120 V, for 30 min.

### 2.5.2. Colony PCR

Colony PCR was used to test correct integration of inserts after Gateway recombination (see Paragraph 3.2). 5 to 10 single colonies per construct were picked, resuspended in 5  $\mu$ L LB, and used as follows: 1  $\mu$ L as template for a colony PCR, 2  $\mu$ L were injected onto a master plate, the remaining 2  $\mu$ L (only for selected colonies) were inoculated in a 24 well plate (sealed with Airpore tape sheet) containing 5 mL LB per well, for o/n cultures at 37 °C, prior plasmids mini-prep. GoTaq® DNA Polymerase (Promega) was used to amplify DNA in 1x Green GoTaq® Reaction Buffer (Promega): a solution ready-to-load onto agarose gel which contains two dyes (blue and yellow) that separate during electrophoresis to monitor migration progress. Reactions contained 0.5 mM dNTPs, 2.5 mM MgCl<sub>2</sub>, 0.5  $\mu$ M primers mixture, 1  $\mu$ L colony suspension (before use, 1  $\mu$ L cells were diluted 1:5, boiled at 98 °C for 10 min and spun for 5 min) and 0.02 U/ $\mu$ L GoTaq® DNA polymerase. Reactions started with an initial denaturation step at 98 °C, for 7 min, followed by 30 cycles of: denaturation (98 °C, for 30 sec), primers annealing (50 °C, for 30 sec), extension (72 °C, for 2 min). A final extension step at 72 °C, for 10 min, was included before cooling to 4 °C. Amplified DNA run on 1 % agarose gels for inspection. One plasmid per construct was sent for sequencing.

### 2.5.3. DNA digestion

DNA digestion is the method used to confirm *appA* cloning. The gene is present in the genome of the host strain, therefore, colony PCR is not recommended, because it would always give an amplification, independently from cloning results.

The restriction enzyme chosen was NruI. This enzyme cleaves the DNA in the sequence 5` TCG | GCA 3`, generating blunt end fragments. Both plasmids pOPINA/B and *appA* contain cleavage sites for NruI. One site is on the plasmids and two sites are on the gene. Cleavage of a correct insert would generate three fragments of 241 bp, 2148 bp and 4104 bp. Instead, empty vectors would only be linearized. Cleavage reaction of the empty vector could be used either as a positive control of the digestion efficacy or as a negative control of the cloning, representing the plasmid without insert. A typical digestion contained 0.5 U/ $\mu$ L NruI, 20 ng/ $\mu$ L plasmids, 1x Buffer 3.1 (NEB – 100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl<sub>2</sub>, 100  $\mu$ g/ml BSA, pH 7.9 @25 °C) in a volume of 20  $\mu$ L. Reactions were incubated at 37 °C, for 1h, before inspection by electrophoresis on 1 % agarose gel.

NruI was also chosen to confirm the insertion of the 15 bacterial phytases cloned in pDEST17. Reactions were set up as before. pDEST contains one restriction site recognised by NruI at position 5750 bp. As the majority of the 15 target phytases genes do not contain cleavage sites, their plasmids will only be linearized by the enzyme. Two phytases differed in this respect: phytase 7 (NruI cuts at position 1420 bp and 5358 bp of the final plasmid generating 2 fragments of 3938 bp and 2024 bp), phytase 15 (NruI cuts at position 732 bp and 5370 bp of the final plasmid generating 2 fragments of 1336 bp and 4638 bp).

#### 2.5.4. Sequencing

Good quality sequencing reads are the most reliable proof of a successful cloning/mutagenesis, as they provide information of the DNA sequence itself. DNA samples were prepared and sent for automated Sanger sequencing by Eurofins Genomics. Results were then compared with the expected sequences to highlight mismatches.

## 2.6. Methods of enzymes purification

*Lysis of bacterial cells.* The preliminary step of most purifications described in this thesis was mechanical cell lysis to allow the collection of the soluble fraction in which proteins were suspended. Cell lysis was carried out by use of either a French Press (3-4 cycles, 16000 psi) or a cell disruptor (2 cycles – 30000 psi). The soluble and insoluble fractions were then separated by centrifugation at 40000 xg, at 4 °C, for 1 hour.

*Immobilized metal affinity chromatography – first cycle.* The first step of the purification of His-tagged proteins was by IMAC, a procedure based on the high-selectivity binding between a protein carrying a 6His-tag and a nickel-NTA resin. The experiment can be organised in four steps: 1) resin equilibration, 2) sample application, 3) resin wash, 4) protein elution. In these studies, HiTrap IMAC HP (1 mL, GE Healthcare) or Superflow Cartridge columns (5mL, Quiagen) were used. The first step consisted in the wash of the metal-affinity resin with a buffered solution containing low concentrations of imidazole, a molecule able to interact with Ni<sup>2+</sup> with high affinity. This allowed the removal of any protein non-specifically bound to the Ni<sup>2+</sup>-NTA resin (columns were used multiple times) prior sample application and to

avoid further non-specific binding during the first stages of sample application. This buffer is referred to as Buffer A.

The resin was now ready for protein binding. The soluble fraction obtained from cell lysis was applied to the column at a flow rate of 0.4-1 mL/min. Untagged proteins should pass through, the 6His-tagged POI should bind to the resin. Normally many proteins can interact with Nickel in this step. For this reason, the resin was washed again with Buffer A, to detach proteins which are weakly bound until stable UV signal was obtained.

The final step was elution of the POI. A gradient of increasing concentration of imidazole (0.0 - 0.5/1 M) was used to displace proteins according to binding affinity. This high imidazole buffer is referred to as Buffer B. The composition of Buffer A and Buffer B used for the purification of POIs is reported in **Table 2.6.1**.

**Table 2.6.1. Compositions of buffers used in IMAC.**

Buffer A and Buffer B used for the purification by 6His/Ni-NTA affinity chromatography of the heterologous proteins overexpressed in this work.

Protein	Buffer A	BufferB
<i>B</i> MINPP	50 mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.8, 300 mM NaCl, 20 mM imidazole	50 mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.8, 300 mM NaCl, 500 mM imidazole
AppA periplasmic	50 mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.5, 300 mM NaCl, 20 mM imidazole	50 mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.5, 300 mM NaCl, 500 mM imidazole
AppAc cytoplasmic	20 mM Tris pH8.0, 300mM NaCl, 10mM imidazole	20 mM Tris pH8.0, 300mM NaCl, 1M imidazole
p1-p16 MINPPs	50mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.5, 300 mM NaCl, 5 mM imidazole, 0.1% Triton, 10% glycerol	50mM NaH <sub>2</sub> PO <sub>4</sub> pH 7.5, 300 mM NaCl, 500mM imidazole
3C-precision protease	50 mM Tris-HCl pH 7.4, 500mM NaCl, 20mM imidazole	50 mM Tris-HCl pH 7.4, 500mM NaCl, 500mM imidazole

### 2.6.1. Purification of His-tagged proteins

*3C-protease cleavage of the His-tag and dialysis.* When proteins were purified for use in crystallization trials, the His-tag was cleaved to avoid its interference in crystal packing. The protein used for the removal was a His-tagged 3C-Precision Protease. This protein cannot auto-proteolyze its tag, therefore protease was removed from the sample by means of a second IMAC step. The cleavage reaction was carried out during o/n dialysis at 4 °C, in 50 mM Tris-HCl pH 7.4, 500 mM NaCl, 20 mM imidazole, in a Bio Design Dialysis Tubing (D10), 8 kDa cut-off. The enzyme was added at a 1:40 ratio to the total purified POI. In this step, proteins with a tendency to precipitate were diluted to reduce protein-protein interactions.

*Second immobilized metal affinity chromatography.* An additional IMAC step was carried out to separate cleaved proteins from the un-cleaved ones and the 3C-protease. The flow through of sample application contained the POI. Buffers used in this procedure are optimised for the purification of 3C-protase. They are described in **Table 2.6.1.**

*Gel filtration.* This chromatographic method divides proteins according to their size. A concentrated protein sample was applied to an equilibrated column packed with inert porous beads which accommodates small particles more easily than large ones, making their transit through the column longer and strictly dependent on their size and molecular weight. Typically, a GF column (HiLoad 15/60 Superdex 75 gel filtration column) was equilibrated applying 3 CV at a flow rate 0.4 mL/min. Following this, the sample was concentrated (spin concentrators were used with a MWCO of 10 or 30 kDa depending on the size of the POI) prior injection to the column. Separation occurred over a volume of 120 mL. The buffers used for each protein are listed in **Table 2.6.1.1.**

**Table 2.6.1.1. Compositions of buffers used in gel filtration.**

Buffer used for the separation of heterologous proteins purified by gel filtration. Proteins molecular weight (MW) and isoelectric point (pI) are reported.

Protein	MW (kDa)	pI	Buffer
BIMINPP-cleaved	55.93	4.95	20 mM HEPES, 150 mM NaCl pH 7.4
AppA untagged periplasmic	44.66	5.95	200 mM NaAcetate, 150 mM NaCl pH 4.5
AppA-cleaved cytoplasmic	44.82	5.95	200 mM NaAcetate, 150 mM NaCl pH 4.5 or 50 mM Tris-HCl pH 7.5
P8-cleaved	47.52	6.19	20 mM HEPES, 150 mM NaCl pH 7.4
P10-cleaved	50.23	8.54	200 mM NaAcetate, 150 mM NaCl pH 4.5
P15-cleaved	46.81	6.33	20 mM HEPES, 150 mM NaCl pH 7.4

## 2.6.2. Purification of un-tagged periplasmic *EcAppA*

*Extraction of the periplasm.* Following protein over-expression, cell pellets were collected by centrifugation at 5500 xg, at 4 °C, for 20 min. To extract the periplasm and to liberate the *E. coli* spheroplasts, the cell wall was broken. This

procedure involved the creation of a gradient concentration of sucrose across the outer membrane. Then, the cell wall was broken by osmotic pressure and periplasm was released.

Cell pellets, weighed carefully, were resuspended in 4 mL ice-cold sucrose solution (20 mM Tris-HCl pH 8.0, 25 % w/v sucrose, 5 mM EDTA) per gram of cells. After 15 min incubation on ice, the excess-sucrose solution was removed by centrifugation (8500 xg, 4°C, 20 min) and collected as control - cells should be maintained intact at this stage. Pellet were used in subsequent steps. They were dissolved in 4-5 mL of ice-cold sucrose-free solution (5 mM MgCl<sub>2</sub>, 1/2 tablet protease inhibitor cocktail (Roche), 40 µL of Lysozyme solution (15 mg/mL)) per gram of cells and incubated on ice for 30 min. The resulting solution contained the released periplasm and was collected by centrifugation at 8500 xg, for 20 min, 4 °C.

*Salting-out of protein in ammonium sulphate.* Differences in solubility of proteins in high ionic strength solutions were used to separate the enzyme of interest. Ammonium sulphate, the salting-out agent, was slowly added to the periplasmic solution in a concentration equal to 40 % saturation, at 4 °C, with stirring (magnetic stirring), and incubated for 30 min [53]. Precipitated proteins were separated by centrifugation (20000 xg, 4 °C, for 30 min). The procedure was then repeated by increasing the concentration of ammonium sulphate in the soluble fraction to 50 %, 60 %, 70 % and 80 % of saturation. Samples of the supernatants and the precipitated proteins were stored for further analysis.

*Solubilisation in glycine buffer and desalting.* *EcAppA* has the quality of being stable and soluble at very low pH. Taking advantage of this characteristic, the protein was resolubilized from pellets in 30 mL of 0.5 M glycine-Cl, pH 1.8. Pellets were dissolved by pipetting and incubated at 37 °C, 180 rpm, for 30 min, before a further separation of soluble and insoluble fractions by centrifugation at 4 °C, 5000 xg, for 20 min. To remove ammonium sulphate and to prepare the sample for cation-exchange chromatography, the protein solution was desalted, and buffer exchanged using a PD-10 column in 50 mM acetic acid, pH 4.5.

*Cation-exchange chromatography.* Ion Exchange Chromatography (IEC) is a method for the separation of protein by their charge. *EcAppA* is a protein very active and stable at low pH, thus, these conditions were chosen for IEC prior to activity assays. The isoelectric point of *EcAppA* is 5.95: in buffers of lower pH, the protein

would be positively charged, therefore cation exchange chromatography was chosen. *EcAppA* was purified using a 1 mL SP-HP column (GE Healthcare) equilibrated in 0.5 M acetic acid, pH 4.5. The sample was applied at a flow rate of 0.7 mL/min, and circulating through the column for 1 h, at 4 °C. Elution was performed by applying a gradient 0 – 1 M NaCl, over 30 CV. To further purify the sample, gel filtration was carried out in 0.2 M NaAcetate pH 4.5, 0.15 M NaCl, following the procedure explained in **Paragraph 2.6.1**.

## 2.7. Biochemical and biophysical techniques of phytases characterisation

### 2.7.1. Phosphatase activity assay - monitoring pNPP hydrolysis

In this colorimetric assay, the substrate is *para*-nitrophenyl phosphate (*p*NPP). The method is designed to make use of the ability of phosphatases to catalyse the reaction of hydrolysis of *p*NPP to generate *p*-nitrophenol. The latter is transformed in alkaline conditions to *p*-nitrophenolate, a chromogenic molecule of yellow colour that has an absorbance maximum at 405 nm. pNPP ( $\geq 95\%$  Premium Quality, Sigma) 10 mM concentration was used in all the assays. Enzyme concentrations and buffers varied depending on the type of assay and the enzyme involved. Reactions were incubated at 37 °C and stopped with 1 M NaOH. Absorbance was measured at 405 nm, after 10 min. Control reactions of buffer only, substrate only and enzyme only were set up in parallel.

### 2.7.2. Phytase activity assay – determination of released phosphate by molybdenum blue method

This assay allows the determination of the free phosphate released by hydrolysis of IP<sub>6</sub> by molybdenum blue reaction. The molybdenum blue reagent is an aqueous solution containing molybdenum (VI), a strong acid (in this case sulphuric acid) and a reductant. The reaction which leads to the production of molybdenum blue, occurs in two steps: the formation of 12-molybdophosphoric acid (favoured at pH 0-1), a Keggin structure around the orthophosphate anion (the PO<sub>4</sub><sup>2-</sup> group is caged inside a 12 MO<sub>6</sub> units, which are linked to one another by bridging oxygens) and the reduction of this heteropoly acid to form molybdenum blue [54]. The absorbance of

molybdenum blue is measured at 700 nM and it is proportional to the Pi concentration. A typical calibration curve shows that the assay is in the linear range from 10  $\mu$ M to 2.5 mM Pi.

Phytic acid dipotassium salt ( $\geq 95\%$  Premium Quality, Sigma) was used as substrate. Enzyme concentrations and buffers vary depending on the type of assay and the enzyme involved. Reactions of 50/100  $\mu$ L were stopped after 15 min by the addition of equal amounts of a freshly prepared solution made of 4 parts of reagent A (12 mM ammonium molybdate tetrahydrate, 5.4% saturated sulfuric acid (98%)) and 1 part of reagent B (0.4 M iron(II)sulphate heptahydrate plus a few drops of saturated sulfuric acid (98%)). Absorbance was measured at 700 nm, after 30 min. Control reactions of buffer only, substrate only and enzymes only were set up simultaneously as well as a calibration curve of increasing concentration of Pi.

### **2.7.3. HPLC separation of inositol polyphosphates**

1 mM D-*myo*-inositol 1,2,3,4,5,6-hexakisphosphate, dodecasodium salt (*Zea mays*, Merck, 99% pure) was used as a substrate. Enzyme concentrations and buffers varied between assays. Reactions were stopped at the desired time intervals by boiling samples at 100 °C, for 10 min. Samples were diluted 5x before injection. The IPs standards used were generated by the hydrolysis of IP<sub>6</sub> in 1 M HCl, 120 °C for 24 hours. The HPLC system consisted of a first pump for sample injection (Jasco PU-2089 I Plus – Quaternary inert Pump) connected in series to two CarboPAC PA200 columns (3x50mm, 3x250mm) in which IPs species were efficiently separated (enantiomers however cannot be resolved) before reaching a chamber in which they were chaotically mixed with a reagent (0.1 % Fe(NO<sub>3</sub>)<sub>2</sub>, 2% HClO<sub>4</sub>), which is injected by a second pump (Jasco PU-1585 Intelligent HPLC Pump). This allows UV absorbance detection at 290 nm (range 1.28 nm, Jasco UV 1575 Intelligent UV/Vis detector – 16  $\mu$ L cell). Samples were separated in a methane sulfonic acid gradient (0 - 0.6 M), flow rate 0.4 mL/min, with water as a counter eluent, reagents were injected at a flow rate of 0.2 mL/min. The total run time for each sample was 50 min: 25 min of gradient, 14 min of 0.6 M methane sulfonic acid, 11 min of water. The peaks area was calculated by integration using the software provided by Jasco (ChromNAV, version 1.19.01).

#### 2.7.4. Differential scanning calorimetry

Differential scanning calorimetry is a technique that can be used to study folding/unfolding transitions of macromolecules [55]. Protein folding is a spontaneous process, generally characterized by a small Gibbs free energy change. It starts during translation and is determined by the amino acid sequence of a protein (Anfinsen's dogma). In aqueous solutions, folding is driven mainly by the hydrophobic effect that serves to decrease the entropy of the surrounding solvent. Hydrogen bonds, electrostatic interactions, van der Waals (s dispersion forces) and S-S bridges also contribute significantly to protein stability and domains movements/vibrations.

To unfold a protein does not only mean to break all the intramolecular forces listed above: the transition is chaotropic, it leads to an increase in the hydration shell disorder, hence an increase of the system entropy. For these reasons the unfolding is energy consuming, and in differential scanning calorimetry energy is provided in the form of heat. Capillary DSC instruments are made of two cells: a reference cell and a sample cell. Solutions are carefully injected to avoid air bubble formation and a constant pressure is applied. Cells are gradually heated, and their temperature is measured at small intervals throughout the process. As proteins require energy to unfold, the additional energy (cal/°C) provided by the heater to the sample cell, to maintain it at an equal temperature to the reference cell, is registered, giving rise to a peak of heat absorbance during unfolding. After subtraction of the baseline it is possible to determine the melting temperature which corresponds to the maximum of heat-capacity. Various parameters can be extrapolated from this experiment: the enthalpy of the transition is obtained by integrating the area under the peak, the difference between the heat-capacity of the native and unfolded states gives an estimate of the temperature dependence of the entropy and enthalpy functions which can be considered an estimation of the temperature dependence of protein stability. The sharpness of the peak of heat absorption correspond to the Van't Hoff enthalpy. A ratio between the enthalpy of transition and the Van't Hoff enthalpy equal to 1 is characteristic of two-state transitions in which there are no partially folded intermediates. A ratio that differs from unit suggests that the transition is more complex [56].

*Protocol.* A VP-DSC (Microcal Inc.) was used for all calorimetry experiments in this work. Initially, 20 buffer readings were taken to build the thermal history of the instrument. The last runs were used as baseline in subsequent data analysis. The temperature gradient was set to 10-110 °C with a scans rate of 200 °C/h which assured

high sensitivity without excessive sharpness of peaks. A pre-scan of 5 min was added at the beginning of each read. At the end of a temperature ramp, following which the sample cell was cooled to 20 °C, the lid was opened to be cleaned before sample loading. The cleaning procedure included rounds of air (1 min), 10 mL dH<sub>2</sub>O interspersed in air, 2 mL 20 % Decon solution, 25 mL buffer. 350 mL of sample solution were loaded. The pressure sensor was removed, and the lid was closed until a pressure of at least 30 psi was established. The procedure needed to be completed at least 5 to 10 min before the start of the pre-scan to allow the re-equilibration of the cells.

## 2.8. X-ray crystallography

X-ray crystallography is a technique that allows the modelling of a three-dimensional structure of a protein. Information of the position of atoms are reconstructed from diffraction patterns collected through monochromatic electromagnetic radiation of protein crystals at the wavelength of X-ray.

### 2.8.1. Protein crystals and their growth

Protein crystals are highly organised heterogeneous aggregates. They are variable in size, though an average optimum at which to aim in crystallization trials could be set around 200  $\mu\text{m}^3$ . The solvent content of the crystal lattice can vary between 26-90 % [57]. Solvent can take up ordered and unordered positions in crystals: electron densities of ordered molecules are identifiable through X-ray diffraction experiments, instead the contributions of unordered molecules account for background noise. The significant amount of solvent content allows small compounds, such as substrate analogues, inhibitors, cofactors, etc, to diffuse into the crystal lattice, thus allowing the study of protein-ligand interaction by X-ray crystallography. However, high hydration, is correlated with a decrease in resolution of diffraction images. Weak forces such as Van de Waals, salt bridges and hydrogen bonds (lattice energy 5-10 kcal/mol), stabilise surface protein-protein interactions in crystals packing. As a result, protein crystals are very fragile and must be handled with care, on the other end this also assures that folding is mostly unaltered, for example, enzymes are generally still active in a crystalline environment.

The fundamental unit of a crystal lattice is called the asymmetric unit and is organised by symmetry operators into a unit cell. Translations of the unit cell into the

three-dimensional space generates the crystal lattice. Unit cells can be classified into space groups based on their symmetry operators. 230 space groups have been identified in nature, but only 65 are compatible with biological macromolecule because of their chirality. Nevertheless, looking at RCSB PDB statistics we can observe that almost half of them fall into three space groups: orthorhombic  $P2_12_12_1$  and monoclinic  $P2_1$  and  $C2$  (<http://www.rcsb.org/>, Berman, JohnWestbrook [58]). Crystal lattice can also be described by Miller indices  $h, k, l$ , that identify the family of parallel planes which cut the axis  $a, b, c$ , of the unit cell at positions  $a/h, b/k, c/l$ .

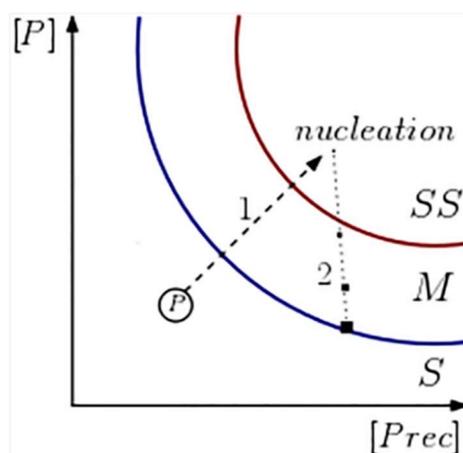
Once the dimensions of the unit cell and the symmetries are known, it is possible to calculate the probable number of molecules ( $Z$ ) contained in it, using the relationship:

$$VM = \frac{V_{cell}}{Z} * MW$$

where  $VM$  is the Matthews volume (it can vary between 1.7 and 3.5 Å<sup>3</sup>/Da),  $V_{cell}$  is the volume of the unit cell and  $MW$  is the molecular weight of the protein [59].

Crystal formation is reached by decreasing protein solubility. To maximize the probability of success the purity of protein solutions should be as high as possible (>98 % (w/w)). In this work the method used for protein crystallization is vapour diffusion. Vapour diffusion experiments are set up as closed systems containing a drop of protein solution, mixed with precipitants dispersed onto an elevated position and an excess of the same precipitant solution (typical volume 100x higher and concentration usually 2x higher than the drop) in a well underneath. The drop solution and well solution are in contact only through air.

To compensate for the high concentration of precipitant in the well solution, the solvent in the drops tends to evaporate and to diffuse into the precipitant solution of the well. The result is the gradual increasing, over time, of protein concentration and precipitant in the drop (**Figure 2.8.1.1 - 1**). When protein concentration reaches its solubility limits, proteins enter a metastable phase in which nucleation becomes more thermodynamically favoured. In fact, the high energy barrier, which must be overcome before nucleation, is only exceeded when the protein solution is supersaturated. However, if supersaturation is too high, there will be the formation of too many nuclei, which will not grow into protein crystals of the required size for X-ray diffraction [60]. When crystals start to grow, the concentration of free protein in solution decreases. When the concentration of protein reaches the solubility limit, crystals growth stops (**Figure 2.8.1.1. - 2**).



**Figure 2.8.1.1. graph representing the crystallization process.**

Proteins are soluble in a protein dependent range of concentrations (S area), outside of which, they precipitate. In the figure, solubility limit is represented by a blue curve. Typically, in a vapour diffusion experiment, protein concentration (P) and precipitant concentration ([Prec]) gradually increase over time (1). When the solubility limit is reached, the protein solution becomes metastable (M area). In this phase, protein can start to nucleate, but the probability of nucleation increases the more they approach supersaturation (SS area). The supersaturation curve is depicted by a red line. If the degree of supersaturation rises too far proteins can precipitate as amorphous aggregates. If nucleation begins, the protein concentration in solution decrease due to depletion, and crystals grow in size (2). The growth of crystals stops when protein concentration in the drop reaches the solubility curve. It is important that the number of nuclei formed is neither too high nor too small, the ideal situation is reached in the metastable zone close to the supersaturation curve.

The factors which influence protein solubility are pH, ionic strength, temperature, dielectric constant of the solvent, and the ratio between the concentration of protein and precipitant. pH determines the distribution of surface charges which stabilize protein-protein interaction and crystal packing. Ionic strength favours crystallization when low (a process called salting-in) or high (salting-out). At low ionic strength, protein solubility decreases as temperature decreases. At high ionic strength, protein solubility decreases as temperature increases. Crystallization can also be achieved by reducing the dielectric constant of the solution as in this condition the electrostatic fields generated by charged groups are less shielded by the solvent, favouring protein-protein interactions. Deionised water has a high dielectric constant (80) but can be reduced by the addition of organic solvents like glycerol, ethylene glycol or polyethylene glycols (PEGs). PEGs are commonly used not only for their ability to decrease the dielectric constant and to create hydrogen bonds with water, but also for their 'excluded volume' effect in liquid solution.

In the work described by this thesis, crystallization was attempted by varying all the parameters above in vapour diffusion experiments. To explore crystallization space pre-made 96-well commercial screens were used. They included Structure

Screens 1 and 2, JCSG-*plus*<sup>™</sup>, PACT *premier*<sup>™</sup>, MIDASplus<sup>™</sup> (Molecular Dimensions). To improve crystal quality and size: the identified conditions of crystallization were optimized in a matrix of varying percentage of each component, available crystals were used for seeding, the final buffer of the protein was exchanged to one that enhanced solubility.

Protein crystallization trials were set up either manually or by using the protein crystallization Robot OryxNano (Douglas Instruments), which allows the setup of sitting-drop, micro batch, high-throughput crystallization screening in 96 wells plates. It can also be used to set up drops containing crystal seeds (**Figure 2.8.1.2**)



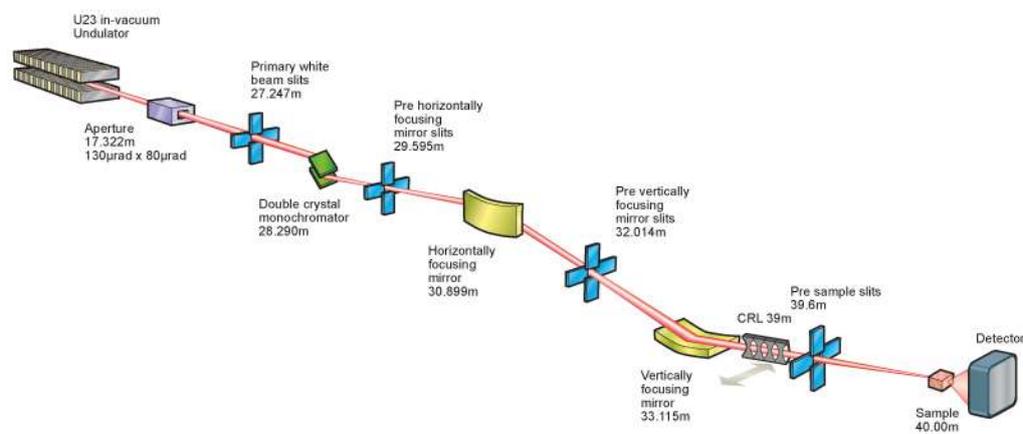
**Figure 2.8.1.2. Robot OryxNano, Douglas Instruments.**

## **2.8.2. X-ray diffraction experiments**

The electromagnetic waves used in macromolecular crystallography are in the range of X-rays. They are chosen because of the need of a wavelength  $\lambda \approx 1 \text{ \AA}$ , small enough to be able to discriminate between atoms placed at distances equal to the length of a covalent bond. Monochromatic beams are used in protein diffraction. They can be generated in-house; however, X-ray diffraction experiments are carried out mostly at synchrotron facilities which deliver high beam intensity (millions of times brighter than the light produced with conventional sources and 10 billion time brighter than the sun) with small deviation resulting in better quality of the spots (smaller, sharper) on diffraction images. High beam intensity enables operators to collect good quality data also when diffracting small crystals (e.g. diameter 10-20  $\mu\text{m}$ ) or crystals with big unit cell and in general allows to collect complete datasets with shorter exposure. On beamlines, constant cryogenic conditions are ensured. They help

to decrease radiation damage of the crystals caused by free radicals generated by the radiolysis of water, which can disrupt the crystal lattice and modify the proteins in it, reducing the quality of the data collected. Cryogenic condition also reduces the thermal motion of proteins improving the quality of the data collected. At synchrotron facilities wavelength can easily and accurately be changed (Diamond Light Source I03 range: 0.6 - 2.48 Å and I04 range: 0.69 - 2.066 Å) and this allows, for example, to collect anomalous data (SIRAS/MIRAS, SAD/MAD), but also it can be used to identify the nature of metals in proteins.

Monochromatic beams are generated directing the synchrotron light coming from an aperture in the storage ring through consecutive optical filters by selective absorption in double crystal monochromators and using focusing mirrors (**Figure 2.8.2.1**).



**Figure 2.8.2.1. configuration of beamlines I03 and I04 at Diamond Light Source synchrotron (Didcot, UK).**

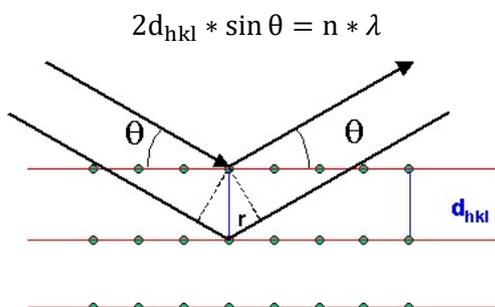
Diffraction data were collected on these beamlines.

Due to the high intensity of the synchrotron radiation, cryogenic temperatures of 100 K are used in the experiments to reduce crystal damage. This condition also allows to prevent the freezing of the solvent content in the form of crystalline ice, which could damage the crystal lattice or produce additional scattering. To protect crystals, before cooling to liquid nitrogen temperatures, they were soaked in solutions containing cryo-protectants such as glycerol, ethylene glycol or PEG 400, which favour the formation of a vitreous amorphous glass-like structure that envelopes the crystals.

In an X-ray diffraction experiments, crystals were recovered on loops from cryo-cooled pucks, vitrified and ready for data-collection. They were then manually aligned to the centre of the X-ray beam with the help of a single or multi-axis goniometer. Following this, 3 to 5 sample images were taken at 45° rotation along the  $\Omega$  axis, at long exposure (0.1-0.2 s) and medium resolution of 2 Å. These images were then automatically indexed by the ISPyB image processing software pipelines

iMosflm [61], EDNA [62] and XIA2 [63]. They provide information on probable space-group, unit cell dimensions and, after evaluation of crystal orientation, mosaicity, diffraction spots shape and signal-to-noise ratio. They also suggest data-collection strategies (data collection resolution, rotation range, oscillation width and exposure time) which can be adjusted by the operator as required.

Diffraction is the physical phenomenon produced when an incident radiation hit electrons associated with atoms in a crystal. Electrons absorb energy and then reemit it. Energy is reemitted in discrete directions defined by Bragg's law (**Figure 2.8.2.2**), which states that positive diffraction is the sum of the electromagnetic waves, reflected by the reticular planes of Miller indices (h, k, l), whose difference of optical path ( $2 d_{hkl} \sin\theta$ ) is equal or multiple of the wavelength ( $\lambda$ ).



**Figure 2.8.2.2. Graphic description of Bragg's law.**

Reticular planes are coloured in red and atoms by green spots. Distance between reticular planes ( $d_{hkl}$ ) is indicated by a blue line. Incident radiation of angles  $\theta$  is represented by black arrows. 'r' is equal to  $d_{hkl} \sin\theta$ . The extra path of the lower beam, whose length is twice 'r' correspond to a multiple of the wavelength  $\lambda$ .

The scattering of an atom will depend on the number of electrons possessed. Its scattering power is described by the atomic scattering factor  $f_j$ , resulting from the ratio between the amplitude of the wave scattered by the atom and the amplitude of the wave scattered by a single electron.

$$f_j(\theta, B) = f_j^0 e^{\left[ \frac{-B_j(\sin^2 \theta)}{\lambda^2} \right]}$$

The atomic scattering factor  $f_j^0$  decreases with increasing  $\theta$  because, as  $\theta$  increases, contributions between different regions of the electronic cloud tend to cancel. It also depends on the temperature factor,  $B_j$ , which describes the standard deviation of the atomic position due to temperature-induced vibrations.

The spots on a diffraction image represent the radiation emitted by a family of reticular planes with Miller indices h, k, l that are in the condition of positive reflection according to Bragg's law. Their intensity is described by the following equation:

$$I = k\lambda |\mathbf{F}_{hkl}|^2 \frac{V_{\text{cryst}}}{V_{\text{cell}}}$$

where  $k$  is a proportionality coefficient,  $\lambda$  is the wavelength of the incident radiation,  $V_{\text{cryst}}$  is the crystal volume,  $V_{\text{cell}}$  is the volume of the unit cell and  $\mathbf{F}_{hkl}$  is the structure factor. The structure factor can also be represented in the complex plane as a vector with magnitude  $|\mathbf{F}_{hkl}|$  and phase  $\alpha_{hkl}$  of the wave reflected by the family of reticular planes of indices  $(h, k, l)$ .

$$\mathbf{F}_{hkl} = |\mathbf{F}_{hkl}| e^{i\alpha_{hkl}}$$

The structure factor for each  $(h, k, l)$  is also the result of the scattering contribution of all atoms in the unit cell and it depends on the atom types and on their locations in the unit cell. For this reason, each single spot contains information on all the atoms in the unit cell:

$$\mathbf{F}_{hkl} = \sum N_j = \sum f_j e^{2\pi i(hx_j + ky_j + lz_j)}$$

the equation above sums the contribution to the structure factor of each  $j^{\text{th}}$  atom in the unit cell at positions  $(x_j, y_j, z_j)$ .

From a mathematical point of view, the structure factor is the Fourier transform of the electron density of the unit cell. Therefore, to determine the electron density of the protein  $\rho(x, y, z)$ , the inverse transform of the structure factor in each point of the unit cell needs to be calculated:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l \mathbf{F}_{hkl} e^{-2\pi i(hx + ky + lz)} = \frac{1}{V} \sum_{hkl} |\mathbf{F}_{hkl}| e^{i\alpha_{hkl}} e^{-2\pi i(hx + ky + lz)}$$

$V$  describes the volume of the unit cell. From a diffraction experiment we could measure the absolute values of the structure factors  $\mathbf{F}_{hkl}$ , but not the phases  $\alpha_{hkl}$ . However, they can be estimated by applying several methods, for example SIRAS/MIRAS, SAD/MAD or molecular replacement. The latter is the method used in this thesis (**Paragraph 2.8.4**).

### 2.8.3. X-ray data processing

X-ray images were automatically integrated and scaled [64] by DLS servers using various pipelines: FAST DP [65], DIALS [66], XIA2 3d/3dii [63], multi-XIA2, autoPROC [67].

There is a four-step process leading to the final objective – the transformation of spot intensities into structure factors amplitudes:

Data reduction: indexing (a number of sample images taken at distant angles are selected to be used in spots finding, the 2D spots are transformed in 3D scattering vectors by using the Ewald sphere construction, cell dimensions are identified in real space or in reciprocal space, cell reduction is performed for the identification of which of the 44 Bravais lattices match at best the unit cell), detector and crystal parameter refinement (positional refinement, post refinement), integration (collection of spots intensities by discriminating them from the background noise and correction).

Symmetry check: identification of the Laue group and space group prediction

Scaling and merging: merging partial reflection and symmetry related reflections to form a complete set of reflections, then scaling to reduce experimental errors due to instrumentation or crystal defects.

Truncation: analyse intensity distributions (e.g. twinning analyses, Wilson statistics), and provide structure factors  $|F_{hkl}|$ .

#### **2.8.4. Methods for the solution of protein structures and refinement**

The approach employed to solve the phase problem was molecular replacement. In this method, the phase is found by making use of the coordinates of a homologous protein that it is predicted to be structurally similar to the target. This approach is the favourite when a suitable structural homolog is available because, if it works, it is the quickest and least laborious way to find monomers orientations in the unit cell.

Molecular replacement is performed based on previously collected information. Information on the Laue point group and probable space group, unit cell dimension, protein sequence and MW are combined to predict in advance the number of monomers per unit cell by applying the Matthews equation [59]. The solvent content generally correlates with the resolution of diffraction experiments, the higher the solvent content, the greater the disorder. Spots intensities and amplitudes of the structure factors  $F_{hkl}$  are used in the determination of phase and model refinement. The latter is guided by geometrical restraints borrowed from the literature, in particular from small molecules crystallography. They include information on bond lengths, bond angles, dihedral angles, chirality, planarity, Ramachandran phi and psi, rotamers, C-beta deviations, side-chains restraints, clashes, etc.

MR procedures often utilize a known homologous structure to build search models. What are the characteristics that potentially good models should show? They should allow to obtain a high signal to noise ratio as result of the search. Two factors are the main determinants: the root mean square deviation from the target structure should be the lowest possible and the scattering region that the model can cover should be the largest possible.

In evaluating the best structural ensemble to be used as a model a few principles can be followed. The higher the sequence similarity of the selected model, to the better [68]. However, this is not a general rule because, what matters is structure similarity, which could be influenced by flexible regions and by all conditions that leads to a conformational change (e.g. ligands/cofactors-binding). Rigid-body domain movements can be approached by fragmenting the structure and using each of the fragments for a separate search. This should be chosen as a second option though, as it decreases the coverage of the scattering region and results in a reduced signal-to-noise ratio during the search.

The quality of the model should be the best possible (e.g. to prefer structures with high resolution limit  $< 2\text{\AA}$ , low R-free/R-work, etc.). The higher the model resembles the protein conformation in the crystal, the closer its structural similarity might be to the target structure. The use of ensembles of superposed homologues structures could help in finding a phase solution when only low identity sequences are available [69]. The removal of flexible loops/regions of the model is another route to decrease model/target r.m.s.d.. Unfortunately, this method could result in a decreased coverage of the model and therefore in a reduced signal-to-noise ratio. Also, it is useful, when possible, to preserve areas involved in crystal packing because a complete protein surface, for the model, would allow phasing programs to be more accurate in scoring possible clashes in ensemble placement. Other methods includes the pruning of dissimilar amino acids sidechains in the model [70] (e.g. by using the CCP4 Chainsaw [71] or the Phenix Sculpt [72]) and the removal of areas with low sequence identity when secondary structure predictions differ from model to target.

Molecular replacement is a search procedure that requires the correct positioning of a structural model in the experimental unit cell. If the phase can be found, the model will, after steps of refinement, assume the structural characteristics of the target. The search is a six-dimensional problem as three rotation angles and three translations that define the correct orientation in the unit cell need to be identified.

Methods have been described, that allow a full six-dimensional search, however correct positioning is generally quite hard when handling a high symmetry space group, low resolution data, densely packed unit cells or elongated proteins [73]. Examples of six-dimensional search methods include systematic algorithms such as SOMoRe [74], MPI-FSEARCH [75] or the script by Sheriff, Klei [76] and stochastic procedures. The most recent of them is MR-REX [73], which employs replica-exchange Monte Carlo simulations, while other examples resort to evolutionary searches [77, 78] or reverse Monte Carlo minimisation [79]. Traditional algorithms divide the search into two steps, a 3n-dimensional rotation function and a 3n-dimensional translation function, to make it computationally more sustainable and time-saving. 6n-dimensional methods though, can be more accurate: because all parameters are known at the same time, observed and calculated structure factors can be deduced, and orientations scored directly. MR-REX is an example of a 6n-dimensional search program able to perform better than most common 2x3 methods when low-accuracy starting models are provided as input [73]. Although 6-dimensional search could give solution for simple space groups, the method has still limitations for large space groups containing multiple monomers per asymmetric unit.

Two-step approaches such as MOLREP [80] or Phaser [81] need to rely instead on approximations because structure factors cannot be calculated if orientations are not fully defined. According to the type of assumption they are based on, these approaches can be categorized as Patterson methods (the oldest) and likelihood methods. Also, algorithms that, having defined a sampling grid in the real space, evaluate each transformation, are referred as 'brute force' approaches. Instead, Fast Fourier Transform (FFT) methods allow the user to generate all values on the Fourier grid simultaneously [82, 83]. The latter are generally the quickest.

*Patterson methods.* Through X-ray data processing, the amplitudes of the structure factors  $\mathbf{F}_{\mathbf{hkl}}$  contributing to the equation describing electron density  $\rho(x, y, z)$  are calculated, but no information is available of the phases  $\alpha_{\mathbf{hkl}}$ .

$$\rho(x, y, z) = \frac{1}{V} \sum_{\mathbf{hkl}} |\mathbf{F}_{\mathbf{hkl}}| e^{i\alpha_{\mathbf{hkl}}} e^{-2\pi i(hx+ky+lz)}$$

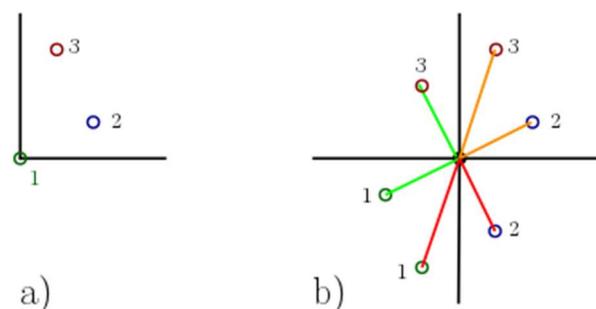
The Fourier transform of the squared structure factors amplitudes  $|\mathbf{F}_{\mathbf{hkl}}|^2$  with phases set to zero:

$$\rho^P(u, v, w) = \frac{1}{V} \sum_{\mathbf{h}} \sum_{\mathbf{k}} \sum_{\mathbf{l}} |\mathbf{F}_{\mathbf{hkl}}|^2 \cos 2\pi(hu + kv + lw)$$

is also named Patterson function and is used in traditional MR as a search method for the phase of a model. This equation can always be solved because it does

not include the phase variable  $\alpha_{hkl}$ . The Patterson function can be depicted like a map of vectors describing all inter-atomic distances in the macromolecule/s, drawn from the origin of a three-dimensional plot (phase  $\alpha_{hkl}$  is set to 0) with the weight of the vectors directly proportional to the scattering of the atoms (**Figure 2.8.4.1**). In fact, the Patterson function  $\rho^P$  is equivalent to the convolution product ( $\times$ ) of the Fourier transform of the structure factors amplitudes  $FT(\mathbf{F}_{hkl})$  and the Fourier transform of the structure factors amplitudes complex conjugate  $FT(\mathbf{F}_{hkl}^*)$ :

$$\rho^P(u, v, w) = FT(\mathbf{F}_{hkl}) \times FT(\mathbf{F}_{hkl}^*)$$



**Figure 2.8.4.1. Example of Patterson calculation in 2D space.**

a) 3 atoms molecule with atoms position 1, 2, 3. b) Patterson map. It represents all the interatomic vectors that can be drawn from starting from each atom and translated to the origin.

By solving the Patterson function, it is possible to draw  $N^2$  peaks for  $N$  atoms, of which,  $N$  peaks will be positioned at the origin of the map, and  $N(N-1)$  will be positioned in the surrounding Patterson space (**Figure 8**). Patterson vectors can be grouped as self-vectors, representing intra-molecular distances, and cross vectors representing the inter-molecular distances (they are generated by crystallographic and non-crystallographic symmetries).

The Patterson function of a small molecule can be solved manually but this is not the case for macromolecules because the number of inter-atomic vectors is too large. In molecular replacement, Patterson functions are calculated for the experimental data and for a model and their r.m.s.d. is minimized to find the orientation of the model in the experimental unit cell.

The search procedure can be divided into four steps. The first is the generation of the molecular transform for the search model (Fourier transformation and Patterson map synthesis of the model). The model is placed in a triclinic cell (P1) large enough to leave a gap between adjacent sets of vector clusters around the origin, to be able to exclude any cross-vector. The second step consists in the generation of the molecular transform for the crystal. This Patterson map will be more complicated, it will include: multiple sets of self-vectors rotated according to the crystal symmetry, overlaps between self-vectors from different origins, cross-vectors deriving from inter-

molecular interactions at unknown translations. To reduce the complexity of the map only a partial area close to the origin will be used in rotation. This region would potentially include the highest number of self-vectors with the lowest number of cross-vectors; the balance is user-defined. The third step uses the Patterson rotation function. The model Patterson is rotated against the experimental Patterson and their correlation coefficient is evaluated at each angle to find which is the one that gives the highest score (a peak in the Patterson map). Experimental data and model data can be evaluated in the Patterson space or by Fourier transform in reciprocal space. The fourth step uses the Patterson translation function. The model structure factors amplitudes for several translation vectors  $\mathbf{t}$  (characterized by the orientation found in the rotation step) are calculated and compared with the experimental data. Cross-vectors are sensitive to translations, self-vectors aren't, also, translation and crystallographic symmetry are strictly correlated. For example, in P1 cells there is no crystallographic symmetry, in this cell no cross-vectors modification can be detected upon translation.

When there are multiple monomers in a unit cell, they are positioned one at a time, leaving fixed the ones for which a solution has been already identified. The Patterson function is not commonly used nowadays, but this method was the first to be developed. In a mathematical comparison with likelihood methods, the Patterson function could be considered as an approximation of the full rotation likelihood function because it is equal to the first term in the Taylor series expansion of the likelihood rotation function. However, in contrast with likelihood methods, in the Patterson function no error estimation is carried out [84, 85].

*Maximum likelihood methods.* In maximum likelihood methods, the evaluation of the best rotation/translation is based on likelihood scores (-log-likelihood). Likelihood is the probability of the observed data, given a specific model orientation and position. The ratio between likelihoods is a measure of confidence and a way to compare two likelihoods, though a ratio produces very small numbers, not suitable to be efficiently stored in a calculator. To by-pass this problem, the log-likelihood gain is calculated instead of the likelihood ratio by summing each log-probability. To simplify the model, reflections are considered independent from one another even if this is not true because they are all influenced by solvent and non-crystallographic symmetry, but the correlation is weak and the approximation is good [86].

Unlike Patterson methods, the phase of the observed structure factors is included in the probability function, even being unknown (nuisance variable), and it

is integrated out at the end of each analysis. Also, likelihood methods estimate experimental and model errors by the central limit theorem (the distribution of an average tend to be Gaussian), which allows the description of the structure factors amplitude distributions as a two-dimensional Gaussian in two-dimensional space that average all the ‘random walks’ taken during structure factors evaluation [84, 86].

As example of likelihood approach, I will describe the *Phaser* [81] pipeline, as this program is used in this thesis. In maximum likelihood methods, as for Patterson method, the search procedure is divided in two main steps: the rotation function (Likelihood Enhanced Fast Rotational Function – LERF) and translation function (Likelihood Enhanced Fast Translational Function – LETF).

In the LERF search function the model is rotated many times on an angular grid (Eulerian angles are used), and maximum likelihood scores are produced for each orientation. The highest are selected to be used in the subsequent translational search. At this stage, calculations are not made on total structure factors, because the rotational component of the phase is known for each position, but not the translational component. For this reason, structure factors cannot be summed. However, it is possible to evaluate each single amplitude to understand which the magnitude of the total structure factors would be. This procedure, statistically, is represented by a ‘random walk’ which results in a two-dimensional Gaussian. Because of the missing translation values, the rotational Gaussian is wider than the translational one. On this value, the -log-likelihood gain (LLG) is calculated (with the phases -nuisance variable- integrated out, a Rice distribution). Rotational searches are particularly difficult for highly symmetrical space groups because multiple orientations can decrease signal to noise ratio.

The translational search is carried out in the same way as the rotational search. One or more selected rotation is translated at grid points in the unit cell space. Because amplitudes and phases are not known for each position, the total structure factors, as before, cannot be summed. Each reflection carries an error, derived from the error in the model and described by a two-dimensional Gaussian. Log-likelihood is calculated on the structure factors in the same manner as for LERF (with the phases -nuisance variable- integrated out).

The best solution from each search procedure are the ones that produce the highest likelihood over the whole dataset of reflections. In *Phaser*, an LLG and a Z-score are provided to the user at the end of the phasing procedure as a way of evaluating the reliability of the result. The LLG (acceptable values typically higher than 150) is calculated by subtracting the likelihood of the model from a likelihood

calculated from atoms chosen at random positions in the unit cell (Wilson Distribution). The Z-score (acceptable values typically higher than 10) is the sum of the standard deviations of the LLG scores calculated on a number of randomly chosen translation and rotation [84].

*Phaser* pipeline includes the following steps:

- Cell content analysis: cell volume, probable space group, protein M.W. are used for the calculation of the Matthews coefficient;
- Anisotropy correction: anisotropic B-factors correction to a final absolute scale;
- Translational NCS correction: Patterson peaks are calculated to search for NCS, twinning analysis, NCS/twinning corrections;
- Rotation function: ensembling (coordinates of the model are converted to structure factors), initial r.m.s.d. calculated on sequence identity, estimation of the expected LLG of the ensemble, rotation function (LERF), the highest peaks from LERF are rescored using the full maximum likelihood rotation function – MLRF [83];
- Translation function: translation NCS, preparation of data for translation function, evaluation of Wilson distribution, check alternative space-groups to be tested, ensembling, translation function (LETf), the highest peaks from LETf are rescored using the full maximum likelihood translation function - MLTF[82];
- Packing: Ca clash test (is there any other model within 2Å from the one in analysis? the number of accepted clashes should be increased for search models with low sequence identity);
- Refinement: rigid body refinement on translations, rotations, B-factors refinement;
- Final refinement: rigid body refinement on translations and rotations, B-factors and model r.m.s.d. refinement.

*Phaser* search is called a ‘tree search with pruning’ because when multiple models have to be positioned, all potential placement discarded for the first component are stored to be used for the second component, and so on [81].

*Refinement.* The procedure of crystallographic refinement consists of the minimization of the difference between the absolute values of the experimental and the calculated structure factors. The coordinates of the model were modified manually in WinCOOT [87] and by automated refinement using the program Refine [88-90]. As

discussed previously, values in bond lengths, bond angles, dihedral angles, chirality, planarity, Ramachandran phi and psi, rotamers, C-beta deviations, side-chains restraints, clashes, etc. are minimized during refinement by rounds of coordinate re-modelling and calculation of an electron density map.

The discrepancies between the calculated structure factors ( $F_{hkl}^{calc}$ ) and the observed structure factors ( $F_{hkl}^{obs}$ ) are estimated by the R factor:

$$R = \frac{\sum_{hkl} \left| |F_{hk}^{obs}| - |F_{hkl}^{calc}| \right|}{\sum_{hkl} |F_{hkl}^{obs}|}$$

R-free value is generally considered to give a more reliable estimation of the quality of the model. R-free is calculated in the same manner as the R factor but by using a subset (5-10 %) of the reflections which have been set aside and not used in the refinement. The closer the model agrees with the experimental data, the lower R-free and R-factor will be.

## 2.9. Molecular docking techniques

Molecular docking is a computational technique used in the prediction of the non-covalent binding of a ligand to a macromolecule. Docking fulfils two main tasks: the ligand pose generation, where the optimal conformations of the ligand and its position in the three-dimensional space are defined, and the scoring of the poses, in which the strength for each binding mode is evaluated [91].

The inaccuracy of the scoring function is one of the biggest limitations in docking. For this reason, multiple scoring function have been developed over the years (force-field based, knowledge-based, empirical-based, machine learning-based, etc.) and new approaches continue to be published. The scoring function implemented in Autodock Vina [92], the program used for docking is derived from the empirical-based X-score [93] and its description can be found in an article by Trott & Olson (2010) [92].

*Protocol.* Virtual screening of protein-ligand interactions was launched for fixed protein models, generated from the X-ray crystal structures of phytases and the torsionally-flexible ligand inositol hexasulphate (used as a control) or inositol hexaphosphate, the substrate. In AutoDockTools 1.5.4, hydrogen atoms were added to the protein structures, the degree of torsional freedom of the ligands was defined and both the models were converted into PDBQT file, the input file format required by Vina. By using the same software, a search space of dimension  $18 \times 22 \times 18 \text{ \AA}^3$  was defined around the active site. A default exhaustiveness index of 8 was used in the search.

## CHAPTER 3

### 3. Probing the catalytic flexibility of the *Escherichia coli* 6-phytase AppA

AppA, the *Escherichia coli* 6-phytase of the histidine phosphatase (HP) family, has been well characterized and successfully engineered for use as an animal feed supplement (mutant sequence not public). Despite the large use of the enzyme, an explanation of its rather rigid preference for the initial site of cleavage of phytic acid (IP<sub>6</sub>) at the 6-phosphate is still lacking. In contrast, multiple inositol polyphosphate phosphatases (MINPPs), also members of branch 2 of the HP superfamily, demonstrate pronounced catalytic promiscuity, showing mixed 1/3-, 5- and 4/6-phytase activities. To help shed light on this difference, the role of the catalytic proton donor residue in AppA has been investigated in comparison with that found in MINPPs. Four AppA active site mutants were generated by site-directed mutagenesis of the wild type HDT amino acid sequence motif containing the presumed proton donor aspartic acid. In so doing, the MINPPs-like HAE sequence motif was introduced and the profiles of inositol polyphosphate (IP) products, generated from hydrolysis of phytic acid by these mutants, were analysed by HPLC.

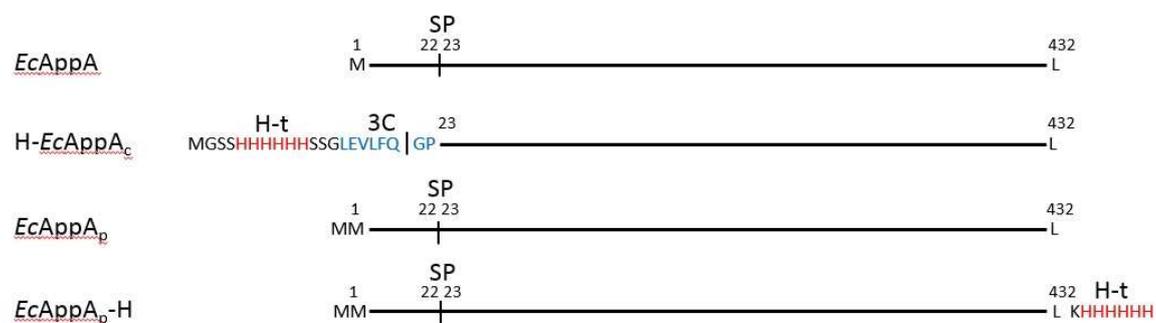
The wild type enzyme generates as major intermediate the IP<sub>5</sub> 4/6-OH species, with a minor IP<sub>5</sub> 1/3-OH intermediate also detected. The HDE mutant (T327E) has a similar IP profile to the wild type but with increased preference for hydrolysis at the 4/6-position. The HAT mutation (D326A) abolishes 6-phytase activity and the mutant displays an increased exclusive 1/3-phytase activity. The HAE double mutant (D326A, T327E) contains a MINPPs-like proton donor motif and shows an enhanced catalytic promiscuity, with diminished 4/6-phytase activity and more prominent 1/3-phytase activity. Finally, the HET mutant (D326E) displays an equal preference for 1/3- or 4/6-phytase activity. Taken together, these results provide the first evidence for the involvement of the proton donor motif in determining the initial site of attack of HP phytases on their IP<sub>6</sub> substrate.

To explain these differences, the structural and enzymatic properties of the mutants were studied. A change in  $K_M$  was observed for all mutants excluding *EcAppA* HET, while a decrease in  $k_{cat}$  of at least 100-fold was displayed by all the mutants apart from HDE that showed a 6-fold loss. The macromolecular X-ray crystal structures solved in complex with the substrate analogue IHS showed a remodelling of the binding pocket in all the mutants. Multiple binding modes for IHS could be

modelled in the electron density of *EcAppA* HET, while the sulphate in position 6 of the inositol ring contacted H17 in the structure of all the other mutants and wild-type, including surprisingly *EcAppA* HAT which displayed a predominant 1/3-phytase activity instead of the 4/6-activity of the wild-type enzyme.

### 3.1. Gene constructs

The gene encoding *Escherichia coli* AppA (*EcAppA*) was amplified from the genome of BL21 (DE3) pLysS and three constructs were cloned into pOPINA or pOPINB vectors. Constructs were designed (1) for the cytoplasmic expression of an N-terminal cleavable His-tag *EcAppA* (**Figure 3.1.1**, H-*EcAppA<sub>c</sub>*), (2) for periplasmic translocation and C-terminal His-tagged (**Figure 3.1.1**, *EcAppA<sub>p</sub>*-H), and (3) for periplasmic translocation, untagged (**Figure 3.1.1**, *EcAppA<sub>p</sub>*).

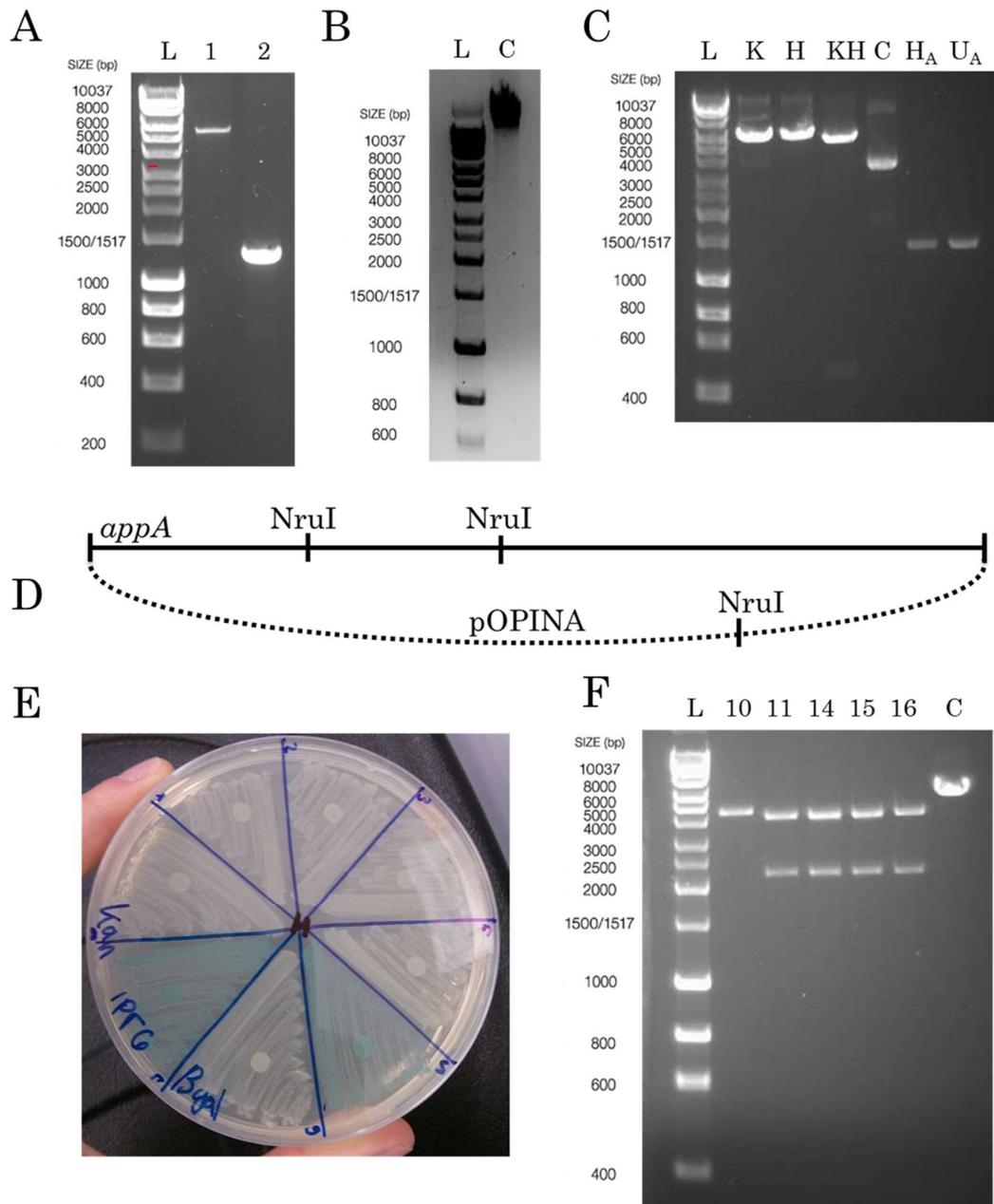


**Figure 3.1.1. *EcAppA* constructs.**

Lines indicate gene length and residues (aa) numbers positions. *EcAppA* is a gene of 432 aa containing a Signal Peptide (SP) for periplasmic translocation (1-22 aa). *appA*, to be translated and retained in the cytoplasm, is fused to an N-terminal cleavable His-tag in the construct H-*EcAppA<sub>c</sub>*. *EcAppA<sub>p</sub>* is full length, therefore, periplasmic targeted.

H-*EcAppA<sub>c</sub>* was cloned by the In-Fusion method (**Chapter 2.4.1**). The agarose gel of the linearized plasmid and the PCR amplified gene (without the region encoding the Signal Peptide for the periplasmic targeting) is shown in **Figure 3.1.2.A**.

*EcAppA<sub>p</sub>* with and without a His-tag were cloned into linearized pOPINA vectors (Oxford protein production facility - UK). *EcAppA<sub>p</sub>* native and C-terminal His-tagged were amplified from the genome of BL21 (DE3) pLysS (**Figure 3.1.2.B**).



**Figure 3.1.2. pOPINA EcAppA cloning. Panel of figures.**

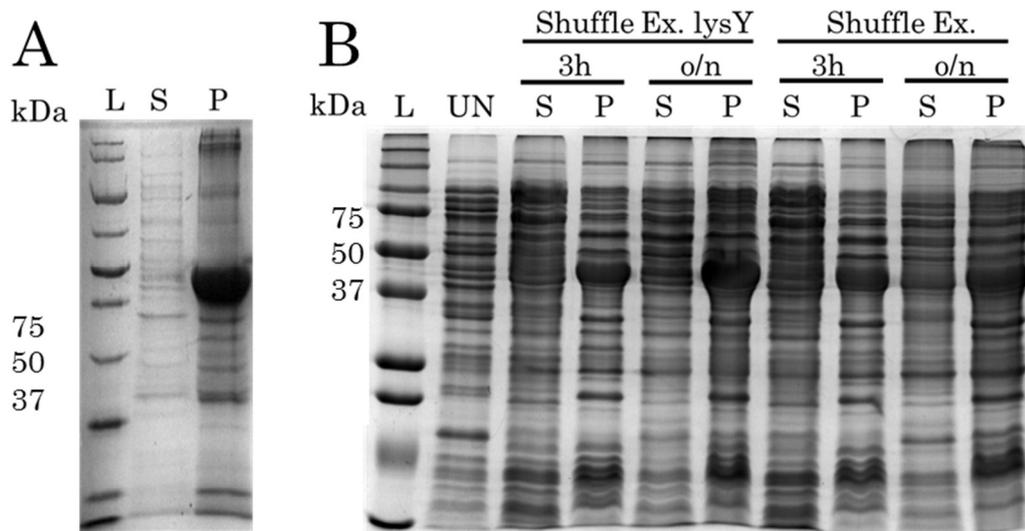
L: HyperLadder 1Kb. **A** - Linearized pOPINB (1 - 5309 bp) and *appA* amplified for insertion on 1% agarose gel (2 - 1230 bp). **B** - 1% agarose gel of the BL21 (DE3) pLysS genome extraction (C). **C** - 1% agarose gel of pOPINA digestion and AppA amplification. HA: amplified C-term uncleavable His-tagged AppA<sub>p</sub> (~1500 bp), UA: untagged AppA<sub>p</sub> (~1500 bp), K: KpnI digestion (5621 bp), H: HindIII digestion (5621 bp), KH: double digestion by KpnI and HindIII (~5300 bp), C: circular pOPINA. **D** - Scheme of NruI digestion of pOPINA/B+insert. *appA* gene is drawn as a solid line, the plasmid as a dotted line. Three cleavage sites are present: one on the plasmid and two on *appA* gene. Cleavage generates three fragments of: 4014 bp, 2148 bp and 241 bp. **E** - Plate of AppAp-H clones. Every slice is prepared by streaking a single white colony from a previous blue/white screening to allow a better visualization of the colour. **F** - 1% agarose gel of the NruI digestion of miniprep plasmids from positive clones. Clones 11,14,15 and 20 are cleaved in three fragments: they contain the correct insert. It was possible to detect a third low MW faint band which is difficult to identify in the image.

Positive clones from transformations of the In-Fusion reactions were selected by blue-white screening (**Figure 3.1.2.E**). A few clones were minipreped and digested with NruI (**Figure 3.1.2.D/F**), which cuts the plasmid twice and the gene once, or KpnI-HF and HindIII-HF, which should be unable to cut the plasmid if insertion is successful. Positive clones were confirmed by sequencing.

## 3.2. Expression trials, solubility and purification

*EcAppA* is characterized by three consecutive disulphide bridges and an additional fourth C-terminal non-consecutive disulphide bridge, the formation of which is catalysed by the periplasmic disulphide isomerase DsbC [94, 95]. This feature renders soluble overexpression of *EcAppA* dependent on the presence of the DsbC enzyme. Soluble protein was obtained only when it was targeted to the periplasm of *E. coli* or in Shuffle Express strains which are engineered to express constitutively DsbC in their cytoplasm.

Before the achievement of this goal, multiple trials were set up with different constructs. First, expression was tested for H-*EcAppAc* in the cytoplasm of Rosetta2(DE3)pLysS and Rosetta-gami2(DE3)pLysS by varying IPTG concentrations (0.1, 0.25, 0.5 mM) and temperature (25, 37 °C). The latter strain carries *trx* and *gor* mutations for enhanced formation of S-S bridges in the cytoplasm. Nevertheless, no soluble protein was obtained (**Figure 3.2.1.A**). The second expression trial involved *EcAppAp-H*, with enzyme targeted to the periplasm of the bacteria. Here, soluble protein was obtained for the o/n overexpression in BL21(DE3), induced with 0.05 mM IPTG at mid-late exponential phase, 37 °C incubation, 180 rpm shaking. However, the result was not a success because the tag limited the solubility of the protein to 0.8 mg/mL. Expression, of the untagged construct, in the same conditions, allowed purification of soluble protein stable during spin-concentration up to ~10 mg/mL. The yield of purified protein was enough to set up crystallization trials and collect diffracting datasets, though the purification procedure was very laborious. This prompted a search for a faster and more productive alternative. Overexpression trials were set-up with the N-terminal cleavable His-tag construct in Shuffle Express T7 (**Figure 3.2.1.B**). This strain has a basal expression of the disulphide isomerase DsbC in its cytoplasm allowing for the formation of non-consecutive disulphide bridges in this compartment.



**Figure 3.2.1. *EcAppA* expression trials and solubility test. Panel of figures.**

H-*EcAppAc* MW: 47 kDa; L: proteins MW ladder; S: soluble fraction; P: pellet; 3h: 3 hours induction; o/n: overnight induction; **A** - 12 % acrylamide SDS-PAGE of soluble and pelleted cell fractions from H-*EcAppAc* overexpression in the cytoplasm of Rosetta 2 (DE3) pLysS, with 0.5 mM IPTG induction at OD 0.6, 37 °C, 180 rpm. No soluble *EcAppA* was detected. **B** - 12 % acrylamide SDS-PAGE of soluble and pelleted cell fractions from H-*EcAppAc* overexpression in the cytoplasm of Shuffle Express T7 and Shuffle Express T7 pLysY, with 0.01 mM IPTG induction at OD ~ 0.9, 37 °C, 180 rpm. Soluble *EcAppA* was detected in Shuffle Express T7, o/n.

*EcAppA* expression in Shuffle Express T7 was previously tested by the strain developers [96]. Here, this host gave positive results and allowed for further studies of *EcAppA* mutagenesis and structural characterisation.

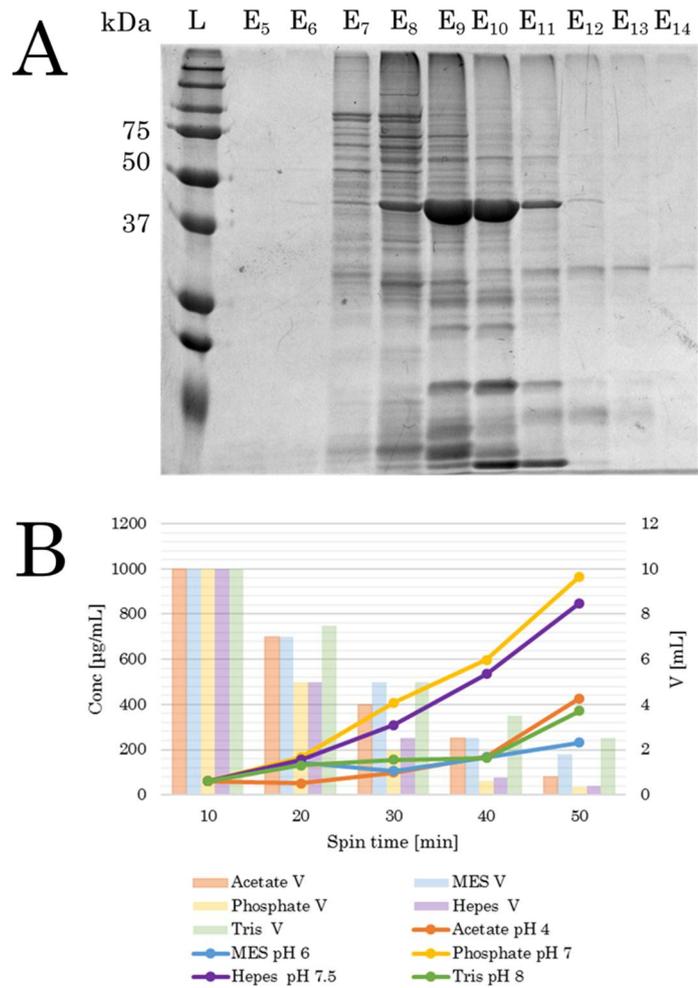
Optimised conditions for expression of the N-terminal cleavable His-tagged *EcAppA* in the cytoplasm of Shuffle Express T7 requires the incubation of the transformed host at 37 °C, 180 rpm shaking until cultures reaches OD of induction (late exponential phase, OD ~ 0.9) with IPTG to be added to a final concentration of 0.01 mM. The protein is expressed o/n. A summary table of the expression trials can be found in **Appendix 8.3.1**.

The first construct to be purified was *EcAppAp-H*. It was lysed and purified by Ni<sup>2+</sup>-NTA IMAC (protocol **Chapter 2.6.1**). Unfortunately, the solubility limit of this construct was 0.8-1 mg/mL. The protein precipitated at higher concentration. Because of this, the construct was considered unsuitable for gel filtration or further purification. A spin concentration experiment was undertaken with multiple buffers to determine optimal conditions for protein solubility. The preferred buffers identified were 50 mM NaH<sub>2</sub>PO<sub>4</sub> pH 7.0, 150 mM NaCl and 50 mM HEPES pH 7.5, 150 mM NaCl (**Figure 3.2.2**). However, after further trials higher concentrations were not reached. For this reason, efforts were redirected to the untagged construct with the

benefit of obviating potential problems in crystallisation arising from the presence of the C-terminal tag.

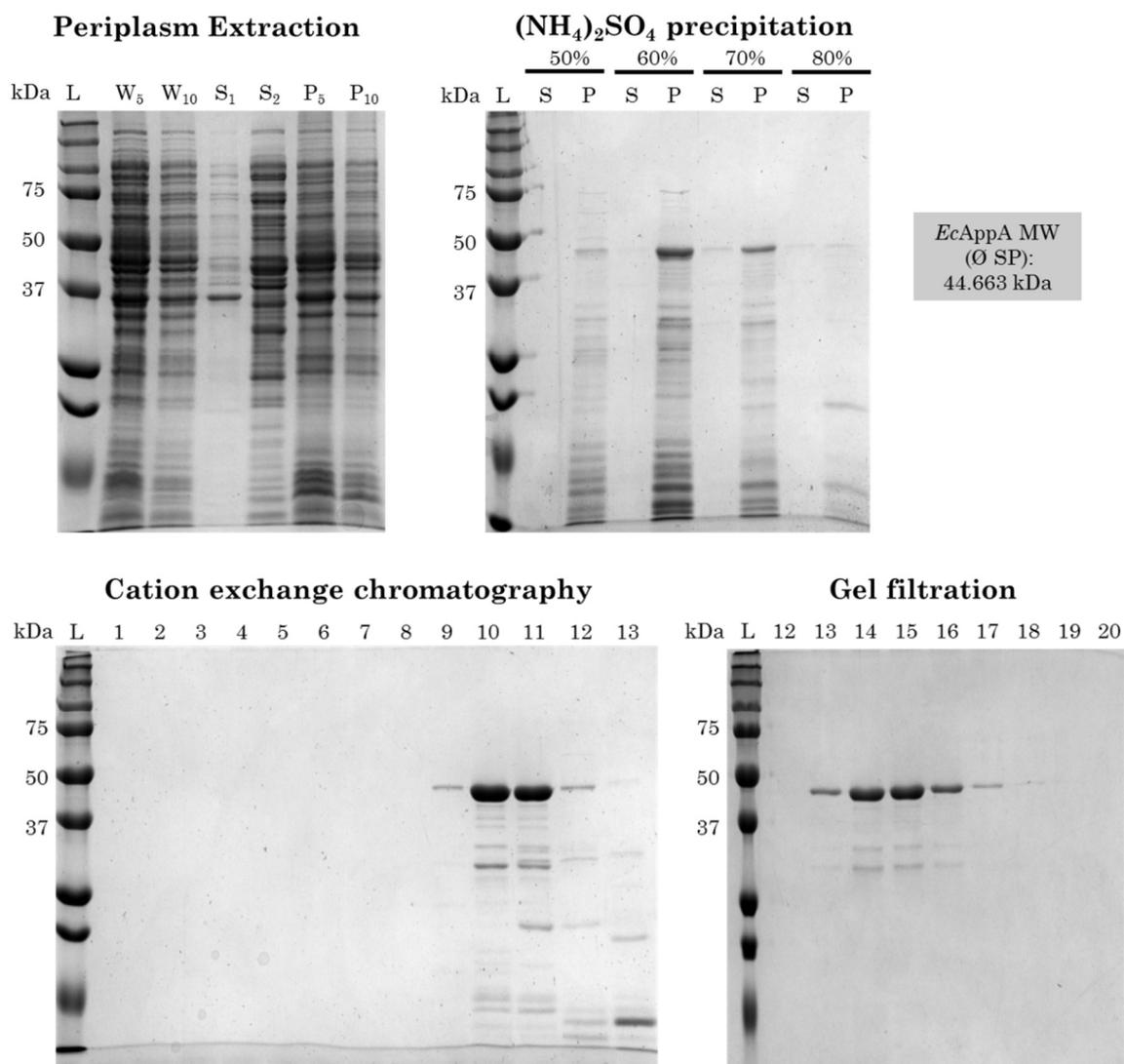
The periplasm-directed untagged *EcAppA* was expressed in the same condition as the C-terminal His-tagged fusion protein. The purification protocol was instead more complex (**Chapter 2.6.2**, Vasudevan, Salim [97]). It included a step of extraction of the *E. coli* periplasm, the salting-out of proteins at increasing concentration of ammonium sulphate [53], the re-solubilisation of *EcAppA* in Glycine buffer pH 1.8, desalting for cation-exchange chromatography and finally gel filtration (**Figure 3.2.3**). The procedure was successful, but being long and tedious, a faster and less laborious method was sought.

A solution to the problem was found when the expression of H-*EcAppAc* in the cytoplasm of Shuffle Express T7 was optimized. The purification protocol (**Chapter 2.6.1**) was straight forward. It included mechanical cell lysis by French Press, a 1<sup>st</sup> Ni<sup>2+</sup>-NTA IMAC purification for the isolation of H-*EcAppAc*, the cleavage of the tag by a His-tagged 3C-protease, a 2<sup>nd</sup> Ni<sup>2+</sup>-NTA IMAC which allowed the collection of the cleaved protein in the FT and a final step of gel filtration. 12 % SDS-PAGE gels of the procedure are reported in **Figure 3.2.4**.



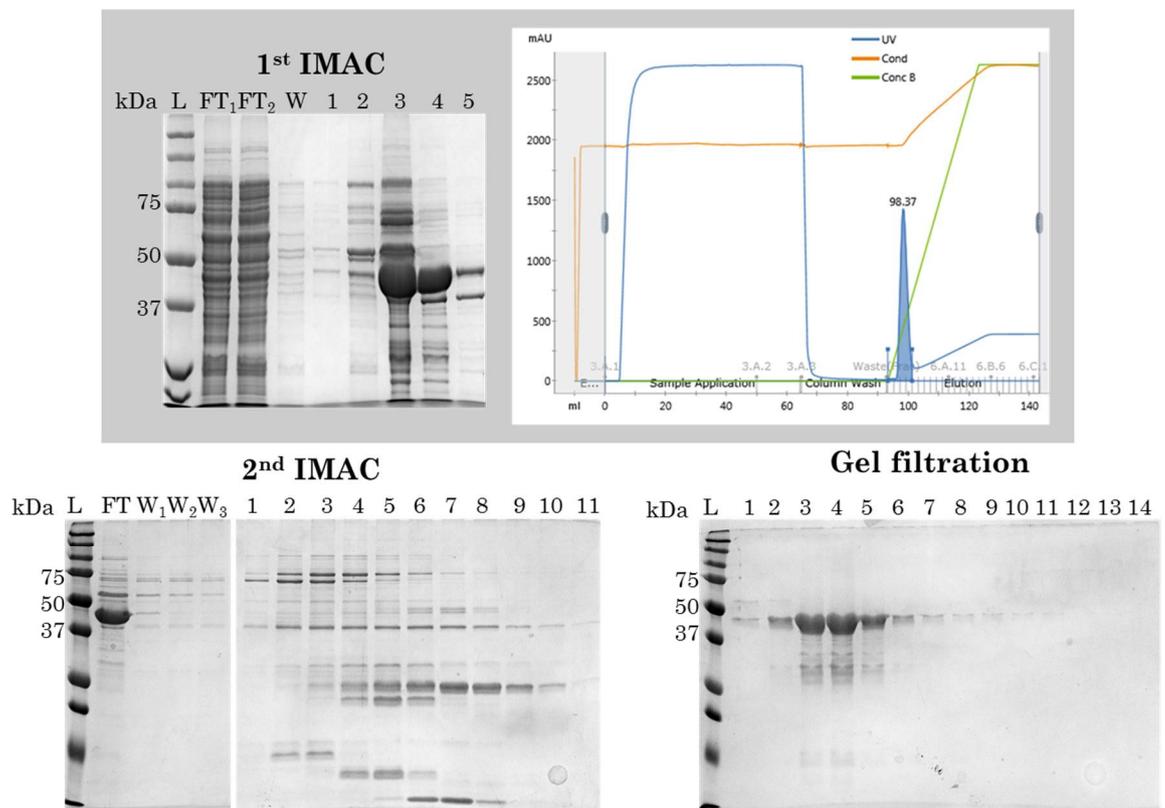
**Figure 3.2.2. Purification of *EcAppA*-H.**

**A** - 12 % acrylamide SDS-PAGE of the IMAC purification, elution fractions 5-14. *EcAppA* (MW ~47 kDa) eluted in fractions 8-11. **B** - Protein stabilization trial. *EcAppA* is diluted in multiple buffers and spin concentrated in cycles of 10 min, 4000 xg, 4 °C. Buffers: Acetic Acid buffer (Ac. Acid pH 4.0), MES buffer (MES pH 6.0), Phosphate buffer (Phosphate pH 7.0), HEPES buffer (HEPES pH 7.5), Tris-HCl buffer (Tris pH 8.0), and is then. x axis: spin time (min), y axis-left: concentration µg/mL, Y axis-right: volume (mL). Lines and dots show the increase in concentration of the protein in each buffer. Bars represent instead the decrease in volume of each protein-containing buffer solution.



**Figure 3.2.3. Purification of EcAppA.**

L: protein ladder. *EcAppA* MW: ~46 kDa, MW of the construct without signal peptide (SP): 44.663 kDa. Protein samples resulting from four purification steps of increasing protein purity are displayed: **extraction of BL21(DE3) periplasm** – W: whole cells, S1: supernatant of the sucrose gradient formation step, S2: periplasm, P: spheroplasts, 5 or 10: 5x and 10x dilutions; **ammonium sulphate precipitation** at increasing saturation (50-60-70-80 %) – S: precipitation of the aliquot collected in the sucrose gradient formation step, P: precipitation of the periplasm; **cation exchange chromatography** – only elution fractions 1-13 at increasing NaCl concentrations are reported; **gel filtration** – only fractions 12-20 are reported.



**Figure 3.2.4. Purification of H-*EcAppAc*.**

L: protein ladder *EcAppA* MW ~46-47 kDa. At the top, 1<sup>st</sup> Ni<sup>2+</sup>-NTA IMAC purification. On the left a 12 % SDS-PAGE is reported: FT – flow through, W – wash, 1-5 elution fractions. On the right a chromatogram of the purification is displayed, it can be divided in three main phases: sample application, column wash and elution. x axis: eluted volume; y axis: units of absorbance, in blue UV absorbance is reported, in green the concentration of eluting buffer (%), in orange conductivity. Bottom left, 2<sup>nd</sup> Ni<sup>2+</sup>-NTA IMAC. A 12 % acrylamide SDS-PAGE is reported. The cleaved protein is found in the flow through (FT), followed by 3 fractions of column wash and the elution of contaminants (fractions 1-11) including the His-tagged 3C-protease. Bottom right, **gel filtration**. A 12 % acrylamide SDS-PAGE is displayed of fractions 1-14. *EcAppA* elutes mainly in fractions 3-5.

### 3.3. Site-directed mutagenesis of active site residues

The constructs for periplasmic expression of *EcAppA* (pOPINA::*EcAppA*) were mutated to introduce four variations to the original proton donor sequence (HDT): D326A (HAT), D326E (HET), T327E (HDE) and the double mutant D326A-T327E (HAE). Mutant sequences to be translated in the cytoplasm are transferred into pOPINB by cloning (construct HIS6-3C-POI).

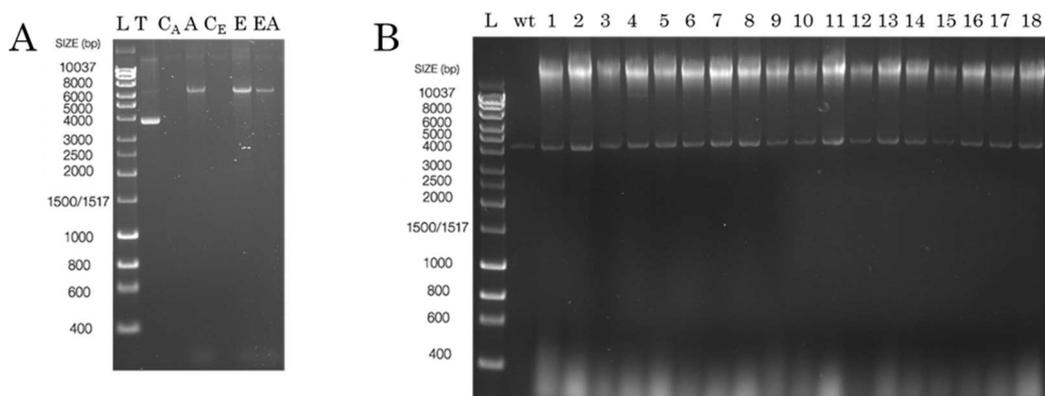
The site-directed mutagenesis protocol included a step of mutant strand synthesis followed by DpnI digestion of the PCR product and transformation. Phusion® High-Fidelity DNA Polymerase was the enzyme chosen for the replication of the mutated plasmids. Reactions were set up to contain 1x HF buffer (which provides 1.5 mM of MgCl<sub>2</sub>), 0.4 mM dNTPs, 4 % DMSO, 0.4 μM of each primer, 0.4

ng/ $\mu$ L of plasmid template and 0.02 U/ $\mu$ L of DNA polymerase. The polymerase chain reactions included an initial denaturation of the double helix at 95 °C for 5 min followed by 13 cycles of: denaturation (95 °C, 5 min), primers annealing (melting temperature of the full-length primers -5 °C, 1 min) and extension of the new amplicons (72 °C, 10 min). After the last cycle, reactions were incubated for a final extension at 72 °C for 30 min before cooling to 4 °C to stop replication. A list of the primers can be found in **Appendix 8.2.3**. Primers were designed according to the method of Liu and Naismith [98], taking into account *E. coli* codon usage. An SDS-PAGE of an optimized amplification is shown in **Figure 3.3.1.A**.

Amplicons were treated with DpnI. The endonuclease degraded parental DNA recognized for its methylations or hemi methylation on target sequences 5'-Gm6ATC-3'. The digestion was carried out adding 2  $\mu$ L DpnI (10 U/ $\mu$ L) in 50  $\mu$ L reaction and incubating the DNA sample at 37 °C for 2 h.

Vectors were transformed and amplified in XL-10 Gold ultracompetent cells (Agilent Technologies). The transformation protocol was modified slightly from the standard protocol (**Paragraph 2.3**) to include a step in which cells were treated with  $\beta$ -mercaptoethanol.

To check that full-length plasmids were amplified, colonies were lysed, and plasmid length analysed (**Figure 3.3.1.B**, protocol **Paragraph 2.5.1**). A few positive clones were grown o/n for plasmid amplification, purified through miniprep (Quiagen) and NruI digested (**Paragraph 2.5.3**). Two plasmids of the right length per mutant were sent for sequencing.

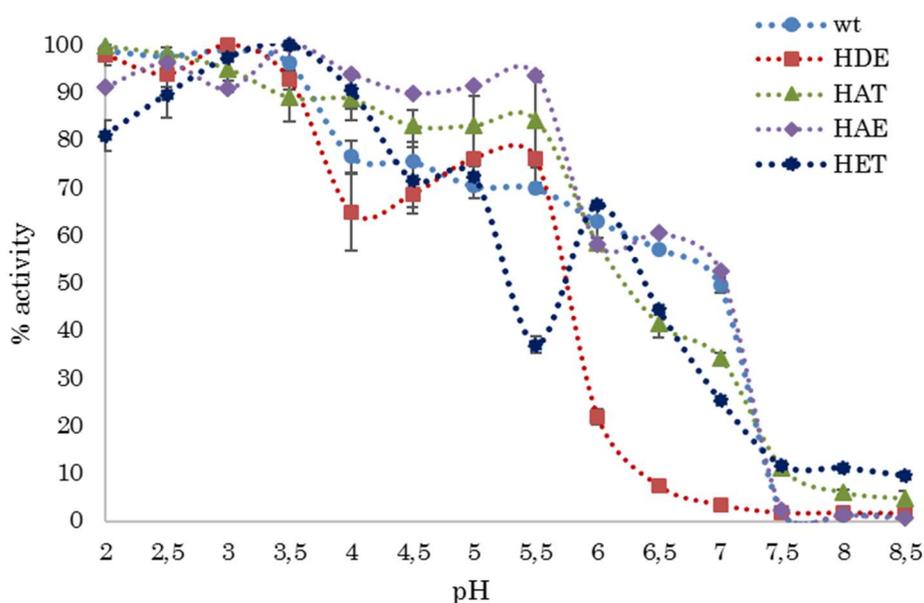


**Figure 3.3.1. Site-directed mutagenesis of *EcAppA***

L: HyperLadder 1 kb. **A** - 1 % agarose gel of the mutagenesis PCR. T: template; C<sub>A</sub>: control of the reaction for the amplification of D326A – water replace the template; A: reaction for the amplification of D326A; C<sub>E</sub>: control of the reaction for the amplification of T327E – water replace the template; E: reaction for the amplification of T327E; EA: reaction for the amplification of D326A+T327E. **B** - 1 % agarose gel of the plasmid length detection screening. wt: miniprep wild-type plasmid. 1-18: DNA content of colonies 1-18. All colonies contain a plasmid of correct length. Colony 4 display a second band, around 8-10 kbp. The top signal corresponds to genomic DNA.

### 3.4. pH profiles

The ability of enzyme to degrade IP<sub>6</sub> was tested in buffers at increasing pH range (0.2 M Glycine-Cl pH 2.0/1.5/3.0/3.5, 0.15 M NaCl; 0.2 M NaAcetate pH 4./4.0/5.0/5.6, 0.15 M NaCl; 0.2 M Bis-Tris pH 6.0/6.5/7.0, 0.15 M NaCl, 0.2 M Tris-Cl pH 7.5/8.0/8.5, 0.15 M NaCl), **Figure 3.4.1**. Reactions were set up in triplicate at fixed concentrations of enzymes (*EcAppA* wt/HDE: 7.5 nM, and mutants: 1  $\mu$ M) and substrate (IP<sub>6</sub>: 1 mM). Enzymes were incubated at room temperature for 10 min before inactivation. Solutions were left to develop their colour for 30 min at room temperature before measuring Abs<sub>700</sub>. Triplicate reads of buffers-only solutions were taken as well as a phosphate calibration curve to check that Abs registered was in the linear range for the assay. Further information on the protocol can be found in **Chapter 2.7.2**.



**Figure 3.4.1. pH profile of *EcAppA* wild-type and mutants.**

x axis: pH, y axis: % activity. Enzyme colour scheme: light-blue circles – wild-type, magenta squares – HDE mutant, green triangles – HAT mutant, purple rhombus – HAT mutant, dark blue suns – HET mutant. Buffer only Abs<sub>700</sub> were subtracted from raw samples Abs<sub>700</sub>, then, the latter were averaged. The % of averaged samples Abs<sub>700</sub> to the maximum average were calculated and plotted above with the respective standard deviations. pH profiles vary between mutants, though the region of activity maximum can be identified for all around pH 2.0-3.5.

The pH profile of *EcAppA* wild-type is in line with the data reported by Rodriguez et al. (1999) [99], where maximum of activity is found at pH 2.5, followed by a gradual decrease at increasing pH, reaching inactivation at pH 8.5. 60% activity is registered at pH 4.5 by Rodriguez et al. vs 72% in my assay (**Figure 3.4.1**). Also, my sample is inactivated at pH 7.5, while Rodriguez's *EcAppA* retains 20% of its

activity at this pH. Results differ markedly from the profiles reported by Golovan et al. (2000) [100], which shows an optimum of activity at pH 4.5 and an early inactivation at pH 6.5.

Mutations influenced the pH profile of *EcAppA*. Despite all enzymes showing an optimum of activity at low pH, between 2 and 3.5, differences were observed at pH 5.5 and above. *EcAppA* HDE is early inactivated at pH 6.0, with only 20% of activity retained. The profile of this mutant slightly resembles *AnPhyA* trends [99]. *EcAppA* HET has a drop in activity at pH 5.5 (35%) and recovers at pH 6 (60%) before a slow decrease to a minimum at pH 7.5-8.5 (10%). The activity of *EcAppA* HAT and HAE mutants is constant at pH range 2.0 - 5.5, but thereafter *EcAppA* HAT gradually reaches inactivation at pH 8.0, while *EcAppA* HEA profile shows a steep drop at pH 7.5, much like the wild-type enzyme.

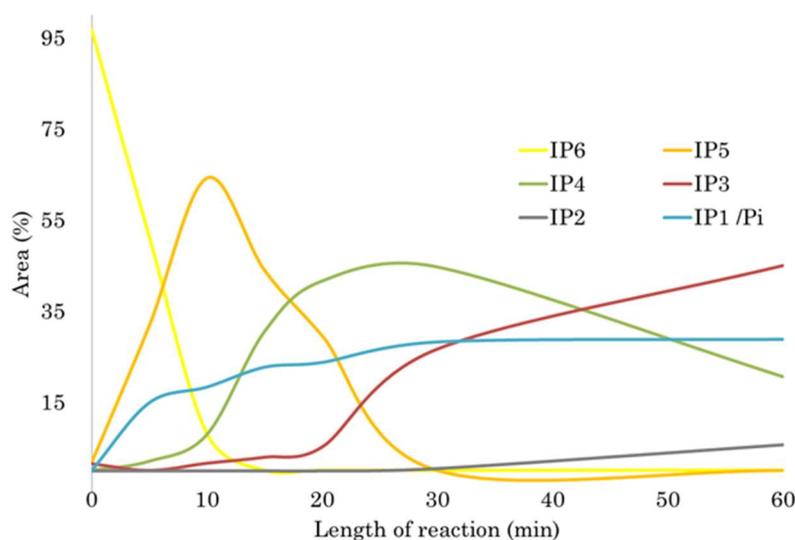
Some of the fluctuation detected has a large standard error which does not allow to differentiate them from the wild-type profile. An example is the relative minimum at pH 4.0 of *EcAppA* HDE. However, some traces have very small error bars, like the one of *EcAppA* HET, which would suggest a reliability of the result. To support these data, additional tests using overlapping buffers for each pH would give a more comprehensive view of the behaviour of the enzymes and would show how buffer choices can influence activity. In fact, pH profile of other enzymes showed discrepancies in % activity between buffers at the same pH (**Paragraph 6.6.1**).

### 3.5. Analysis of the product of IP<sub>6</sub> hydrolysis by HPLC

To evaluate catalytic promiscuity of wild-type *EcAppA* and mutants in attacking IP<sub>6</sub>, their products of hydrolysis were separated by HPLC (protocol in **Chapter 2.7.3**). Lengths of reactions and enzymes concentration were iteratively modified to stop IP<sub>6</sub> cleavage at comparable stages (e.g. IP<sub>6</sub> % of depletion was held constant for each mutant, to compare processivity of IP<sub>5</sub> removal).

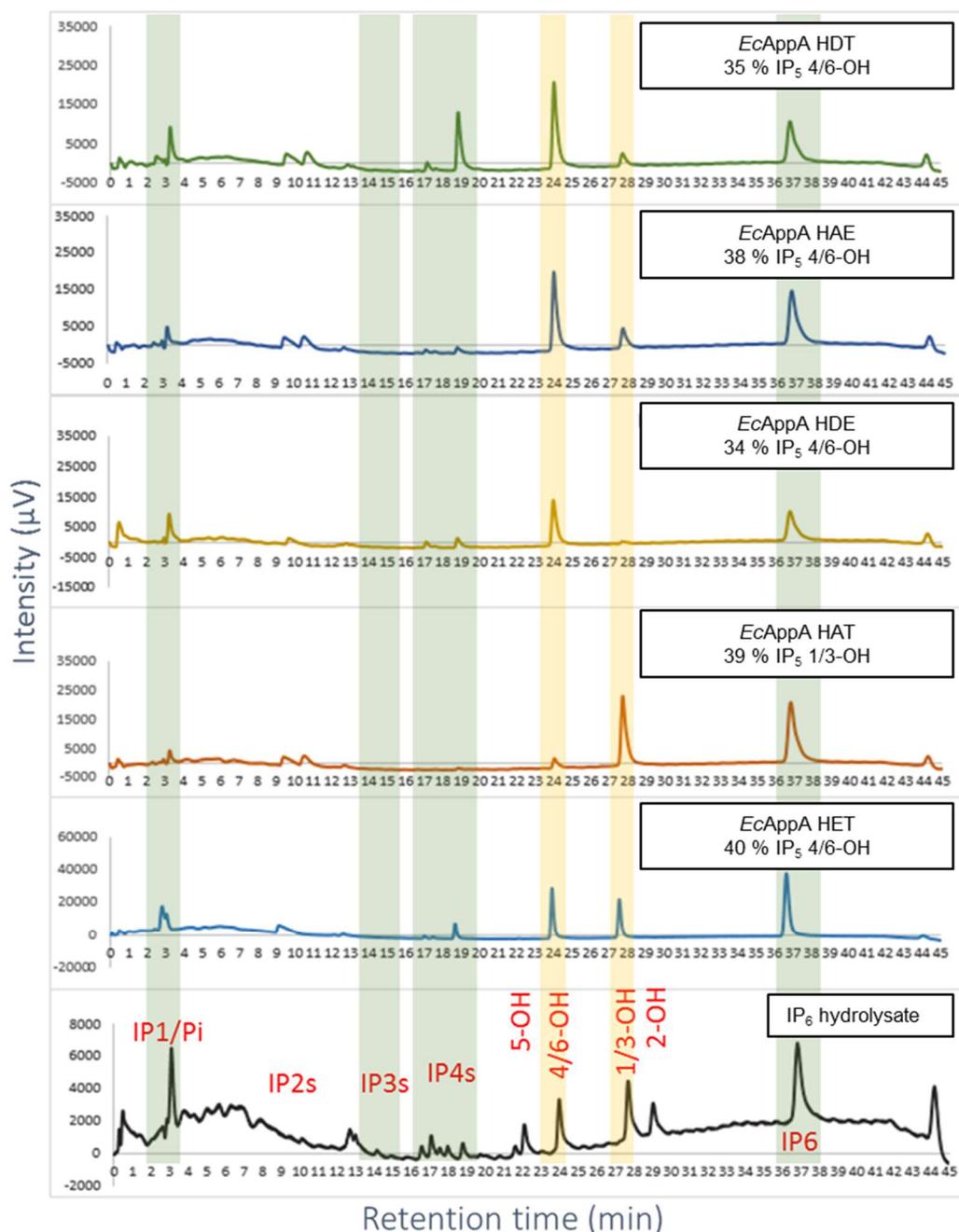
Typically, sequential hydrolysis of IP<sub>6</sub> by *EcAppA* leads to the generation of one predominant IP<sub>5</sub> species I(1,2,3,4,5)P<sub>5</sub> (also referred as 6-OH IP<sub>5</sub>) accounting for the 87% of the produced IP<sub>5</sub> (with the remaining 13% being I(2,3,4,5,6)P<sub>5</sub>, also called 1-OH IP<sub>5</sub>). Further attack yields to I(2,3,4,5)P<sub>4</sub>, than I(2,4,5)P<sub>3</sub> or I(1,2,4)P<sub>3</sub>, I(1,2)P<sub>2</sub> or I(2,5)P<sub>2</sub> or I(4,5)P<sub>2</sub> and finally I(2)P<sub>1</sub> or I(5)P<sub>1</sub> [35]. Results for the wild-type enzyme are consistent with former work.

Mutant enzymes proved to sequentially hydrolyse IP<sub>6</sub> (an example in **Figure 3.5.1**), however, their positional specificity of attack varied in relation to the mutation introduced. *EcAppA* HAE showed a slightly increased production of 1/3-OH IP<sub>5</sub> (22 % of the total IP<sub>5</sub>), while *EcAppA* HDE decreased the production of 1/3-OH IP<sub>5</sub> to 6 %. *EcAppA* HAT, surprisingly revealed a shift in stereospecificity achieving a predominant production of 1/3-OH IP<sub>5</sub> equivalent to 89 % of total IP<sub>5</sub>. *EcAppA* HET, instead, proved to be more promiscuous releasing 47 % of 1/3-OH IP<sub>5</sub> and 53 % of 4/6-OH IP<sub>5</sub> (**Figure 3.5.2**).



**Figure 3.5.1. Sequential hydrolysis of IP<sub>6</sub> by *EcAppA* HAE.**

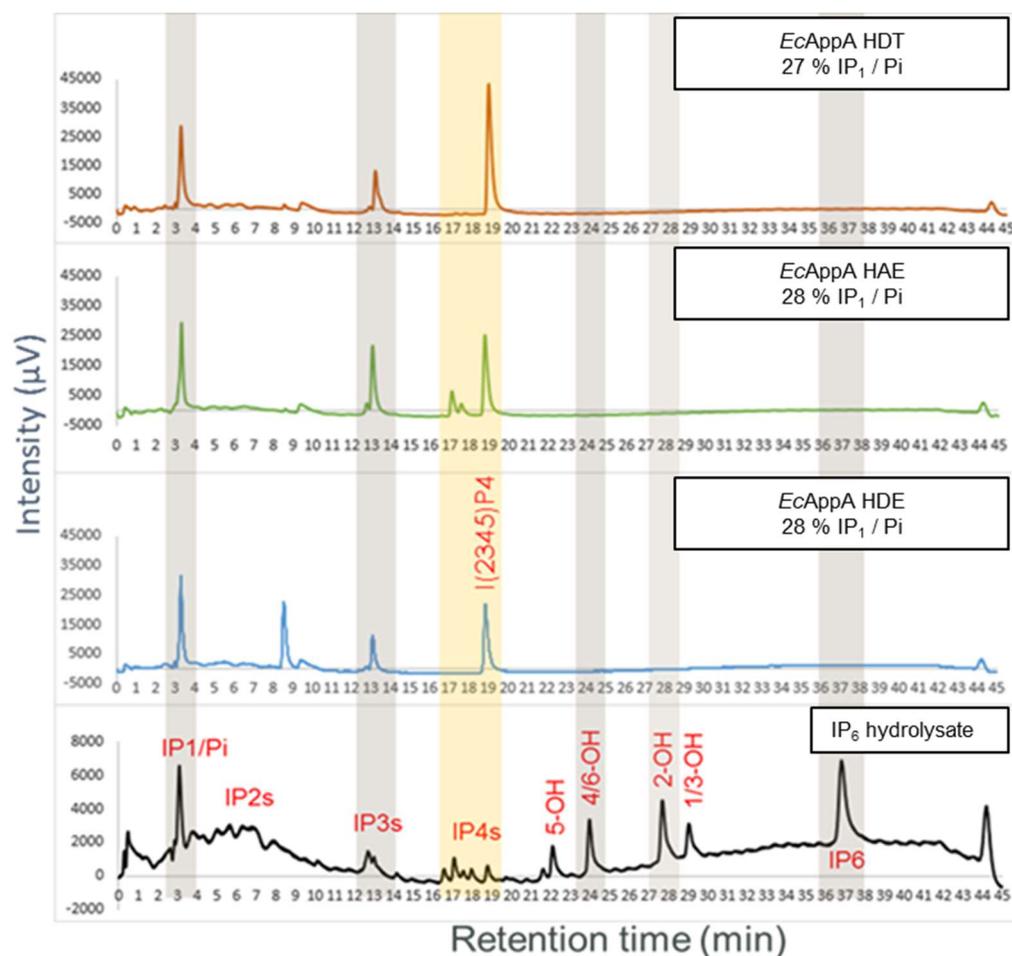
x axis: length of incubation (min). y axis: % peak area. The substrate, IP<sub>6</sub> (in yellow) is completely hydrolysed to IP<sub>5</sub> (in orange) in about 10 min. When IP<sub>6</sub> has almost totally disappeared, IP<sub>5</sub> cleavage begins. When IP<sub>5</sub> is processed to IP<sub>4</sub>, the latter becomes a substrate.



**Figure 3.5.2. HPLC chromatogram of reactions stopped when total IP<sub>5</sub> peaks area is equal to 34-40 % of total IPs.**

x axis: retention time (min) pH, y axis: intensity ( $\mu\text{V}$ ). The enzyme proton donor triplet and the predominant IP<sub>5</sub> peak area (%) are reported on the top right corner of each chromatogram. From the top, the chromatograms of *EcAppA* wild-type, *EcAppA* HAE, *EcAppA* HDE, *EcAppA* HAT, *EcAppA* HET and an IP<sub>6</sub> hydrolysate standard are displayed. A clear shift in stereospecificity from the production of 6-OH IP<sub>5</sub> to the accumulation of a 1/3-OH IP<sub>5</sub> is shown by the mutant *EcAppA* HAT. The *EcAppA* HET mutant instead displays a promiscuous production of either 4/6-OH IP<sub>5</sub> or 1/3-OH IP<sub>5</sub>.

Differences in stereospecificity of attack on IP<sub>6</sub> resulted in the downstream generation of a varied populations of IP<sub>4</sub> species. For example, *EcAppA* HAE showed increased production of 1/3-OH IP<sub>5</sub> (among IP<sub>5</sub>s) and a consequent change in profile of IP<sub>4</sub>s (Figure 11).



**Figure 3.5.3.** HPLC chromatogram of reactions stopped when total Pi/IP<sub>1</sub> peaks area is equal to 28 % of total IPs.

x axis: retention time (min) pH, y axis: intensity (µV). The enzyme proton donor triplet and the predominant IP<sub>5</sub> peak area (%) are reported on the top right corner of each chromatogram. From the top, the chromatograms of *EcAppA* wild-type, *EcAppA* HAE, *EcAppA* HDE and an IP<sub>6</sub> hydrolysate standard are displayed. It is possible to notice the appearance of secondary IP<sub>4</sub> peaks produced by *EcAppA* HAE.

### 3.6. Enzyme kinetics studies

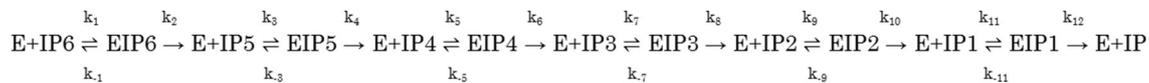
Enzyme kinetics is one of the most powerful tools for evaluation of mechanisms of catalysis and substrate specificity. The Michaelis-Menten equation,

$$v_0 = \frac{V_{max} * [S]}{K_M + [S]}$$

where  $v_0$  is the initial rate,  $V_{max}$  is the maximum rate,  $K_M$  is the Michaelis constant equal to the substrate concentration at  $1/2 V_{max}$  and  $S$  is the substrate

concentration, was devised in 1913 by Leonor Michaelis and Maud Menten and it is used in all the enzyme kinetics calculation throughout this thesis.

However, not all the assumptions on which MM equation is based, can be easily applied to this family of enzymes. In fact, if we formalise a typical phytase reaction of IP<sub>6</sub> hydrolysis as:



and we assume that:

$$k_{-1} \gg k_2, k_{-3} \gg k_4, k_{-5} \gg k_6, k_{-7} \gg k_8, k_{-9} \gg k_{10}, k_{-11} \gg k_{12}$$

in conditions of great excess of the primary substrate, IP<sub>6</sub>, over enzyme, the steady-state assumption on which the MM equation is based:

$$\frac{d[EIP_6]}{dt} = 0$$

is valid only at the very first stages of reaction, after the transient phase, but when IP<sub>6</sub> is still present in great excess over the first product, IP<sub>5</sub>, as the latter will be used as a substrate itself:

$$\frac{d[EIP_6]}{dt} \gg \frac{d[EIP_5]}{dt}$$

The condition cannot be easily achieved and checked, particularly at the substrate range 25-400 μM, due to the limitations of sensitivity the phosphate detection assay and its non-continuous (end-point) nature. We should also keep in mind that, despite the initial rate of reaction being under evaluation, the values obtained may not reflect the hydrolyses of the sole substrate, IP<sub>6</sub>, but of the hydrolyses of a mixture of IP<sub>6</sub> and multiple IP<sub>5</sub> products, whose rate constants may differ from the rate constant of phytate.

Reactions of 100 μL were set up in triplicate at fixed concentrations of enzymes (*EcAppA* wild-type and *EcAppA* HDE: 7.5 nM and the rest of the mutants: 1 μM) and increasing concentration of substrate (25, 50, 100, 200, 400, 600, 800, 1200 μM). They were incubated for 5, 10, 15, 20, 25, and 30 min at room temperature. The buffer chosen was the most used in the literature for this enzyme: 0.2 M NaAcetate pH 4.5, 0.15 M NaCl [97, 101-104]. Reactions were inactivated by addition of molybdenum blue reagent in equal part and the Abs<sub>700</sub> was measured after 30 min incubation with the stopping reagent. Data were processed with the 'nls' function provided in R (<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nls.html>), that determines the nonlinear least-squares estimates of the parameters of a nonlinear model. In this analysis, the non-linear model is the MM equation. The goodness of the model's fit is given by residual errors and t-test. Results are reported in **Table 3.6.1**.

Comparable  $K_M$  values were obtained for the wild-type enzyme and *EcAppA* HET, while the other mutants displayed an increase in binding affinity.  $K_M$  was 5-fold lower for the proton donor-less HAT and HAE double mutant. This may arise from difficult release of the hydrolysed substrate. A 1.7-fold decrease in  $K_M$  was measured for the HDE mutant. The  $k_{cat}$  was also strongly affected in particular for the mutants HAT and HAE which displayed a 970-fold and an 817-fold decrease in turnover number, followed by HET with an 80-fold decrease. The mutant with highest turnover number was *EcAppA* HDE with only a 6-fold reduction of  $k_{cat}$ . Among the mutants, *EcAppA* HDE displayed the highest catalytic efficiency, albeit 3.5-fold lower than the wild-type. The other remaining enzymes HET, HAT and HAE showed respectively an 85-fold, a 200-fold, a 170-fold decrease in  $k_{cat}/K_M$ .

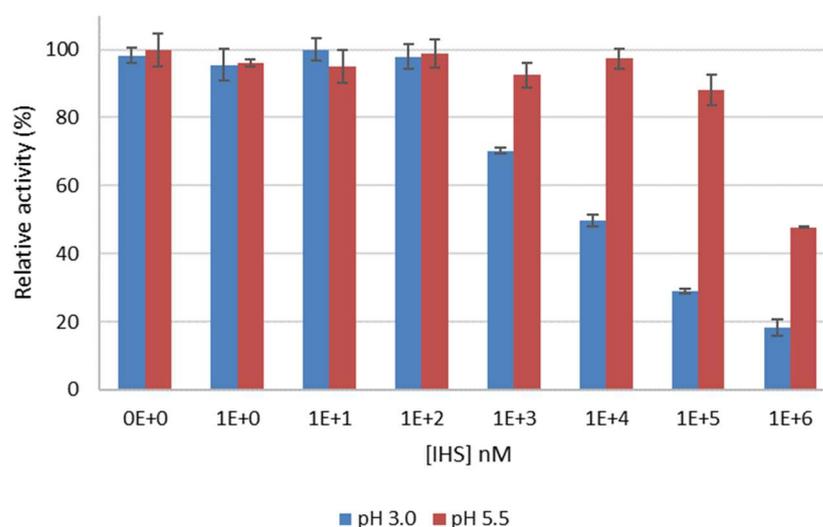
**Table 3.6.1. Kinetic parameters for the hydrolysis of inositol hexasulphate.**

Reactions were incubated for 5 min before inactivation.  $V_{max}$  is expressed in  $\mu\text{M}\cdot\text{min}^{-1}$ ,  $K_M$  in  $\mu\text{M}$ ,  $k_{cat}$  in  $\text{min}^{-1}$ , the turnover number  $k_{cat}/K_M$  in  $\text{min}^{-1}\mu\text{M}^{-1}$  and phytase units FTU in  $\text{U}\cdot\text{mg}^{-1}$ . Standard errors, the t-test values on the goodness of MM equation fit, t-value and P-value, are also reported.

Protein	E ( $\mu\text{M}$ )	Parameters	Values	St. Error	t-value	Pr(> t )
<i>EcAppA</i> w-type	7.5E-03	$V_{max}$	78	3	24.25	3.23E-07
		$K_M$	161	3		
		$k_{cat}$	10209	421	7.03	4.15E-04
		$k_{cat}/K_M$	64	9		
		FTU	228	9		
<i>EcAppA</i> HAT	1.0	$V_{max}$	10.5	0.4	24.60	2.96E-07
		$K_M$	32	7		
		$k_{cat}$	10.5	0.4	4.40	4.58E-03
		$k_{cat}/K_M$	0.32	0.07		
		FTU	0.24	0.01		
<i>EcAppA</i> HET	1.0	$V_{max}$	128	6	20.68	8.33E-07
		$K_M$	171	28		
		$k_{cat}$	128	6	6.12	8.72E-03
		$k_{cat}/K_M$	0.7	0.1		
		FTU	2.8	0.1		
<i>EcAppA</i> HDE	7.5E-03	$V_{max}$	12.7	0.2	72.16	4.77E-10
		$K_M$	92	5		
		$k_{cat}$	1689	24	17.66	2.12E-06
		$k_{cat}/K_M$	18	1		
		FTU	37.7	0.5		
<i>EcAppA</i> HAE	1.0	$V_{max}$	12.4	0.3	42.55	1.13E-08
		$K_M$	33	4		
		$k_{cat}$	12.5	0.3	7.66	2.59E-04
		$k_{cat}/K_M$	0.38	0.05		
		FTU	0.28	0.01		

### 3.7. Inhibition by inositol hexasulphate (IHS)

In many cases, IHS is reported to act as competitive inhibitor of HP phytases e.g. *E. coli* AppA (PDB id: 4TSR), *A. niger* PhyA (PDB id: 3k4q), *B. thetaiotaomicron* MINPP (PDB id: 4fdv) and others. This molecule is structurally identical to inositol hexaphosphate (IP<sub>6</sub>), with sulphate groups substituting phosphates, thus making it a good compound to be used in macromolecular crystallography to mimic the binding of the substrate. Soaking of *EcAppA* crystals with IP<sub>6</sub> was unsuccessful, probably because phytate is quickly hydrolysed. Attention was turned to IHS, but at pH close to neutrality (the pH of crystallisation), IHS did not bind to *EcAppA* to the extent of resolvable electron density. Successful binding was achieved instead at low pH (pH 3.5), as shown by the inhibition assay (**Figure 3.7.1**). The enzyme is tightly inhibited by the competitive inhibitor which displays an IC<sub>50</sub> (the inhibitor concentration able to decrease initial velocity by 50%) of ~10 μM.



**Figure 3.7.1. Inhibition assay by IHS.**

x axis: IHS concentration [nM], y axis: relative activity (%). In blue, the activity of reactions set up in Glycine buffer pH 3.0 is reported, while in red NaAcetate buffer pH 5.5 was used. When the concentrations of substrate and IHS in solution are equal (1 mM) relative activity is 80 % reduced: the inhibitor has a tighter binding than the substrate.

Reactions of 50 μL were set up in triplicate at fixed concentrations of enzymes (7.5 nM) and substrate (IP<sub>6</sub>: 1mM). Serial dilutions of IHS at concentrations: 0 nM, 1 nM, 10 nM, 100 nM, 1 μM, 10 μM, 100 μM, 1 mM were mixed to two buffered solutions of pH 3.0 (0.2 M Glycine-Cl, 0.15 M NaCl) and pH 5.5 (0.2 M NaAcetate, 0.15 M NaCl). Reactions are incubated at 37 °C for 30 min before inactivation by addition of molybdenum blue reagent and Abs<sub>700</sub> was measured after 30 min incubation with the stopping solution.

### 3.8. Crystallisation with inositol hexasulphate

Macromolecular structures of *EcAppA* were solved by Lim et al. (2000) and by Wu et al. (2015) for the wild-type enzyme apo, its complex with tungstate or IP<sub>6</sub> and for the H39A mutant in complex with IHS (**Table 3.8.1**). *EcAppA* crystallization and X-ray crystallography were employed again in this study for the identification of the structural changes determining differences in positional stereospecificity of IP<sub>6</sub> attack between wild-type enzyme and mutants.

**Table 3.8.1. *EcAppA* macromolecular structures deposited in the PDB.**

Structure	Operator	Summary
<i>EcAppA</i> : IHS PDB id: 4tsr	Wu <i>et al.</i> (2015)	Res.: 2.07-25.00 Å Space group: P2 <sub>1</sub> R-free: 0.22 R-work: 0.17
<i>EcAppA</i> apo PDB id: 1dkl	Lim <i>et al.</i> (2000)	Res.: 2.30-25.00 Å Space group: P2 <sub>1</sub> R-free: 0.20 R-work: 0.16
<i>EcAppA</i> : Hg <sup>2+</sup> PDB id: 1dkn	Lim <i>et al.</i> (2000)	Res.: 2.40-25.00 Å Space group: P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> R-free: 0.27 R-work: 0.22
<i>EcAppA</i> -tungstate: Hg <sup>2+</sup> PDB id: 1dko	Lim <i>et al.</i> (2000)	Res.: 2.38-25.00 Å Space group: P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> R-free: 0.23 R-work: 0.21
<i>EcAppA</i> H39A: IHP, Hg <sup>2+</sup> PDB id: 1dkp	Lim <i>et al.</i> (2000)	Res.: 2.28-25.00 Å Space group: P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> R-free: 0.23 R-work: 0.19
<i>EcAppA</i> H39A: IHP, Hg <sup>2+</sup> PDB id: 1dkq	Lim <i>et al.</i> (2000)	Res.: 2.05-25.00 Å Space group: P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> R-free: 0.26 R-work: 0.22
<i>EcAppA</i> : Hg <sup>2+</sup> PDB id: 1dkm	Lim <i>et al.</i> (2000)	Res.: 2.25-25.00 Å Space group: P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> R-free: 0.27 R-work: 0.22

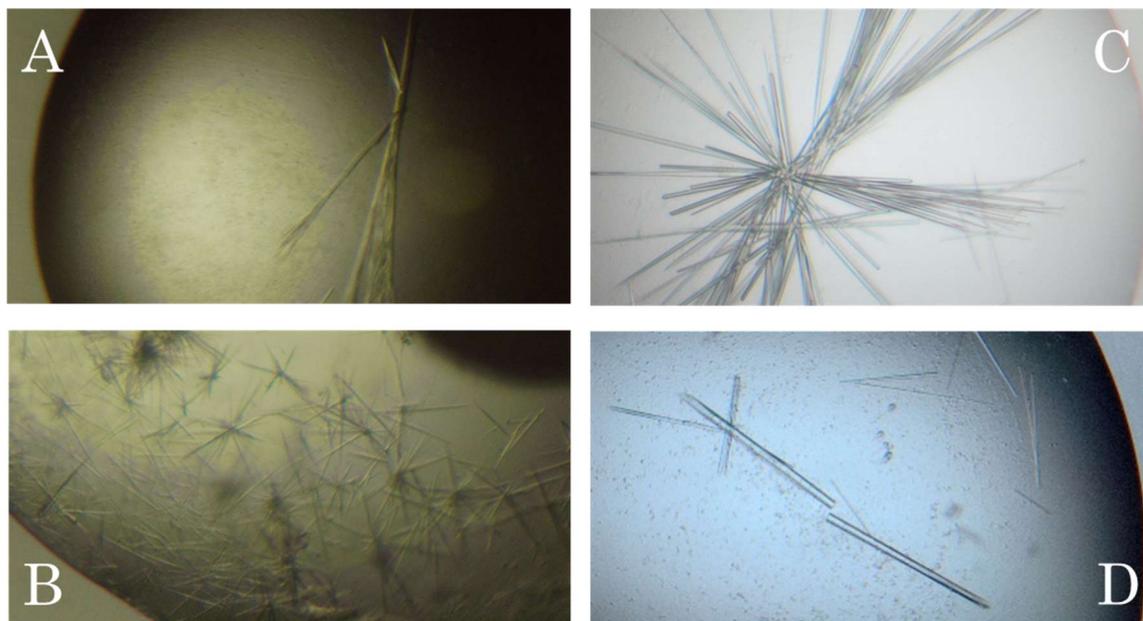
The aim of this work was to capture enzyme:IHS binding states for the mutants, in order to compare their X-ray crystal structures with that of the wild-type. *EcAppA* wild-type and mutants crystallized via sitting drop vapour diffusion at an optimal concentration of 2.0 mg/mL (in drop). At first, the crystallisation condition adopted by Wu et al (2015) was used (0.1 M HEPES pH 7.5, PEG 3350 (18 %, or 20 %, or 22 %), PEG 8000 (2 %), 0.01 or 0.02 M NiCl<sub>2</sub>) in 46-well plates trials manually set-up and incubated at 16 or 25 °C. Each drop (1 µL) contained equal quantities of

enzyme (0.5  $\mu$ L) and precipitant solution (0.5  $\mu$ L), each well was filled with 300  $\mu$ L precipitant solution. The protein buffer was 0.2 M NaAcetate pH 4.54. The crystals obtained, however, were very thin needle, often growing from the same nucleation centre (“broom shape”). Despite this, they diffracted at a resolution of 2 Å, but they were hard to harvest and, in particular, to soak with IHS (**Figure 3.8.1.A and B**). In an attempt to improve crystal morphology and to obtain an enzyme:IHS complex, micro-seeding was employed in co-crystallisation plates at a final drop concentration of substrate analogue of 1mM. The protein buffer was exchanged for 0.1 M HEPES pH 7.5, 0.15 M NaCl, yielding larger, fragile single crystals of plate shape. Despite being fragile, they were harvested, soaked with 5 mM IHS in well solution diluted with 30 % glycerol, snap frozen and X-ray diffraction data were collected. The resolution improved reaching 1.8 Å, though, no enzyme:inhibitor complex was detected bound in the active site cleft.

A further trial was set up, this time searching for new crystallisation conditions in the commercial crystallization matrices Structure Screen I/II and JCSG Plus. Crystallographic 96 wells trays (Molecular Dimensions, Newmarket, UK) were set up using an OryxNano protein crystallization robot (Douglas Instruments Ltd) for enzyme crystallization, again, via sitting drop vapour diffusion in the presence of the substrate analogue inositol hexasulphate (IHS). Seeding was also performed via OryxNano. Two drops of 0.6  $\mu$ L per crystallization condition were set up containing, respectively, equal quantities of enzyme-inhibitor solution and precipitant solution (0.3  $\mu$ L x 2), or enzyme-inhibitor solution (0.3  $\mu$ L), precipitant solution (0.2  $\mu$ L) and seed stock (0.1  $\mu$ L). Each well was filled manually with 50  $\mu$ L precipitant solution. The plates were incubated at 25 °C. Protein concentration was 1.6 mg/mL and the buffer used was 50 mM Tris-HCl pH 7.5, 0.15 M NaCl. Final concentration of IHS in the protein solution was 1 mM. Crystals grew in multiple conditions and morphologies (**Figure 3.8.1.C and D**). The best for quality of diffraction and stability were thicker, needle-shaped single crystals grown in cobalt (II) chloride hexahydrate 5 mM, cadmium chloride hemi(pentahydrate) 5 mM, magnesium chloride hexahydrate 5 mM, nickel(II) chloride hexahydrate 5 mM, 0.1M HEPES pH 7.5, 12 % w/v PEG 3350 (highest resolution: 1.19 Å) or 0.1M MES pH 6.5, PEG 8000 (18 % w/v), 0.2 M calcium acetate hydrate (highest resolution: 1.37 Å).

To increase the chances of success in capturing the enzyme:IHS complex, a soaking experiment was also performed at low pH (pH 3.0-3.5), at which, the strongest enzyme inhibition was measured. Different durations of crystal incubation in the soak solution were also tested. The optimal soak in both growth conditions was the

following: 20 % PEG 3350, 9 mM IHS in 0.6 M glycine pH 3.0, 20 % glycerol for 10-25 min. It ensured a change in pH by an increased buffer strength. IHS was maintained at its maximum soluble concentration (9 mM at 16 °C) by diluting a 42 °C preincubated 30 mM stock solution of substrate analogue. Glycerol was used as cryoprotectant in combination with PEG 3350, whose increased concentration, was aimed to prevent crystals dissolution. Crystals were harvested on LithoLoops and stored in liquid nitrogen. Diffraction data collection was carried out at the Diamond Light Source (DLS) (Oxfordshire, UK) on beamlines i03 and i04.



**Figure 3.8.1. *EcAppA* crystals.**

**Pictures A and B:** results from initial crystallisation trials using the growth condition of Wu et al (2015, PDB is: 4tsr). **Pictures C and D:** improved crystal morphology obtained by changing crystallisation conditions, cross-seeding, and co-crystallisation with the substrate analogue IHS. The crystal of figure D was cut to allow the harvesting of the two halves.

### 3.9. X-ray crystal structure determination

A total of 50 datasets for *EcAppA* wild-type and mutants were collected at 100 K: 12 at resolutions between 1.2-1.6 Å (**Figure 3.9.1**), 21 between 1.6-2.2 Å, 12 between 2.2-3.0 Å and the remainder with lower resolution (**Table 3.9.1**). Of these datasets, 15 were of *EcAppA* wild-type or mutants in complex with IHS. Eleven contained Pi or an unknown compound bound to the active site, and 16 were apo-protein structures. The X-ray diffraction images were automatically integrated and scaled. Datasets with the best diffraction statistics were chosen for structure solution. Diffracted crystal had all a P<sub>2</sub><sub>1</sub> space group and a cell close to the average dimensions of a: 64 Å, b: 48 Å, c: 55 Å,  $\alpha$ : 90 °,  $\beta$ : 101 °,  $\gamma$ : 90 °. Crystals contained one molecule per asymmetric unit. The phase problem was solved through molecular replacement.

The *EcAppA* mutant complex with IHS by Wu *et al* (2015) (**Figure 3.9.2.A**) was edited to remove water molecules, ions, ligand and was used as a search model in phasing by PHASER [81]. Phasing procedure was straightforward and provided acceptable solutions which were subsequently manually remodelled using WinCoot [87] and refined with phenix.refine [89]. The ligand .cif files used in refinement of the structure was obtained from the WinCoot dictionary. Ligand occupancies were refined. Also, in the refinement of low resolution structures, weight optimisation (X-ray/stereochemistry, X-ray/ADP) and secondary structure restraints were used.

The substrate analogue was usually found with its sulphate groups in a 5-equatorial/1-axial conformation. However, in the structure solved for the mutant *EcAppA* HET, a molecule in a conformation of 5-axial/1-equatorial was also detected at the entrance of the active site, just above the 5-equatorial/1-axial IHS molecule, and both interacting with Lys46 (**Figure 3.9.2.B**). In all the structures, IHS can be modelled with the monosulphoester group at position 6 of the inositol ring pointing towards the catalytic histidine His39 (**Figure 3.9.2.C**). However, the density for the IHS molecule bound to the active site of *EcAppA* HET in 5-equatorial/1-axial conformation is not well defined. The enzyme might be able to host the substrate analogue in multiple orientations. In fact, the monosulphoester groups at position 3 or 6 of the inositol ring of IHS were modelled in apposition to His39. Occupancies and temperature factors for the conformers were refined (**Table 3.9.2**). Temperature factors for the structure of the E:IHS complex of *EcAppA* HET and a ligand omit single difference Fourier map for the E:IHS complex of *EcAppA* HAE are reported in **Figure 3.9.2.C**. Ramachandran statistics for all non-Pro/Gly residues and refinement statistics for a few selected structures are reported in the **Table 3.9.2**.

Protein	Nr. Structure	Nr. Structure per Res.			IHS bound	Pi bound	apo-protein
		1.2-1.6 Å	1.6-2.2 Å	2.2-3.0 Å			
AppA w-type	7		4	3	1	1*	3
AppA D326A	20	3	7	6	4	3+7*	2
AppA D326E	10	5	3	2	1		9
AppA T327E	6		5		5		
AppA D326A T327E	7	4	2	1	4		2
Total	50	12	21	12	15	11	16

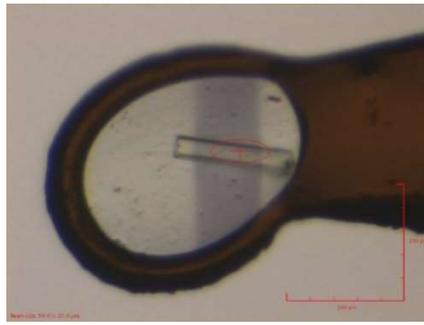
**Table 3.9.1. Summary of the macromolecular structures collected for *EcAppA* wild-type and mutants.**

50 crystal structures were solved: 12 at resolution (1.2-1.6 Å), 21 at resolution (1.6-2.2 Å), 12 at resolution (2.2-3.0 Å), structures with lower resolution are not included. 15 data-sets were collected for enzyme:IHS complexes, 16 for apo enzymes, the rest were enzyme:Pi complexes or enzymes covalently linked with unknown intermediates.

Data set	Type	Res (Å)	Ligand	Ligand occ. (%)	R-work (%)	R-f ree (%)	Ramachandran analysis (%)
AppA_3_1_3	wt	49.8- 1.85	Pi	67	17.38	21.60	97.77 - 1.49 - 0.74
APPA_14_1A_2	HAT	64.89 - 2,08	Pi	79	19.52	23.10	98.73 - 1.27 - 0.00
APPA_16_1_2	HAT	50.05 -1,85	H17-Pi-H303	100	18.60	21.54	99.02 - 0.98 - 0.00
APPA_20_3_5	HAE	62.77 - 1,92			21.75	25.93	98.98 - 1.02 - 0.00
AppA_31_1_2	HET	49.53 - 1,37			17.53	19.52	98.99 - 1.01 - 0.00
AppA_36_2_2	HET	26.30 - 2,56	IHS IS6	59 / 41 48 - 100	20.32	25.95	96.27 - 3.48 - 0.25
AppA_38_1_2	wt	38.36 - 1,76	IHS	58	17.74	20.58	98.75 - 1.25 - 0.00
AppA_38_2_2	wt	38.21 - 1,99			21.18	25.34	98.48 - 1.27 - 0.25
AppA_40_1_2	HAT	41.13 - 1,41	IHS	97	16.30	19.17	98.99 - 0.76 - 0.25
AppA_40_4_2	HAE	41.32 - 1,51	IHS	90	16.46	18.57	98.99 - 1.01 - 0.00
AppA_41_1_2	HAE	41.09 - 1,19	IHS	83	18.69	19.97	98.99 - 0.76 - 0.25
APPA_46_1_4	HDE	64.30 - 1.86	IHS	100	20.24	24.21	97.77 - 1.74 - 0.50

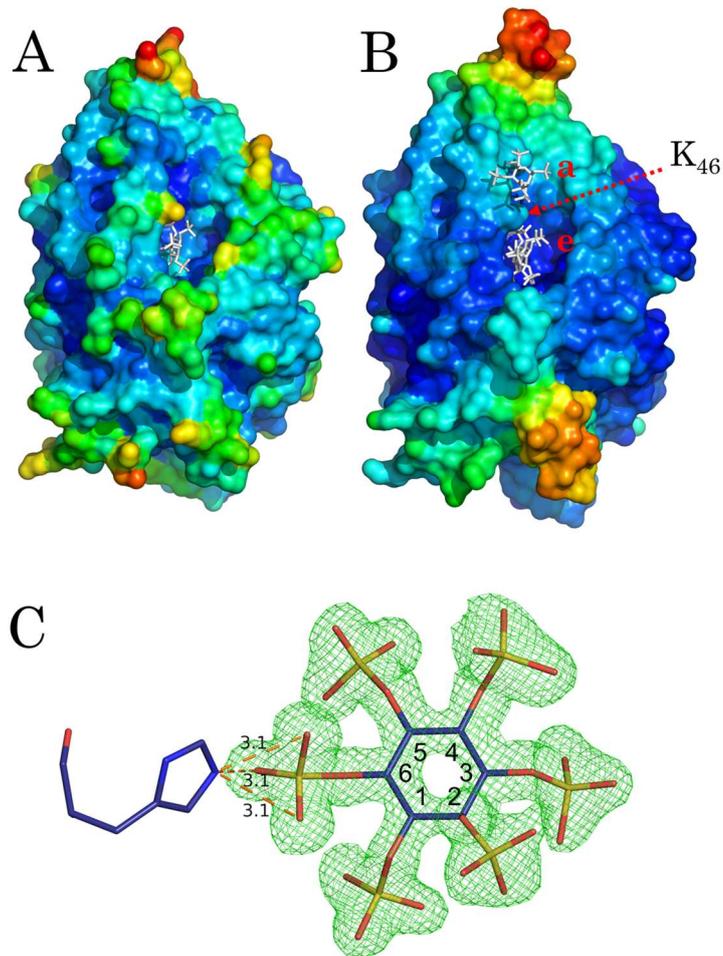
**Table 3.9.2. Refinement statistics of selected *EcAppA* wild-type and mutant structures.**

Resolution is given for low res. beam - high res beam. Ligand occupancies (%), R-work and R-free (%) are reported. Ramachandran % of non-Pro/Gly residues that fall in energetically favoured or allowed regions and outliers for the backbone  $\phi$  and  $\psi$  are displayed (favoured - allowed - outliers).



**Figure 3.9.1. EcAppA:IHS crystal on LithoLoop after diffraction.**

The picture of the left captures a crystal of around 100x20x10  $\mu\text{M}$  and space group P21 which diffracted at a high resolution of 1.19 $\text{\AA}$ . The scale, in red, indicate 100  $\mu\text{m}$ .

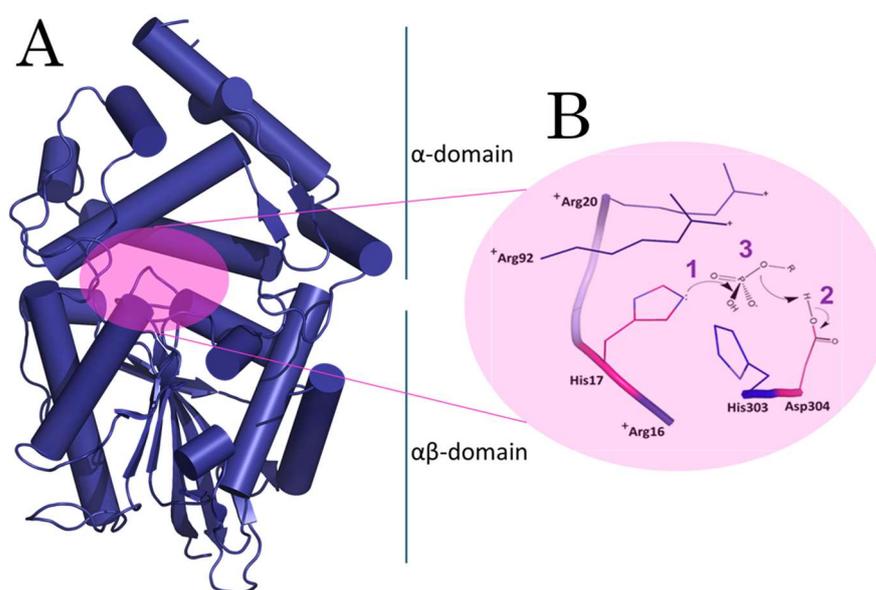


**Figure 3.9.2. EcAppA PDB id: 4TSR, EcAppA HET:IHS complex and omit single-difference Fourier map of the substrate analogue IHS in EcAppA HAE mutant.**

**A/B)** Surface representations of the structures 4TSR (PDB ID) and EcAppA HET:IHS complex, coloured by temperature factors (gradient blue to red from low to high temperature factors). Structures are aligned and the substrate analogue, IHS is displayed by white sticks. **A)** IHS binds in a 5-equatorial/1-axial sulphate conformation. **B)** IHS binds in a 5-axial/1-equatorial sulphate conformation (**a**) and in a 5-equatorial/1-axial sulphates conformation (**e**). K46 contributes to the coordination of the two binding modes. **C)**  $F_o - F_c$  omit map of the density surrounding IHS (contour level  $\sigma = 3.0$  r.m.s.d.) in EcAppA HAE mutant and the corresponding models of the substrate analogue. The catalytic histidine His17 makes contacts to the sulphate group bound at C6 of the inositol ring of IHS.

### 3.10. Structural analysis

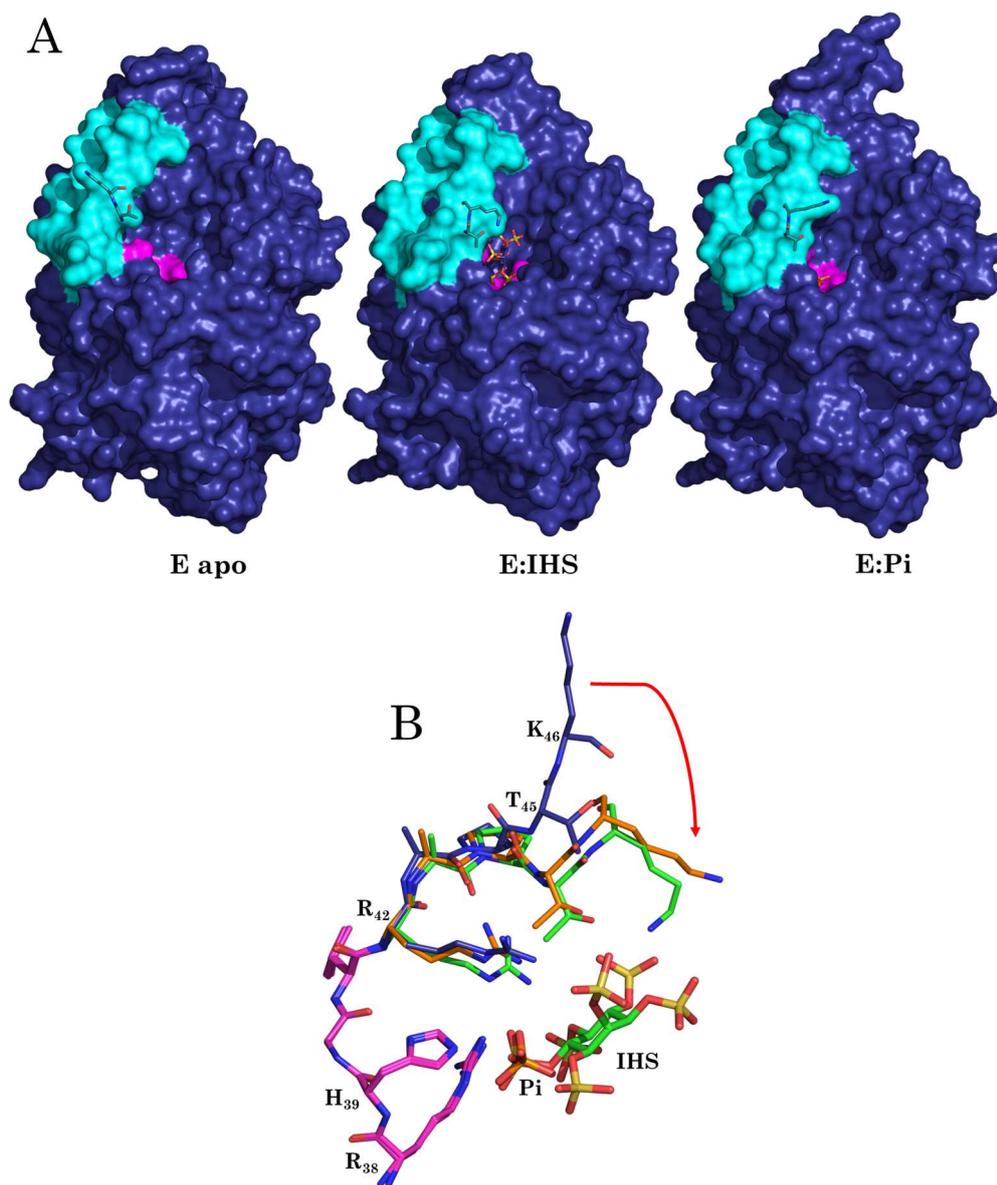
The active site of the histidine phosphatase (HP) phytase, *EcAppA*, lies between two structural domains: an  $\alpha$ - and an  $\alpha/\beta$ -domain (**Figure 3.10.1.A**). When  $\text{IP}_6$  enters the active site cavity, His39 initiates hydrolysis by attacking one of the Pi groups of the substrate (**Figure 3.10.1.B, step 1**). Asp326 acts as a proton donor (**Figure 3.10.1.B step 2**) allowing the release of  $\text{IP}_5$  (or lower intermediate species) and the formation of a phosphohistidine intermediate (**Figure 3.10.1.B step 3**). Asp326 then acts as a general base in the dephosphorylation of His39 and release of inorganic phosphate [24-27, 31].



**Figure 3.10.1. Structural overview of *EcAppA*.**

A) The overall structure of the enzyme comprises an  $\alpha$ -domain and an  $\alpha/\beta$ -domain. The active site cleft is positioned at their interface. B) A close-up of the catalytic residues in the E:IHS form (PDB id: 4tsr). Hydrolysis occurs in 3 steps: 1 – nucleophilic attack of His39, 2 – proton donation by Aps326 and substrate release, 3 – formation of a phopho-histidine intermediate and regeneration of the enzyme by donation of the Pi group to water or a different acceptor.

The binding of IHS involves the rearrangement of the region facing the active site of the loop R42-T48 ( $\alpha$ -domain) to allow residues R42, T45 and K46 to make contacts with the substrate analogue [24]. **Figure 3.10.2.A** displays the overall conformational change of the loop (in light-blue) from the apo enzyme form (PDB id: 1dkl) to the enzyme:IHS bound form (E:IHS) and to the enzyme:Pi bound form (E:Pi). The amino acids displaying the largest displacement (15.5 Å) are T45 and K46 which “close” the active site cleft over the substrate (displayed by sticks). **Figure 3.10.2.B** presents a close-up of the region of the active site involved in the movement.



**Figure 3.10.2. Structural overview of conformational changes in *EcAppA*.**

**A)** Surface representation of the apo-enzyme (PDB ID: 4TSR), the enzyme:IHS complex AppA\_41\_1\_2, the enzyme:Pi complex AppA\_3\_1\_3. The loop which undergoes a conformational change is displayed in light-blue; the catalytic core is magenta coloured; Thr45 and Lys46 are represented by sticks. **B)** close-up of the active site region involved in the movement. Fixed region is coloured magenta for all the three structures; the moving residues are coloured in blue (apo-enzyme), in orange (enzyme:Pi) or in green (enzyme:IHS).

The catalytic core includes: **R38** – which establishes H-bonds with the sulphate groups at position C6 and C5 of IHS, with the proton donor D326, with D241 and D325; the catalytic histidine **H39** – that lies within 3 Å from the sulphate group at position C6 of IHS; **R42** – which establishes H-bonds with the sulphate groups at position C6 and C5 of IHS; **R135** – which establishes H-bonds with sulphate groups at position C6 and C1; **H325** and the proton donor **D326** [31] – which lies 2.5 Å from an hydroxyl group of the sulphate bound at position C5 of the inositol ring of IHS.

It is probable that additional amino acids have a role in determining substrate orientation and, potentially, its positional specificity of hydrolysis. To better study the active site, residues were grouped in binding pockets within 6 Å from each sulphate group of IHS in the structure PDB ID: 4TSR. A pocket was built around each sulphate of IHS, as represented in **Figure 3.10.3**. Binding pockets were named A to F clockwise, starting from the sulphate group in position C6 of inositol ring of IHS, which makes contacts with the Nε2 of the nucleophilic histidine His39. Pocket A is the pocket containing residues involved in acid-base catalysis.

The number of amino acids in each pocket are unevenly distributed, in fact, pocket A, B and C, which coordinates the sulphate groups in position C6, C5 and C4 of the inositol ring of IHS, contain 10, 13 and 11 residues, respectively, while pocket D, E, F, which correspond to the sulphate in position C3, C2 and C1, contain 3, 5 and 8 residues. Amino acids can contribute to more than one pocket. This is the case for example of T23, which lies within 6 Å from all the sulphate groups of IHS and it is involved in the conformational change of the R42-T48 loop. K46 is included in pockets C, D, E and undergoes a large conformational change upon IHS binding. K46 is the only residues outside the catalytic core able to establish a hydrogen bond with IHS at the sulphate group in position C3 of the inositol ring. Moreover, in *EcAppA* HET, it coordinates a second IHS molecule bound to the enzyme. These facts taken altogether seem to suggest an involvement of K46, an amino acid exposed on an intrinsically flexible region of the protein surface, in the first recruitment of the substrate before its orientation in the active site cleft.

In the solution of *EcAppA* wild-type and mutant structures, pocket B was identified as prone to chelate ions by establishing 4/5 coordinations. Amino acids involved in the binding were H272 (1 coordination), D347 (2 coordinations), Thr349 (2 coordinations) and, less often, the proton donor Asp326 (1 coordination). Coordinations have an average length of 2.4 Å. An ion or water is present in this area in the majority of the solved structures. Ions were predicted to be Ca<sup>2+</sup>, Mg<sup>2+</sup>, Ni<sup>2+</sup>, K<sup>+</sup>, Cd<sup>2+</sup>, depending on the composition of the crystallisation condition used and on the information elaborated by phenix.refine [90]. Ions were proved to interfere with *EcAppA* activity [103]. This might happen when ions are chelated in this active site pocket and, therefore, able to interact with the proton donor to tune enzyme activity by favouring or contrasting the release of the product.

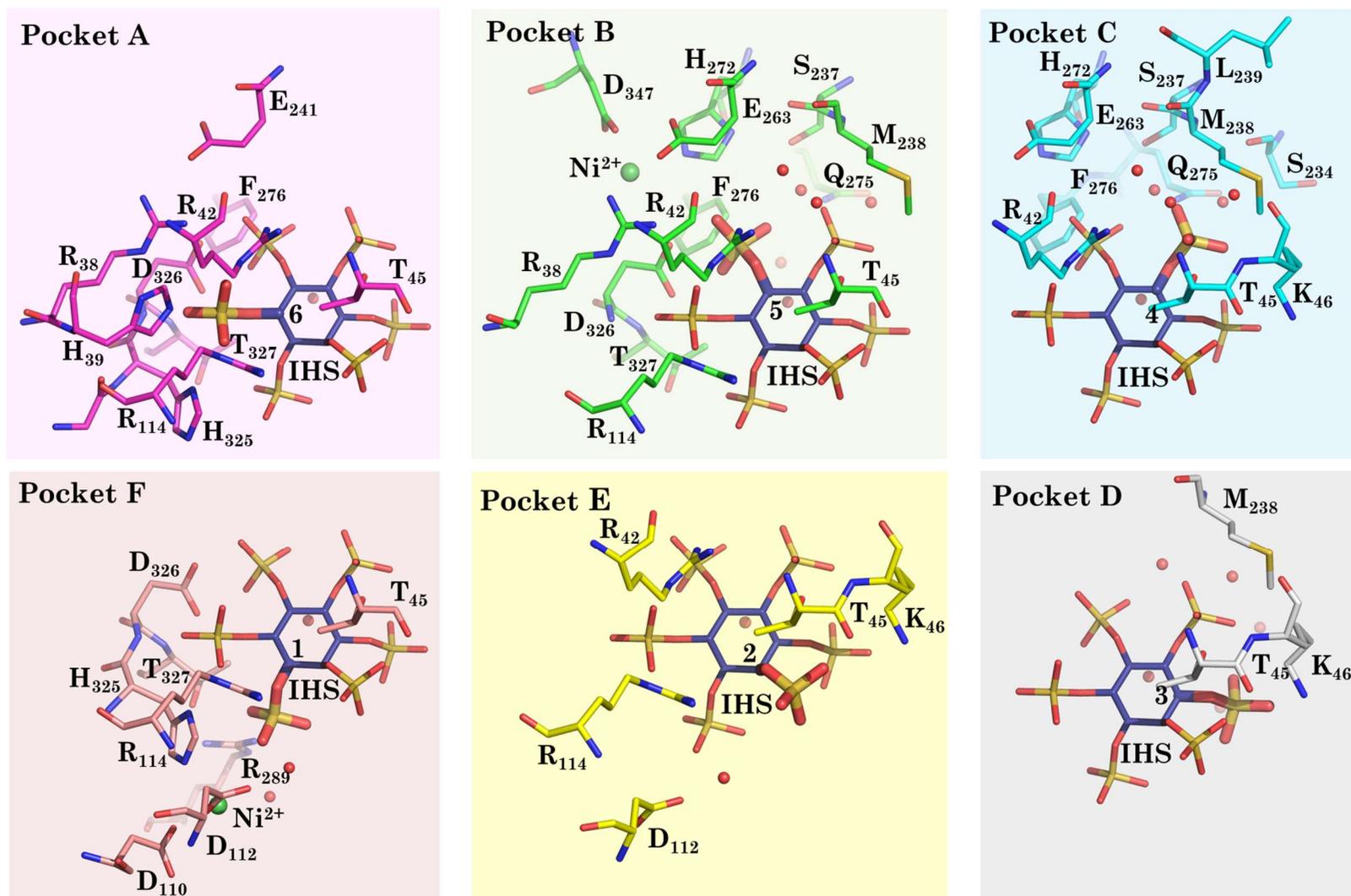


Figure 3.10.3. Active site pockets of *EcAppA*.

### Figure 3.10.3. Active site pockets of *EcAppA*.

Each binding pocket contains residues within 6 Å from the sulphate group of IHS (highlighted by wider stick radius). Many residues are present in multiple pockets. The numbering of carbon residues of the inositol ring follows an anticlockwise order while active site pockets are named clockwise. Water molecules are displayed by red dots, Ni<sup>2+</sup> ions by green dots.

After this preliminary analysis, the effect of mutations in the proton donor motif was evaluated. Representatives of the structures of the complex E:IHS for the wild-type and each mutant were chosen based on the quality of X-ray data and structure solution statistics. Their superposition by DynDom revealed that no conformational change of the overall structure occurred as consequence of mutations. An r.m.s.d. of the whole protein best fit of 0.58 Å for *EcAppA* HET (the widest), 0.26 Å for *EcAppA* HAT and HAE, 0.24 Å for *EcAppA* HDE was calculated. Inspection of residues in the active site cavity of E:IHS complexes and apo-structures confirmed the reorientation of the  $\alpha$ -domain residues H39-T48, R114, E241, H272, F276, H325, D326 upon IHS binding (**Figure 3.10.3**, Pocket A).

A comparison between the binding site pockets of the collected E:IHS complexes of *EcAppA* wild-type and mutant revealed changes in size and polarity which result in shifts of the substrate analogue orientation in the active site cleft.

*EcAppA* HDE, characterized by initial production of IP<sub>5</sub> in ratio 94 % 4- or 6-OH IP<sub>5</sub> and 6 % 1- or 3-OH IP<sub>5</sub> (**Figures 3.5.2** and **3.10.5**), presents slightly larger pockets A and B because of the loss of the side-chain non-polar -CH<sub>3</sub> group of T23. Instead, pocket F decreases in size and its polarity increases by virtue of addition of the glutamate at position 327. IHS was bound to this mutant directing its sulphate in position 6 towards the nucleophilic histidine (**Figure 3.10.4**, *EcAppA* HDE).

The double mutant *EcAppA* HAE, which generates 78 % of 4- or 6-OH IP<sub>5</sub> and 22 % of 1- or 3-OH IP<sub>5</sub> during IP<sub>6</sub> hydrolysis (**Figures 3.5.2** and **3.10.5**), displays even larger pockets A and B with an overall change in polarity because, except the catalytic core histidine and arginines, only the backbone nitrogen of A326 is left in these pockets to interact with IHS (2.9 Å distance). However, E327 is found in multiple conformations in the structures collected, either facing the substrate from pockets A and B (at 2.7 Å and 3.3 Å from the sulphate group in position C1 or C5 of the IHS inositol ring respectively – the monosulphoester at C6 points towards H39) or directed at the back of pocket F (at 3.8 Å from the sulphate group in position C1). These multiple orientations of the proton donor and its ability to contact two sulphate groups reshaping one side or the other of the active site cavity may lie behind the increased ability of the mutant to produce 1- or 3-hydroxy IP<sub>5</sub> intermediates (**Figure 3.10.4**, *EcAppA* HAE).

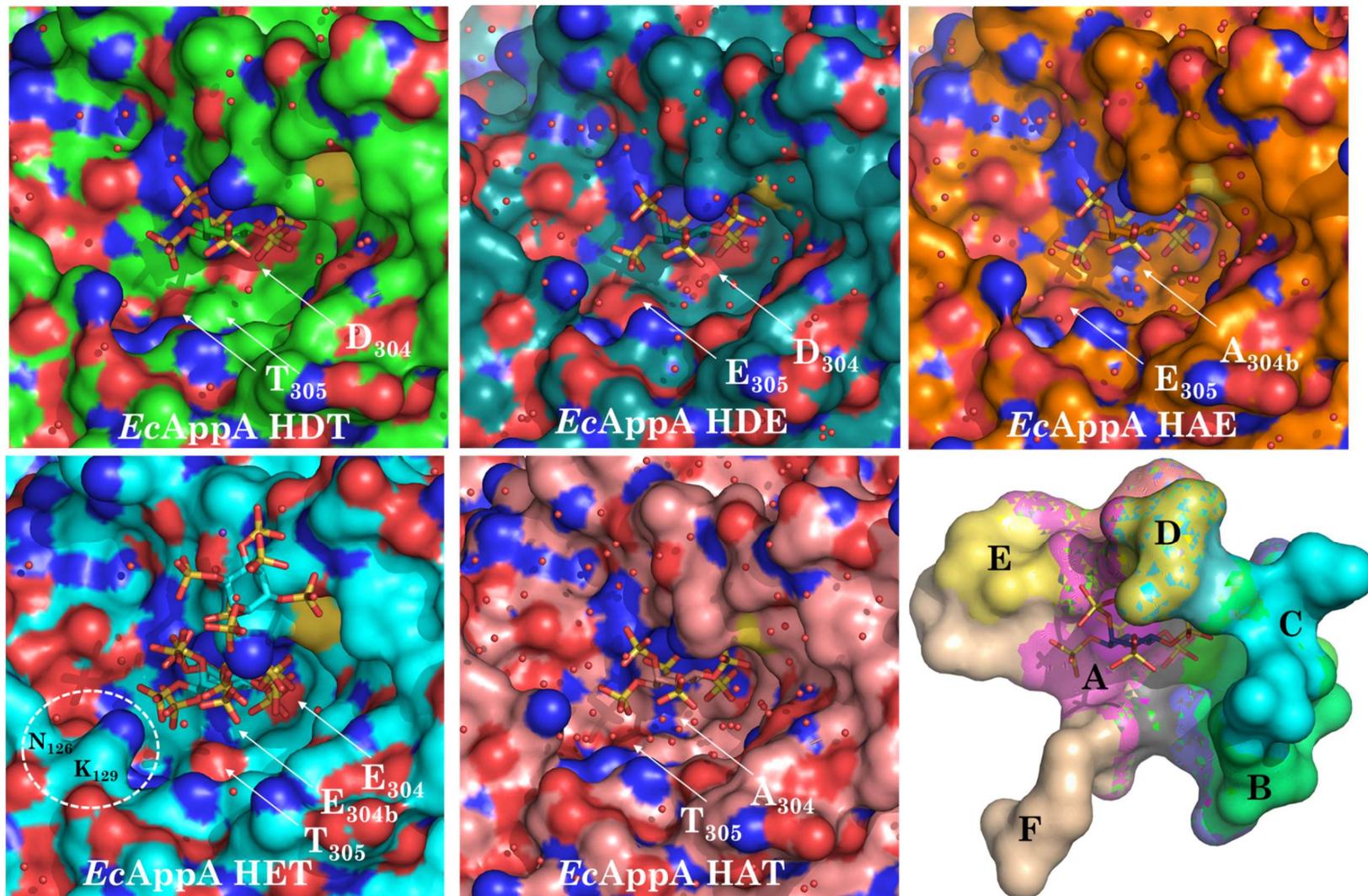


Figure 3.10.4. Active site representation of the E:IHS complex for w-t and mutants.

**Figure 3.10.4. Active site representation of the E:IHS complex for wild-type and mutants.**

The surfaces of enzymes active sites are represented here in the five panels. Red patches correspond to oxygen, blue to nitrogen, yellow to sulphur, and carbons are coloured by enzyme. Red dots are water molecules. IHS is represented by sticks. On the bottom right, the active site pockets volumes for the wild-type enzyme are given. Areas with multiple fragmented colours are shared between pockets.

The mutant *EcAppA* HET, which produces an almost equal amount of 4- or 6-OH IP<sub>5</sub> and 1- or 3-OH IP<sub>5</sub> during IP<sub>6</sub> hydrolysis (**Figures 3.5.2 and 3.10.5**), shows the greatest reshaping of the active site, something which may allow IHS to bind in multiple conformation. In the X-ray structure solved for this mutant, IHS is modelled to direct its C6-sulphate towards the catalytic histidine. The enzyme displays larger A and B pockets because of the absence of the aspartate side chain in the surrounding of IHS. In fact, glutamate position its carboxylic group at the back of the pockets establishing only weak interactions with the sulphate in position C5 of IHS (3.9 Å from IHS). On the other end, pocket F is reduced in size because Asn126 and Lys129 are redirected: they now lie at a distance of 5.3 Å and 4 Å, respectively, to the sulphate of pocket F. Thr327 and Asp90 also appear to be closer to the substrate analogue. A change in polarity occurs in pockets B and F because of the distal position of the carboxylic group of Glu326 and the introduction of the amino group of Lys129. All these changes seem to have produced overall an active site able to accommodate multiple binding modes and to ensure proton donation by Glu326 or possibly other residues of the active site (e.g. Asp90, Ser215) (**Figure 3.10.4, *EcAppA* HET**).

The mutant *EcAppA* HAT, characterized by a production of 11 % of 4- or 6-OH IP<sub>5</sub> and of 89 % of 1- or 3-OH IP<sub>5</sub> (**Figures 3.5.2 and 3.10.5**), shows an enlargement of pocket A and B as well, again, due to the absence of the aspartic acid side chain. The loss produces a change in polarity of these pockets because, similarly to the double mutant, the only hydrogen bond left between the proton donor triplet and IHS is between the backbone nitrogen of Ala326 and the sulphate group of IHS in pocket A (2.9 Å distance). The other pockets remain unchanged, however, the binding of the substrate (which points its sulphate group bound to C6 of the of IHS towards His17) is shifted. This allows a closer interaction with Thr327 (3.1 Å distance to IHS) and with two water molecules, one coordinated by Ser215 (2.8 Å distance to H<sub>3</sub>O<sup>+</sup>) and the other coordinated by His250 (2.5 Å distance to H<sub>3</sub>O<sup>+</sup>), Asp327 (2.9 Å distance to H<sub>3</sub>O<sup>+</sup>) and Thr327 (2.8 Å distance to H<sub>3</sub>O<sup>+</sup>). The first water molecule lies at a distance of 2.9 Å from the sulphates in pockets B and C. The second water is positioned at 3.7 Å from the sulphate in pocket B. Thr327 and these two water molecules, in particular the one coordinated by Ser215, could potentially be acting as proton donor in place of Asp326 (**Figure 3.10.4, *EcAppA* HAT**).

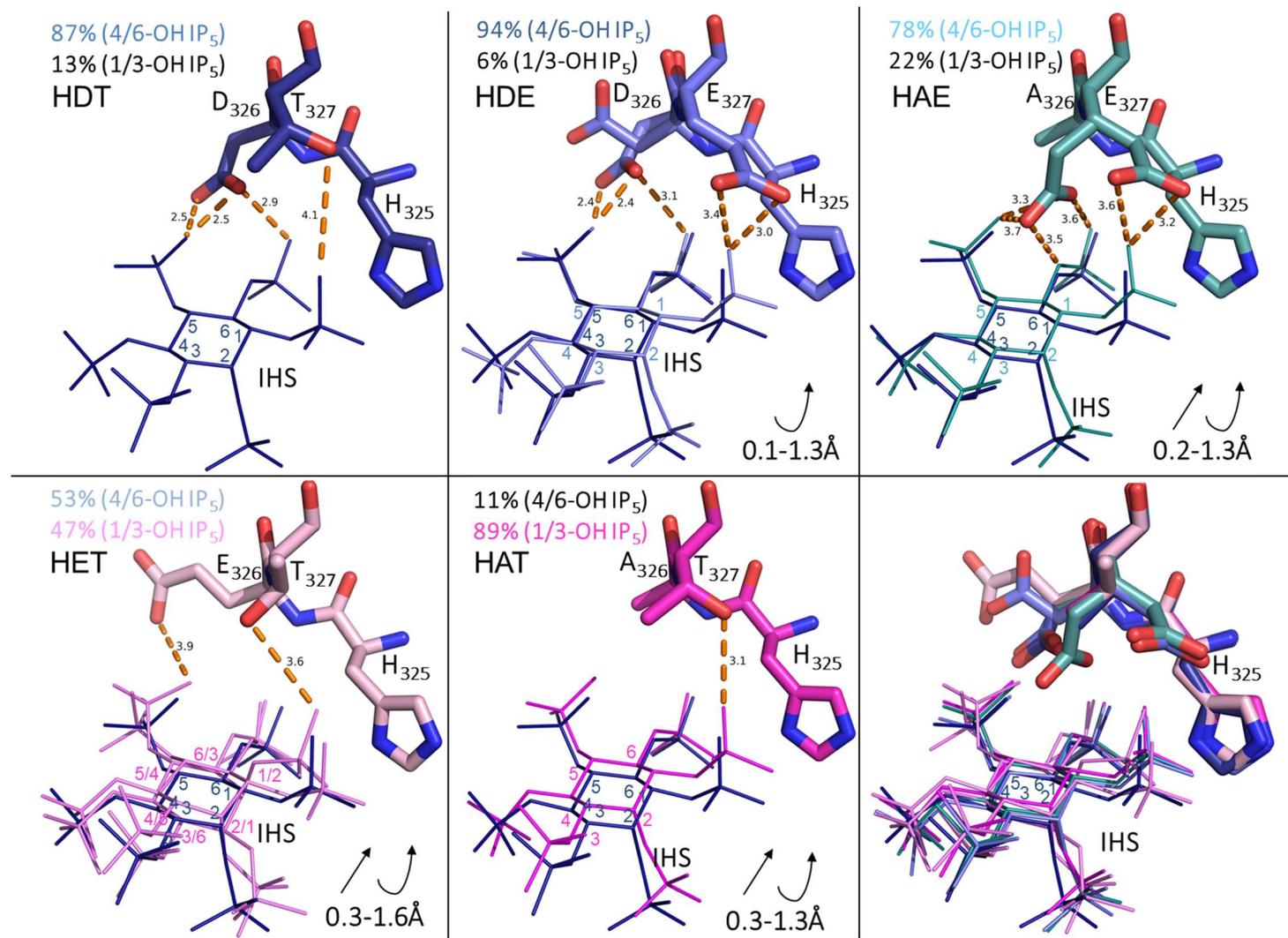


Figure 3.10.5. Shifts in binding of IHS to the wild-type and mutant enzymes.

**Figure 3.10.5. Shifts in binding of IHS to the wild-type and mutant enzymes.**

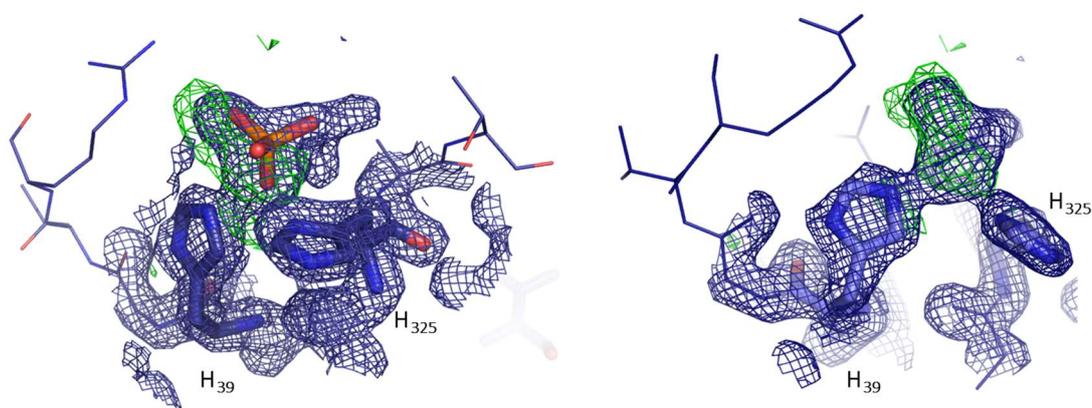
The picture displays the binding of IHS to the all the proton donor triplets HDT (wild-type), HDE, HAE, HET, HAT and the superposition of all binding poses. IHS is coloured in dark blue, when bound to the wild-type enzyme. This orientation is superposed to the IHS binding poses found in *EcAppA* mutants (blue to pink gradient depending on prevalence of IP<sub>5</sub> 4/6-OH or 1/3-OH production, respectively). IHS carbon numbering is reported for each pose following the colour scheme of IHS. For each enzyme, the % of IP<sub>5</sub> peaks produced in IP<sub>6</sub> hydrolysis is reported. Straight and curved arrows represent, respectively, translational and rotational movement. Minimum and maximum atoms shifts are indicated. Atoms distances are reported by dashed lines.

A closer view of the IHS binding modes in each mutant macromolecular structure is given in **Figure 3.10.5**. The shift in substrate binding appear to vary from a minimum of 0.1 Å to a maximum of 1.1 Å per phosphate group. At first glance, this change may not be considered significant. In particular, when noting that the IHS pose does not vary together with the changes in positional specificity of hydrolysis of the enzymes. Indeed, in the active site of all the *EcAppA* wild-type and mutant structures (including *EcAppA* HAT, which displays a preference for the hydrolysis of the phosphates at position 1 and/or 3) is present an IHS directing its phosphate group at position 6 of the carbon ring towards the catalytic histidine. Doubts on the substrate-mimicking power of the IHS analogue molecule could therefore arise. However, docking studies show that IP<sub>6</sub> tend to adopt the same conformation as IHS in the active site of all the structure of mutants:IHS complexes, where the latter was removed previous docking. This suggests that IHS could be considered an optimal analogue of IP<sub>6</sub>. Possible reasons behind the unexpected preference of all the mutants for a binding through the C6-sulphate of IP<sub>6</sub> to the nucleophilic histidine may be: (1) this binding mode is energetically preferred in protein crystals, but for enzymes in solution, positional specificity of IP<sub>6</sub> hydrolysis is changed; (2) an alternative mechanism of product release occur in the HAT mutant, where the preferred binding pose of IP<sub>6</sub> is unchanged, but hydrolysis could only take place when other, maybe less favoured, binding pose occur in solution, possibly because an alternative proton donor comes at reach.

Despite there may not be a relation with IP<sub>5</sub> peaks population and IHS binding pose, the change in polar contacts formation and/or breakage due to the introduction of mutations in the proton donor area seems, in fact, in line with enzyme kinetics (**Figure 3.10.5** and **Table 3.6.1**). The wild-type proton donor triplet allows the formation of polar contacts between Asp326 and the sulphate in pocket F (2.5 Å) and A (2.9 Å). In the *EcAppA* HDE mutant, the monosulphoester group of pocket F has shorter interactions with Asp326 (2.4 Å) and this is in line with the high turnover number of this enzyme in comparison with the other mutants. Instead, in *EcAppA*

HAE and HAT, the proton donor to IHS distance widen and the turnover number drops: probably the proton donor triplet is no longer able to efficiently participate in acid-base catalysis leading, as a consequence, to a decrease in  $K_M$  due to difficulties of the mutants in the substrate release [31].

The former suggestion that the higher affinity of *EcAppA* HAT for substrate may reflect poorer release of product (the formation of the phospho-enzyme intermediate is the rate-limiting step [31]) is supported by the presence, in multiple structures (seven), of a bridge linking an unknown ligand to the catalytic histidine His39 and the histidine of the proton donor triplet His325 (**Figure 3.10.6**). This ligand, which apparently seemed covalently bound to the enzyme, could represent a transition state of reaction. In fact, peak height suggests the presence of a phosphorous-containing compound linked to H39 and H325, with a shape compatible with an acetyl phosphate (**Figure 3.10.6**). Acetyl phosphate concentration is estimated to reach 3 mM in the cytosol of *Escherichia coli* wild-type cells [105], thus it proposed that the compound may interact and bind mutant *EcAppA* during overexpression. Multiple attempts to identify the ligand by mass spectrometry were unsuccessful.



**Figure 3.10.6. Possible enzyme-substrate intermediate of reaction**

The figures show electron density (blue hatching – double difference, green hatching – positive single difference Fourier map) bridging the catalytic histidine H39 and the histidine of the PD motif H325. This is observed in 7 structures of *EcAppA* HAT (HAT). The image on the left gives a transverse view while the image on the right a frontal view. A phosphate bound is modelled in the image on the left.

### 3.11. Discussion

The proton donor motif and the RHGxR motif are highly conserved areas of the active site of HP2. Residues acting as proton donors in this class are aspartic acid (HDx motif), as in *EcAPPA*, and glutamic acid (HAE motif), which is characteristic of MINPPs phytases such as *BtMINPP*. To test whether it was possible to engineer MINPPs-like catalytic promiscuity into *EcAppA*, the MINPPs HAE proton donor (PD) was engineered into the *E. coli* enzyme. Four mutants were generated: the proton donor-less *EcAppA* HAT (D326A), *EcAppA* HET (D326E), *EcAppA* HDE (T327E) and the proton donor swap mutant *EcAppA* HAE (D326A/T327E).

Expression trials of *EcAppA* have again confirmed that DsbC plays a role in the formation of the non-consecutive disulphide bridge of *EcAppA* [95]. Cytoplasmic expression was obtained only when using a Shuffle Express T7 strain that constitutively expresses DsbC in this compartment. Also, in one of the constructs, the C-terminal of *EcAppA* was modified by the addition of an His-tag. The fusion-protein had limited solubility and was characterised by early precipitation, linking one more the C-terminal with protein stability. In fact, the removal of the C-terminal and its disulphide bridge, enhances thermal stability. It was proposed that *EcAppA* begin to collapse starting from this end, when increasing the temperature [94].

Stereospecificity of myo-inositol hexakisphosphate dephosphorylation proved to follow sequential hydrolysis in the wild-type enzyme as reported by Greiner et al (2000) [35]. In order, phosphates on carbons 6/1/3/4/5 are hydrolysed. HPLC profiles of the IPs generated by mutants show that mutations influence positional stereospecificity in IP<sub>6</sub> hydrolysis, with a marked change for the Asp326 proton donor mutants *EcAppA* HAT (proton donor-less), which produces a predominant 1/3-OH IP<sub>5</sub> peak, and *EcAppA* HET which shows to produce 1/3-OH and 4/6-OH IP<sub>5</sub> in equal quantities. On the contrary, the single mutant *EcAppA* HDE becomes more 4/6-OH IP<sub>5</sub> positionally stereoselective. However, the proton donor swap mutation of *EcAppA* HAE almost restores the 4/6-OH IP<sub>5</sub> predominant product profile characteristic of the wild-type enzyme, although a larger array of IP<sub>4</sub> species is produced by the mutant.

Enzymes characterisation reveals that a catalytic toll is payed as consequence of mutations. The most affected mutants are *EcAppA* HAT and the double mutant *EcAppA* HAE which suffer a 200-fold drop in catalytic efficiency with 1000-fold reduced turnover number and 5-fold increased binding affinity. The mutant *EcAppA* HET maintains the binding affinity of the wild-type protein but has an 80-fold

decrease in turnover number. The enzyme with the least altered profile is *EcAppA* HDE with a 3.5-fold decrease in catalytic efficiency.

In line with Ostanin and Etten (1993) [31], a role again is suggested for D326 in proton donation, for the release of the product, and for the formation of the phosphoenzyme intermediate. The statement is supported by the fall in turnover number and the increase in binding affinity for the mutants, but also, possibly, by the structure solution of an enzyme-substrate intermediate, which could be obtained only when proton donation is impaired. Instead, a D326E mutation appear to restore the wild-type binding affinity for this substrate (although with an impaired turnover number), unlike what found when hydrolysing lower phosphorylated compounds by this mutant [31]. When replacing aspartate with glutamate, there is only a partial complementation of function, probably because the longer side chain constraints the positioning of the carboxyl group affecting the volume of the binding pocket and may impede proton donation to smaller phosphorylated compounds.

Clearly, the initial hypothesis that a proton donor swap could be the switch from a selective to a promiscuous IP<sub>6</sub> hydrolysis is false. However, the single mutation D326E results in a more promiscuous enzyme towards IP<sub>6</sub> hydrolysis. To investigate the reasons behind this result, X-ray crystal structures of mutants in complex with IHS are solved. The substrate analogue binds the wild-type enzyme through its sulphate in position 6 of the carbon ring, in line with *EcAppA* stereospecificity traits [35]. However, the result disagrees with the findings of Lim et al. (2000) [24]. They solved an *EcAppA* H39A structure in complex with phytate, in which, the substrate points its 3-phosphate towards the alanine mutant. It may be possible that the mutation of the catalytic histidine altered IP<sub>6</sub> binding [24].

The use of IHS instead of IP<sub>6</sub> could be considered a limitation of this thesis, however, it is the best compromise found. The substrate is quickly processed at low pH and it would not be found even in crystals soaked at basic pH (pH 9.5), where *EcAppA* is presumably inactive. Docking studies also showed that IHS and IP<sub>6</sub> employ similar binding poses in the mutant structures (data not shown). Therefore, IHS could be considered a good mimicking compound.

An additional binding mode was found for IHS in the mutant *EcAppA* HET. The compound binds K46 (3.0 Å), Q49 (2.9 Å) and L50 (3.0 Å) in a 5-axial/1-equatorial conformation coordinating a potassium ion. K46 lies within 3.0 Å from both IHS conformers, in fact a 5-equatorial/1-axial molecule also lies in the active site of this enzyme. K46 and the surrounding loop are known to undergo a local conformational change upon substrate binding [24]. Data could suggest that this flexible area,

exposed on the enzyme surface, could be involved in the first recruitment of the substrate facilitating its entrance in the active site cleft. Mutagenesis studies on the residues are needed to test its role in substrate binding.

Overall, the X-ray crystal structures seem to suggest that non-positional specificity may be the outcome of differences in shapes and polarity of the active site pockets. In particular, active sites comparisons suggest that an enlargement of pocket B allows the accommodation of multiple IHS binding poses (e.g. for *EcAppA* HAT and HET), while the decrease in size of pocket F, by the insertion of E327 (e.g. *EcAppA* HDE) or the redirection of Q126 and K129 (e.g. *EcAppA* HET), leads to an increase of 4/6-OH IP<sub>5</sub> production. Further proves need to be collected to test these conjectures.

These results could be a starting point for the development of a more promiscuous commercial *EcAppA*, which would possibly allow to degrade IP<sub>6</sub> to completion. It would be interesting to study the effect of an exchange of active site between *EcAppA* and a target MINPP, to test the transferability of promiscuity between these sub-branches of HP2s using the same strategy as Lehmann et al (2000) [106]. They transferred residues from *Aspergillus niger* NRRL 3135 phytase to a synthetic phytase (phytase-1), and they successfully demonstrated that a transfer of favourable catalytic properties from one phytase to another is possible by swapping active sites. Another approach would be of saturation mutagenesis of pocket B and F. This method would allow to understand to which extent mutations in these pockets could favour catalytic promiscuity. However, *E. coli* AppA phytase has already been optimise for thermostability and catalytic activity. For this reason, it would also be optimal, in future studies, to mutagenize the commercial enzyme (sequence not available), because any active site mutation effect on catalysis would be affected by the surrounding chemical environment.

# CHAPTER 4

## 4. *Bifidobacterium longum* MINPP and its domain motion during the catalytic cycle

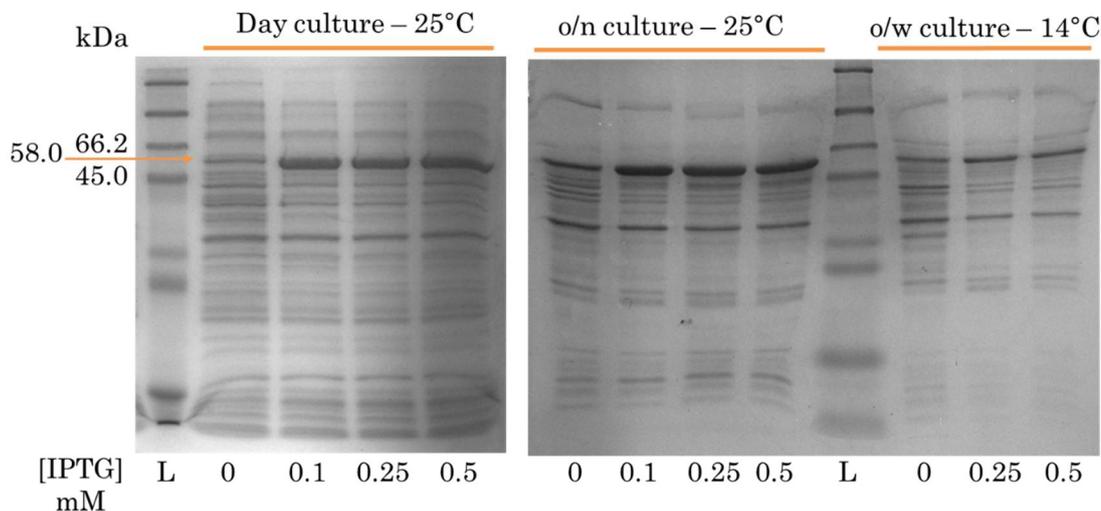
Among phytases of the histidine phosphatase family, for which relevant crystal structure data are available (**Table 5.2.1**, labelled by <sup>a</sup>), no domains movements have been detected upon substrate or substrate analogues binding. Therefore, these enzymes are presumed to lack inherent structural flexibility throughout catalysis. In this work, a crystal structure of a MINPP from *Bifidobacterium longum* subsp. *infantis* (*B*MINPP), has been solved. The bacterium is a Gram-positive human gut commensal and it is known for its positive role in the early development of the infants' gut [107]. Its phytase, firstly discovered by [21], has been chosen because of its low amino acid sequence identity (~23 %) from the only previously known MINPP found in Bacteria, the *Bacteroides thetaiotaomicron* MINPP [41]. Also, *B*MINPP is cell-wall anchored, unlike *Bt*MINPP, which is known to be packed in outer membrane vesicles and targeted to interact with intestinal epithelial cells.

The crystal structures I solved during my PhD, taken together with previously determined data, provide snapshots along the catalytic cycle of the enzyme: 1) the apo-protein [40]; 2) a model for the substrate-bound complex (the complex with inositol hexasulphate, IHS, a substrate analogue); 3) a catalytic intermediate (*B*MINPP E401Q-phosphohistidine intermediate); and 4) the product complex (*B*MINPP in complex with inorganic phosphate - [42]). These structures reveal a large domain motion during catalysis. *B. longum* MINPP is therefore the first HP phytase seen to undertake such movement.

### 4.1. Expression

The MINPP gene of *B. longum* was supplied by Vicente Monedero (IATA-CSIC, Spain) and cloned in a truncated form (deleted of the signal peptide and the C-terminal sortase dependent cell wall-anchoring) by Arthur Li into pOPINF [40], a high copy number plasmid (500-700 copies per cell), T7 expression system which is IPTG-inducible. Rosetta 2 (DE3) pLysS cells were employed for protein production. Expression trials were set up to identify the best condition of expression (**Figure 4.1.1**). Large scale expression of 2 L (500 mL x 4 flasks) were set up in LB broth containing ampicillin (100 µg/mL), inoculating 5 mL o/n culture. The cultures were

incubated with shaking (180 rpm) at 37 °C until OD 0.6, when they were induced with 0.5 mM IPTG. Cells expressed *B/MINPP* o/n at 25 °C, with shaking (180 rpm) and, after 17h, they were harvested by centrifugation at 4 °C, 5500 xg, for 20 min. Pellets were resuspended in 60 mL Lysis Buffer (50 mM NaH<sub>2</sub>PO<sub>4</sub> pH 7.8, 300 mM NaCl, 20 mM imidazole), snap frozen in liquid nitrogen, and stored at -80 °C.



**Figure 4.1.1. Overexpression trials of *B/MINPP* in Rosetta (DE3) pLysS.**

*B/MINPP* MW is 58 kDa. The protein was expressed at 25 °C or 14 °C, for 3 h or o/n or o/w, at IPTG concentrations: 0, 0.01, 0.25, 0.5 mM.

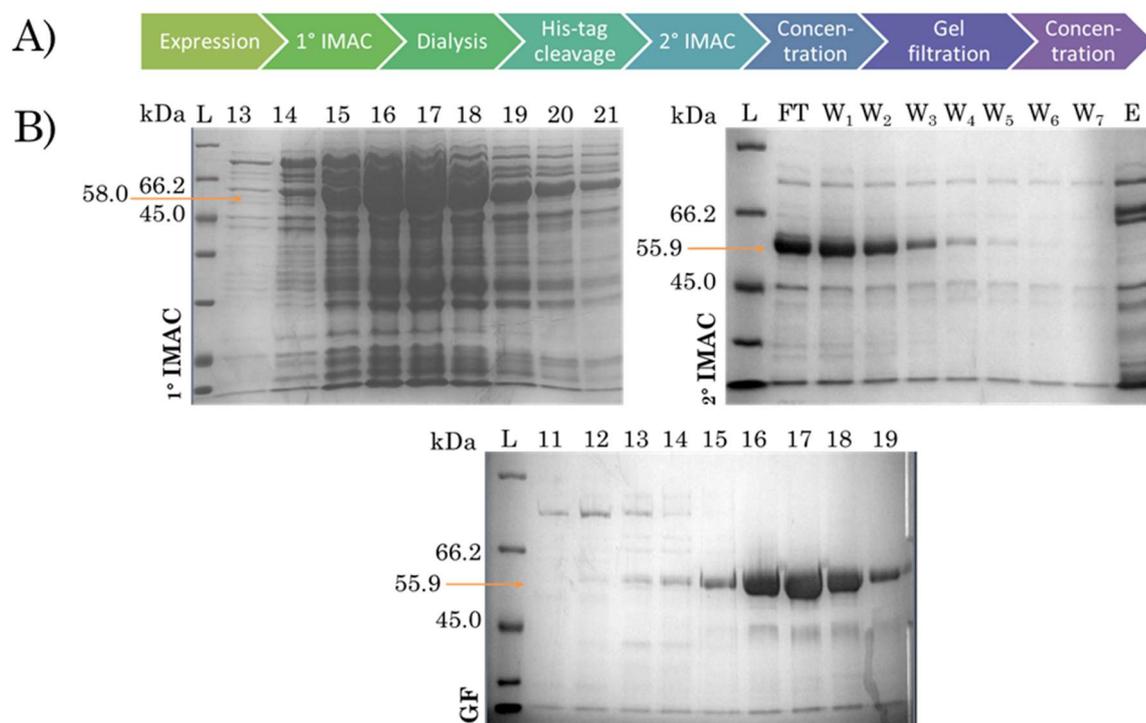
## 4.2. Purification

Cells were thawed for the separation of the soluble fraction (lysis by French press and centrifugation). Protein purification was obtained by metal affinity chromatography using a cleavable His-tag fused to the N-terminal of *B/MINPP*. The cell extract was recirculated (flow rate 4 mL/min) in a pre-equilibrated 5 mL Ni-NTA affinity column (Superflow Cartridge, Quiagen), for 1 h. Chromatography was performed following standards methods described in **Chapter 2.6.1**. Eluted proteins were analysed by 12 % acrylamide SDS-page gel (**Figure 4.2.1**) to assess their purity grade. The first IMAC step, gave a yield of total protein of 20-30 mg, however, the protein purity was not estimated to be higher than 30 % at this stage.

The second step of purification consisted of the cleavage of the tag (o/n dialysis at 4 °C for the incubation with His-tagged 3C-protease) and the separation by IMAC of the cleaved protein (collected from the flow through). Metal affinity chromatography was performed as described in **Chapter 2.6.1**. The eluted fractions were again inspected on a 10 % acrylamide SDS-PAGE (**Figure 4.2.1**) which showed

that this step was very effective for the achieving of high protein purity (estimated as higher than 80 %).

The final purification step was performed by gel filtration (GF) as described in **Chapter 2.6.1**. Two cycles of GF were carried out loading 3 mL samples to a HiLoad 15/60 Superdex 75 gel filtration column. The fractions corresponding to the peaks of protein absorbance were inspected on an 8 % acrylamide SDS-PAGE gel (**Figure 2**). A sample purity higher than 98 % was achieved with a yield of 5.4 mg/L of protein.



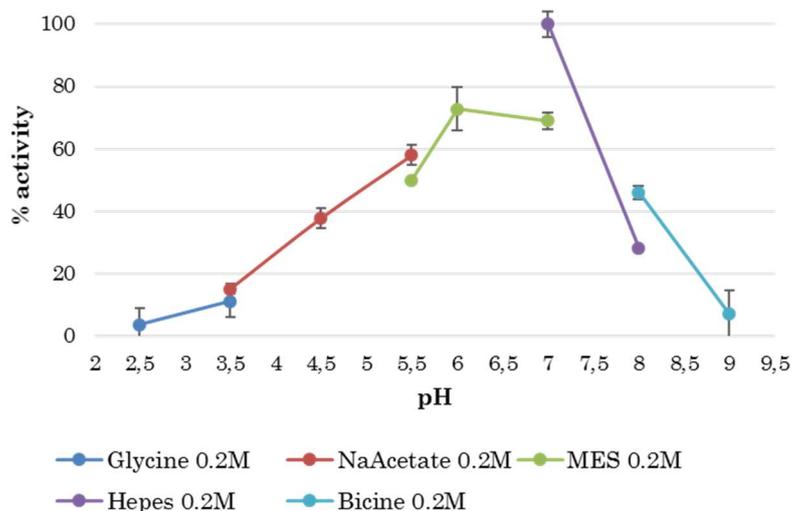
**Figure 4.2.1. Purification of *BIMINPP*.**

A) The top graphic describes the order of steps in purification: a first IMAC followed by dialysis and 3C-protease cleavage of the His-tag, a second IMAC for the separation of the cleaved protein, concentration, gel filtration and a second concentration in preparation for crystallization screenings. B) The three SDS-PAGE reported depict three steps of protein purification: the first IMAC (top left), the second IMAC (top right) and the gel filtration (bottom). Protein MW: 58 kDa – before His-tag cleavage, 55.9 kDa – after His-tag cleavage. Fractions are labelled by order of elution (e.g. 13, 14, 15, etc.) of or by stage in purification and number (e.g. flow through – FT, wash 1, 2, etc. – W1, W2, etc., elution – E).

### 4.3. Characterisation

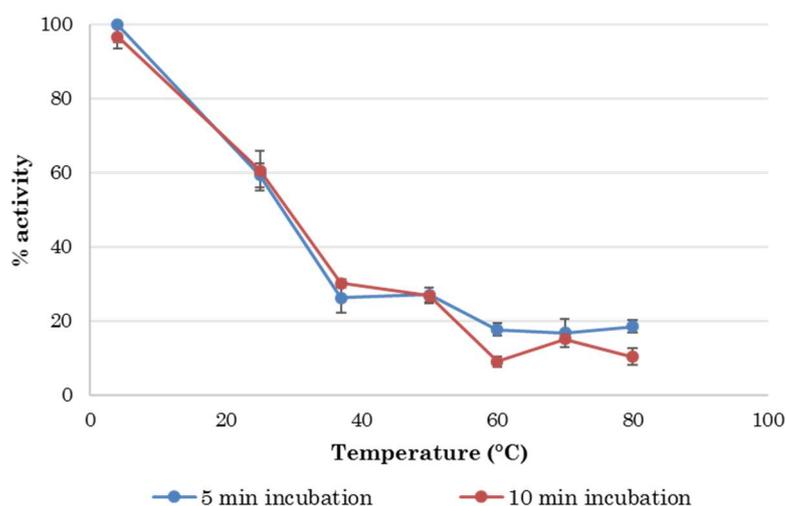
The profile of enzymatic activity as a function of pH was tested at 37 °C in triplicate using the following buffers: pH 2.5/3.5 - 0.2 M Glycine-Cl, 0.15 M NaCl, pH 3.5 / 4.0 / 5.0 / 5.5 - 0.2 M NaAcetate, 0.15 M NaCl, pH 5.5 / 6.0 / 6.5 / 7.0 - 0.2 M MES 0.15 M NaCl, pH 7.0 / 8.0 - 0.2 M HEPES, 0.15 M NaCl, pH 8.0 / 9.0 - 0.2 M Bicine, 0.15 M NaCl. The phosphate released was quantified by monitoring the absorbance

of its complex with molybdenum blue reagent at 700 nm (**Chapter 2.7.2**). 50  $\mu$ L reactions were set up in triplicate at fixed concentrations of enzyme (100 nM) and substrate (IP<sub>6</sub> - Sigma Premium Quality: 1 mM) and incubated at 37 °C for 15 min before inactivation by addition of molybdenum blue reagent. Solutions were left to develop for 30 min at room temperature before measuring absorbance. Triplicate reads of buffer-only solutions were taken as well as a phosphate calibration curve to check that the absorbance registered was in the linear range. Results are reported in **Figure 4.3.1**.



**Figure 4.3.1. pH profile of B/MINPP.**

x axis: pH; y axis: % activity. Maximum of activity was identified at pH 7 in HEPES buffer. The same result was not obtained in MES buffer which seems to negatively affect enzyme activity.



**Figure 4.3.2. Test of B/MINPP recovery after heating.**

x axis: Pre-incubation temperature (°C); y axis: % activity. The assay is carried out in 0.2 M HEPES pH 7, 0.15 M NaCl, with 5 mM IP<sub>6</sub>. Reactions proceeded for 15 min, at 37 °C, after 5 or 10 min pre-incubation at the temperatures: 4, 37, 50, 60, 70, 80 °C. Optimal of activity was obtained when protein samples were left at 4 °C before reaction.

A recovery after heating experiment was also set up (**Figure 4.3.2**). Protein samples were pre- incubated at 4, 37, 50, 60, 70 and 80 °C for 5 or 10 min, before injection to a reaction mixture of 5 mM IP<sub>6</sub>, 0.2 M HEPES pH 7.0, 0.15 M NaCl. They were incubated at 37 °C for 15 min. Reactions were set up in triplicate. Experimental conditions varied from the experiment of Tamayo-Ramos, Sanz-Penella [21] where thermal stability was tested at pH 5.5 [21].

Temperature and pH profiles, substrate specificity and inhibition by Ca<sup>2+</sup> reported by Tamayo-Ramos et al (2012) are significantly different to the results here described, especially regarding thermal stability [21]. Recovery after heating tests showed in my experiments a 20 % residual activity after incubation for 10 min at 80 °C in contrast to their 45 % residual activity after incubation at 80 °C for 15 min. This discrepancy may be a result of the usage of a different buffer system and pH.

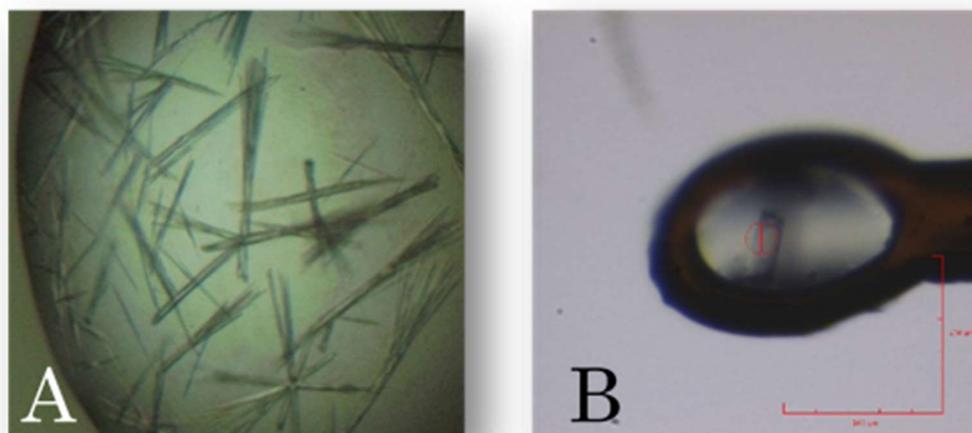
#### 4.4. Crystallisation of the complex *B*/MINPP:IHS and *B*/MINPP E401Q

The structure of apo *B*/MINPP was solved by Arthur Li, UEA [40]. To gain further insight into the catalytic mechanism of this enzyme, co-crystallization trials of *B*/MINPP were set up in the presence of inositol hexasulphate.

Wild-type *B*/MINPP was purified and concentrated to 9.2 mg/mL. Six crystallographic 96 wells trays (Molecular Dimensions, Newmarket, UK) were set up using an OryxNano protein crystallization robot (Douglas Instruments Ltd) enzyme crystallization via sitting drop vapour diffusion in the presence of the substrate analogue inositol hexasulphate (IHS). Three commercial screens were tested: Structure - Screen I and II (Molecular Dimensions), JCSG-plus (Molecular Dimensions) and PACT premier (Molecular Dimensions). Plates were set up for incubation at two temperatures, 4 °C and 16 °C. Each drop (0.5 µL of volume) contained equal quantities of enzyme-inhibitor solution (0.25 µL) and precipitant solution (0.25 µL), each well was filled manually with 50 µL precipitant solution. The protein buffer was 20 mM HEPES, NaCl 150 mM, NaOH pH 7.4. The final amount of protein in solution after the incubation with the substrate analogue and before the dilution in the crystallographic screens was 8.3 mg/mL. The final concentration of IHS in the protein solution was 1 mM.

Needle shaped crystals grew in a wide range of conditions. Eleven crystals were harvested on LithoLoops, cryo-protected by adding 30 % glycerol or 30 % ethylene glycol to the mother liquor solution and stored in liquid nitrogen. Diffraction

data collection was carried out at the Diamond Light Source (DLS) (Oxfordshire, UK) on the beamline i03.

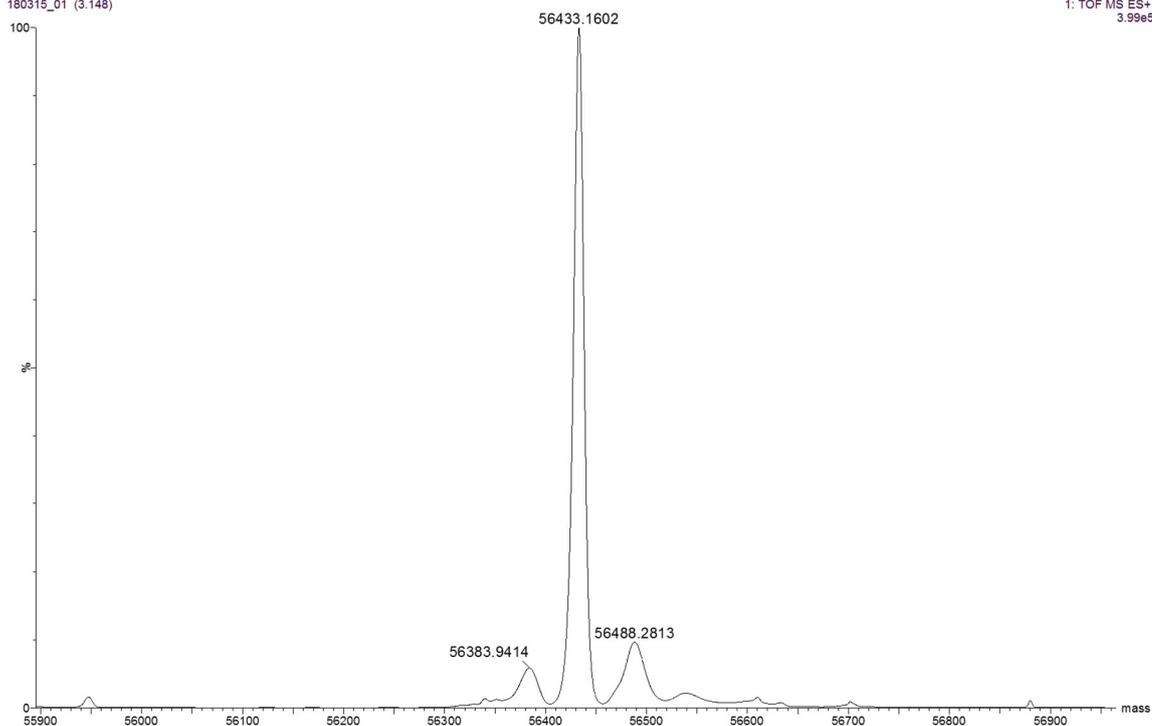


**Figure 4.4.1. Crystals used in diffraction experiments.**

**A-** Image of the drop containing six crystals that were used in diffraction experiments. Precipitant conditions were: 10 % w/v PEG 1000 10 % w/v PEG 8000. **B-** Cryo-loop mounted on the i03 beamline at DLS (Oxfordshire – UK).

The proton donor-less mutant E401Q *B*/MINPP (**Figure 4.4.1**) (cloning performed by Monika Zietek, (UEA)) was also crystallized with the aim of improving the resolution of the structure of the a phospho-histidine intermediate. This intermediate is presumed to persist because of the mutation that inactivates the proton donor residue E401Q. A part of the purified sample was sent for mass spectrometry analysis while the rest was used to set up crystallization trials. The conditions which produced diffracting crystals of apo *B*/MINPP were used. These were 0.1 M MES pH 6.5, 0.01 M zinc chloride, 14-16-18 % PEG 6000 at a protein concentration of 9 mg/mL (protein buffer: 20 mM HEPES pH 7.4, 0.15 M NaCl).

The intact mass of the protein was measured by positive electrospray ionization time-of-flight mass spectrometry (**Figure 4.4.2**). The expected protein mass of the cleaved enzyme (N-term His-tag removed) is 56349.89 Da, however mass spectrometry analysis revealed three main protein fragments of 56383.94, 56433.16 and 56488.28. The major peak is around 83,27 Da bigger than the predicted protein MW, a difference which is close to the MW of phosphate adduct (80,99 Da). Conformational changes in a protein can influence the solvent-accessible surface area ( $A_s$ ), from which the average protein charge depend.  $A_s$  is directly proportional to the MW of the protein (Marsh & Teichmann (2011) –  $A_s = 4.84 * MW^{0.760}$ ). Deviations from this behaviour can be indicative of structural disorder and conformational flexibility [108].



**Figure 4.4.2. Time of flight mass spectrometry positive electrospray ionisation of a purified *BIMINPP* sample.**

x axis: mass (Da), y axis: % on the total sample. Three main protein population are represented, respectively at 56383.14 Da, 56433.16 Da and 56488.28 Da. The arrow highlights the possible phosphohistidine enzyme intermediate of reaction.

## 4.5. X-ray crystal structure determination

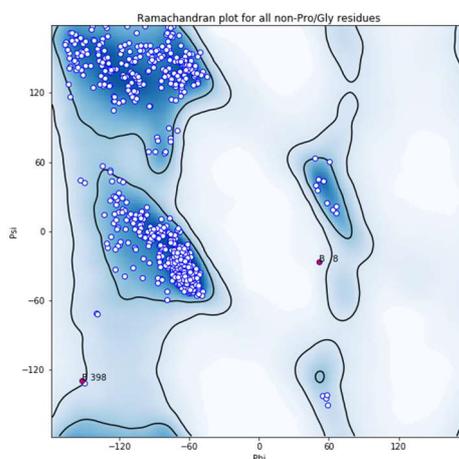
X-ray experiments resulted in the solution of multiple *B*/MINPP structures by the efforts of three people: Arthur Li, Monika Zietek and Isabella Acquistapace. **Table 4.5.1** gives a summary of all the collected structures. My work focused on the solution of the molecular structure of *B*/MINPP in complex with the substrate analogue IHS and the improvement of the structure of a phosphate-histidine intermediate in the mutant *B*/MINPP E401Q.

**Table 4.5.1. Solution of *B*/MINPP macromolecular structures.**

Structure	Operator	Summary
<i>B</i> /MINPP apo	Arthur Li, UEA	Res.: 1.65-46.53 Å Space group: P1 R-free: 0.18 R-work: 0.15
<i>B</i> /MINPP E401Q : phosphate	Monika Zietek, UEA	Res.: 1.71-68.23 Å Space group: P1 R-free: 0.19 R-work: 0.15
<i>B</i> /MINPP E401Q phospho-histidine (His45)	Isabella Acquistapace, UEA	Res.: 2.40-68.86 Å Space group: P1 R-free: 0.24 R-work: 0.18
<i>B</i> /MINPP : IHS	Isabella Acquistapace, UEA	Res.: 1.84-70.57 Å Space group: P2 <sub>1</sub> R-free: 0.24 R-work: 0.20

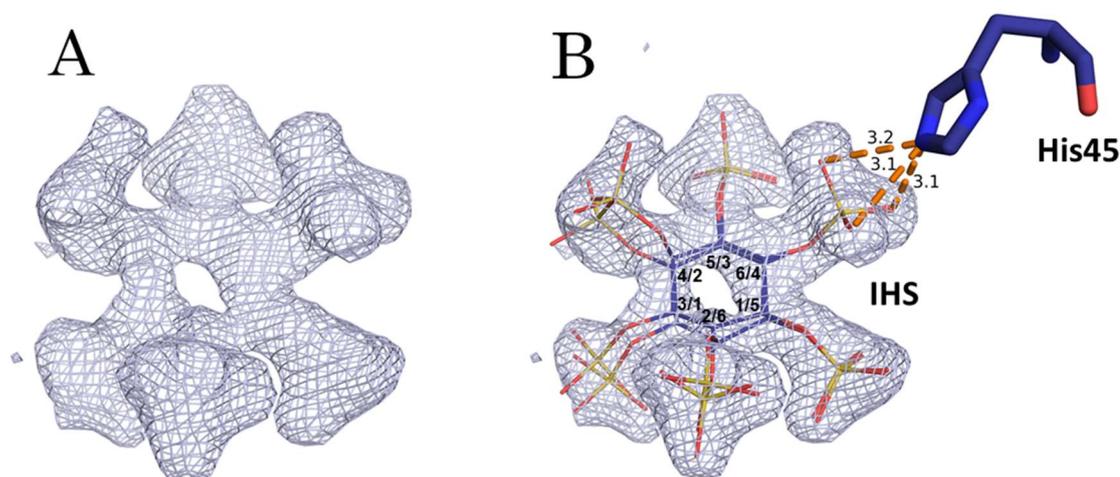
The experiments, which aim was the capturing of the enzyme-substrate analogue complex, resulted in twelve datasets, collected at 100 K, from six cryo-protected crystals. The X-ray diffraction images were automatically integrated and scaled and the dataset with the best diffraction statistics was chosen for structure solution (**Table 4.5.1**). The diffracted crystal grew in 10 % (w/v) PEG 1000 and 10 % (w/v) PEG 8000 and was cryo-protected by the addition of 30 % glycerol to the mother liquor. It contained two molecules per asymmetric unit. The phase problem was solved through molecular replacement. The monomeric *B*/MINPP apo-structure [40] was edited to remove water molecules, Zn<sup>2+</sup> ions and was used as a search model in phasing by PHASER [81]. However, it was not possible to find the solution as the enzyme underwent a conformational change upon binding of the substrate analogue. For this reason, two separate ensembles, one representing the  $\alpha$ -domain and the other the  $\alpha/\beta$ -domain, were created and given as inputs for a subsequent search. The resolution limits were also tuned, with the high-resolution limit decreased to 4 Å. Using these parameters the phasing procedure provided an acceptable solution which

was subsequently manually remodelled using WinCoot [87] and refined with phenix.refine [89]. The ligand .cif file used in refinement was obtained from the WinCoot dictionary. The Ramachandran plot for all non-Pro/Gly residues is shown in **Figure 4.5.1** and refinement statistics given in the **Table 4.5.2**.



**Figure 4.5.1. Ramachandran plot for all non-Pro/Gly residues Phi and Psi angles.** R8 and A398 are reported as outliers. The graph was generated by MolProbity [109].

The substrate analogue bound to the active site of the enzyme in two different orientations. These present the monophosphoester group at either position 4 or 6 of the inositol ring pointing towards the catalytic histidine His45. Occupancies and temperature factor for both the conformers have been refined and a ligand omit single difference Fourier map is displayed in **Figure 4.5.2**.



**Figure 4.5.2. Omit single-difference Fourier map of the substrate analogue IHS.**

Contour level is set to  $\sigma = 3.0$  r.m.s.d. **A)**  $F_o - F_c$  omit map of the density surrounding IHS; **B)**  $F_o - F_c$  omit map of the density surrounding IHS and the corresponding models of the substrate analogue. Two conformers seem to fit the density map. The catalytic histidine His45 makes contacts to the sulphate groups bound to C4 or C6 of the inositol ring (carbon positions are labelled as “conformer 1/ conformer 2”).

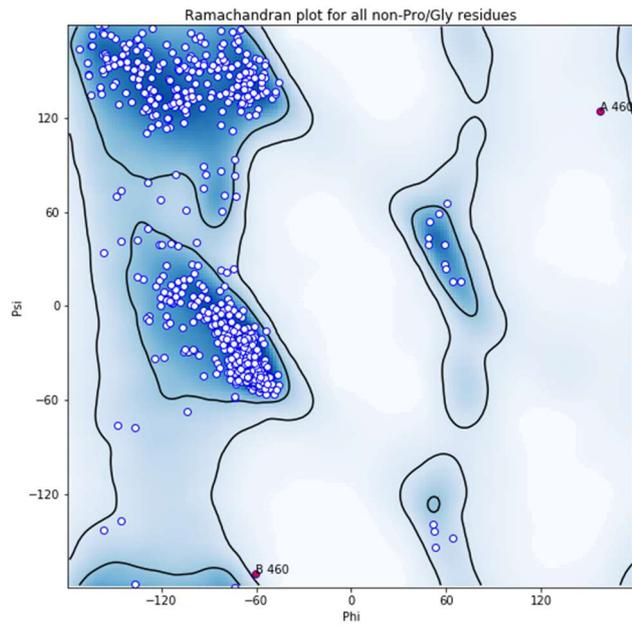
**Table 4.5.2. Data collection and refinement statistics for the *B*MINPP:IHS complex.**

Statistics for the high-resolution bin are in brackets.

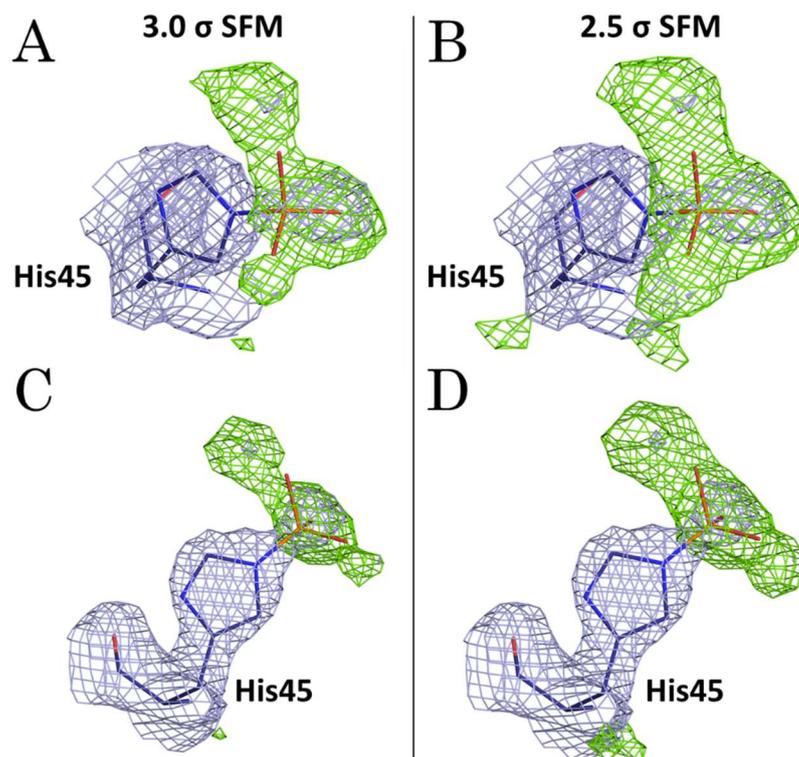
<b>Protein</b>	<b><i>B</i>MinPP: IHS</b>
<b>Data Collection</b>	
Wavelength (Å)	0.97960
Space group	P 1 21 1
<u>Cell parameters:</u>	
a , b , c (Å)	70.8, 106.4, 76.5
α, β, γ (°)	90.0, 112.6, 90.0
Resolution limit (Å)	70.57- 1.84 (1.91 - 1.84)
R merge	0.159
(I)/sd(I)	4.6 (1.6)
Completeness (%)	99.65 (98.9)
Multiplicity	3.4 (3.3)
Overall temperature factor (Å <sup>2</sup> )	10.898
<b>Refinement Statistics</b>	
Protein monomers per asymmetric unit	2
Total atoms	9374
Water molecules	1447
R work	20.85%
R free	23.81%
<u>Ramachandran Analysis (%):</u>	
Most favoured	98.1%
Outliers	0.2%
<u>RMS deviations:</u>	
Bonds (Å)	0.004
Angles (°)	0.70
Planes (Å)	0.004
<u>Mean Atomic B-value (Å<sup>2</sup>)</u>	16.99
Macromolecule	15.16
Ligands	15.85
Solvent	26.95
<u>Occupancies (%)</u>	
Chain A – conformers A, B	54, 46
Chain B – conformers A, B	56, 44

*B*MINPP E401Q structure was also solved in space group P1 by molecular replacement. Phase was obtained by searching with three separate ensembles: an apo-*B*MINPP monomer [40], a phosphate bound *B*MINPP monomer [42] and a potential phosphohistidine-containing *B*MINPP monomer [42]. A solution was found with two copies of *B*MINPP E401Q in the asymmetric unit corresponding to phosphate- (however no density correspondent to phosphate was detected) and NEP-containing monomers. The structure was solved using the same procedure as for the *B*MINPP:IHS complex. The Ramachandran plot for all non-Pro/Gly residues is shown in **Figure 4.5.3** and refinement statistics are given in the **Table 4.5.3**.

The catalytic histidine, His45, is present in the polypeptide chains as in apo form (chain A) or in the phosphohistidine form (chain B). Occupancy and temperature factor for the phosphohistidine residue have been refined and a phosphate-omitted single difference Fourier map is displayed in **Figure 4.5.4** (contour levels of 2.5 or 3.0 r.m.s.d).



**Figure 4.5.3.** Ramachandran plot for all non-Pro/Gly residues Phi and Psi angles. The 460 of chain A and B are reported as outliers. The graph was generated by MolProbity [109].



**Figure 4.5.4.** Omit single-difference Fourier map of the phosphate portion on NEP45. Single difference Fourier density Maps (SFM) are coloured in green. Double difference Fourier density maps are coloured in blue and their contour level is set to  $\sigma = 1.5$  r.m.s.d. **A)** Top view, Fo-Fc omit map is set to  $\sigma = 3.0$  r.m.s.d.. **B)** Top view, Fo-Fc omit map is set to  $\sigma = 2.5$  r.m.s.d.. **C)** Side view, Fo-Fc omit map is set to  $\sigma = 3.0$  r.m.s.d.. **D)** Side view, Fo-Fc omit map is set to  $\sigma = 2.5$  r.m.s.d..

**Table 4.5.3. Data collection and refinement statistics for the structure of *BIMINPP* E401Q.**

Statistics for the high-resolution bin are in brackets.

<b>Protein</b>	<b><i>BIMinPP</i> E401Q</b>
<b>Data Collection</b>	
Wavelength (Å)	0.97628
Space group	P 1
<u>Cell parameters:</u>	
a , b , c (Å)	55.51, 88.99, 72.99
α, β, γ (°)	73.12, 71.30, 78.99
Resolution limit (Å)	50.4- 2.40 (2.48 – 2.40)
R merge	0.212
(I)/sd(I)	2.3
Completeness (%)	94.6 (88.8)
Multiplicity	3.4 (3.3)
Overall temperature factor (Å <sup>2</sup> )	25.50
<b>Refinement Statistics</b>	
Protein monomers per asymmetric unit	2
Total atoms	8117
Water molecules	286
R work	18.47%
R free	23.89%
<u>Ramachandran Analysis (%):</u>	
Most favoured	96.8%
Outliers	0.2%
<u>RMS deviations:</u>	
Bonds (Å)	0.006
Angles (°)	0.80
Planes (Å)	0.005
<u>Mean Atomic B-value (Å<sup>2</sup>)</u>	
Macromolecule	28.31
Ligands	45.06
Solvent	34.77
<u>Occupancies (%)</u>	
Chain B – NEP	96

## 4.6. Structural analysis

The structure of *B*/MinPP, like all HP, consists of two domains, an  $\alpha/\beta$ -domain and an  $\alpha$ -domain (**Figure 4.6.1.A**). While the  $\alpha/\beta$ -domain is conserved in the family, the  $\alpha$ -domain is protein-specific. *B*/MINPP  $\alpha$ -domain is larger than the average *Ec*AppA-like phytases  $\alpha$ -domain and is characterized by a loop that spans above the active site (**Figure 4.6.1.C/D/E**).

The catalytic core of *B*/MINPP consists of the nucleophilic His45 residue, three arginine residues (R44, R48, R148) which coordinate the substrate during catalysis, and the proton donor triplet HAE (H339 to E401). **Figure 4.6.1.B** gives a representation of the cleft architecture apo-enzyme form [40] and allows a closer look to the catalytic core residues. To note that R48 and R142 are oriented towards different directions from R44 or the catalytic histidine, H45.

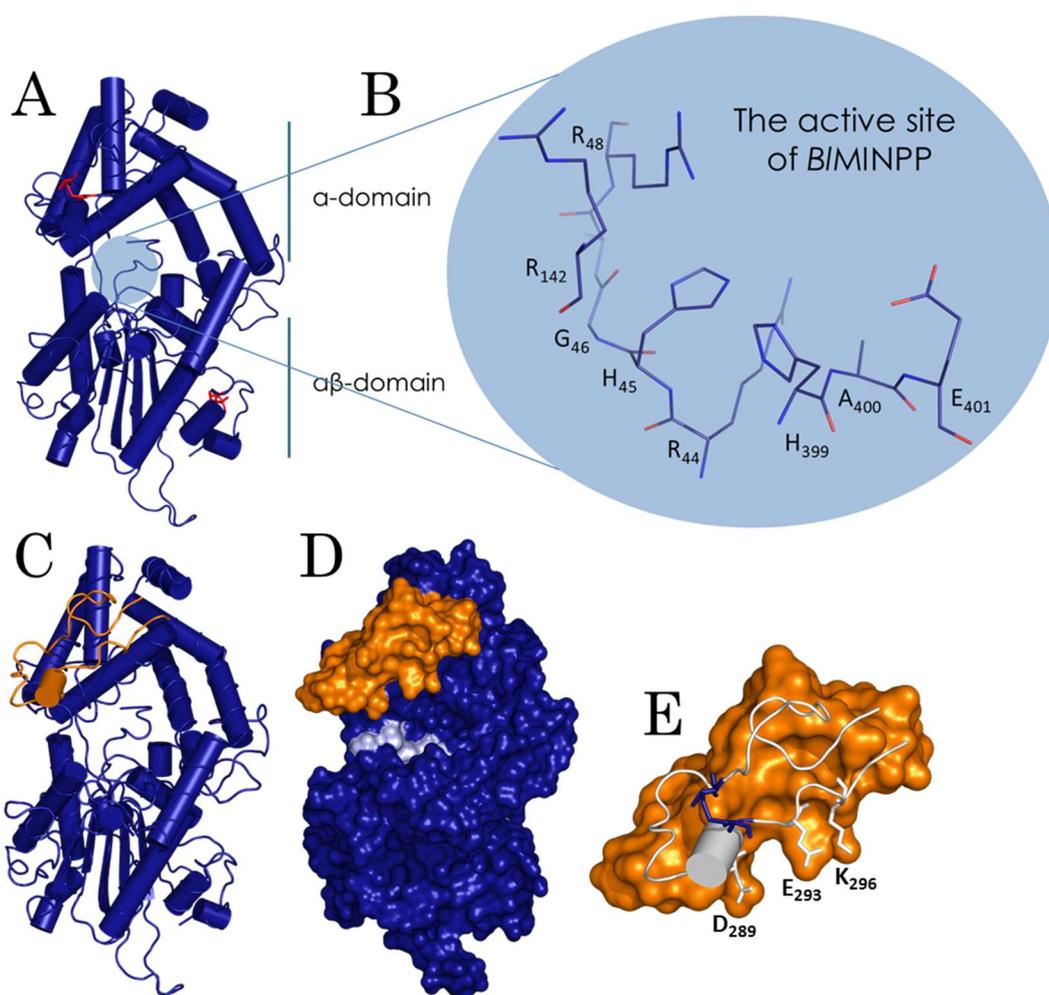
The structure of the enzyme-IHS complex provides a model for the enzyme-substrate interaction and reveals the binding of substrate analogue IHS in two orientations, the first with the sulphate group at position C4 of the inositol ring oriented towards His45 and the other with C6 in this position. Interactions of IHS with the catalytic core residues are the same for both the conformers of IHS. Promiscuity of binding increases the far we shift our view away from the nucleophilic histidine to other active site specificity pockets, which can accommodate multiple orientations of the sulphate groups of IHS (**Figure 4.6.2**). The presence of conformational disorder (i.e. multiple bound conformations) in the IHS ligand is consistent with the reduced positional stereospecificity of this enzyme relative to that observed for the more specific canonical HP phytases.

The sulphate groups lying in the catalytic pocket (on C4 or on C6), make close contacts with: the catalytic histidine H45 (2.9-3.1 Å distance), R41, R48 and R142 (2.7-3.0 Å distance) and the proton donor triplet (H399 - 2.9 Å, A400 - 2.8 Å). This is the result of a reorientation of those residues upon binding of the substrate analogue, with the larger rearrangement occurring for R48 (6.1 Å shift) and R148 (8.0 Å shift) (**Figure 4.6.2**). The carboxylate group of the presumed proton donor, E401, presents one of its oxygens to the centre of the inositol ring. The active sites are essentially identical between the two monomers in the asymmetric unit, with an exception being the orientation of the sidechain of E401 (**Figure 4.6.2**).

A rearrangement of the active site residues is also detected upon binding of inorganic phosphate [42] (**Figure 4.6.3.4**). Catalytic core residues points toward phosphate in a similar way to the one detected for the enzyme:IHS complex. On the

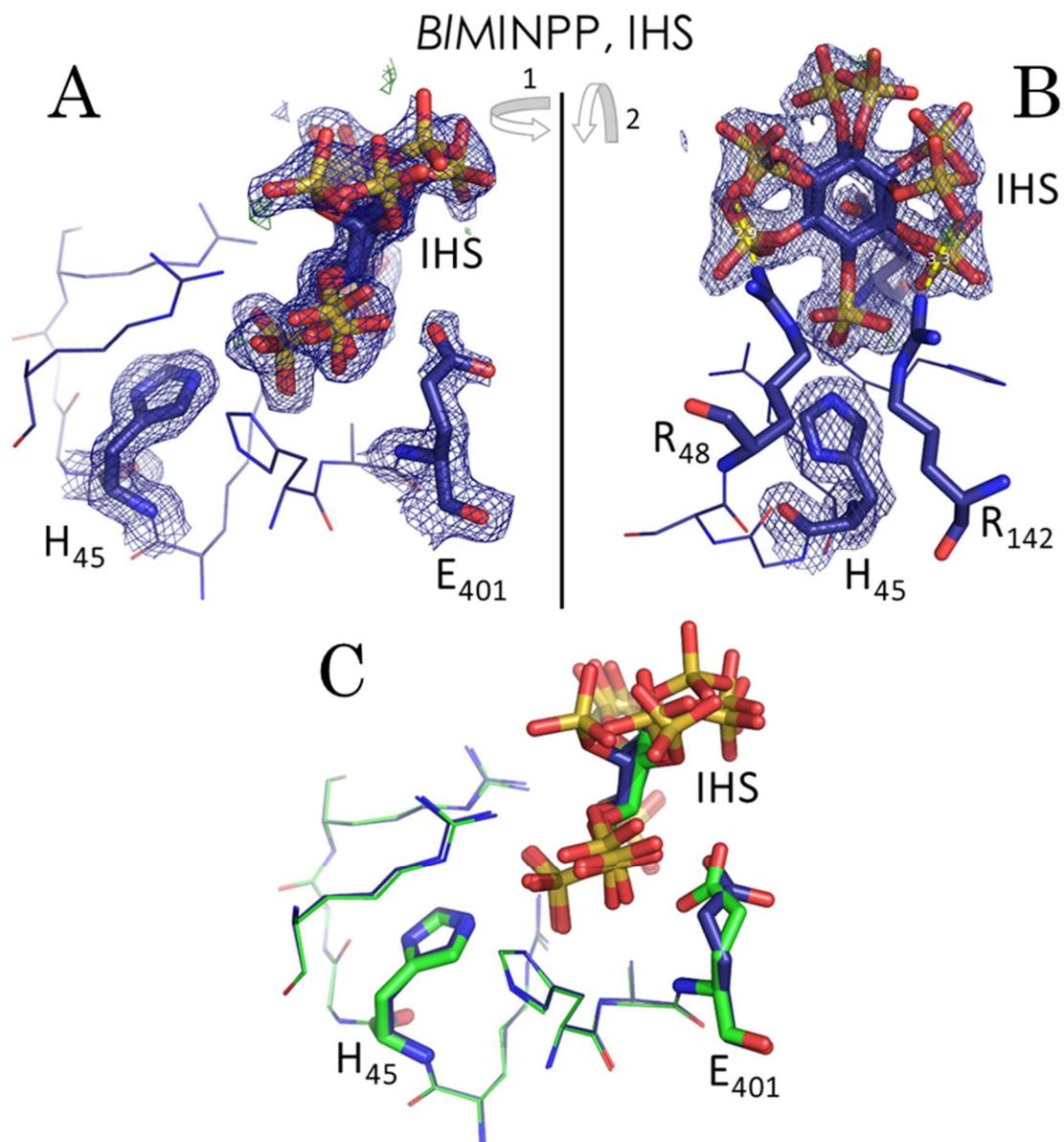
contrary, when the nucleophilic histidine is in a phosphorylated form, the catalytic core residues are only partially reorganised towards the oxygen groups of  $\text{PO}_3^{2-}$  (**Figure 4.6.3.3**).

Snapshots of the catalytic cycle are described by these movements of the active site residues (**Figure 4.6.3**). The enzyme is first in its apo-form (**Figure 4.6.3.1**); the catalytic core arginine residues then reorient to stabilize substrate binding (**Figure 4.6.3.2**); nucleophilic attack by the catalytic histidine occurs and the dephosphorylated intermediate of the first stage of reaction exits the active site, leaving a phosphorylated enzyme (**Figure 4.6.3.3**); the active site is then regenerated by donation of the phosphate to water or an alternative acceptor molecule (**Figure 4.6.3.4**).



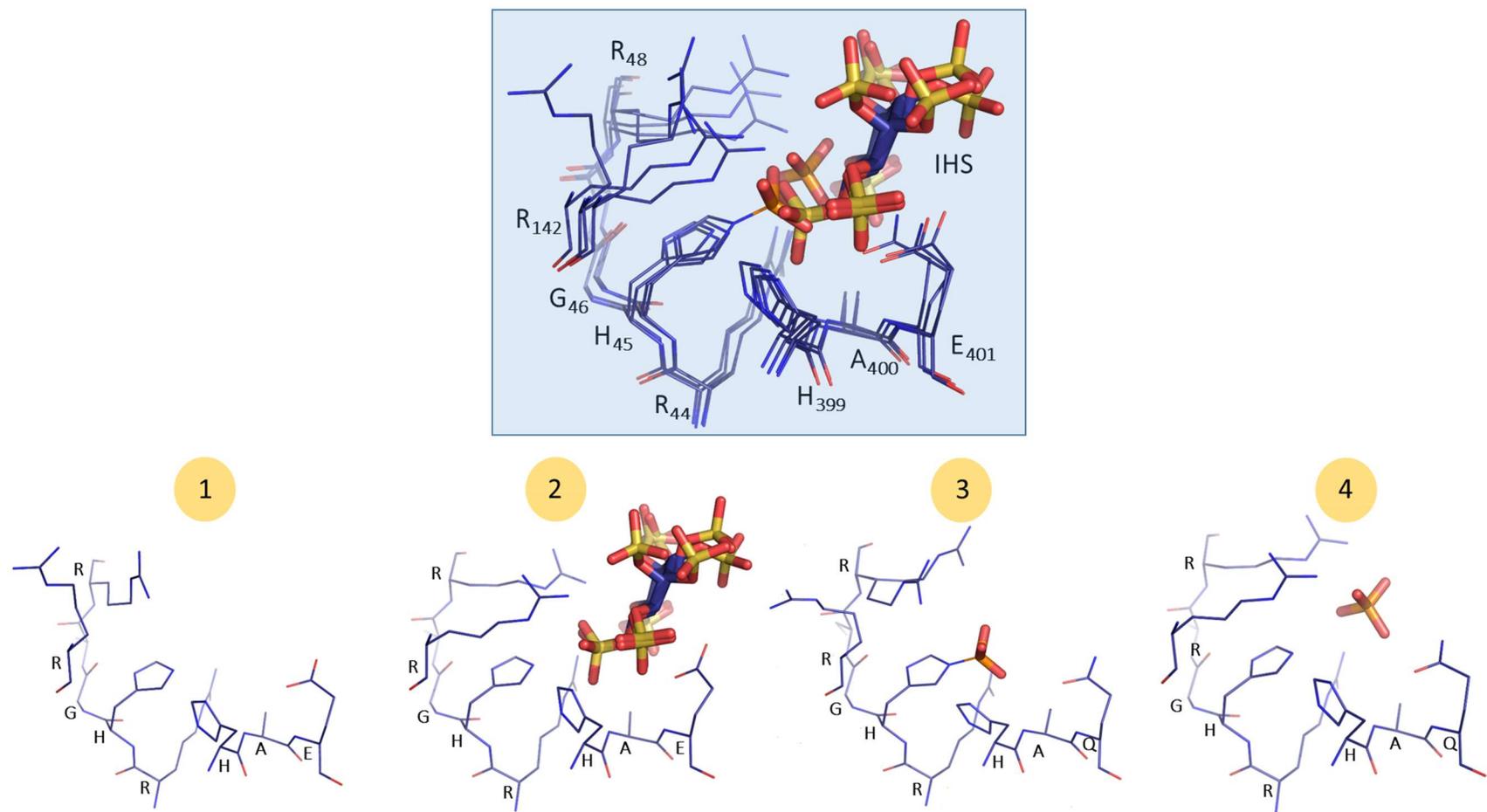
**Figure 4.6.1. Structural overview of *BMINPP*.**

**A)** The overall structure of the enzyme [40] comprises an  $\alpha$ -domain and an  $\alpha/\beta$ -domain. They frame the active site cleft which is positioned at their interface. Disulphide bridges are displayed by red sticks. **B)** A close-up of the core catalytic residues captures their 3D array in the enzyme apo-form. **C)** A newly identified loop of the  $\alpha$ -domain, characteristic of *BMINPP* is coloured in orange. **D)** Surface representation of *BMINPP* and the loop which span over the active site. The loop is coloured in orange while the catalytic core is in light-blue. **E)** A close-up of the loop displays in sticks the residues that face the active site cleft.



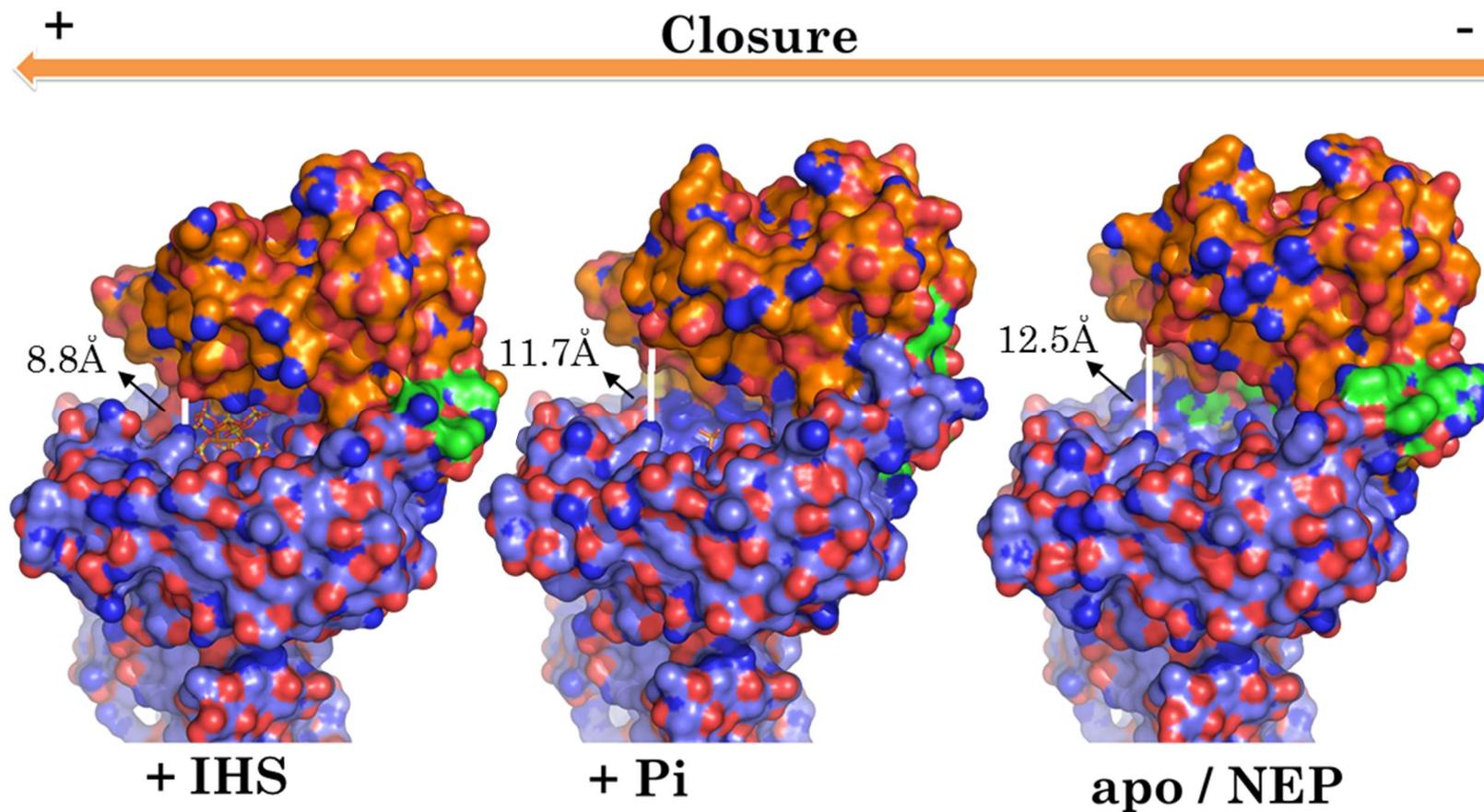
**Figure 4.6.2. Active site configuration of *B/MINPP* in the IHS bound form.**

Side (A) and top (B) views of *B/MINPP* active site in its interaction with IHS. Electron density maps are displayed for the catalytic histidine and IHS:  $2|F_o| - |F_c|$  in blue ( $\sigma = 1.5$  r.m.s.d.) and  $|F_o| - |F_c|$  in green ( $\sigma = 3.0$  r.m.s.d.). (C) Superposition of the active sites of the two monomers present in the asymmetric unit of the *B/MINPP*:IHS complex. These are identical except for the orientation of the proton donor residue Glu401.



**Figure 4.6.3. Snapshots of the catalytic cycle of *B/MINPP*.**

The top image is a superposition of the four active site states that shows larger rearrangements for R48, R142 and E401. At the bottom, each active site conformation is presented, providing ‘snapshots’ of the catalytic cycle of *B/MINPP*: 1) Apo-enzyme [40], 2) *B/MINPP* substrate analogue-bound complex, 3) Phospho-enzyme intermediate, 4) Enzyme phosphate-bound complex [42].



**Figure 4.6.4. *B/MINPP* conformational changes during the catalytic cycle.**

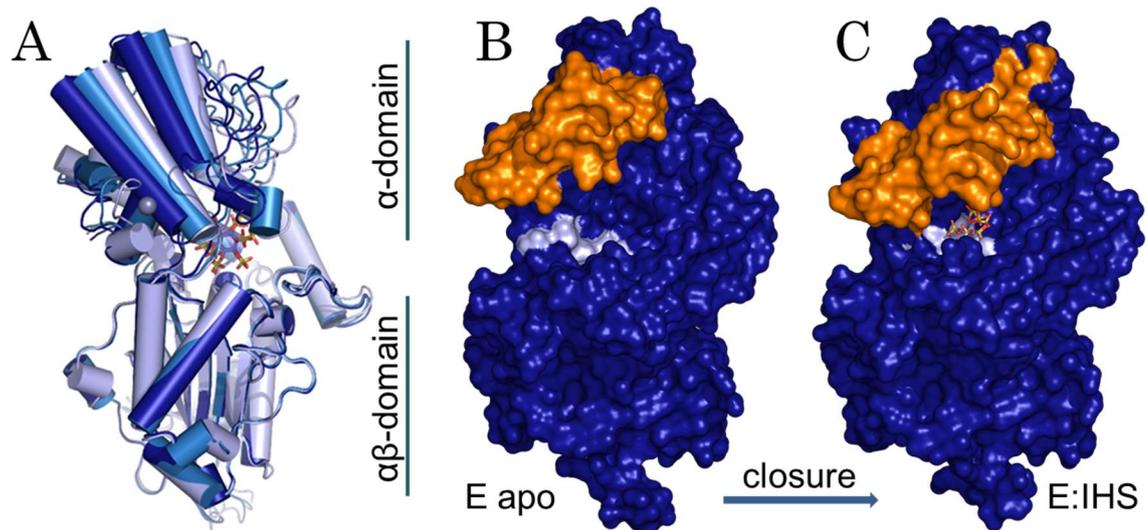
Three surface representations of *B/MINPP* are displayed: substrate analogue-bound form (left: +IHS), phosphate-bound form (centre: +Pi, Zietek [42]) and apo or phosphohistidine form (right: apo / NEP, [40]). The hinge region, identified by DynDom [110], is coloured in green; the fixed domain ( $\alpha/\beta$ -domain) is purple; the moving domain ( $\alpha$ -domain) is orange. To visually quantify the extent of closure, the distances between K192 ( $\alpha/\beta$ -domain) and E293 (U-loop -  $\alpha$ -domain) are reported.

These conformational changes, however, are not limited to the active site, like the small loop conformational changes of *EcAppA* (**Figure 3.10.2**), but rather they propagate to the  $\alpha$ -domain in its whole (**Figure 4.6.4**, **Figure 4.6.5.A**). When the substrate analogue binds to the active site, a closure of the  $\alpha$ -domain towards the  $\alpha/\beta$ -domain occurs, removing solvent from the catalytic cleft, and stabilizing the binding of remote phosphate groups through interaction with a *B/MINPP*-characteristic  $\alpha$ -domain loop (**Figure 4.6.5.B/C**). Residues of the loop in close proximity to the ligand are K296, which achieves a minimum distance of 3.9 Å to IHS, and E293 which lies within 5 Å from the peripheral sulphate groups of the substrate analogue. The binding of phosphate, the product of hydrolysis, results instead only in a partial  $\alpha$ -domain closure (**Figure 4.6.5.A**).

An analysis with DynDom [110] revealed a domain motion with an r.m.s.d. of whole protein best fit between apo and IHS-bound form of 2.22 Å and a fixed domain best fit of 0.98 Å. The region involved in the motion is a portion of the  $\alpha$ -domain, while the whole  $\alpha/\beta$ -domain and the remainder of the  $\alpha$ -domain undergo a limited shift and are therefore considered 'fixed domain'.

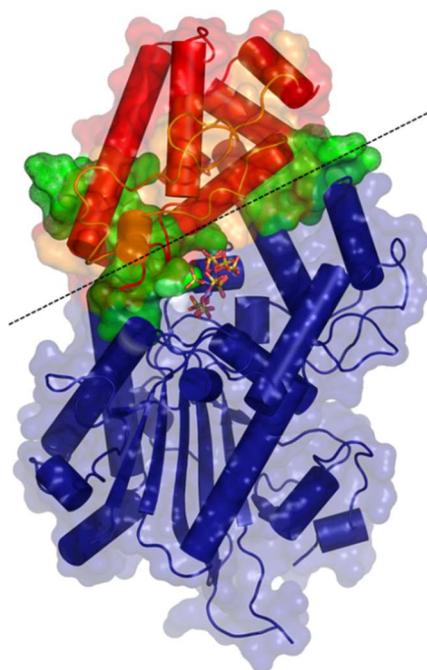
The program identified an overall rotational movement of 17.5 ° of the apo enzyme upon substrate binding, corresponding to 91 % closure of the moving domain (**Figure 4.6.6**, in red and orange) over the fixed domain (**Figure 4.6.6**, in blue) in the IHS-bound conformation of *B/MINPP*. The phosphate-bound form exhibits a lesser degree closure (68.6 % vs 91 %) and a rotation of 10°. The percentage measure of the degree of closure can be defined from the square of the projection axes on the closure axis.

Hinge residues were also identified (**Figure 4.6.6**, in green). Because interdomain screw axes [111] are located in the proximity of bending residues (see centre of rotation, **Figure 4.6.6**), these amino acids are considered to be acting as mechanical hinge with the interdomain screw axis as effective hinge axes. Two mechanical hinges are identified, one on the left of the interface between fixed- and moving-domain, and the other on the right (**Figure 4.6.6 and 4.7.1**). This example is in line with the description of the interactions driving domain motion by Hayward [112].



**Figure 4.6.5. Conformational changes of *B/MINPP*.**

**A)** In blue: apo-protein and phosphoenzyme intermediate; in turquoise: phosphate-bound enzyme, in light-blue: substrate analogue-bound enzyme (IHS mimicks the substrate  $IP_6$ ).  $\alpha$ -domain and  $\alpha/\beta$ -domain separation is also displayed. **B)** Surface representation of *B/MINPP* apo and its characteristic U-loop over the active site. The loop is coloured in orange while the catalytic core is highlighted in light-blue. **C)** Surface representation of *B/MINPP*:IHS. The enzyme is in a closure state. U-loop is coloured in orange while the catalytic core is highlighted in light-blue.



**Figure 4.6.6. Domain movement topology in *B/MINPP*:IHS.**

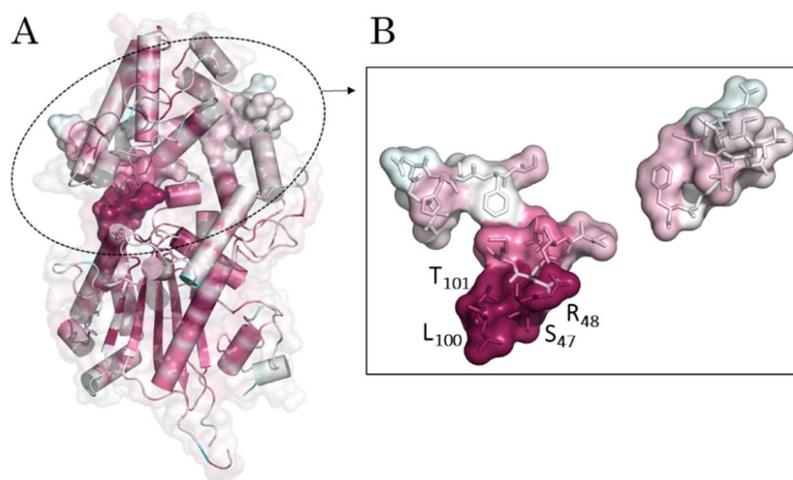
Polypeptide chain is coloured depending on the areas of movement defined by DynDom [110]. A dotted line is drawn on the centre of rotation. The fixed domain is coloured in blue, while the moving domain is red with the exclusion of the *B/MINPP*-characteristic U-loop, in orange. Hinge residues are displayed as a green surface. IHS is represented by sticks.

## 4.7. Conservation of domain flexibility

The hinge is the area most involved in domain movement. Hinge residues were identified through the algorithm implemented by DynDom (**Figure 4.6.6**, **Figure 4.7.1** and **Table 4.7.1**). An analysis of the conservation of these residues among bacterial MINPPs was performed. Bacterial MINPPs representative enzymes were collected following the protocol outlined in **Chapter 5**. A .pdb file containing the conservation scores for each *B*/MINPP residue was generated using the program ConSurf [113]. Hinge residues were mapped on the structure of the *B*/MINPP:IHS complex and they are displayed in **Figure 4.7.1.A**.

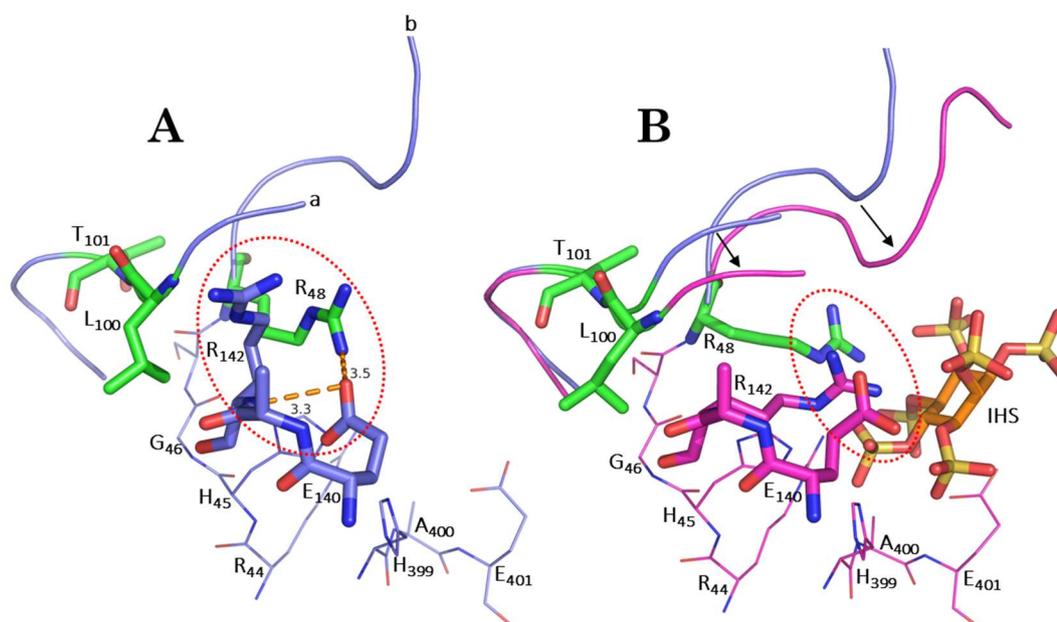
Among the identified amino acids some active site residues can be found (**Table 4.7.1**). They are S47/R48, two residues that are part of the RHGxRxP motif, and K100/T101 and amino acids belonging to the conserved GxLTx<sub>2</sub>G motif, commonly found in HP2 (**Chapter 5**). These active site residues are the only amino acids that both contributes to the hinge and are highly conserved in HP2 (**Figure 17.B, magenta**). The conservation of the remainder hinge residues decreases as we move away from the active site. Hinge residues with the largest psi° and phi° change are D317 and M318 (**Table 4.7.1**). They are in direct contact with S47 (RHGxR motif) and Q103 (GxLTx<sub>2</sub>G motif). Their change in orientation allows the propagation of the movement to the back of the  $\alpha$ -domain.

A closer inspection of the hinge residues suggests that the domain movement originates from a shift in R48 and R142 upon substrate binding (in the same manner as it happens for the small loop movement in *EcAppA*). To contact the substrate analogue, R48 loses a salt bridge interaction with E140, and reorients, together with R142, towards IHS (**Figure 4.7.2**). R44, R48 and R142 establish multiple polar contacts with the substrate analogue. The backbone rotation of R48, propagated to its downstream loop, resulting in a partial closure of the domain (**Figure 4.7.2**). The displacement of R48 and R142 causes a shift in GxLTx<sub>2</sub>G motif which also contributes to propagate the movement. Loops downstream hinge residues (termed **loop a** and **b**) allow the shift of the back (**loop a, Figure 4.7.2**) and the front (**loop b, Figure 4.7.2**) regions of the  $\alpha$ -domain.



**Figure 4.7.1. Mechanical hinge residues of *BMINPP* coloured by conservation.**

Colour gradient: from low conserved amino acids, in light-blue, to highly conserved amino acids, in purple. **A)** *BMINPP* coloured by conservation, the hinge is framed in a circle of dashes. **B)** Hinge close-up, highly conserved residues are highlighted.



**Figure 4.7.2. R48 rotation and downstream loop motion in IHS binding.**

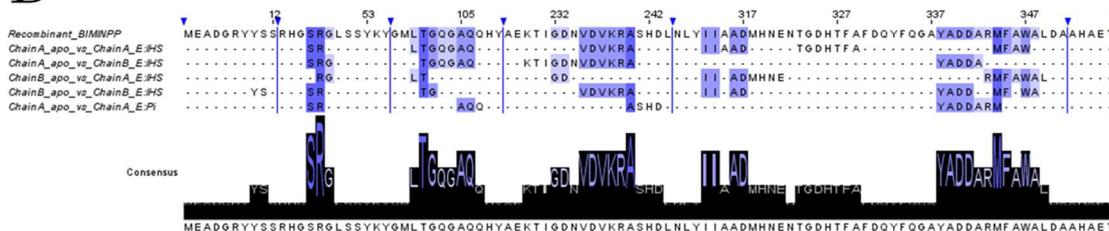
Loops that propagate the domain motion are indicated in the apo-enzyme by “a” (modulates the movement of the back region of the  $\alpha$ -domain) and “b” (modulates the movement of the front region of the  $\alpha$ -domain). Hinge residues are coloured green. Polar contacts are represented by dashes. Dotted lines are drawn around R48, E140 and R142 to highlight their reorientation **A)** A representation of the configuration of conserved hinge region in the apo enzyme structure. The image shows the polar contacts between R48, E140 and R142. **B)** A representation of the conserved hinge region in the IHS bound structure. IHS is represented by orange sticks with the phosphate group on C6 of the inositol ring pointing towards the catalytic histidine. R48 and R142 are redirected to coordinate the substrate analogue. The shift of loop a and b is displayed by overlapping the configuration of the apo enzyme form (in blue) and the IHS bound form (in violet). Movement is highlighted by arrows.

Table 4.7.1. Summary of the results of the DynDom analysis.

A

Comparisons	rmsd whole protein best fit (Å)	Rmsd fixed domain best fit (Å)	Rotation angle (°)	Closure (%)	Res change psi <sup>1</sup> (i) / phi <sup>2</sup> (i+1)	Dynamic contacts
ChainA_apo vs ChainA_E:IHS	2.06	0.63	17.5	91.7	2.5(S47) / -9.7(R48) -3.5(M99) / -6.7(L100) 7.8(L100) / -15.6(T101) 13.3(N234) / 8.8(V235) -10.4(V235) / 1.5(D236) -8.7(A240) / 4.6(S241) 10.3(A315) / -1.6(A316) 16.7(A316) / -27.2(D317) 28.2(D317) / -22.9(M318) -13.9(N322) / 22.7(T323) 25.5(T323) / 9.3(G324) -18.8(G324) / -12.7(D325) 16.5(D325) / 9.7(H326) -10.9(H326) / 8.6(T327) 2.3(F328) / -10.3(A329) 73.4(A329) / -57.8(F330)	Mixed movement 98->142; 142->98.
ChainA_apo vs ChainB_E:IHS	2.12	0.70	18.1	82.3	6.0(L100) / -7.7(T101) 6.8(T101) / -3.9(G102) -4.4(D233) / -7.8(N234) 19.1(N234) / 9.8(V235) -9.0(V235) / -0.9(D236) -5.7(D236) / 4.3(V237) -12.5(A240) / 6.1(S241) -7.5(Y338) / 11.6(A339) -15.6(A339) / -8.3(D340) 14.9(D340) / 15.2(D341) 11.0(D341) / 3.7(A342) 3.7(A342) / 14.9(R343)	Hinge Movement 98->142; 142->98; 99->141; 302->352/348; 306->348; 306->349; 349->306.
ChainB_apo vs ChainA_E:IHS	1.22	0.36	9.8	100	3.4(M99) / -11.7(L100) 8.1(L100) / -5.3(T101) 4.7(Y312) / -5.9(I313) 15.1(A316) / -25.5(D317) 25.3(D317) / -20.4(M318) -10.5(H319) / -3.1(N320) 6.7(N320) / -6.0(E321) -15.6(A342) / 11.4(R343) -3.2(M344) / 9.4(F345) -8.1(A348) / 14.0(L349) -5.4(L349) / 5.2(D350)	Shear Movement 98->142; 142->98; 99->141/145; 230->234; 234->230; 234->231; 231->234; 237->231.
ChainB_apo vs ChainB_E:IHS	1.17	0.53	8.8	94.2	-10.9(Y10) / 3.5(S11) 1.8(S11) / 9.6(S12) -6.0(S47) / 11.9(R48) 8.8(N234) / 11.5(V235) -11.6(V235) / 0.3(D236) -11.5(A240) / 8.4(S241) 19.3(A316) / -26.7(D317) 21.5(D317) / -19.3(M318) -7.3(A337) / 4.8(Y338) -5.7(Y338) / 9.6(A339) -18.3(A339) / -9.6(D340) 20.9(D340) / 8.1(D341) 1.7(D341) / 12.0(A342) -14.1(R343) / 8.7(M344) -11.2(M344) / 14.4(F345) -6.4(F345) / 4.1(A346) -10.6(A346) / 13.6(W347) -4.6(A348) / 7.3(L349)	Hinge Movement 98->142; 142->98; 99->141/145; 106->320; 320->106; 302->232/352; 306->349; 315->311/319/328.
Average_apo vs E:IHS	1.79	0.55	8.8 - 18.1	82.3 - 100		
ChainA_apo vs ChainA_E:Pi	1.24	0.46	10	69.8	8.1(S47) / -10.8(R48) 8.0(A339) / 3.6(D340) -11.8(D340) / 5.5(D341)	Hinge Movement 98->142; 142->98; 142->99; 99->142; 99->141; 302->237; 237->302; 302->232/238/348; 258->238; 300->238; 301->238; 306->348; 306->349; 349->306; 310->349.
ChainB_apo vs ChainA_E:Pi	0.37	-	-	-		

B



**Table 4.7.1. Summary of the results of the DynDom analysis.**

**A-** Summary table of DynDom results. R.m.s.d. of protein best fit, r.m.s.d. of fix domain best fit, rotation angle (°) and % closure is reported for each pair of structure comparison. Residues change in psi (°) and phi (°) of a residue (i) and its downstream adjacent (i+1) are listed for residues that have an overall change in psi and phi higher than 10 °. Dynamic contacts are also indicated as well as the type of movement identified. **B-** Residues identified as part of the hinge are aligned to the recombinant *B/MINPP* sequence for each pair of structure comparison. A consensus sequence has been created to highlight the residues that has been identified in multiple structure comparisons. R48 has been recognised as part of the hinge in all DynDom analysis, followed by S47, T101, A240 and M344.

It is important to note that a consequence of the  $\alpha$ -domain movement in *B/MINPP* is also the shift of a characteristic loop (U-loop, **Figure 4.6.1** and **Figure 4.6.5**) over the active site cleft close to the substrate (the minimum distance to IHS of the loop residue Lys296 is 3.9 Å). This loop is probably the trigger for the more extensive domain movements in this enzyme than the small loop conformational changes that can be detected in *EcAppA*. In fact, the movements of *B/MINPP* seem to be initiated in the same region as seen in the *E. coli* phytase, however the area lies behind the *B/MINPP* U-loop, which is in turn linked to the back of the  $\alpha$ -domain (when the same orientation as **Figure 4.6.5** is adopted). This is the reason why small structural changes, confined in *EcAppA* to active site region, propagate to involve almost the whole  $\alpha$ -domain of *B/MINPP*.

Due to the important role of the U-loop in  $\alpha$ -domain movement, its conservation in other MINPPs was analysed. A pool of bacterial MINPPs sequences was collected and clustered. This is described in **Chapter 5.4**. These sequences were also inspected for the presence/absence of a homologue region of the characteristic *B/MINPP* loop (named the U-loop). As results of this analysis, the U-loop was found in only a small number of bacterial MINPPs, including limited members of *Actinobacteria*,  $\beta/\gamma$ -*Proteobacteria* and *Firmicutes*. A tree was inferred by maximum-likelihood of representative bacterial MINPPs (for information see **Chapter 5.5.4**) and annotated to display enzymes containing a region homolog to the U-loop (**Figure 4.7.3**).

Based on the high conservation of residues primary involved in domain motion, the mechanism is thought to potentially be shared among a large number of enzymes of the HP2 family. However, no other structural data collected for this family of show any domain opening/closure. It seems possible that this large conformational change is strictly associated with the presence of the U-loop of *B/MINPP*, which allows additional residues to contribute to the substrate coordination and stereospecificity during catalysis.

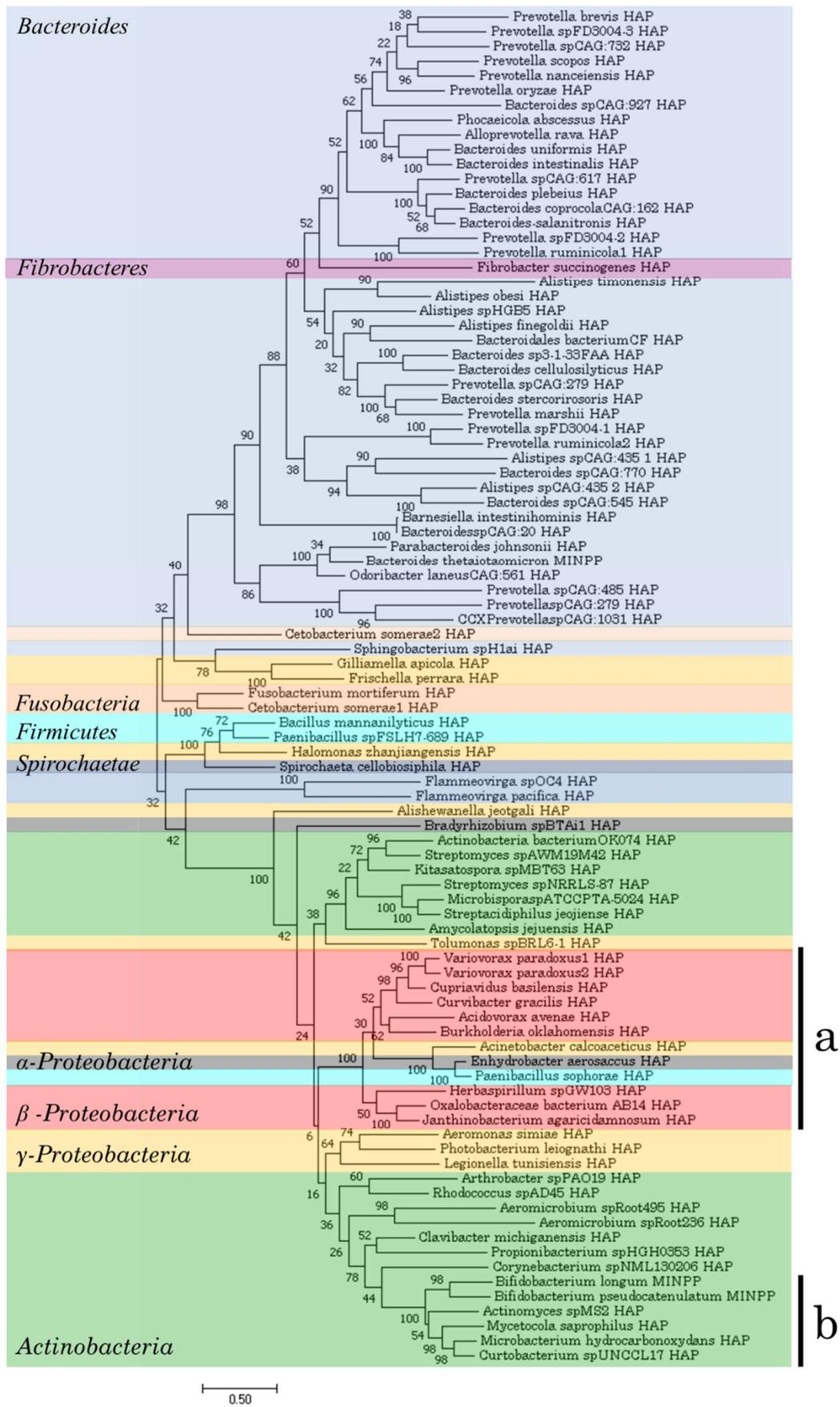


Figure 4.7.3. Molecular Phylogenetic analysis by Maximum Likelihood method inferred for bacterial MINPPs.

**Figure 4.7.3. Molecular Phylogenetic analysis by Maximum Likelihood method inferred for bacterial MINPPs.**

The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The enzymes in which a U-loop homolog region was found are indicated by the letters a and b.

The orientation of the amino acids of the catalytic core in the binding of IHS is conserved among all the HP2 phytases structure collected to date, while the remaining sulphate (mimicking phosphate) groups are coordinated in an enzyme-specific fashion, making fewer interactions with the enzyme the farther they lie from the catalytic histidine. This characteristic HP2 mode of substrate binding is achieved only in the closed conformation of *B*/MINPP. However, the positioning of the U-loop in the closed structure prevents the access of the substrate to the active site cleft, therefore explains for a significant domain motion to allow substrate binding.

The *B*/MINPP:IHS binding mode of the catalytic core is essentially identical to that seen in other member of the MINPPs family, therefore no difference in the mechanism of catalysis is suggested by this work. However, this does not exclude *a priori* the existence of an intrinsic flexibility in a wider range of MINPPs. It is useful to consider briefly the potential reason why no X-ray crystal structures of a HP2 in an open conformation has been solved. A possibility is that a closed conformation in the HP2 phytase structures may be induced and maintained by weak crystal lattice forces. Such an enzyme form may potentially reduce entropy by solvent depletion and/or the achievement of a closer packing of monomers with a decrease in the overall thermal motion. This would not be the first example in which weak crystal forces can support closed conformations. It was the case, for example, in the close configuration of the C-lobe of apolactoferrin in absence of iron binding [114] or in the structure of the bacteriophage T4 lysozyme wild-type [115]. However, further analysis would be required to prove or reject the hypothesis.

*B*/MINPP is a phytase anchored to the cell wall of *Bifidobacterium longum*, and is thus exposed to the varying conditions to which its niche, the gut, is subjected. It may be performing catalysis in a quite viscous environment characterised by limited molecular diffusion, at varying pH which depends upon the temporary localisation of the bacteria in the gastro-intestinal tract. Also, it probably interacts with an assorted set of phosphomonoesters (e.g. inositol phosphates, casein phosphopeptides) . The closure/opening mechanism could ensure efficient diffusion of the substrate to the active site cleft and a selective binding to phytate, its single substrate. In fact, as shown by Tamayo-Ramos, Sanz-Penella [21], the enzyme is particularly active on IP<sub>6</sub> but displays a 50-fold decrease in activity to its second

preferred substrate, acetyl-phosphate (2.3 % activity - **Figure 20**). Activity towards lower phosphorylated compounds was not tested. The absence of the second arginine of the catalytic core motif RHGxR in HP1 phosphatases and its full conservation in the HP2 [26] particularly links these residues to substrate specificity. Full domain closure may be induced by a specific binding of IP<sub>6</sub> to Arg48 which cannot be obtained with any other monophosphoester or phosphate-containing compound (as an induced-fit mechanism). It is interesting to note that the phosphate containing structure solved by Zietek [42] undergoes, for example, only a partial domain closure (**Figure 14**).

**Table 4.7.2. Activity of bifidobacterial phytases on different phosphorylated substrates. [21].**

Substrate (1 mM)	% phytase activity (mean $\pm$ SD) <sup>a</sup>	
	PhypA	PhylA
Phytate	100 $\pm$ 0.25	100 $\pm$ 0.34
<i>p</i> -Nitrophenyl-phosphate	0.9 $\pm$ 0.42	1.05 $\pm$ 0.11
Phosphoenolpyruvate	ND	ND
Acetyl-phosphate	6.7 $\pm$ 1.8	2.3 $\pm$ 0.45
AMP	ND	0.4 $\pm$ 0.11
ADP	ND	ND
ATP	ND	ND
Glyceraldehyde-3-phosphate	ND	ND
Glucose 1-phosphate	ND	ND
Glucose 6-phosphate	ND	ND
Fructose 1-phosphate	ND	ND
Fructose 1,6-bisphosphate	4.9 $\pm$ 0.1	1.45 $\pm$ 0.45

<sup>a</sup> The reactions were carried out in triplicate in 100 mM acetate buffer, pH 5.5, at 50°C. ND, no activity detected.

## 4.8. Discussion

*B*/MINPP is the first HP phytase seen to undergo a significant conformational change during catalysis. This begs the question as to why this movement occurs and whether there are other MINPPs phytases which follow this mechanism.

The residues involved in forming the mechanical hinge and the screw axis about which domain rotation takes place have been identified and their degree of conservation among MINPPs has been established. The identification of the hinge about which motion occurs suggests that domain movement is initiated by the binding of IHS (mimicking IP<sub>6</sub>) to the highly conserved active site residues Arg48 and Arg142. Substrate-binding induced conformational changes in these residues propagates to the moving domain.

The essentiality of domain movement in *B*/MINPP has been considered. The conformational change may take place to allow additional residues on the U-loop (conserved only in a limited number of MINPPs from *Actinobacteria*, *β/γ-Proteobacteria* and *Firmicutes*) to contribute to substrate coordination and stereospecificity during catalysis. Hinge residues that contributes to the catalytic core of *B*/MINPP are conserved among MINPPs. The enzyme is proposed to achieve IP<sub>6</sub> hydrolysis following the general mechanism of catalysis of the HP2 family, for these reasons an intrinsic flexibility of MINPPs as a family should not be excluded *a priori*. In addition, *B*/MINPP is a cell wall anchored enzyme facing a variety of environmental conditions; in this context the closure mechanism would ensure efficient diffusion of the substrate and selective binding of phytate to the active site.

The data presented in this chapter reveals a previously unknown conformational change involved in *B*/MINPP catalysis, which may contribute to the substrate specificity profile of this enzyme. In principal, this may allow the identification of alternative candidate MINPPs enzymes which employ a mobile  $\alpha$ -domain during catalysis by sequence database search. Particularly interesting would be the generation of a U-loop knock-out and the mutation of the loop residues involved in enzyme:substrate interactions (Asp289, Glu293, Lys296) to gain insights into the U-loop functionality.

# CHAPTER 5

## 5. Genome mining of positionally non-stereoselective MINPPs

This chapter describes a procedure employed to identify representative MINPPs sequences in bacterial genomes, leading to the selection of 16 candidate enzymes for enzymatic and structural characterization. The family of clade 2 histidine phosphatases (HP2) was analysed and its members classified according to (1) positional stereospecificity of IP<sub>6</sub> hydrolysis, (2) proton donor type (the involvement of the proton donor in determining positional specificity was discussed in **Chapter 3**) and (3) phylogenesis. Its members were further organised in four main groups (A) HP2 non-phytases, (B) *E. coli* AppA-like bacterial phytases, (C) MINPPs and (D) fungal phytases. Conserved regions shared across the HP2 family, or unique to each subgroup (A, B, C or D) were identified and mapped onto the available macromolecular crystal structures to determine their possible function(s). These sequence motifs were then used in the analysis of bacterial MINPPs sequences mined by database search.

Sixteen candidate sequences were selected for characterization (**Chapter 6**) on the basis of their active site composition, which showed a variety of traits potentially involved in determining positional stereospecificity of IP<sub>6</sub> hydrolysis in MINPPs.

The genome-mining procedure involved the following four main steps:

1. Identification of a set of HP2 "standard sequences": enzymes whose activity (and, when possible, stereospecificity) has already been characterized experimentally;
2. The analysis of standard sequences to detect sequence patterns potentially underlying differences in protein function;
3. The collection of hypothetical HP2 sequences through runs of BLAST [1, 2] searches using as queries the bacterial MINPPs sequences among the standards;
4. Comprehensive sequence analysis to detect interesting features to be further explored through the biochemical characterisation of selected enzymes.

## 5.1. Collection of standard sequences

Branch 2 of the family of histidine phosphatases (His\_Phos\_2 family, branch 2, PfamID: PF00328) is a group with a common structural architecture – a conserved  $\alpha/\beta$ -domain and a more diverse  $\alpha$ -domain – but with functional drifts between its components [26]. Among its members can be found not only phytases but also enzymes that hydrolyse primarily different substrates e.g. glucose-1 phosphatase or lysophosphatidic acid phosphatase. Thus, it becomes important, when searching specifically for MINPPs sequences, to be able to distinguish them from enzymes that may have different substrate specificity or different positional specificity in IP<sub>6</sub> hydrolysis.

Several branches 2 HP enzymes (HP2) were collected through a literature search and subsequently analysed to highlight the unique features of each functional subgroup. Automatically annotated HP enzymes were excluded and only sequences whose functionality has already been characterized experimentally were reported here as 'standard sequences'. The collected enzymes are listed in **Table 5.1.1**. They were grouped as: enzymes that are not primarily phytases (group A), sequential phytases (group B and D: bacterial and fungal phytases, respectively) and MINPPs phytases (group C). Sequential phytases were furthermore categorized for their preferential site of attack as 3- or 6-phytases. Also, their X-ray crystal structures, when available, were used for the structural comparisons of the catalytic cleft (**Paragraph 5.3**).

**Table 5.1.1. List of HP-clade2 sequences whose activity has been confirmed experimentally.**

These sequences are referred as 'standard' throughout Chapters 5 and 6. They were used in the identification of MINPPs-unique conserved sequence patterns. Standards are divided in four groups: group A – non-phytases, group B – bacterial phytases, group C – MINPPs and group D – fungal phytases. To each enzyme is assigned a letter as identifier of the group and a number as identifier of the enzyme (e.g. a5, b10). For each enzyme, GenBank or UniProt ID and references are provided. Superscripts a, b, c, d and e indicate respectively: HP of known three-dimensional structure, HP that are primarily 6-phytases, HP that are primarily 3-phytases, HP that are primarily 5-phytases and promiscuous phytases.

Group	Nr°	Sequences	GenBank/UniProt ID
Group A: non- phytases	6	<b>a1-</b> <i>Escherichia coli</i> glucose-1-phosphatase [1] <b>a2-</b> <i>Homo sapiens</i> lysophosphatidic acid phosphatase [2] <b>a3-</b> <i>Homo sapiens</i> testicular acid phosphatase precursor [3] <b>a4-</b> <i>Homo sapiens</i> lysosomal acid phosphatase isoform 1 precursor [4] <b>a5-</b> <i>Homo sapiens</i> prostatic acid phosphatase [5] <b>a6-</b> <i>Francisella tularensis</i> HP [6]	KYQ65676 <sup>a</sup> BAA89311 <sup>a</sup> NP_149059.1 NP_001601  AAA60021 <sup>a</sup> OCQ70881 <sup>a</sup>

Group B: bacterial phytases	16	<b>b1- <i>Escherichia coli</i> HP phytase AppA [7]</b> <b>b2- <i>Yersinia pestis</i> HP phytase [8]</b> <b>b3- <i>Citrobacter amalonaticus</i> HP phytase [9]</b> <b>b4- <i>Yersinia intermedia</i> HP phytase [10]</b> <b>b5- <i>Citrobacter braakii</i> HP phytase [11]</b> <b>b6- <i>Obesumbacterium proteus</i> HP phytase [12]</b> <b>b7- <i>Yersinia kristensenii</i> HP phytase [13]</b> <b>b8- <i>Buttiauxella sp.GC21</i> HP phytase [14]</b> <b>b9- <i>Pseudomonas syringae</i> HP phytase [15]</b> <b>b10- <i>Burkholderia sp.a13(2014)</i> HP phytase [16]</b> <b>b11- <i>Shigella sp.CD2</i> HP phytase [17]</b> <b>b12- <i>Enterobacter agglomerans</i> or <i>Erwinia herbicola</i> or <i>Pantoea agglomerans</i> HP phytase (unpublished)</b> <b>b13- <i>Pectobacterium carotovorum</i> subsp. <i>carotovorum</i> or <i>Erwinia carotovora</i> subsp. <i>carotovora</i> HP phytase [19]</b> <b>b14- <i>Dickeya paradisiaca</i> HP phytase [20]</b> <b>b15- <i>Klebsiella pneumoniae</i> HP phytase [21]</b> <b>b16- <i>Hafnia alvei</i> HP phytase [22]</b>	A0A017IGE1 <sup>a,b</sup> ABU98780 ABI98040 ABI95370.1 AAS45884.1 <sup>a,b</sup> AAQ90419 ABX75421 <sup>a,b</sup> ABX80238.1 AAN77879.1 BAQ94585.1 CCA94903.1 ADZ99368.1  ABY76184.1  ABW76125.1 AAL59319.1 <sup>a,c</sup> AFG25721.1 <sup>a,b</sup>
Group C: MINPPs	13	<b>c1- <i>Bifidobacterium longum</i> MINPP [23]</b> <b>c2- <i>Bacteroides thetaiotaomicron</i> MINPP [24]</b> <b>c3- <i>Bifidobacterium pseudocatenulatum</i> MINPP [25]</b> <b>c4- <i>Triticum aestivum</i> MINPP [26]</b> <b>c5- <i>Hordeum vulgare</i> MINPP [27]</b> <b>c6- <i>Rattus norvegicus</i> MINPP [28]</b> <b>c7- <i>Mus musculus</i> MINPP [29]</b> <b>c8- <i>Homo sapiens</i> MINPP, isoform1 [30]</b> <b>c9- <i>Lilium longiflorum</i> MINPP [31]</b> <b>c10- <i>Gallus</i> MINPP [32]</b> <b>c11- <i>Dictyostelium discoideum</i> MINPP1 [33]</b> <b>c12- <i>Drosophila melanogaster</i> MINPP 1a [34]</b> <b>c13- <i>Drosophila melanogaster</i> MINPP2 [35]</b>	B7GTVO <sup>a,e</sup> WP_057099050 <sup>a,e</sup> COBTR1 <sup>e</sup> A0FHBO A0FHA7 O35217 Q9Z2L6 Q9UNW1 Q0GYS1 <sup>d</sup> NP_989975 XP_638245 NP_524109 NP_511055
Group D: fungal phytases	14	<b>d1- <i>Agrocybe pediades</i> HP phytase [36]</b> <b>d2- <i>Peniophora lycii</i> HP phytase [37]</b> <b>d3- <i>Ceriporia sp. CBS 100231</i> HP phytase [38]</b> <b>d4- <i>Trametes pubescens</i> HP phytase [39]</b> <b>d5- <i>Aspergillus fumigatus</i> HP phytase [40]</b> <b>d6- <i>Myceliophthora thermophila</i> HP phytase [41]</b> <b>d7- <i>Thermomyces dupontii</i> HP phytase [42]</b> <b>d8- <i>Penicillium oxalicum</i> <i>Myceliophthora thermophila</i> HP phytase [43]</b> <b>d9- <i>Trichoderma reesei</i> HP phytase [44]</b> <b>d10- <i>Aspergillus niger</i> HP phytase [45]</b> <b>d11- <i>Schwanniomyces capriottii</i> HP phytase [46]</b> <b>d12- <i>Cyberlindnera fabianii</i> HP phytase [47]</b> <b>d13- <i>Wickerhamomyces anomalus</i> or <i>Pichia anomala</i> HP phytase [48]</b> <b>d14- <i>Debaryomyces castellii</i> HP phytase [49]</b>	Q96VTO <sup>b,c</sup> CAC48195.1 <sup>b</sup> CAC48164.1 <sup>b,c</sup> CAC48234.1 <sup>b,c</sup> AHZ62778.1 <sup>c</sup> AAB52508.1 <sup>c</sup> AAB96873.1 <sup>c</sup> AAL55406.1 <sup>c</sup>  ACQ15319.1 <sup>c</sup> CAA78904.1 <sup>a,c</sup> ABN04184 BAH58739.1 CBI71332.1  AEN51860.1 <sup>c</sup>

<sup>a</sup> known three-dimensional structure; <sup>b</sup> HP 6-phytases; <sup>c</sup> HP 3-phytases; <sup>d</sup> HP 5-phytases; <sup>e</sup> promiscuous phytases.

[1] Dassa, Marck [116] [2] Li, Dong [117], [3] [118], [4] [119], [5] Etten [27], [6] Singh, Felts [120], [7] Lim, Golovan [24], [8] [121], [9] Luo, Huang [122], [10] Huang, Luo [123], [11] Kim, Kim [124], [12] Zinin, Serkina [125], [13] Fu, Huang [126], [14] Shi, Huang [127], [15] Cho, Lee [128], [16] Graminho, Takaya [129], [17] [130], [19] [131], [20] [132], [21] [133], [22] Ariza, Moroz [134], [23, 24] Tamayo-Ramos, Sanz-Penella [21], [25] Stentz, Osborne [41], [26, 27] Dionisio, Holm [45], [28] Craxton, Caffrey [36], [29,30] Chi, Tiller [22], [31] Mehta, Jog [23], [32] Cho, Choi [39], [33] Cho, King [38], [34,35] Cheng and Andrew [47], [36-39] Lassen, Breinholt [135], [40] Liu, Huang [34], [41] Lee, Cho [136], [42] [137], [43] Lee, Cho [136], Lee, Choi [138], [44] [139], [45] Oakley [140], Phillippy and Mullaney [141], [46] [142], [47] [143], [48] Joshi and Satyanarayana [144], [49] Boze, Aumelas [13].

## 5.2. Grouping of standard sequences

MINPPs contain the same catalytic conserved residues as other HP2 enzymes: specifically, the “RHGxRxP” motif, the “R” and the “Hxx” proton donor motif. The most common proton donor motif in MINPPs is the “HAE” triplet, which differs from the “HDx” triplet of *E. coli*-like and yeast phytases. Despite this, MINPPs with different proton donor motif have been identified [22, 45, 47] and, for this reason, the feature cannot be considered a MINPPs signature. Usually, MINPPs are distinguished by phylogenesis because they group as a family in a separate branch from other HP [22, 23, 41, 45]. However, multiple ways of HP2 classification were taken into consideration in the analyses of this family of enzymes.

Standard HP2 enzymes were processed in two stages:

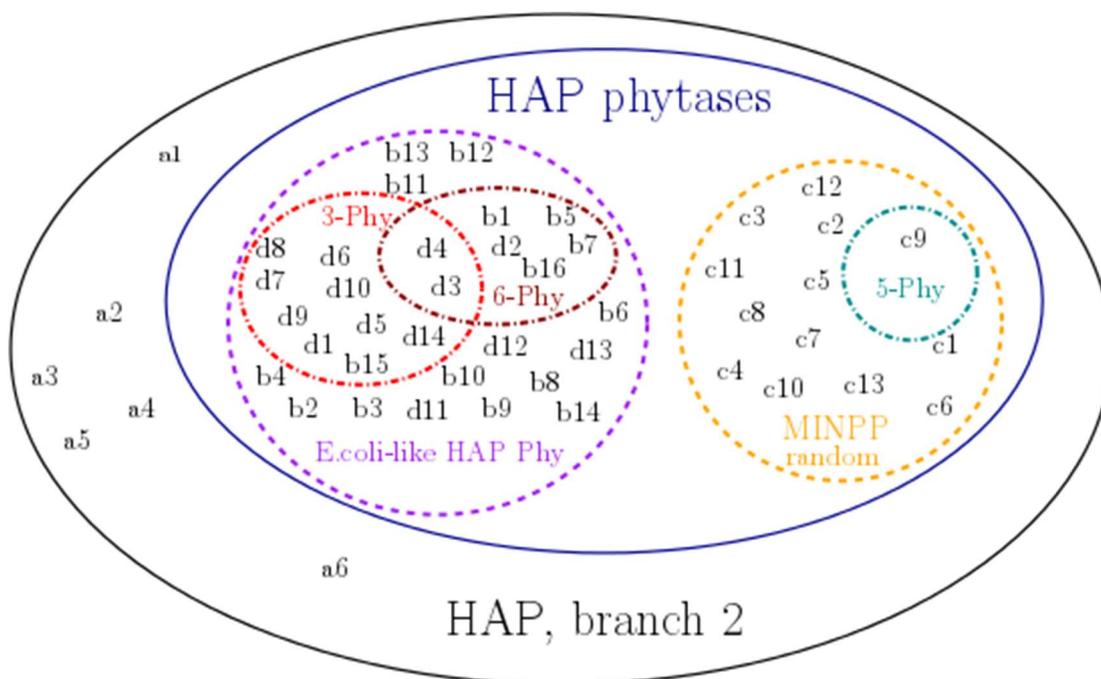
1. they were grouped: by substrate specificity, by proton donor, by phylogenesis;
2. each group is analysed and compared to the others in order to highlight group-unique features by multi-sequence alignments and inspection of the available X-ray crystal structures.

### 5.2.1. Substrate specificity

The first clear divergence in the family of HP2 is functional: its members can hydrolyse a variety of phosphomonoesters and not all have phytase activity.

Standard enzymes listed in **Table 5.1.1** are represented in a graphic of sets (**Figure 5.2.1.1**) which groups them accordingly to their substrate specificity. The family of HP2, includes: HP that are not phytases, i.e. enzymes that hydrolyse primarily substrates other than phytic acid, such as glucose-1-phosphate - a1 [145] or lysophosphatidic acid - a2 [117], and HP phytases (**blue set**) [26]. HP phytases can be furthermore divided into MINPPs phytases (**orange set**), with the tendency to hydrolyse almost indiscriminately five of the six phosphate groups on IP<sub>6</sub> [21, 41], in fact, only two phytases that cleave the axial phosphate at position 2 of the carbon ring have been yet identified: *Debaryomyces castellii* CBS 2923 and *Schwanniomyces occidentalis* HP [13], and “*E. coli*-like” HP phytases (**purple set**), which seem to process IP<sub>6</sub> groups in a sequential order [13, 133-135, 139, 141, 146-148]. An exception to the rule is the *Lilium longiflorum* MINPP which displays preferential 5-phytase activity (c9, **cyan set**). The set of “*E. coli*-like” HP phytases includes bacterial and fungal phytases. They can be organized in two further subsets: 3-phytases (**red set**)

and 6-phytases (**brown set**), in accordance to their preferential site of attack. The products of hydrolysis of enzymes that are not included in the red and brown sets have not been yet characterized by HPLC chromatography.



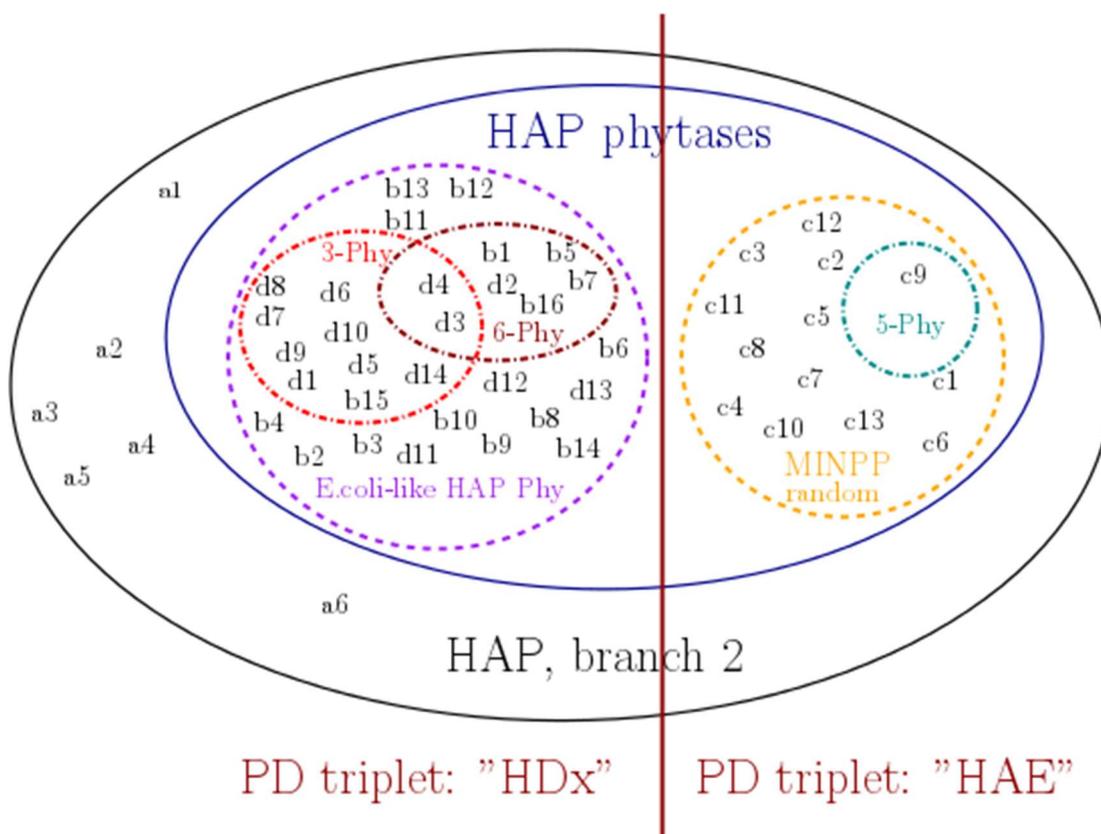
**Figure 5.2.1.1. Sets representation of HP, branch 2.**

The **black set** contains the whole family of HP, branch 2. The **blue set** includes HP phytases. They can be classified by positional specificity of IP<sub>6</sub> hydrolysis in: 1) MINPPs, **orange set**, hydrolysis is almost indiscriminate with the exception of lily pollen MINPP 5-phytase – c9 (E.C.3.1.3.72), **cyan set**; 2) *E.coli*-like HP phytases, **purple set**, mostly classified in 3-phytases (E.C. 3.1.3.8), **red set**, and 6-phytases (E.C.3.1.3.26), **brown set**. Enzymes outside the 3-Phy and 6-Phy sets have not been yet characterized for positional specificity. This set includes bacterial and fungal phytases.

## 5.2.2. Proton donor

The core catalytic residues “RHGxR”, containing the catalytic histidine, seem to be conserved in all the members of this family. The proton donor motif (PD) though, diverges in MINPPs from the classical “HDx” triplet seen in HPs (**Figure 5.2.2.1**). The most common PD in this set is a glutamic acid found in an “HAE” motif (**Figure 5.2.2.1**). A few exceptions whose activities have been confirmed are: *Drosophila melanogaster* MIPP1 (sequence “HST”) and MIPP2 (sequence ”HSG”) [47], and *Dictyostelium discoideum* MIPP2 (sequence “HSE”) [38]. Alternative PD were identified in *Danio rerio* MIPPb (sequence ”HGE”), *Anopheles gambiae* MINPP (sequence “HST”), and *Cryptococcus neoformans* MIPP (sequence “HEV”) [47]. The “HST” and “HSG” substitutions are in contrast with the hypothesis of the essentiality of a carboxylic acid group in the active site for an HP-catalysed reaction to take place [27, 29, 31, 33], though, it could be possible that: 1) the catalytic rates of these

enzymes are impaired in comparison with “HD/Ex” or “HxD/E” counterpart; 2) in these enzymes a Glu or an Asp residue may be inserted at another position in the active site cleft; 3) proton donation may occur through an alternative mechanism.



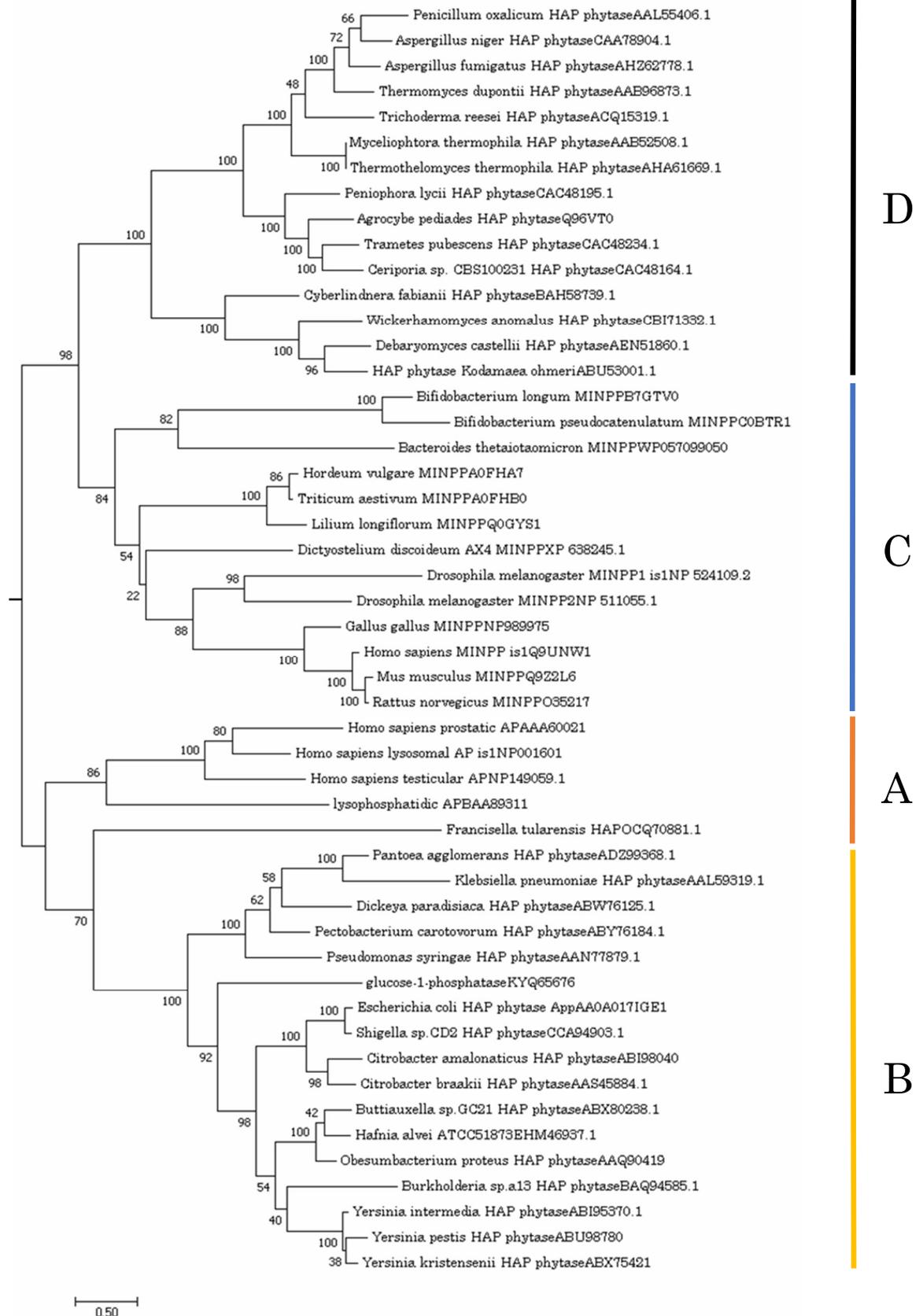
**Figure 5.2.2.1. Sets representation of PD motifs distribution in HP, branch 2.**

The same sets as in **Figure 5.2.1.1** are displayed, a **brown line** marks the division between phytases containing the “**HDx**” proton donor motif and the “**HAE**”-containing MINPPs. Only exceptions to the rule are *Drosophila melanogaster* MINPP1 and 2 (c12 and c13), which display an “**HST**” and “**HSG**” and the *Dictyostelium discoideum* MINPP2 (c11), “**HSE**”-containing. The **black set** contains the whole family of HP, branch 2. The **blue set** includes HP phytases. They can be classified by positional specificity of IP<sub>6</sub> hydrolysis in: 1) MINPPs, **orange set**, hydrolysis is almost indiscriminate with the exception of lily pollen MINPP 5-phytase – c9 (E.C.3.1.3.72), **cyan set**; 2) *E.coli*-like HP phytases, **purple set**, mostly classified in 3-phytases (E.C. 3.1.3.8), **red set**, and 6-phytases (E.C.3.1.3.26), **brown set**. Enzymes outside the 3-Phy and 6-Phy sets have not been yet characterized for positional specificity. This set includes bacterial and fungal phytases.

### 5.2.3. Phylogenesis

The evolutionary history of the collected standard sequences was inferred using the Maximum Likelihood method based on the JTT matrix-based model [149]. The tree with the highest log likelihood (-39216.72) is shown in **Figure 5.2.3.1**. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.0297)). A bootstrap analysis was inferred from 50 replicates [150]. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The analysis involved 50 amino acid sequences. All positions with less than 25% site coverage were eliminated. That is, fewer than 75% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 535 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [151].

The resulting phylogenetic tree is clearly divided in two branches: one includes fungal phytases and MINPPs, the other includes bacterial phytases and acid phosphatases of diverse function. Prostatic, lysosomal, testicular and lysophosphatidic acid phosphatases (**branch A**) appear to be closer to bacterial HP2 than to MINPPs or fungal phytases. In fact, they are clustered in a separate branch and share a probable common ancestor with the bacterial enzymes. The same happens for the unique gene of *Francisella tularensis* [120]. The branch of bacterial enzymes (**branch B**) is instead split in three: two branches of bacterial phytases that evolved separately from each other and an additional branch representing *E. coli* glucose-1-phosphatase (92% of trees). MINPPs (**branch C**) seem to have evolved in parallel with the organisms that host them: bacterial MINPPs belong to a separate clade from eukaryotic enzymes, plant MINPPs cluster in a clade distinct from animal enzymes, the amoeba *D. discoideum* MINPPs is clustered in a separate branch from *Arthropoda* or *Chordata* enzymes. MINPPs are the only HP2 enzymes shared among all kingdom of life with the exception of *Archaea*: their rate of evolution and the data collected *in vivo* suggest an essential role for these proteins, at least in *Eukaryotes*, and an ancient origin for these particular HPs family members as proposed by Stentz et al (2014) [41]. Two quite distinct branches of fungal phytases (**branch D**) have been identified with high confidence (100% of trees), one of the two is forked in turn into two further branches.



**Figure 5.2.3.1. Molecular Phylogenetic analysis by Maximum Likelihood method.**

**A)** histidine phosphatases with diverse functionality; **B)** bacterial phytases and *E. coli* glucose-1-phosphatase; **C)** MINPPs **D)** fungal phytases. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

## 5.3. Analysis of standards

### 5.3.1. X-ray crystal structures

The X-ray crystal structures available for members of HP2 were inspected and used as reference for the manual curation of multiple sequence alignments of the standard sequences and to define substrate binding pockets. The available structures analysed for each group are:

- Group A: *Homo sapiens* lysophosphatidic acid phosphatase type 6 in complex with L-(+)-tartrate (PDB id: 4job, **Figure 5.3.1.1 - A.a**), malonate (PDB id: 4joc), Tris (PDB id: 4jod); *Homo sapiens* prostatic acid phosphatase apo (PDB id: 1cbi) or in complex with L-(+)-tartrate (PDB id: 2hpa, **Figure 5.3.1.1 - A.b**), phosphate (PDB id: 1nd6), alpha-benzylaminobenzylphosphonic acid (PDB id: 1nd5).
- Group B: *Citrobacter brakii* HPPHy (PDB id: 3xhc); *Enterobacter cloacae* HPPHy (PDB id: 3ntl); *Escherichia coli* AppA apo (PDB id: 1dkl) or in complex with inositol hexasulphate (PDB id: 4tsr, **Figure 5.3.1.1 - B.a**), phytate (PDB id: 1dkp), tungstate (PDB id: 1dko); *Escherichia coli* glucose-1-phosphatase in complex with glucose-1-phosphate (PDB id: 1nt4, **Figure 5.3.1.1 - B.b**); *Hafnia alvei* HPPHy apo (PDB id: 4ars) or in complex with inositol hexasulphate (PDB id: 4aro) or L-(+)-tartrate (PDB id: 4aru); *Klebsiella* sp. ASR1 HPPHy apo (PDB id: 2wnh) or in complex with sulphate (PDB id: 2wni, 2wu0); *Yersinia kristensenii* HPPHy (PDB id: 4arv); *Legionella pneumophila* HPPHy in complex with L-(+)-tartrate.
- Group C: *Bacteroides thetaiotaomicron* MINPPs in complex with phosphate (PDB id: 4fdt) or inositol hexasulphate, (PDB id: 4fdu, **Figure 5.3.1.2 - C.a**); *Bifidobacterium longum* subsp. *infantis* MINPPs apo, in complex with inorganic phosphate or inositol hexasulphate (**Figure 5.3.1.2 - C.b**).
- Group D: *Aspergillus awamori* HPPHy in complex with sulphate (PDB id: 1qfx); *Aspergillus fumigatus* HPPHy apo (PDB id: 1qwo, 1ska, 1skb) or in complex with inorganic phosphate (PDB id: 1sk8, 1sk9, **Figure 5.3.1.2 - D.a**); *Aspergillus niger* PhyA apo (PDB id: 1ihp, 3k4p) or in complex with inositol hexasulphate (PDB id: 3k4q); *Debaryomyces castellii* HPPHy apo (PDB id: 2gfi, **Figure 5.3.1.2 - D.b**).

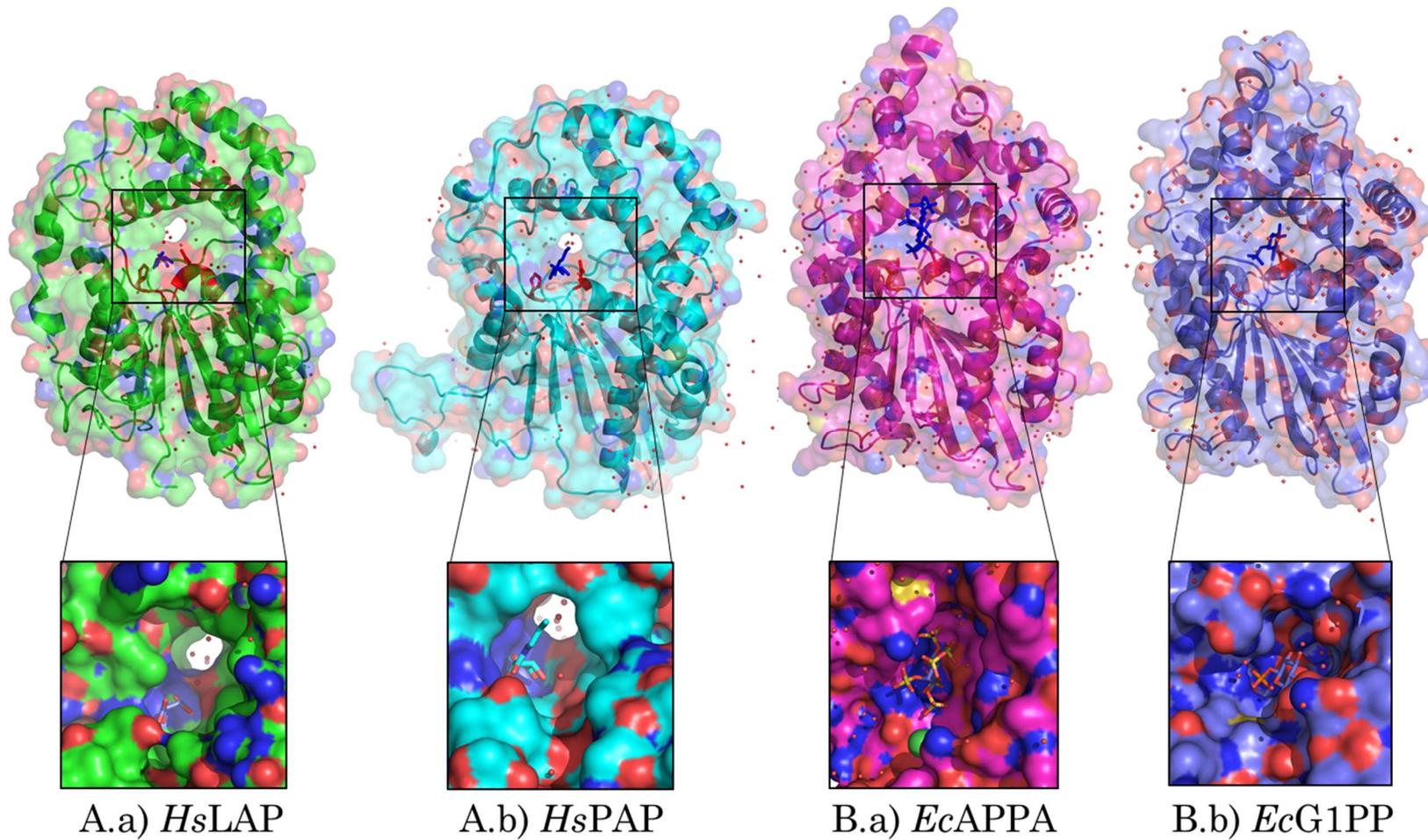


Figure 5.3.1.1. Crystal structures of standard sequences of class A and B. page 139

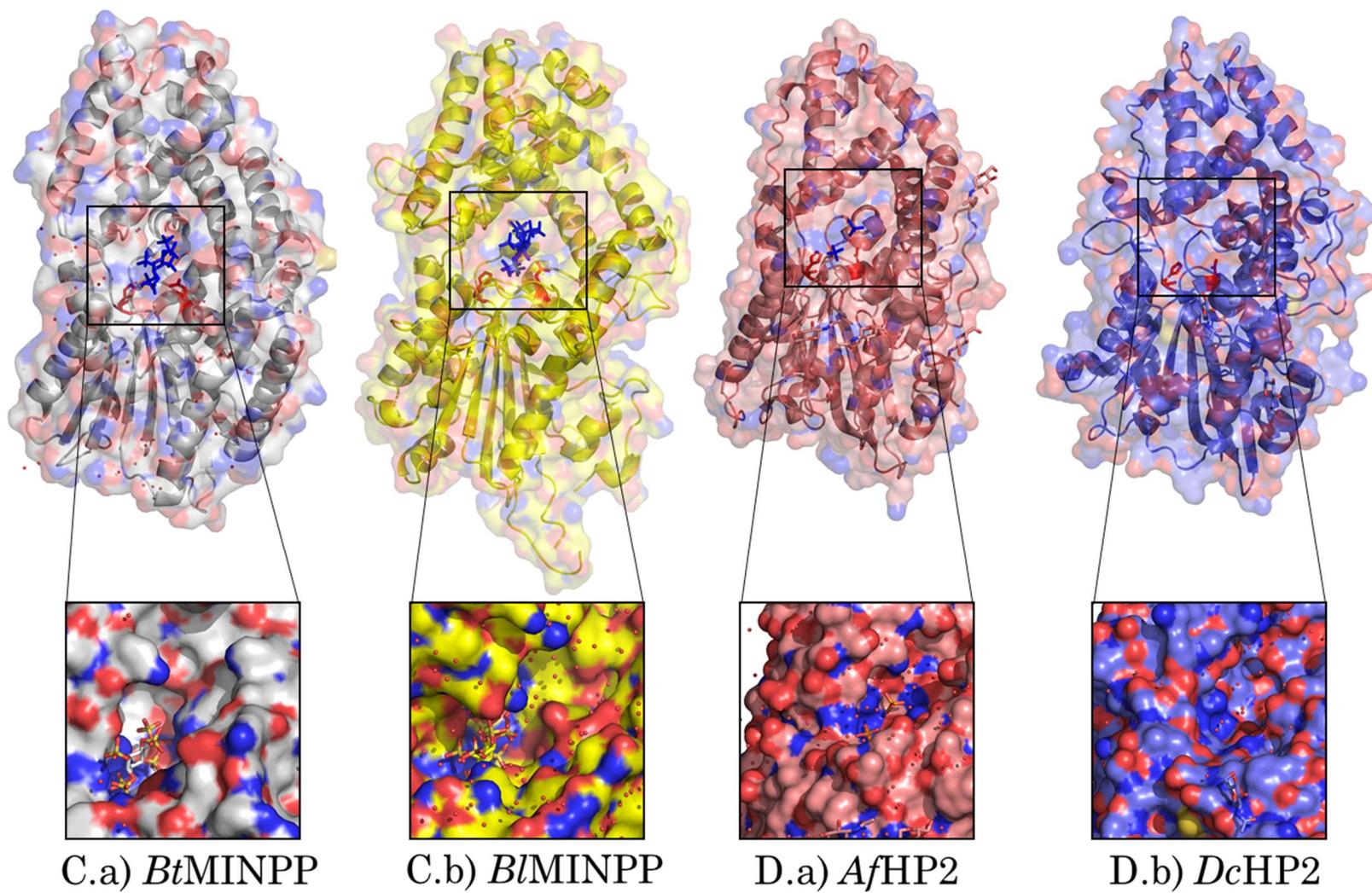


Figure 5.3.1.2. Crystal structures of standard sequences of class C and D. page 139

**Figure 5.3.1.1 and 5.3.1.2: Crystal structures of enzymes corresponding to standard sequences of class A, B, C and D.**

Surface and tertiary structure (cartoon) of selected enzymes are displayed, as well as a close-up of the active site cleft. Ligands are shown as sticks and waters as dots. **A.a)** *Homo sapiens* lysophosphatidic acid phosphatase type 6 (*HsLAP*) in complex with L-(+)-tartrate (PDB id: 4job); **A.b)** *Homo sapiens* prostatic acid phosphatase (*HsPAP*) in complex with n-propyl-tartramic acid (PDB id: 1nd6); **B.a)** *Escherichia coli* AppA (*EcAppA*) in complex with inositol hexasulphate (PDB id: 4tsr); **B.b)** *Escherichia coli* glucose-1-phosphate phosphatase (*EcG1PP*) in complex with glucose-1-phosphate (PDB id: 1nt4); **C.a)** *Bacteroides thetaiotaomicron* MINPPs (*BtMINPP*) in complex with inositol hexasulphate (PDB id: 4fdu); **C.b)** *Bifidobacterium longum* susp. *infantis* MINPP (*BtMINPP*) in complex with inositol hexasulphate (unpublished); **D.a)** *Aspergillus fumigatus* HPPhy (*AfHP2*) in complex with phosphate (PDB id: 1sk8); **D.b)** *Debaryomyces castellii* HPPhy (*DcHP2*) apo (PDB id: 2gfi).

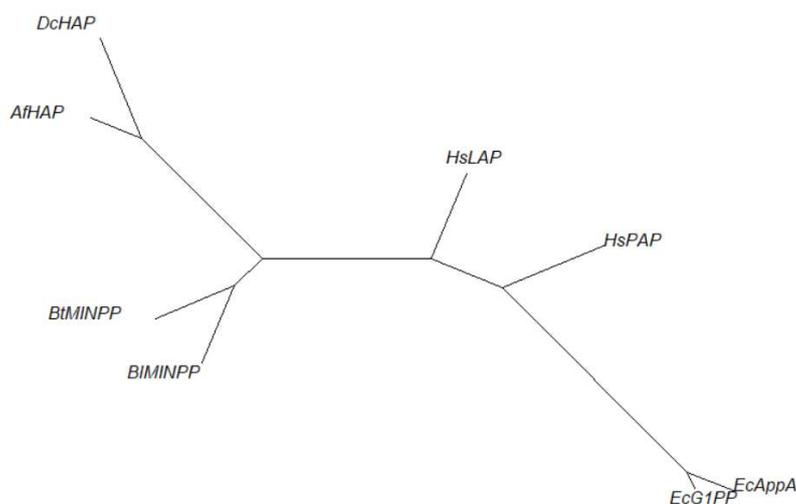
HP phytase structures are made up of two domains: a conserved  $\alpha/\beta$ -domain and a more diverse  $\alpha$ -domain with the active site lying in between the two. The review of Rigden [26] gives more insight into domains architectures among the family. In particular, the clade 2 of HPs has a conserved minimal scaffold with enzyme-specific insertions and deletions. The active site residues involved in catalysis are the conserved nucleophilic histidine of the motif RHGxRxP, where the two arginines have a role in orienting the substrates in the optimal position for hydrolysis. A second histidine of the motif “HD/xE/x” contribute to restore the protonation of the catalytic histidine after the first step of catalysis and a glutamic or aspartic acid residues presumably donates a proton to the substrate reducing the free energy barrier in cleavage. Other residues contribute in defining size, shape and local charge of the active site regions affecting substrate specificity and positional specificity of cleavage. This topic will be discussed further in **Chapter 6**.

When looking at the global enzyme structures it is noticeable that members within the same group (A, B, C, or D) are more structurally related than to other HP2 members, this evidence partially validates the phylogenetic tree of the family reported in **Chapter 5.3**. An interesting finding is that the shape and charge distribution of the active sites seem generally not conserved and enzyme-specific. Also, it must be highlighted in structures of group A (HP that are not phytases, **Figure 5.3.1.1 – A**), the presence of a water tunnel that reaches the back of the active site, which is absent in the other HP phytases and glucose-1-phosphatase (**Figure 5.3.1.1/2 – B, C, D**). Comparing the latter with *EcAppA*, both group B members, it is possible to note a smaller catalytic cleft in glucose-1-phosphatase than in the *E. coli* phytase, in line with their differences in substrate specificity (**Figure 5.3.1.1 – B**). In group C, *BtMINPP* (in its closed conformation) shows a very small active site in comparison with *BtMINPP* (**Figure 5.3.1.2 – C**). In fact, *BtMINPP*, unlike *BtMINPP*, undergoes

a conformational change during its four stages of catalysis and it need to open to allow the entrance of the substrate in the active site cleft (**Chapter 4**). Instead, fungal phytases structures (group D) seem to resemble more closely MINPPs structural features than other HP2 members (**Figure 5.3.1.2.D**).

As mentioned before, in human lysophosphatidic acid phosphatase (**Figure 5.3.1.1, A.a**) and prostatic acid phosphatase (**Figure 5.3.1.1, A.b**) a flexible loop extension, which acts as a closing lid at the back of the active site, is positioned to allow the formation of an open water tunnel, unlike other HP2 enzymes. In the structures of bacterial HP2 the lid is instead in a “close” conformation, shielding the back of the active site. It is interesting to note that, in contrast, fungal HP2 and MINPPs have a longer loop insertion which partially folds as an  $\alpha$ -helix making interactions with the  $\alpha$ -domain and anchoring the loop on the side of a big water tunnel. These enzymes though, have introduced an alternative loop/helix lid at their N-terminal, effectively ensuring the complete closure of the tunnel. It would be interesting to study the dynamic behaviour of these proteins to investigate the movements of the lids which could have multiple roles in catalysis. An alternative approach would be the expression of “lid-deleted” enzymes to inspect the effects of lid-removal in catalysis.

To have an idea of the overall structural differentiation of HP2, sample structures from each group were pairwise aligned using FATCAT. R.m.s.d and raw similarity scores were stored (**Table 5.3.1.1**). Similarity scores were rescaled to build a matrix of distances between structures. The matrix was used to draw a dendrogram by neighbour joining in R (**Figure 5.3.1.3**). The structural similarity dendrogram is in line with the phylogenetic tree inferred for HP2.



**Figure 5.3.1.3. Structural similarity dendrogram of HP2.**

Proteins are identified by the same acronyms as in **Table 5.3.1.1**. The dendrogram was obtained from a matrix of structural distances. The tree was generated by neighbour joining using R.

**Table 5.3.1.1. Matrix of the root mean square deviations calculated from the pairwise alignment of eight standard HP2 proteins.**

Enzymes groups: group A – *Homo sapiens* lysophosphatidic acid phosphatase (**HsLAP**) and prostatic acid phosphatase (**HsPAP**); group B – *Escherichia coli* AppA (**EcAppA**) and glucose-1-phosphatase (**EcG1PP**); group C – *Bifidobacterium longum* subsp. *infantis* MINPP (**BIMINPP**) and *Bacteroides thetaiotaomicrom* MINPP (**BtMINPP**); group D – *Debaryomyces castellii* HPPhy (**DcHP**) and *Aspergillus ficuum* HPPhy (**AfHP**). For each protein r.m.s.d. is reported (in brackets) as well as the raw similarity score produced by FATCAT.

(r.m.s.d) raw score	<b>DcHP</b>	<b>AfHP</b>	<b>BIMINPP</b>	<b>BtMINPP</b>	<b>HsLAP</b>	<b>HsPAP</b>	<b>EcG1PP</b>	<b>EcAppA</b>
<b>DcHP</b>	0	(2.59) 760.50	(3.09) 505.57	(3.05) 633.19	(3.13) 491.18	(2.97) 527.09	(3.05) 484.80	(3.03) 492.57
<b>AfHP</b>	(2.59) 760.50	0	(3.04) 555.44	(3.04) 636.10	(3.02) 502.00	(3.04) 540.98	(3.13) 514.58	(3.24) 502.29
<b>BIMINPP</b>	(3.09) 505.57	(3.04) 555.44	0	(2.50) 761.88	(3.35) 460.32	(3.08) 485.05	(3.24) 455.68	(3.04) 467.20
<b>BtMINPP</b>	(3.05) 633.19	(3.04) 636.10	(2.50) 761.88	0	(3.23) 479.63	(3.11) 540.25	(2.99) 522.98	(3.18) 497.65
<b>HsLAP</b>	(3.13) 491.18	(3.02) 502.00	(3.35) 460.32	(3.23) 479.63	0	(3.00) 624.91	(3.08) 511.10	(3.11) 523.14
<b>HsPAP</b>	(2.97) 527.09	(3.04) 540.98	(3.08) 485.05	(3.11) 540.25	(3.00) 624.91	0	(3.04) 517.92	(3.08) 571.09
<b>EcG1PP</b>	(3.05) 484.80	(3.13) 514.58	(3.24) 455.68	(2.99) 522.98	(3.00) 624.91	(3.04) 517.92	0	(2.32) 897.24
<b>EcAppA</b>	(3.03) 492.57	(3.24) 502.29	(3.04) 467.20	(3.18) 497.65	(3.08) 511.10	(3.08) 571.09	(2.32) 897.24	0

### 5.3.2. Identification of conserved motifs and their function

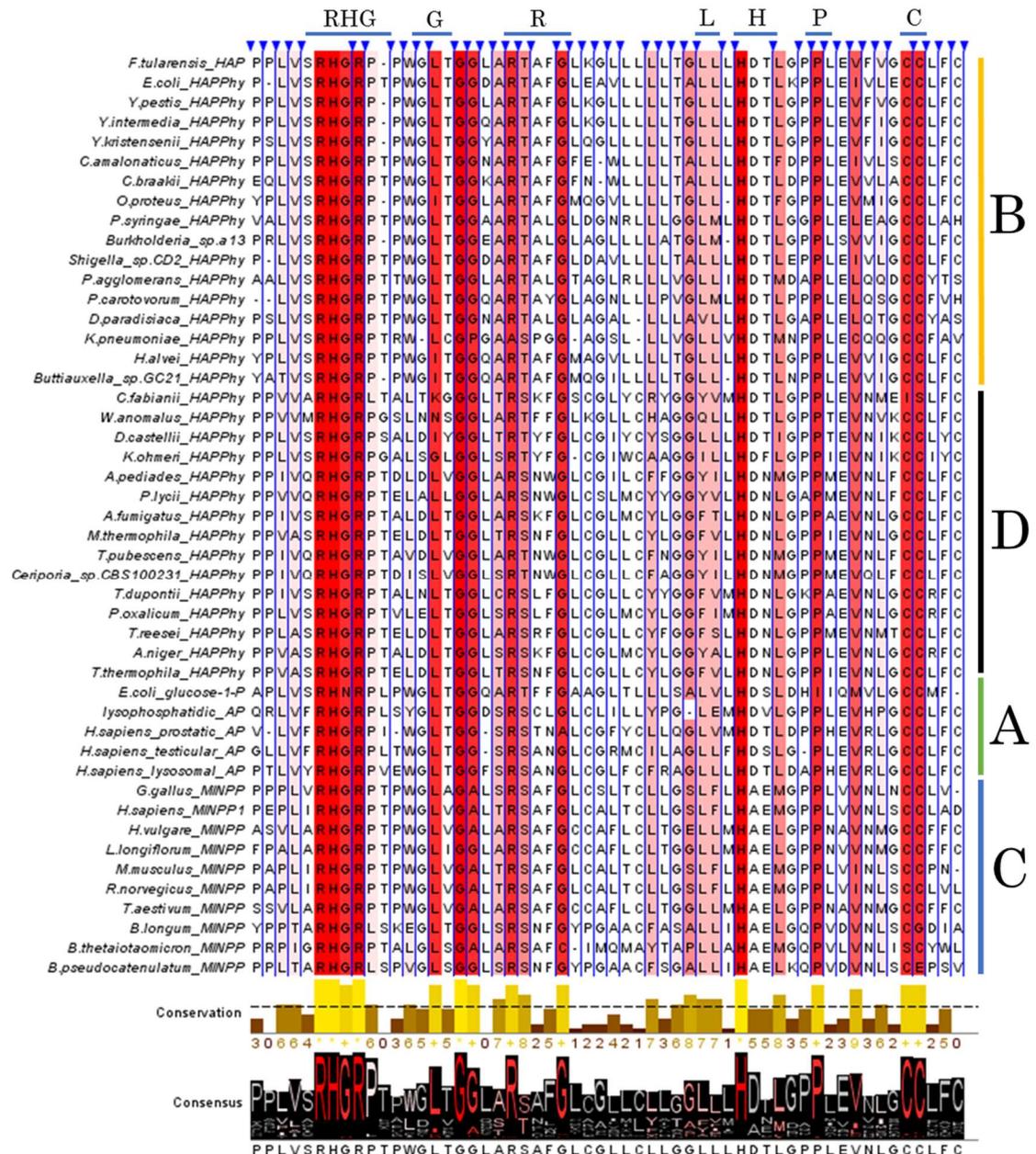
The family of HP2 was divided in the four groups (A, B, C, D), based on the phylogenetic analysis of standard sequences inferred in **Chapter 5.2.3**. Conserved patterns, either shared by the whole family or group-specific, were searched by multiple-sequence alignments and were used for the identification of novel MINPPs.

Multi-sequence alignments were performed with Clustal omega [152], MUSCLE [153], Kalign [154], T-coffee [155]. The MUSCLE algorithm seemed to generate better global alignments for this class of proteins, where other software failed in handling the sequences displaying the lowest similarity; for this reason, MUSCLE was chosen to be used in this work. Alignments quality was also estimated by GUIDANCE2 [156]. Low quality regions were manually edited using Jalview 2 [157] or ESPrpt3 [158], when needed, to match the structural data.

The first alignment included all the standard sequences collected in the HP2 family and highlighted the conserved motifs that are shared by HP2 (**Figure 5.3.2.1**).

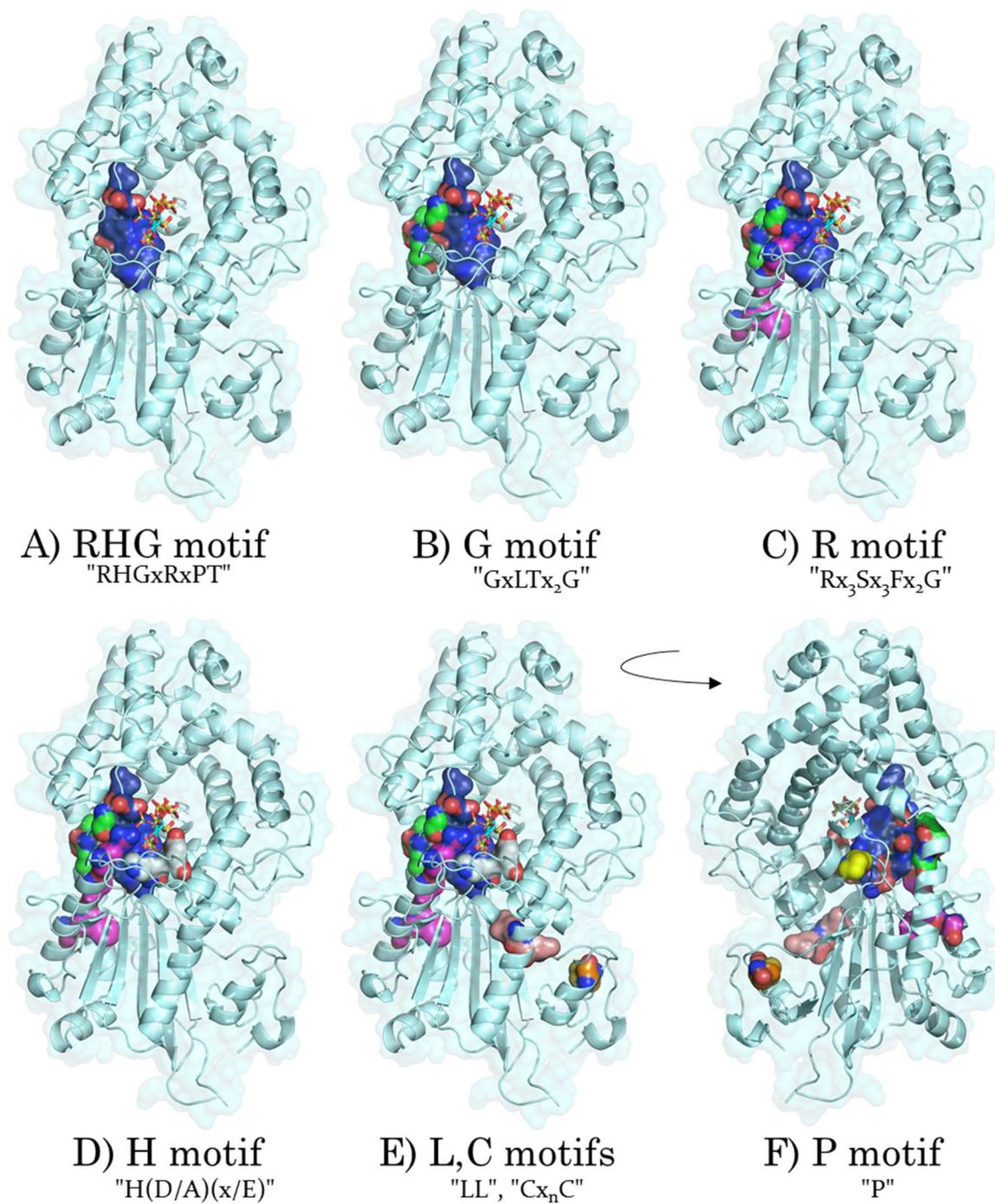
Motifs are mapped in **Figure 5.3.2.2** onto the *Bt*MINPP structure. They are: **(1)** the catalytic core “RHG<sub>x</sub>R<sub>x</sub>PT” (**Figure 5.3.2.2 - A**, RHG motif, displayed as blue surface); **(2)** the “G<sub>x</sub>LT<sub>x</sub>2G” motif (**Figure 5.3.2.2 - B**, G motif, displayed as green surface), which is involved in HP2 conformational changes (**Chapter 4**) and whose carbonyl group of the first glycine is able to contact either the second arginine of the RHG motif or the arginine of the R motif (2.9 Å distance); **(3)** the “R<sub>x</sub>3(S/T)<sub>x</sub>3F<sub>x</sub>2G” motif (**Figure 5.3.2.2 - C**, R motif, displayed as violet surface), which lies on an  $\alpha$ -helix and contains an arginine involved in substrate binding; **(4)** the “LL” motif (**Figure 5.3.2.2 - E** – L motif, displayed as pink surface), which is positioned in a hydrophobic pocket at the interface between  $\alpha$ - and  $\alpha/\beta$ -domain; **(5)** the proton donor sequence “H(D/A)(x/E)” (**Figure 5.3.2.2 - D**, H motif, displayed as grey surface); **(6)** a conserved proline (**Figure 5.3.2.2 - F**, P motif, displayed as yellow surface), whose backbone is involved in the stabilisation of the first arginine residue of the RHG motif; **(7)** and an N-terminal disulphide bridge (**Figure 5.3.2.2 - E**, CC motif displayed as orange surface), although, the latter is less conserved in bacterial MINPPs. The majority of the HP2 conserved motifs are either part of the catalytic core or interact with amino acids responsible for catalysis.

Multiple-sequence alignments of enzymes belonging to each phylogenetic group A, B, C, D were run in parallel to search for group unique-motifs (**Appendix 8.4.1**). Regions identified as MINPPs-specific (**Figure 5.3.2.3**) were six. **(1)** The “(Y/H)<sub>x</sub>2(S/T)<sub>x</sub>3Y” motif (**Figure 5.3.2.4 - A**, Y motif, displayed as green surface) is part of the N-terminal lid. The tyrosine residues help in the stabilisation of the loop, while the carbonyl group of Ser/Thr is involved in an interaction with the first arginine of the RHG motif (2.9 Å distance) and the hydroxy group lies at 3.7 Å from the ligand. **(2)** Residues “R<sub>x</sub>7LF” (**Figure 5.3.2.4 - B** – S motif, displayed as cyan surface) are probably conserved for reason of structural stabilization. The Arg residue faces the N motif of N-terminal lid loop and “LF” the  $\beta$ -sheet of the  $\alpha/\beta$ -domain. **(3)** The “RFFD/YHFK” region (**Figure 5.3.2.4 - C** – F motif, displayed as violet surface) is part of the active site and makes contacts with the substrate analogue. An evidence of the involvement of the arginine residue of the RFFD motif in *Bt*MINPP positional stereospecificity was obtained by substituting the residue with an aspartate [41], a mutation that abolished the 4/6-OH IP<sub>5</sub> production in the enzyme (wild-type *Bt*MINPP is able to cleave all the inorganic phosphate groups bound to the inositol ring of IP<sub>6</sub> except C2). **(4)** The “(F/Y)<sub>x</sub>2(K/R)” (**Figure 5.3.2.4 - D** – S motif, displayed as yellow surface) is also part of the active site cleft. In fact, the lysine residue can make weak ion pairs with the substrate analogue (3.9 Å distance).



**Figure 5.3.2.1. Multi-sequence alignment of all standard sequences.**

Protein names are reported on the left of the alignment. On the right the group division is annotated: A – non-phytase HP2, B – bacterial HPPHy, C – MINPPs and D – fungal HPPHy. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 50% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 50% conservation. The figure was created in Jalview [157]. Motifs with 100% conservation are the “RHGxRxPT” containing the catalytic histidine (named RHG motif), a further “R” which is involved in the substrate binding (named R motif), the “H” in the proton donor motif (“H motif”). High conservation is also observed in other regions: the “G motif” involved in HP2 conformational changes, “L motif”, “P motif” and “C motif”.



**Figure 5.3.2.2. Structural representation of HP2 conserved motifs in *B/MINPP*.**

*B/MINPP* is displayed by a light-blue surface overlapped by its cartoon representation. IHS is drawn by sticks and coloured by atom. Motifs surfaces are highlighted by the use of a colour scheme, as follow: **A)** RHG motif of sequence "RHGxRxPT", blue surface; **B)** G motif of sequence "GxLTx<sub>2</sub>G", green surface; **C)** R motif of sequence "Rx<sub>3</sub>Sx<sub>3</sub>Fx<sub>2</sub>G", violet surface; **D)** H motif of sequence "H(D/A)(x/E)", grey surface; **E)** L and C motifs of sequence "LL" and "Cx<sub>n</sub>C", pink and orange surface respectively; **F)** P motif of sequence "P", yellow surface. Figure A), B), C), D) and E) display a front-view of the enzyme while figure F) a back-view, the change is highlighted by a turning-arrow.

(5, 6) The “P(M/F/Y)<sub>x</sub>AN” loop (Figure 5.3.2.2 - F and 5.3.2.4 - E – P and N motif, displayed as yellow and pink surfaces) span across the back of the catalytic core establishing backbone hydrogen bonds with the catalytic histidine (2.7 Å) and the first arginine of the RHG motif (3.0 Å). The carboxylic group of the latter also lies at 3.5 Å from the carbonyl group of the proline and at 3.8 Å from the sulphate of the methionine. (7) The “V<sub>x2</sub>N<sub>x</sub>LE” (Figure 5.3.2.4 - F – E motif, displayed as orange surface) makes contacts with the Y motif, the S motif, the P motif and the N motif. Overall, the MINPPs-unique motifs identified are either positioned in the active site cleft and able to contribute in substrate orientation or correlated to the presence of the N-terminal active site lid structure.

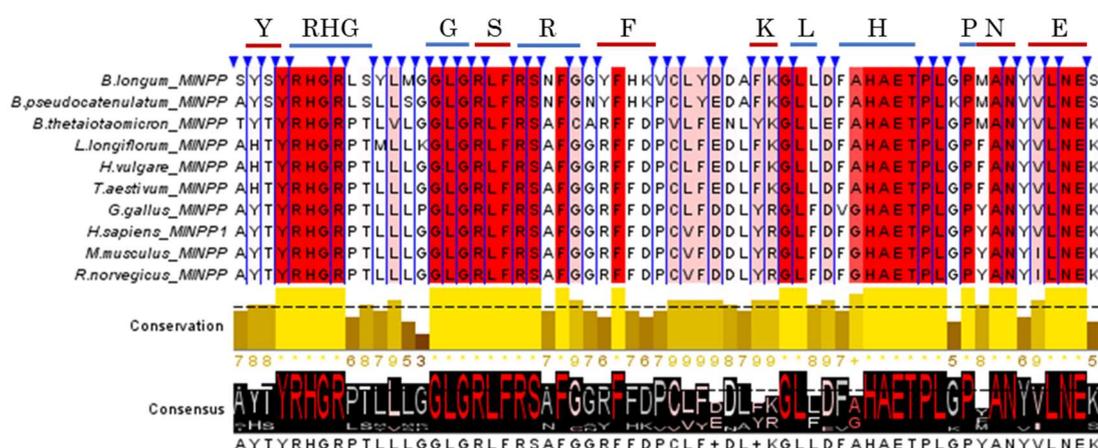
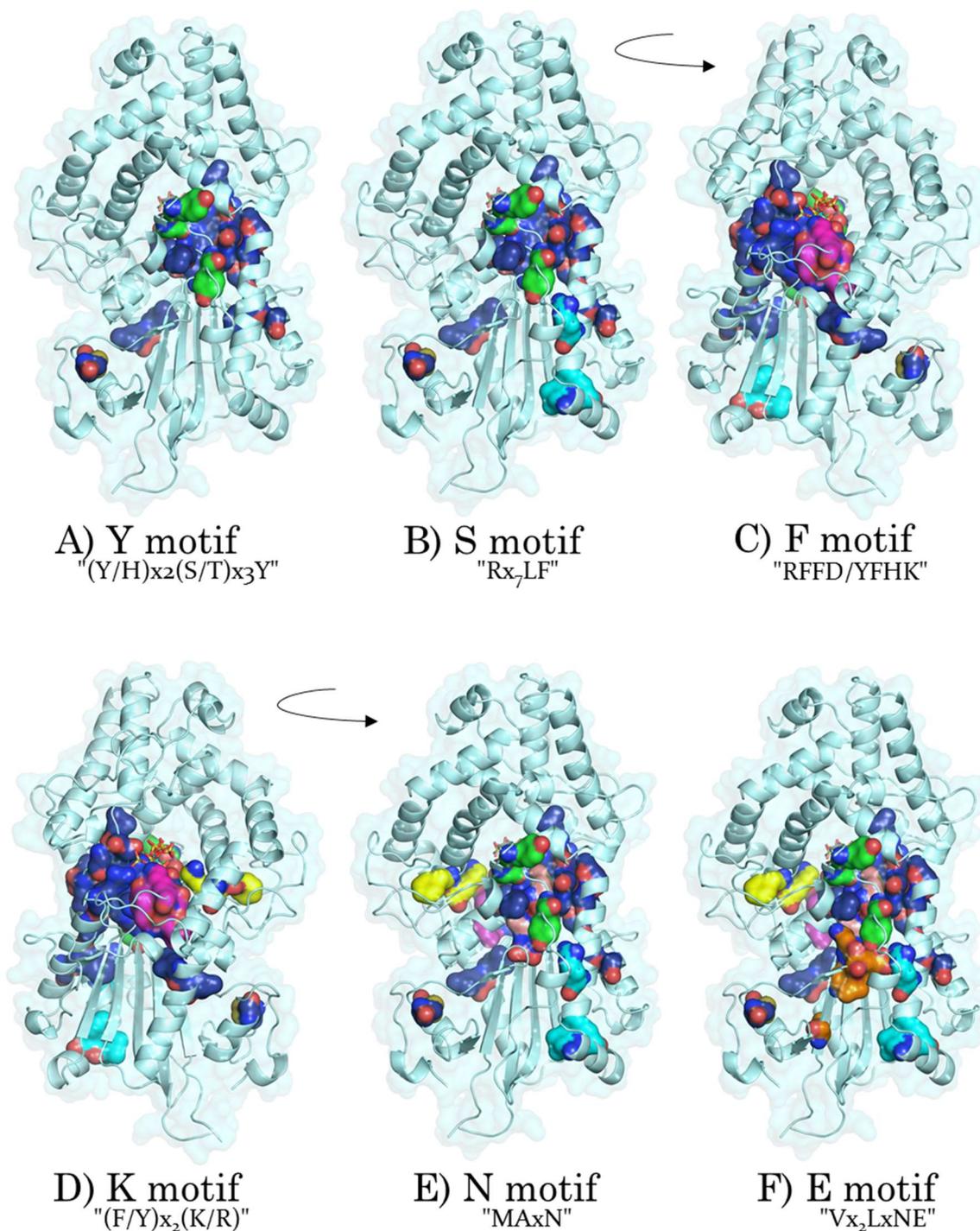


Figure 5.3.2.3. Multi-sequence alignment of MINPPs standard sequences.

Protein names are reported on the left of the alignment. MINPP: multiple inositol polyphosphate phosphatase. Only regions of consensus higher than 70% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157]. HP2 motifs are highlighted in blue and named as RHG, G, R, L, H and P. MINPPs-unique motifs are indicated in red and named as Y, S, F, K, N and E.



**Figure 5.3.2.4. Structural representation of MINPPs conserved motifs in *B/MINPP*.**

*B/MINPP* is displayed by a light-blue surface overlapped by its cartoon representation. IHS is drawn by sticks and coloured by atom. Motifs surfaces are highlighted by the use of a colour scheme, as follow: HP2 conserved regions highlighted in **Figure 5.3.2.2** are represented by a blue surface; **A)** Y motif of sequence “(Y/H)<sub>x2</sub>(S/T)<sub>x3</sub>Y”, green surface; **B)** S motif of sequence “R<sub>x7</sub>LF”, light-blue surface; **C)** F motif of sequence “RFFD/YFHK”, violet surface; **D)** K motif of sequence “(F/Y)<sub>x2</sub>(K/R)”, yellow surface; **E)** N motif of sequence “MA<sub>x</sub>N”, pink surface; **F)** E motif of sequence “V<sub>x2</sub>L<sub>x</sub>NE”, orange surface. Figure C), D) display a front-view of the enzyme while figure A), B), E) and F) a back-view, the change is highlighted by a turning-arrow.

## 5.4. Collection of novel MINPPs sequences and clustering

Amino acid sequences of previously uncharacterized bacterial proteins were mined through three BLAST searches among the RefSeq databases of (1) *Bifidobacterium* excluding *Bifidobacterium longum*, (2) *Actinobacteria* excluding *Bifidobacterium*, (3) *Bacteria* excluding *Actinobacteria*, all using *Bifidobacterium longum* MINPP as query [1, 2]. The idea of organising the sequences in 3 group of expanding taxonomic diversity had the aim of allowing the evaluation of sequence- and (possibly) stereospecificity-increasing divergence.

### *Bifidobacterium longum* MINPP

  <i>Bifidobacterium</i>	 <i>Bifidobacterium longum</i>
  <i>Actinobacteria</i>	 <i>Bifidobacterium</i>
  <i>Bacteria</i>	 <i>Actinobacteria</i>

When *Bacteroides thetaiotaomicron* MINPP was the query to search for novel MINPP among *Bacteria*, the sequences pulled down were the same as using *B*MINPP. Results were also used as queries, but they did not produce new hits. When HP positional specific phytases (*Escherichia coli* AppA-like phytases) were used as queries, none of the resulting sequences matched the ones of previous searches, showing again that the two groups, MINPPs and positional-specific HP, are quite distinct. An alternative search was performed by HMMER [159, 160] <https://www.ebi.ac.uk/Tools/hmmer/> in multiple databases, results were comparable.

The sequences collected were selected by coverage and by e-value (cut-off:  $1 \cdot 10^{-04}$ ). Redundant and partial sequences were removed. The remaining proteins were pooled and clustered by genus. Each group was multi-sequence aligned using MUSCLE [153] and the resulting percentage identity matrices (pim) were used to cluster sequences by identity score (cut-off identity used: 50%).

### *Bifidobacterium longum* MINPP

  <i>Bifidobacterium</i>	 <i>Bifidobacterium longum</i>
	sequences: 15, cluster: 1
  <i>Actinobacteria</i>	 <i>Bifidobacterium</i>
	sequences: 95, clusters: 18
  <i>Bacteria</i>	 <i>Actinobacteria</i>
	sequences: 267, clusters: 72

15 non-redundant sequences were collected among *Bifidobacterium* (taxid: 1678), excluding *Bifidobacterium longum* (id:216816). Their identity scores in the

percentage identity matrix (pim) were all higher than 50%. The sequences show large regions of 100% identity to *B. longum*. The *Bifidobacterium* sequence with lowest identity (53.8%) belongs to *B. pseudocatenulatum* MINPP, an enzyme that has already been expressed and characterized in our lab (unpublished data). An analysis of its predicted active site residues showed the presence of only two substitutions, D317N and E321D, which seem not to have perturbed stereospecificity, in fact both the enzymes show a promiscuous attack on IP<sub>6</sub> with a predominance of an I(12356/12345)P<sub>5</sub> peak (unpublished data). Because no significant changes were detected in the active site of these 15 proteins, they were not analysed any further.

95 non-redundant sequences were collected among *Actinobacteria* (taxid:2011754), excluding *Bifidobacterium* (taxid: 1678). Their identity scores in the percentage identity matrix exceeded the minimum cut-off of 50%. Therefore, they were clustered by genus and % identity in 18 groups.

267 non-redundant sequences were collected among *Bacteria* (taxid:2), excluding *Actinobacteria* (taxid:2011754). Their identity scores in the percentage identity matrix exceeded the minimum cut-off of 50%. Therefore, they were clustered in 72 groups.

**Table 5.4.1. Number of clusters and proton donor distribution in *Genus* for which multiple sequences were mined.**

*Genus* for which a single sequence was collected are excluded from the table. The number of sequences identified (Nr° Seq.), the clusters generated (Nr° Cl.), the number and the type of proton donor triplets (Nr° PD) identified are reported in the table.

<i>Genus</i>	Nr° Seq.	Nr° Cl.	Nr° PD
<i>Acidovorax</i>	3	1	1 - HAE
<i>Acinetobacter</i>	3	1	1 - HAE
<i>Aeromonas</i>	14	1	1 - HAE
<i>Alishewanella</i>	2	1	1 - HAQ
<i>Alistipes</i>	14	4	1 - HDV/G/T
<i>Bacteroidales</i>	2	1	1 - HDG/S
<i>Bacteroides</i>	69	13	8 - HA/GE, - HDG/T/V/Y - HES/T
<i>Burkholderia</i>	21	1	2 - HA/GE
<i>Cetobacterium</i>	2	1	1 - HAE
<i>Curvibacter</i>	2	1	1 - HAE
<i>Flammeovirga</i>	2	2	2 - HAE - HST
<i>Herbaspirillum</i>	2	1	1 - HAE
<i>Janthinobacterium</i>	4	1	1 - HAE
<i>Paenibacillus</i>	4	2	1 - HAE
<i>Photobacterium</i>	3	1	1 - HAE
<i>Prevotella</i>	44	17	7 - HAE - HDG/S/T - HEM/TV

From each *Actinobacteria* or *Bacteria* cluster (total of 90 clusters between the two groups) a representative sequence was arbitrary chosen. Representatives were further analysed for “Pfam matches”, they were multi-aligned, and their phylogenetic trees were inferred.

A diverse set of potential MINPPs were found in *Bacteria* (**Table 5.4.1**). Diversity was particularly accentuated in the *Bacteroides* and *Prevotella* genus, for whose 13 and 17 clusters of identity >50% were produced. These enzymes showed to possess a plethora of divergent proton donor triplets, though all containing an aspartic or glutamic acid residue. On the contrary, in *Aeromonas* and *Burkholderia* genus were identified respectively 14 and 21 sequences but they could all be clustered in a single group of identity >50%. Only *Alishewanella*, *Arsukibacterium* and *Flammeovirga* displayed the apparent proton donor-less triplets HAQ or HST.

## 5.5. Selection of MINPPs of interest

The process of selection included two phases:

1. Evaluation of sequence divergence and search for outliers (in other words, for proteins that may not be MINPPs) by multiple-sequence alignment and phylogenesis inference. Comparison with the previously determined MINPPs-unique motifs.
2. Active site prediction based on the binding pockets of enzymes of known structure and multiple-sequence alignments. Selection of interesting sequences for *in vitro* characterisation.

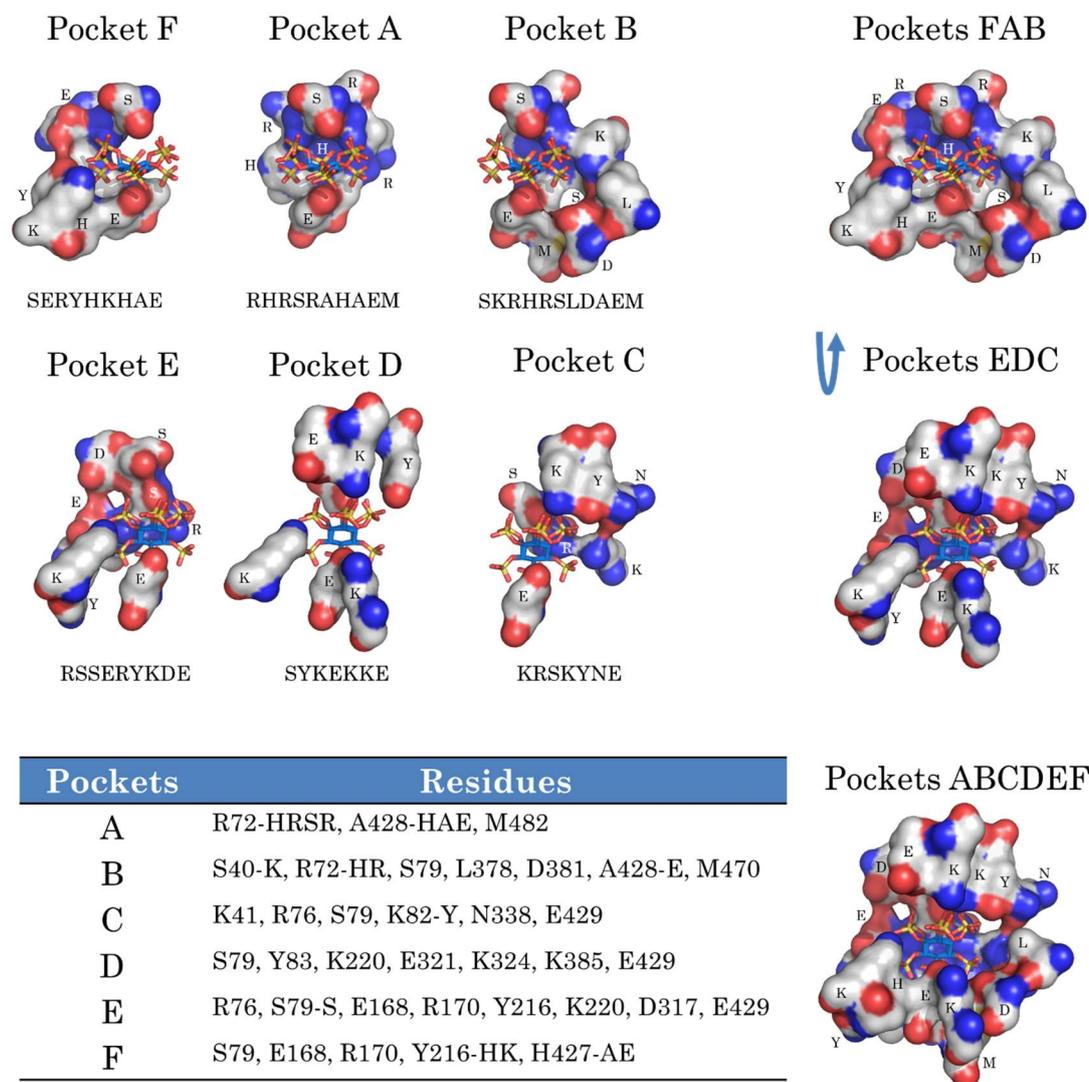
In the following paragraphs the combined results from these two analyses will be described.

### 5.5.1. Definition of active site pockets

Despite the analysis of HP2 and MINPPs-unique motifs conservation was carried out for whole length MINPPs sequences, only active site composition was used in the selection of sequences of interest. Efforts were concentrated in choosing the sequences that, altogether, the most would represent active site variability in MINPPs.

Active site pockets for the binding of the six phosphates of IP<sub>6</sub> were created based on the structural models of *Escherichia coli* AppA (PDB ID: 4tsr), *Aspergillus niger* PhyA (PDB ID: q) and two bacterial MINPPs, *Bacteroides thataiotaomicron*

(PDB ID: 4ftu) and *Bifidobacterium longum* MINPP (data unpublished) in complex with IHS. Their active sites were superimposed and binding pockets, one for each phosphate of *myo*-inositol hexakisphosphate, were defined including all amino acids within 6Å distance from each sulphate group of the substrate analogue, inositol hexasulphate. Here, the binding pockets defined in *B*/MINPP are reported (**Figure 5.5.1.1.**).



**Figure 5.5.1.1. Active site pockets in *Bifidobacterium longum* subsp. *infantis* MINPP (*B*/MINPP).**

Binding pockets for each phosphate group of *myo*-inositol hexakisphosphate were predicted from the structure of the E:IHS complex of *B*/MINPP. Pockets are named clockwise A, B, C, D, E, F, starting from the pocket in which the nucleophilic attack occur. Surfaces are displayed for each pocket (C: grey, O: red, N: blue, S: yellow), the inhibitor inositol hexasulfate is represented by sticks. Amino acid composition is reported for each pocket. The table also report residues numbering in *B*/MINPP. “Pockets FAB” and “Pockets EDC” represent two halves of the active. “Pockets ABCDEF” instead displays the whole active.

It is interesting to notice the symmetrical composition of Pocket A, in which an oxygen atom lays over (hydroxy group, S79) and under (carboxylic group, E429) the inositol ring, while an arginine at each side bind the groups adjacent to the phosphate coordinated by the catalytic histidine, His73. This may suggest that an important contribution to stereospecificity could be given by the three-dimensional array of the residues in the contiguous pockets which seem to alternate positively and negatively charged side-chains in order to orient the substrate by attraction/repulsion forces.

This model of the catalytic cleft was used to select the regions of the multiple-sequence alignments possibly containing active site residues. The amino acids included in the active site pockets were mapped into the multi-sequence alignments (**Figure 5.5.2.1** and **5.5.3.1**). Most of the active site showed to be conserved and represented by portions of the HP2 or MINPPs-unique motifs described before: Y motif, RHG motif, R motif, F motif, U motif (residues lying on the *B*/MINPP U-loop homologues regions, **Chapter 4.6**), K motif, H motif and N motif. Four residues on the flexible loop downstream to the RHG motif were also mapped in the alignments as well as other three positions characterized by a lower conservation (**Figure 5.5.2.1**). The results of the analysis are below described and report (1) HP2-motif conservation, (2) MINPPs-unique motif conservation and (3) an analysis of the active site configuration in the *Actinobacteria* and the *Bacteria* pools.

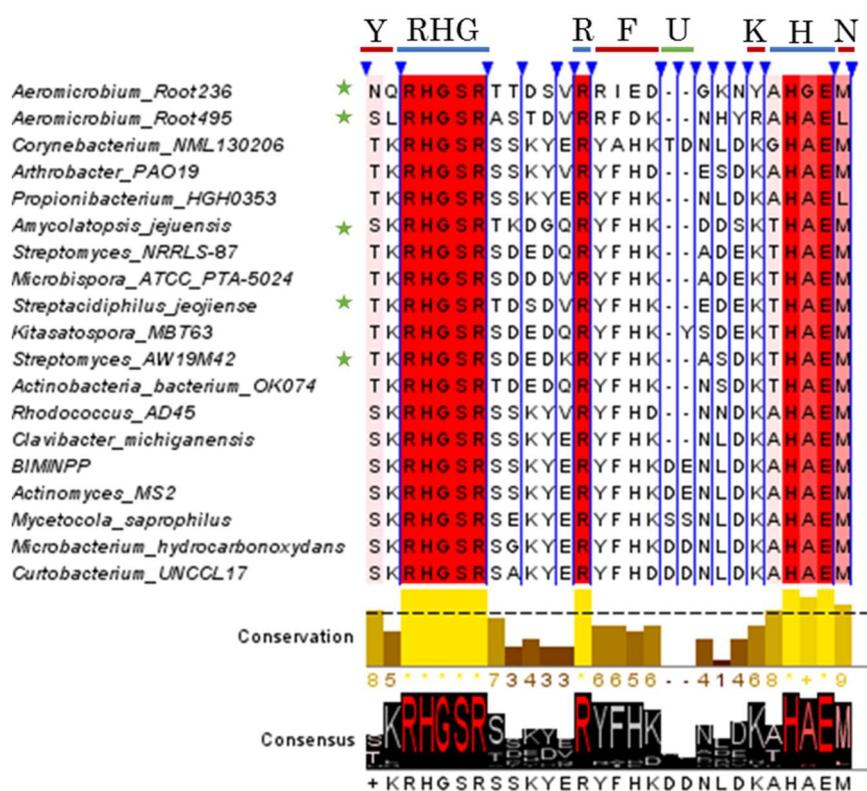
### **5.5.2. Conserved motifs in the active site of *Actinobacteria***

The 18 representative sequences of *Actinobacteria* proved, on the whole, to have a very high conservation of their HP2-typical regions. A multi-sequence alignment of the 18 *Actinobacteria* representative can be viewed in **Appendix, Figure 8.4.2.2**. The alignment of their active site is reported in **Figure 5.5.2.1**.

A sequence from *Aeromicrobium* Root236, showed interesting modifications, an “HGE” catalytic triplet substitutes the classical “HAE” and the conserved proline (P motif), positioned at the back of the active site, was replaced by an arginine.

Two sequences of the *Aeromicrobium* genus, including the one just mentioned, showed to possess unusual active site MINPPs-unique motifs F and K, respectively a “RIED” and a “RFDK” instead of the “RFFD” (characteristic of *Bt*MINPP and eukaryotic MINPPs) or “YFHK” (found in *Bifidobacterium* species) and a Tyr or a Gln in place of the Lys in the K motif. Because these mutations were predicted to significantly affect the polarity of the active site pockets (differences in F motif would affect pockets E and F, differences in K motif would affect pockets B, C and D) both

the *Aeromicrobium* sequences (A. Root236 and Root495) were chosen for further characterisation. *Corynebacterium*, *Arthrobacter* and *Rhodococcus* species also showed modifications in the F motif sequence: a “YAHK” in *Corynebacterium* and a “YFHD” in the other two sequences. It was interesting to note, in fact, that the mutations in the F motif seem to follow an evolutive pattern (**Figure 5.5.2.1**, F motif). The representative *Actinobacteria* sequences vary by stepwise substitutions (highlighted in bold): YFHK, YAHK or YFHD or **RFDK**, **RFFD**, **RIED**. Also, in *Bacteroides thetaiotaomicron* MINPP, R183 in the RFFD motif was proved to be involved in determining enzyme positional specificity of IP<sub>6</sub> hydrolysis [41]. Because of these reasons, the area was analysed as possible hotspot for the determination of positional stereoselectivity traits.



**Figure 5.5.2.1. Multiple sequence alignment of the active site of 18 *Actinobacteria* representative sequences.**

Protein names are reported on the left of the alignment. MINPP: multiple inositol polyphosphate phosphatase. Only amino acids possibly within 6 Å from the substrate are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display a cut-off limits of 70% conservation and the consensus sequence. The figure was created in Jalview (Waterhouse et al., 2009). HP2 motifs are highlighted by a blue line and named as RHG, R and H. MINPPs-unique motifs are annotated by a red line and named as Y, F, K and N. The amino acids lying on *BIMINPP* U-loop homologue regions are highlighted by a green line.

The *Actinobacteria* sequences chosen for synthesis were five. The two *Aeromicrobium* enzymes described at the beginning of this section, *Amycolatopsis jejuensis* MINPP, *Streptacidiphilus joejiense* MINPP and *Streptomyces* AW19M42 MINPP. The latter three sequences were chosen for their variability in the flexible loop downstream the RHG motif. In particular, *Amycolatopsis jejuensis* was chosen for its “TK” pair (present also in *EcAppA*, **Chapter 3**) possibly in close proximity to the substrate.

### 5.5.3. Conserved motifs in the active site of *Bacteria* (excluding *Actinobacteria*)

The *Bacteria* pool, as expected, showed an increased diversity, the conservation of HP2 motifs is indeed lower in comparison with the *Actinobacteria* sequences. MINPPs-unique motifs were found in all the sequences analysed either present in the active site (motifs Y, F, K and N) or involved in the interaction stabilising the N-terminal lid loop (motifs Y, S, N and E). A multi-sequence alignment of the 72 *Bacteria* representative can be viewed in **Appendix, Figure 8.4.2.3**.

Remarkable differences were found in the proton donor triplet (H motif), which displayed multiple variations from the most common “HAE”: the apparent proton donor-less “HST” and “HAQ” motifs, the “HDx” triplets “HD(Y/V/S/T/A/G)”, the “HEX” triplets “HE(T/M/S/V)” and the “HAE” alternative triplet “HGE”. Other noteworthy changes were identified in the P motif, where the proline highly conserved in HP2 was found to be substituted by M/K/S/C/G, while the downstream N motif was conserved. However, sequences carrying variant of the “MAxN” MINPPs-unique motif were also identified. Changes in the Met residues were particularly interesting. In fact, despite the “PMAxN” motif establishes mainly backbone interactions with the catalytic core residues, the Met side-chain is oriented towards the catalytic core in both *B*/MINPP and *Bt*MINPP and its sulphate lies at close distance from all the catalytic core residues (3.8 – 4.5 Å) and within 5.5 Å from the substrate analogue. This residue was found to be substituted with Cys, Lys, Leu but also with the bulky Phe and Tyr (present also in plants and animals MINPPs enzymes). Another MINPPs-unique area of interest was the F motif whose evolution again seemed to be traceable, e.g. in the motifs YFHK, YFMK, YFMN, YYMN, **FYLN** (possible substitutions highlighted in bold).

From the *Bacteria* pool eleven sequences were selected for synthesis carrying “HST”, “HD(T/Y/V/G)”, “HE(T/M)”, “HAQ” and “HAE” proton donor H motifs. Multiple

sequences were chosen from the *Bacteroides* genus because of the high sequence diversity discovered and yet to be explored.

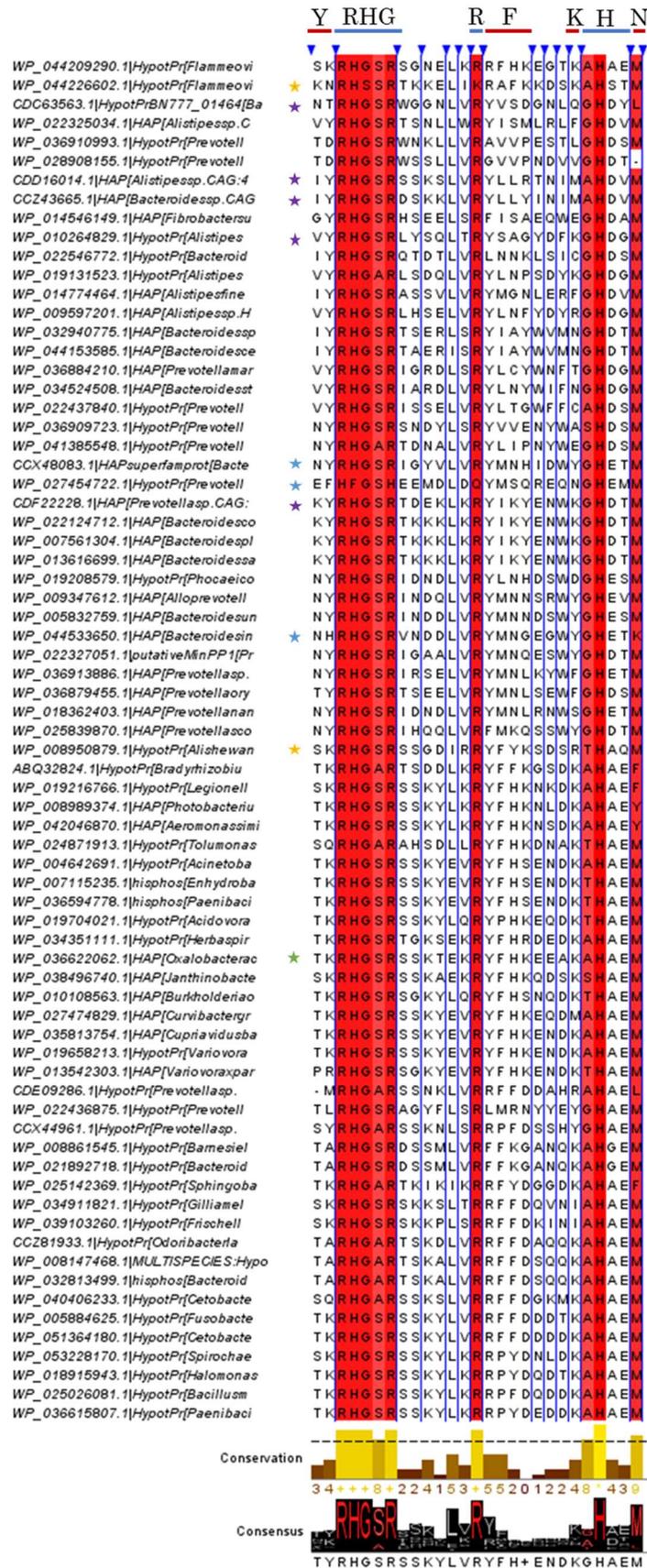


Figure 5.5.3.1. Multiple sequence alignment of the active site of 72 *Bacteria* representative sequences.

### Figure 5.5.3.1. Active site multiple sequence alignment of 72 *Bacteria* representative sequences

Protein GenBank identifications codes are reported on the left of the alignment. Only amino acids possibly within 6 Å from the substrate are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display a cut-off limits of 70% conservation and the consensus sequence. The figure was created in Jalview (Waterhouse et al., 2009). HP2 motifs are highlighted by a blue line and named as RHG, R and H. MINPPs-unique motifs are annotated by a red line and named as Y, F, K, N.

### 5.5.4. Phylogenesis of bacterial MINPPs sequences

The evolutionary history of the collected bacterial MINPPs sequences was inferred using the Maximum Likelihood method based on the JTT matrix-based model [149]. The tree with the highest log likelihood (-70598.92) is shown in **Figure 3**. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.3483)). A bootstrap analysis was inferred from 50 replicates [150]. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The analysis involved 92 amino acid sequences. All positions with less than 25% site coverage were eliminated. That is, fewer than 75% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 482 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [151].

The first characteristic of the tree (**Figure 5.5.4.1**) that come into sight is that enzymes from different genus can populate the same monophyletic group. Looking at branch length (number of substitutions per site), it is possible to observe the presence of a small number of short-branch ancestor sequences, in the centre of the phylogram, including MINPPs from *Fusobacteria*, *Firmicutes*, *Spirochaetae*,  $\gamma$ -*Proteobacteria* and *Bacteroides* and characterised by an “HAE” proton donor motif and an “RxxD” F motif sequence. Two clades are nested from these ancestors, one includes *Bacteroides* and *Fibrobacteres* enzymes and is characterised by the substitution of the “HAE” triplet with an “HDx” or “HEx” motif and the loss of conservation of the F motif. The other includes a variety of species from multiple genus (*Actinobacteria*,  $\alpha$ -,  $\beta$ - and  $\gamma$ -*Proteobacteria*, *Firmicutes*) and is characterised by a conserved “HAE” triplet and a “YFHK” F motif.

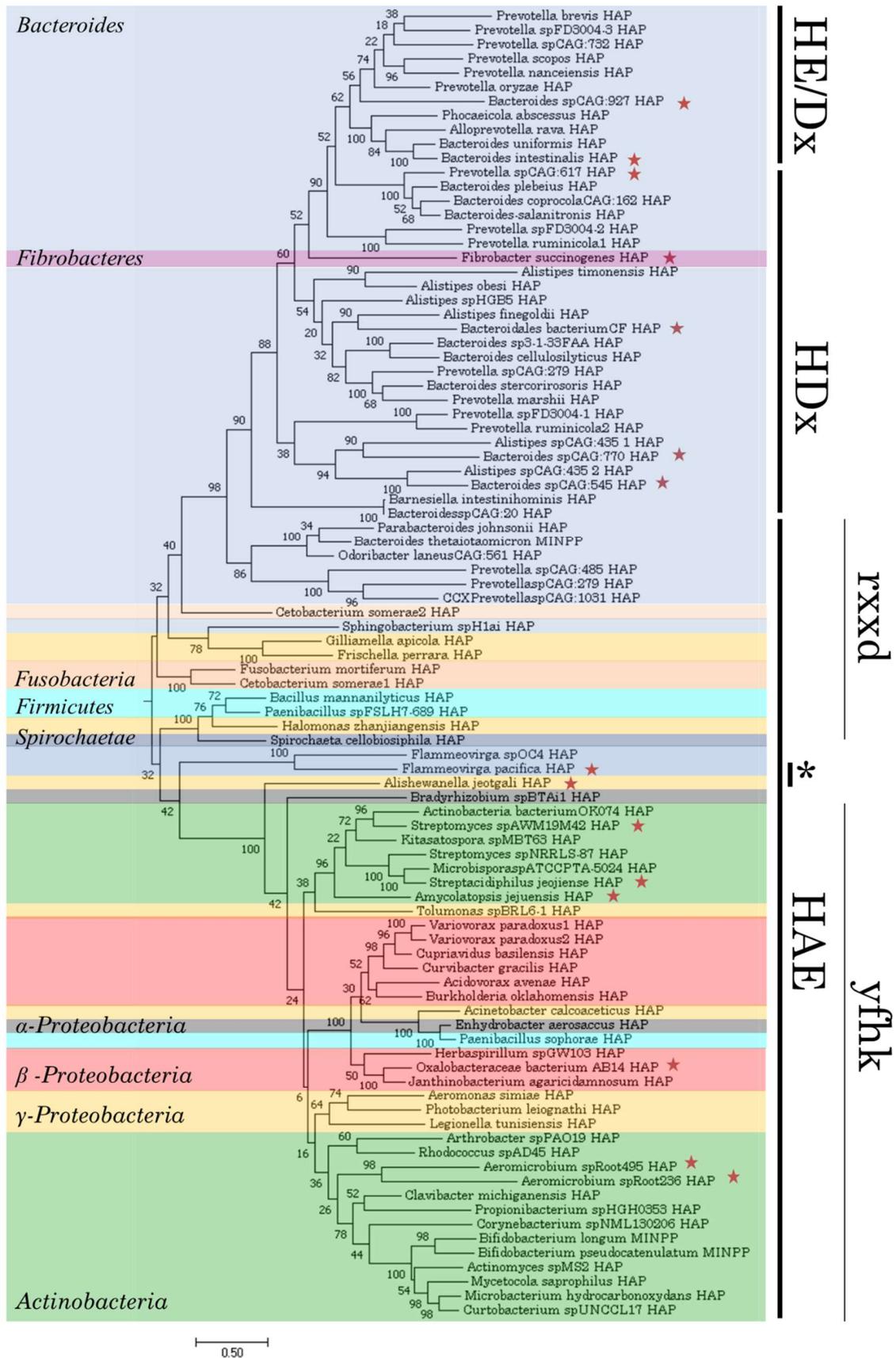


Figure 5.5.4.1. Molecular Phylogenetic analysis by Maximum Likelihood method.

Caption in the next page.

#### Figure 5.5.4.1. Molecular Phylogenetic analysis by Maximum Likelihood method.

The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Stars highlight the sequences that were selected for gene synthesis and further characterization. Bacterial genus follows a colour-scheme where light-blue indicates *Bacteroides*, violet *Fibrobacteres*, orange *Fusobacteria*, cyan *Firmicutes*, blue *Spirochaetae*, grey  $\alpha$ -*Proteobacteria*, red  $\beta$ -*Proteobacteria*, yellow  $\gamma$ -*Proteobacteria* and green *Actinobacteria*. H motif divergence is annotated: the majority of the sequences present an “HAE” which evolved to “HDx” in *Bacteroides* and again to “HEx” in a subgroup of the same genus. Asterisk “\*” highlights the proton donor-less sequences HST and HAQ. with F motif evolution is also reported and shows ancestor sequences characterized by an “rxxd” sequence which evolved to a conserved “yfhk” in on branch and was subjected to faster changes in the other. Capital letters indicate 100% conservation of a residue, lower case indicate the presence of substitutions.

## 5.6. Analysis of the 16 selected sequences prior synthesis

### 5.6.1. Biological sources

In total 16 proteins were selected for biochemical characterization. They come from a wide variety of sources. 60% of the enzymes chosen are found in bacteria colonizing multiple areas of the digestive tract (rumen, gut and faeces) and belonging to the families of *Bacteroides* or *Actinobacteria*. Many phytases were chosen from these genera because of their diversity in sequence and, possibly, structure/function. Three enzymes come from a symbiont of sea squirt (organ not specified) or of the roots of *Arabidopsis thaliana*. Two enzymes are produced by bacteria living in sediments (deep-sea Pacific Ocean and low pH uranium mines), they may be used to scavenge phosphate from phosphomonoester accumulated in the soil, a limiting nutrient for their growth. In particular, in *Flammeovirga pacifica* [161, 162], an organism living at the bottom of the Pacific Ocean, multiple thermostable and pH stable enzymes have already been identified. Example are an arylsulfatase [163], an alpha-amylase [164] and a  $\beta$ -agarase [165]. Other extremophiles considered in the batch of MINPP-containing hosts are *Arsukibacterium* sp. MJ3 [166], which was isolated in a cold alkaline environment (Ikka Fjord), and *Oxalobateraceae bacterium* [167], found in samples collected in a low pH uranium mine sediment (**Table 5.6.1.1**).

The bacterial genomes of standards and novel sequences were also inspected to identify the gene cassettes containing these MINPPs sequences, which were not found to be conserved across the species. Genomic context is not considered to be useful to identify bacterial MINPPs sequences while eukaryotic MINPPs are often found in regions close to adenosine 5'-phosphosulfate kinase (APSK) and ATPase

family AAA domain 1 (ATAD1) genes. For this reason, the possibility of horizontal gene transfer of MINPPs across *Bacteria* could be a possibility. Phylogenetic data support this hypothesis if considered the clades including protein sequences from different genera (100% bootstrap score). An example is the branch that includes *Acinetobacter calcoaceticus* HP (*γ-Proteobacteria*), *Enhydrobacter aerosaccus* HP (*α-Proteobacteria*) and *Paenibacillus sophorae* HP (*Firmicutes*) **Figure 5.5.4.1**.

**Table 5.6.1. Bacterial sources and ecological niches.**

The column on the left reports the number of identification of each protein (1-16) and the bacterial species that translate them. The middle column lists NCBI identifiers. The column on the right describe the source from which the bacteria/transcript was isolated/amplified.

Protein	Prot NCBI id	Source
P1 – <i>Flammeovirca pacifica</i> (1_Fp_noSP)	WP_044226602.1	Pacific ocean, deep-sea sediments <i>Bacteroidetes</i>
P2 – <i>Arsukibacterium</i> sp. MJ3 (2_AMJ3_noSP)	WP_052750238.1	Ikka Fjord – Greenland, cold alkaline environment <i>γ-Proteobacteria</i>
P3 – <i>Prevotella brevis</i> ATCC 19188 (3_Pb_noSP)	WP_027454722.1	Rumen <i>Bacteroidetes</i>
P4 – <i>Bacteroides</i> sp. CAG:927 (4_BCAG927_noSP)	CCX48083.1	Human gut metagenome <i>Bacteroidetes</i>
P5 – <i>Bacteroides intestinalis</i> (5_Bi_noSP)	WP_044533650.1	Human faeces – Japan, Asia <i>Bacteroidetes</i>
P6 – <i>Bacteroides</i> sp. CAG:545 (6_BCAG545_noSP)	CCZ43665.1	Human gut metagenome <i>Bacteroidetes</i>
P7 – <i>Bacteroides</i> sp. CAG:770 (7_CAG770)	CDC63563.1	Human gut metagenome <i>Bacteroidetes</i>
P8 – <i>Prevotella</i> sp. CAG:617 (8_PCAG617_noSP)	CDF22228.1	Human gut metagenome <i>Bacteroidetes</i>
P9 – <i>Bacteroidales bacterium</i> CF (9_BbCF)	WP_014546149.1	Chloroform degrading mixed culture <i>Bacteroidetes</i>
P10 – <i>Fibrobacter succinogenes</i> (10_Fs_noSP)	WP_014546149.1	Rumen cellulolytic bacteria <i>Fibrobacteres</i>
P11 – <i>Oxalobacteraceae bacterium</i> AB (11_Ob_noSP)	WP_036622062.1	Low pH uranium mine sediments – Ronneburg, Germany <i>β-Proteobacteria</i>
P12 – <i>Aeromicrobium</i> sp. Root236 (12_AR236_noSP)	WP_056397321.1	Roots of <i>Arabidopsis thaliana</i> – Cologne, Germany <i>Actinobacteria</i>
P13 – <i>Aeromicrobium</i> sp. Root495 (13_AR495_noSP)	WP_056287230.1	Roots of <i>Arabidopsis thaliana</i> – Cologne, Germany <i>Actinobacteria</i>
P14 – <i>Streptacidiphilus jeojiense</i> (14_Aj_noSP)	WP_030259532.1	Soil from cave – Jeju Island, South Korea <i>Actinobacteria</i>
P15 – <i>Amycolatopsis jejuensis</i> (15_Aj_noSP)	WP_051792931.1	Dried bat dung from cave – Jeju Island, South Korea <i>Actinobacteria</i>
P16 – <i>Streptomyces</i> sp. AW19M42 (16_SAW19M42_noSP)	WP_024494695.1	Sea squirt microbiota – Tysfjorden, Northern Norway <i>Actinobacteria</i>

### 5.6.2. Analysis of the conservation of HAP2 and MINPP motifs

The chosen sequences are characterized by the following HP2 conserved regions (Figure 5.6.2.1):

- The RHG motif is conserved for all with the exception of *Flammeovirga pacifica* p1, which is characterised by an “RHSxR” motif.
- Variation in the G motif includes “Gx(E/D)x<sub>3</sub>G” instead of “GxLx<sub>3</sub>G” for *Arsukibacterium* MJ3 p2, *Oxalobacteraceae bacterium* p11, and *Amicolatopsis jejuensis* p15.
- The conserved arginine of R motif (“R-S-F-G”), able interact with the substrate in the active site cleft, is replaced by a Lys in *Flammeovirga pacifica* p1. The motif does not present a conserved Gly in all the sequences. The following substitutions are also present: Ser is replaced by a Thr in *Amicolatopsis jejuensis* p15, Phe is replaced by a Lys in *Aeromicrobium* Root236 p12, by a Gln in *Bacteroidales bacterium* CF p9, by an Ala in *Prevotella* CAG:617 p8 or by a Glu in *Bacteroides* CAG:927 p4 and *Bacteroides intestinalis* p5.
- The L motif (“LLxD”) is quite conserved. However, a “FL” is found in *Flammeovirga pacifica* p1, “VL” or “LV” are present in *Arsukibacterium* MJ3 p2, *Aeromicrobium* Root236 p12, *Bacteroides* CAG:770 p7, *Bacteroidales bacterium* CF p9 and a “TL” is carried by *Fibrobacter succinogens* p10.

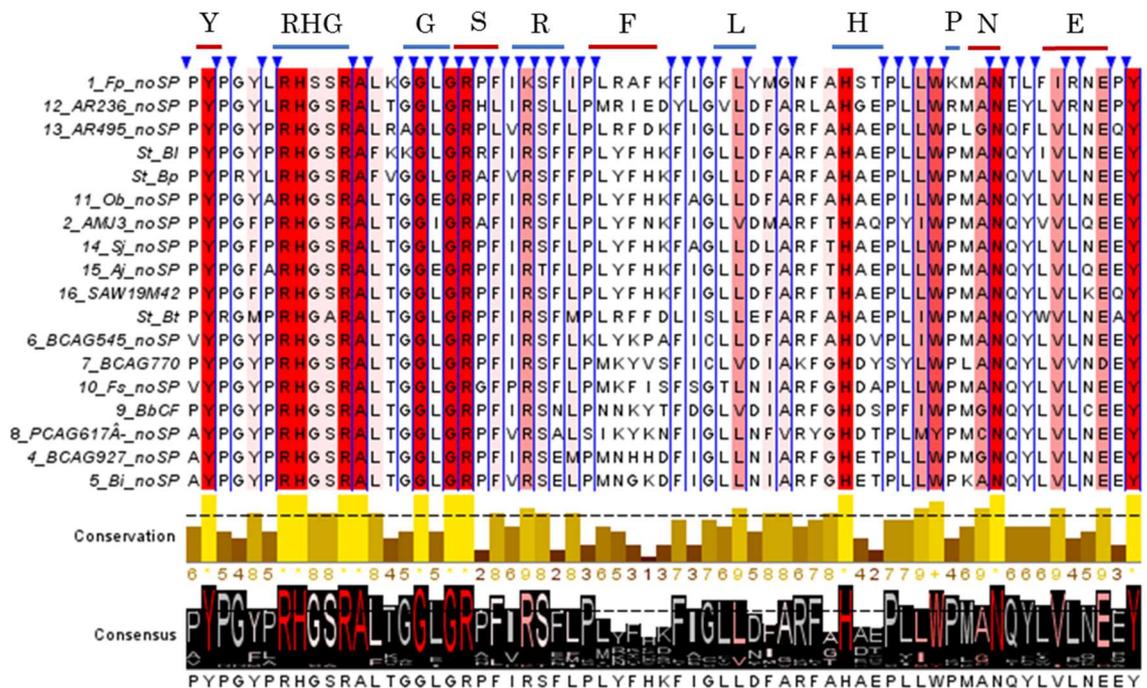


Figure 5.6.2.1. Multi-sequence alignment of selected MINPPs.

### Figure 5.6.2.1. Multi-sequence alignment of selected MINPPs.

Protein identifiers are reported on the left of the alignment. At the top, conserved regions are highlighted and named. Areas of conservation from 70% to 100% are coloured with a red gradient. At the top, HP2 conserved motifs are highlighted by blue lines and MINPPs-unique motifs by red lines. The bottom annotations display cut-off limits (dashed line) of conservation and consensus for each amino acid. The figure was created in Jalview [157]. Areas of high conservation are the RHG motif containing the catalytic histidine (“RHGxR”), the G motif (“GxLx<sub>2</sub>G”) a further R involved in the substrate binding (“R motif”), the H in the proton donor motif (“H motif”), the N of the E motif and a terminal Y which help in stabilizing the C-terminal of the enzymes.

- The H motif (“HAE”) is not conserved. The selected sequences contain an “HST”, “HAQ”, “HGE”, “HET” and “HD(V/Y/A/S/T)”.
- The P and N motifs (P-A-N-ly) are mostly conserved, however, the proline is substituted by a Lys in *Flammeovirga pacifica* p1 and by an Arg in *Aeromicrobium* Root236 p12; Ala is substituted by a Gly in a few proteins or by a Cys in *Prevotella* CAG:617 p8; Gln is fully conserved. The substitutions in *Flammeovirga* and *Aeromicrobium* are particularly interesting because the long side-chain of these amino acids could easily reach the nearby catalytic core residues.

The sequence chosen present variability in some of the MINPPs-unique motifs (Figure 15):

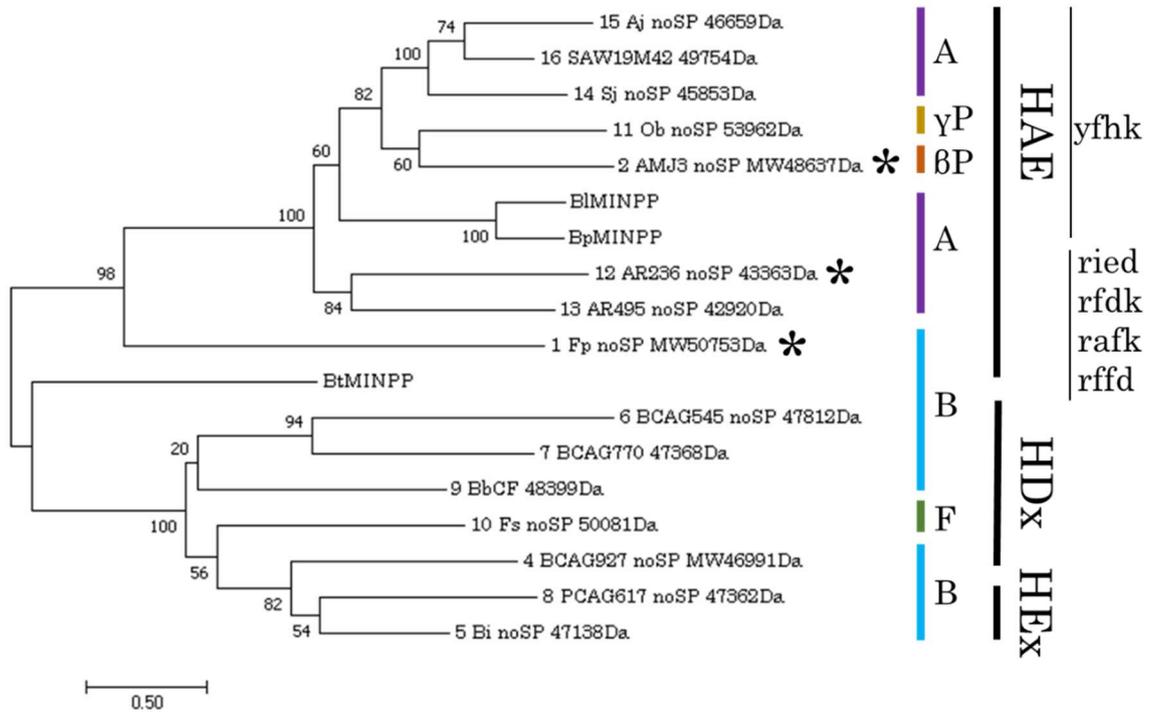
- The Tyr stabilising the N-terminal lid motif (Y motif) is fully conserved as well as the Arg of the S motif (“R-LF”).
- The “YFHK” or “RFFD” of the F motif is not conserved among all the selected sequences. A “YFHK” sequence is only present in *Oxalobateraceae bacterium* p11, *Arsukibacterium* MJ3 p2, *Amicolatopsis jejuensis* p15, *Streptacidiphilus jeojense* p14 and *Streptomyces* W19M42 p16.
- The E motif (“Vx<sub>2</sub>LxNE”) has conserved V and E. Exceptions are *Flammeovirga pacifica* p1 that carries an Ile instead of a Val and *Bacteroides* CAG:770 p7 which contains an Asp instead of a Glu.

### 5.6.3. Phylogenesis

The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model [149]. The tree with the highest log likelihood (-16574.62) is shown in **Figure 5.6.3.1**. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.6523)). The analysis involved 18 amino acid sequences. All positions with less than 25% site coverage were eliminated. That is, fewer than 75% alignment gaps, missing data, and ambiguous bases were allowed at any position. There was a total of 466 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [151].

The tree is divided in two main branches with *Bacteroides thetaiotaomicron* MINPP representing the “ancestor” among the 18 sequences under analysis, in line with the phylogram previously described in **Figure 5.5.4.1**). The sequences are divided in two clades, a monophyletic group that contains MINPPs from *Bacteroides* and *Fibrobacteres* (a genus recently differentiated from *Bacteroides*) and another that includes enzymes of mixed origins: *Bacteroides*, *Actinobacteria*, *beta-* and *gamma-Proteobacteria*.

The evolution of the H and F motif is annotated on the tree in **Figure 5.6.3.1**. It is interesting to note the spectrum of phytases diversity generated among *Bacteroides*, which can count phytases carrying a variety of proton donor triplets (HST, HAE, HDx, HEx) and an even larger number of F motifs.

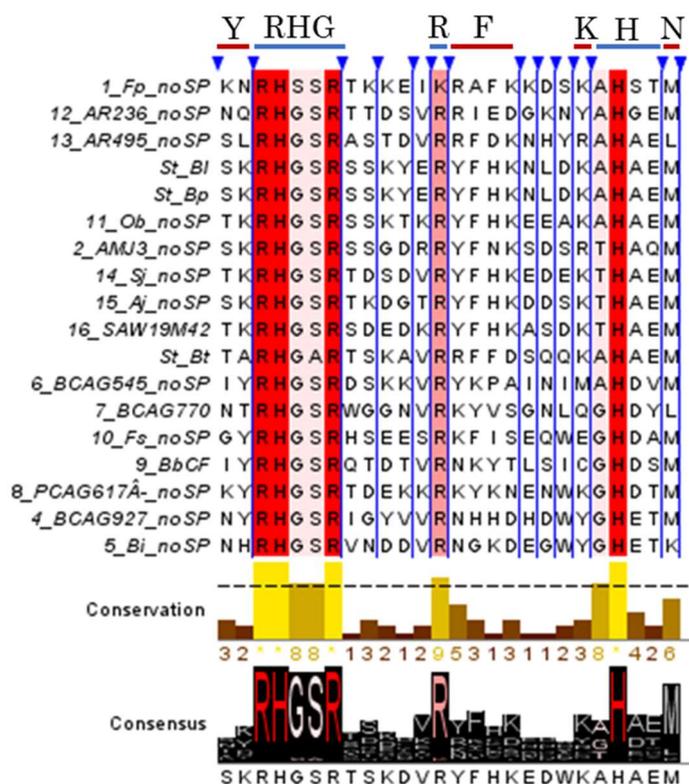


**Figure 5.6.3.1. Molecular phylogenetic analysis by maximum likelihood method of selected MINPPs.**

The tree with the highest log likelihood (-16574.62) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGA7 [151]. A violet line and the letter A indicate sequences from the *Actinobacteria* genus, a dark-yellow line and  $\gamma$ P indicate sequences from the  $\gamma$ -*Proteobacteria* genus, an orange line and  $\beta$ P indicate sequences from the  $\beta$ -*Proteobacteria* genus, a light-blue line and B indicate sequences from the *Bacteroides* genus, a green line and F indicate sequences from the *Fibrobacter* genus. Differences in the proton donor triplet and the F motif are reported on the right side of the tree. Represented proton donors are “HAE”, “HDx”, “HEx” and three proton donor triplets highlighted with an asterisk “\*”: *Arsukibacterium* MJ3 (p2) – “HAQ”, *Flammeovirga pacifica* – “HST” and *Aeromicrobium* Root236 – “HGE”. The F motif is conserved only in “HAE” triplet-containing sequences.

#### 5.6.4. Active-site prediction

The only areas of high conservation among the selected sequences are primarily involved in catalysis. They are the RHG motif, the R motif involved in substrate binding and the H motif containing the proton donor triplet. Enzyme 1\_Fp, from *Flammeovirga pacifica*, is an exception to the rule, displaying an “RHSxR” motif instead of an “RHGxR” and a Lys in place of the Arg included in the R motif (**Figure 5.6.4.1**). The proton donor triplets are not conserved because a criterion for the selection of these enzymes was to represent the diversity of this region. The proton donor included are the “HST”, the “HAQ”, the “H(G/A)E”, the “HD(V/Y/A/S/T)” and “HET”. The F motif diversity is also represented to allow, even if only partially, to trace the evolution of the motif in the family of MINPPs. In fact, this small sample of enzymes, despite chosen to represent as much as possible the diversity of MINPPs active sites, still does not cover the variability of the active sites of the class, as described in **Chapters 5.5.2.1** and **5.5.3.1**.



**Figure 5.6.4.1. Multi-sequence alignment of the active site residues of selected MINPPs.**

Protein id are reported on the left of the alignment. At the top, conservation regions are named annotated. Only amino acids predicted to be among 6 Å from the phosphates group of the substrate are reported here. Areas of conservation from 70% to 100% are coloured with a red gradient. The bottom annotations display cut-off limits (dashed line) of conservation and the consensus sequence. The figure was created in Jalview [157]. Areas of high conservation are the RHGxR region (2) containing the catalytic histidine, an additional R involved in substrate binding (6), and the H of the proton donor motif (12).

### 5.6.5. Genes optimization and synthesis

Gene sequences were optimized for transcription and translation in *Escherichia coli* hosts by using the program GeneOptimizer™ (ThermoFisher Scientific) prior to gene synthesis. This multiparametric program takes into consideration GC content, the presence of Shine Dalgarno sequences, TATA boxes, termination signals, artificial recombination sites, RNA instability motifs, ribosomal entry sites, repetitive sequences, codon usage, secondary structure formation and other additional variables [168, 169]. Gene optimization was essential in this study because preliminary analysis showed that the majority of the sequences chosen had a suboptimal codon usage and GC content for *E. coli* expression (**Table 5.6.5.1**).

**Table 5.6.5.1. % GC content of the 16 MINPPs selected for gene synthesis.**

Proteins are named by their assigned number 1-16. The optimal % GC in *E. coli* K-12 strains is also reported.

Proteins	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	<i>E. coli</i>
% GC	29.5	52.6	54.4	51.8	44.1	50.4	49.5	48.8	42.2	50.2	65.5	69.5	64.3	71.1	62.2	67.6	50.8

Gene optimisation included the removal of the signal peptide for periplasmic translocation in the sequences in which it was present. No disordered N-terminal or C-terminal areas were found in the selected sequences. Three-dimensional modelling and a prediction of disulphide bridges location was also carried out (**Appendix 8.4.3, 8.4.4**).

Protein p3, from *Prevotella brevis*, was an interesting candidate due to its homology with all HP2 and MINPPs conserved regions, except the conserved catalytic “RHGxR” which is substituted by an “HFGSH” sequence and the absence of conservation of the R motif. It would have been interesting to discover the functional differences that such an enzyme could carry. However, this sequence was not sent for synthesis. The selected gene sequences were sent for synthesis, amplified by PCR and cloned into plasmids for protein expression. These procedures will be described in **Chapter 6**.

## 5.7. Discussion

The work described in this chapter highlights MINPPs-unique sequential features that could potentially be linked to enzyme promiscuity of IP<sub>6</sub> hydrolysis. Fifteen sequences, representing the active site diversity of MINPPs, were selected for further characterisation. The aim of this work was to study the overall positional specificity of IP<sub>6</sub> hydrolysis of the class to potentially unveil the determinants behind MINPPs promiscuity. This approach is chosen as an alternative to iterative saturation mutagenesis of the active sites and is meant to be a preliminary step towards further rational design of catalytic promiscuity in the family of HP2.

To identify MINPP-unique features, sequence conservation studies were carried in HP2. On the basis of functional characterisation and phylogenesis, standard enzymes were classified into four groups (A) HP2 non-phytases, (B) *E. coli* AppA-like bacterial phytases, (C) MINPPs and (D) fungal phytases. Regions of high conservation among the whole HP2 family were determined as well as group-unique motifs. These areas were mapped onto macromolecular crystal structures and their role was assigned as being either as structural or functional. Structural motifs were predicted to be involved in conformational changes (HP2 motifs RHG and G), in the positioning of the N-terminal MINPPs lid (motifs Y, S and E) and in the stabilisation of the  $\alpha/\beta$ -domain (HP2 motifs L and C). On the other hand, conserved functional residues were considered to be those located in the active site cleft and were predicted to be either involved in catalysis (RHG, H, P and N motifs) or in the substrate binding and orientation (RHG, R, Y, F and K motifs). The sequence motifs of each class, identified on the basis of a comprehensive literature review of sequential, structural and stereospecificity MINPPs features, could be used for further studies on the HP2 family, in the classification of novel sequences from metagenome studies and in further in rational design.

In this work, conserved MINPPs motifs were used as reference in the analysis of bacterial MINPPs sequences collected by databases search. A genome mining BLAST search was carried out among the RefSeq database to collect novel MINPPs of interest using *Bt*MINPP as query (note that when *Bt*MINPP was used the same results were produced) [1, 2]. The sequences collected were grouped into three clusters of increasing diversity. These included sequences collected from (1) the *Bifidobacterium* genus, (2) *Actinobacteria* excluding the *Bifidobacterium* genus, (3) *Bacteria* excluding *Actinobacteria*. Representative sequences were selected from each of the three clusters. The conservation of HP2 and MINPPs motifs were evaluated as

were their predicted active site composition to highlight variable traits possibly involved in enzyme positional stereospecificity.

*Bifidobacterium* sequences showed very high levels of identity with *Bifidobacterium longum* and *pseudocatenulatum* MINPPs and for this reason they were not considered for further studies. *Actinobacteria* MINPPs representative sequences showed an increased diversity. HP2 motifs were mostly conserved while MINPPs motifs with the highest variations were the active site regions F and K, however, a sequence also presented a substitution in the proton donor triplet (motif H). The evaluation of the active site residues that were not included in HP2 or MINPPs conserved motifs showed interesting amino acid composition in the loop downstream the RHG motif, which is subjected, in *EcAppA* and *B/MINPP*, to conformational changes upon substrate binding and is predicted to be the area involved in first substrate recruitment (**this chapter, sections 3 and 4**). A total of five sequences were selected from this group including the sequence variations above described.

Despite the higher diversity, HP2 and MINPPs motifs were present also in the wider cluster of *Bacteria*. An exception to the rule was a sequence from *Prevotella brevis* which showed to possess an “HFGSH” in place of the “RHGxR” (RHG motif) and an “Q<sub>x3</sub>A<sub>x3</sub>A” instead of a “R<sub>x3</sub>S<sub>x3</sub>F” in the R motif, the remaining motifs were conserved. It would be interesting to further characterise this enzyme to understand if it is a pseudoenzyme (activity lost) or if its active site evolved to acquire new functionalities. The highest variability was found in motifs H, P, N and F. Ten sequences in total were chosen for further characterisation among *Bacteria*. A total of 15 sequences were sent for synthesis, expressed and characterized for their positional stereospecificity of IP<sub>6</sub> hydrolysis.

Finally, phylogenetic inference suggested a subdivision of MINPPs in two major clades. A small number of ancestor sequences were identified from which two enzyme groups branched. One mainly constituted by *Bacteroides* MINPPs, the other containing MINPPs from multiple genera. While ancestral sequences presented a “HAE” proton donor triplet and a “RxxD” F motif, *Bacteroides* enzymes evolved to include different H motifs “H(D/E)x” and present a lower conservation of the F motif. The other clade, on the contrary, displayed a conserved “HAE” triplet and a mostly conserved “YFHK” F motif. It should be noted that the collected sequences homogeneously represent bacterial MINPPs, as an identity cut-off of 50% within each genus was used in the selection of representatives.

Taken as a whole, the work presented in this chapter has allowed the description of novel MINPPs active sites. It has allowed the prediction of the amino acids belonging to the substrate binding pockets for each MINPPs collected. Furthermore, it has allowed the identification of conserved motifs in HP2 and suggested the evolution of MINPPs-unique features among the collected sequences, whose effects on promiscuity will be evaluated in **Chapter 6**.

# CHAPTER 6

## 6. Characterization of 15 novel MINPPs

Fifteen MINPPs from *Bacteria* were selected following genome mining (**Chapter 5**) to be characterised in terms of their positional stereospecificity for hydrolysis of IP<sub>6</sub>. These enzymes displayed a variety of inositol polyphosphate hydrolysis profiles ranging from the very highly positionally-stereospecific proteins 5, 8 and 10 which showed a preference for the hydrolysis of phosphate groups at C4 or C6 of the inositol ring of IP<sub>6</sub>, to the high (proteins 1 and 12), medium (proteins 2, 6, 7 and 13), low (proteins 4, 11, 14 and 15) and very low positionally stereospecific proteins 9 and 16. These results give insights to the sequence determinants of positional stereospecificity in this class of enzymes. The X-ray crystal structures of the complexes of two of these enzymes with IHS were solved and refined at high resolution. One of these is a very highly positionally-stereospecific MINPP (p8) and the other is characterized by low positional stereospecificity (p15). These structures allowed docking experiments with IP<sub>6</sub> to be carried out using “*in silico*” mutated enzymes. In these experiments the active sites were modified in a reciprocal fashion and docked with the substrate in order to identify substitutions which lead to an increase of positional stereoselectivity in p15 and a decrease in p8. In this fashion, promising hotspots for positional stereospecificity were predicted and will be tested experimentally.

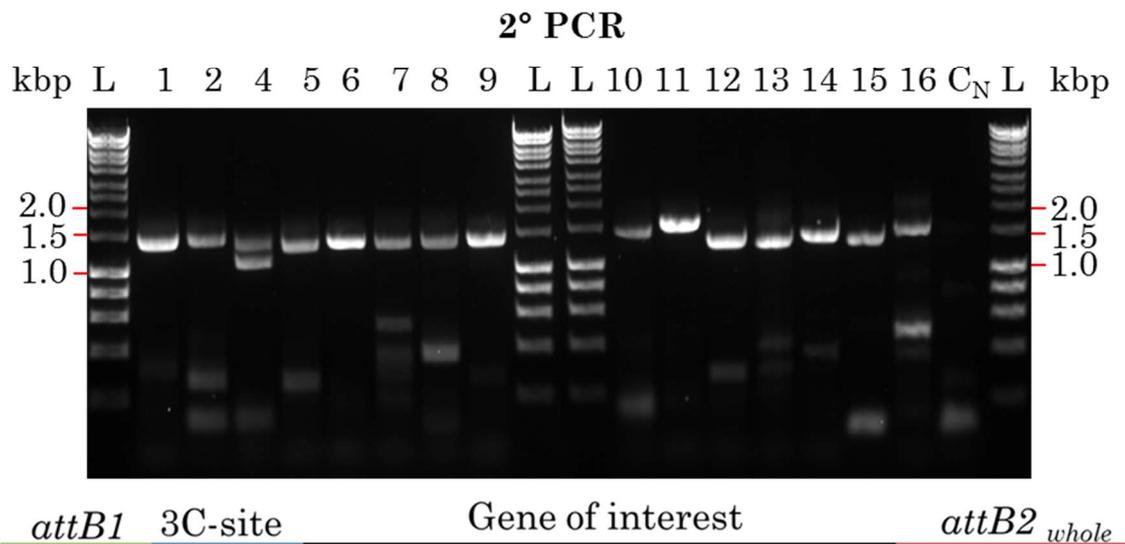
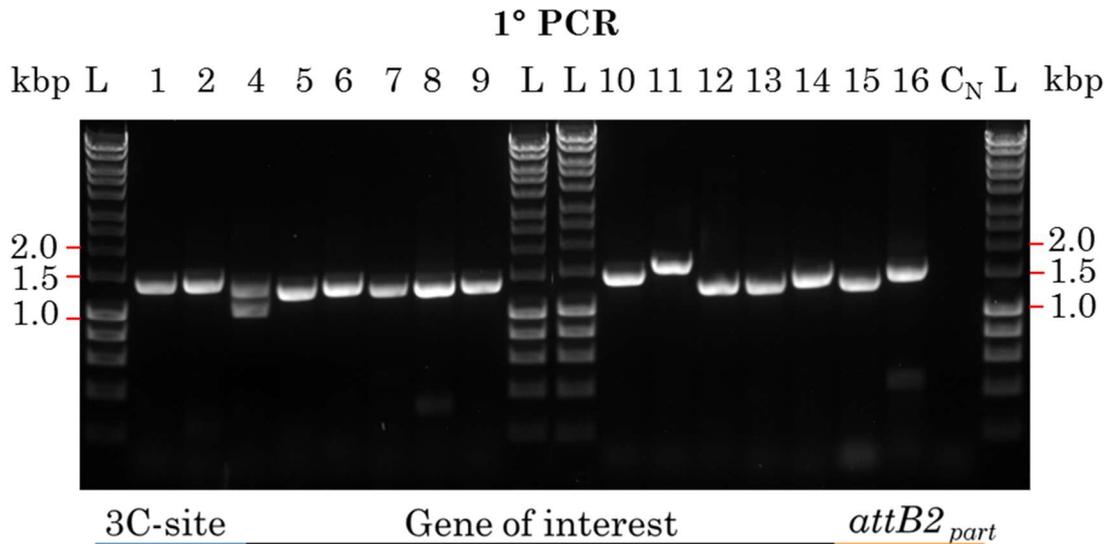
### 6.1. Cloning

The 15 genes encoding for MINPPs of interest were synthesized and cloned using the Gateway method (more details in **Chapter 2**). The PCR for the amplification of the genes of interest was designed in two steps. The first step allowed the addition of a 5' 3C-cleavage site extension and a 3' partial *attB2* extension. The second step amplified the final constructs, which contained a 5' *attB1* + 3C-protease cleavage site extension and a 3' *attB2* extension. Agarose gels of the PCRs set up for the 15 genes are shown in **Figure 6.1.1**. The products of the first PCR were diluted and used as templates for the second PCR without intermediate steps of DNA purification. Both amplifications were carried out with no problems and all genes were amplified with different degree of purity. Unfortunately, two DNA products of close size were obtained for the gene encoding for protein 4. PCR mixes were

incubated with the donor vector pDONR207 in the presence of BP clonase for their integration in the plasmid. Reactions were transformed in Stellar cells and colonies were obtained for all the genes (**Figure 6.1.2**). Six colonies per construct were screened in the search of the desired inserts. The screening procedure included for each clone to set-up of a colony PCR, a back-up master plate and an over/night culture in prevision of plasmids minipreps (**Figure 6.1.3**).

Plasmids that produced an amplicon of the correct size in the step of colony PCR were minipreped and sent to Eurofin Genomics for the sequencing of their insert. When the correct integration was confirmed, the genes were transferred by LR reaction into the destination vector pDEST17. LR reactions were again transformed into containment cells for screening of correct recombination. Correct insertion of the gene of interest into the destination vector pDEST17 was tested by the digestion of minipreped plasmids using the restriction enzyme NruI.

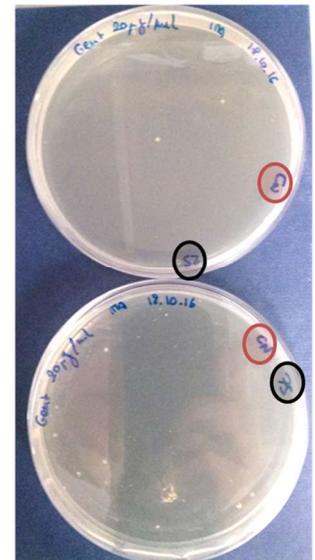
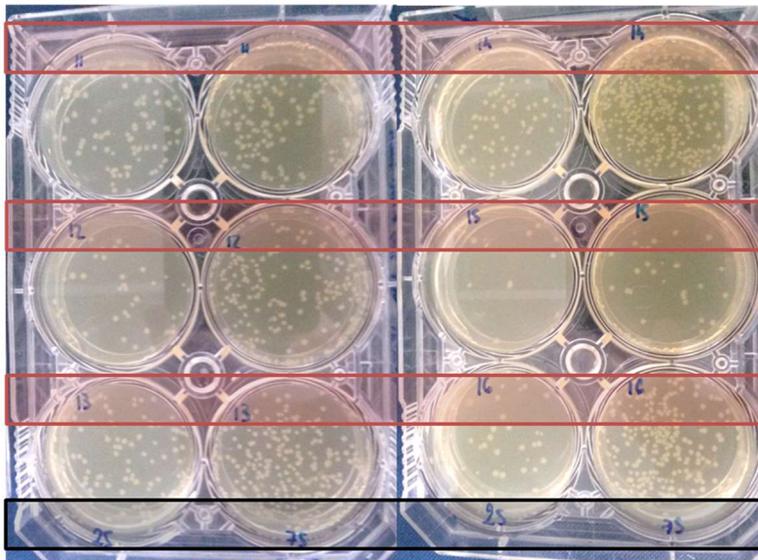
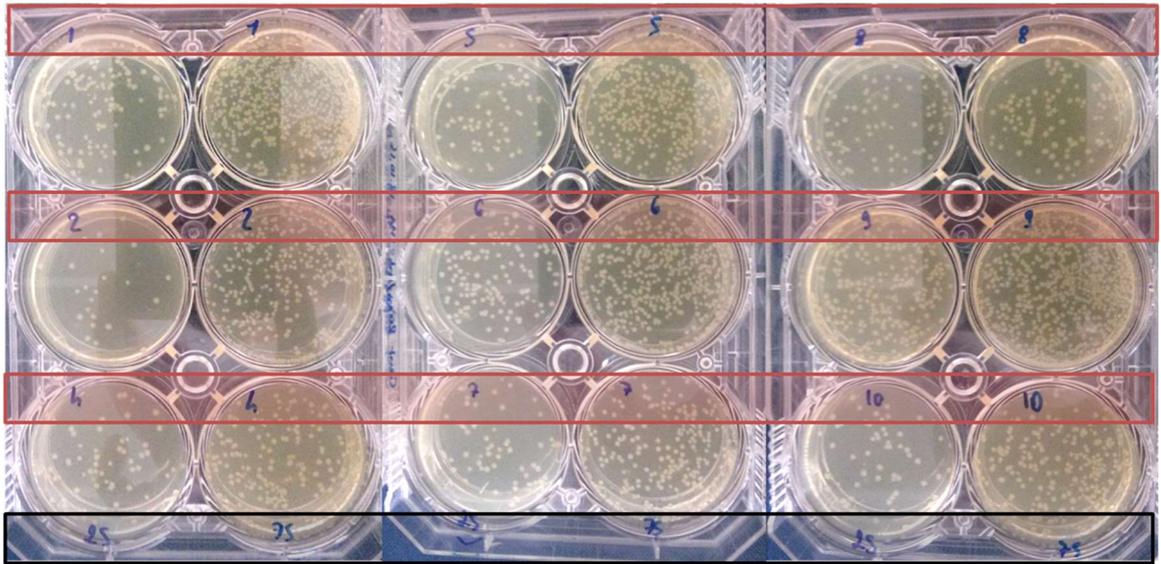
The first cycle of cloning resulted in the integration into pDEST17 of the genes encoding for p1, p6, p9, p11, p12, p13, p14, p15. The remaining genes were successfully integrated in a second cycle of cloning, which was set up using the same protocol but, in addition, the products of the 2<sup>nd</sup> PCR amplification of the entry clone were gel purified. Also, the PCR amplification of the gene encoding for p4 was repeated using a slightly higher melting temperature (52 °C instead of 51°C) and a larger number of constructs were screened by using the method of cell lysis and plasmid length detection illustrated in **Chapter 2**.



Genes	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	1° ext	2° ext
Length (bp)	1302	1308	1233	1218	1236	1245	1236	1245	1272	1326	1482	1200	1197	1284	1248	1380	42	46

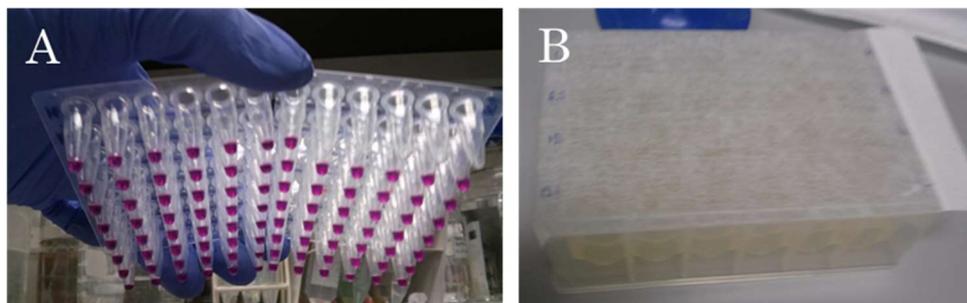
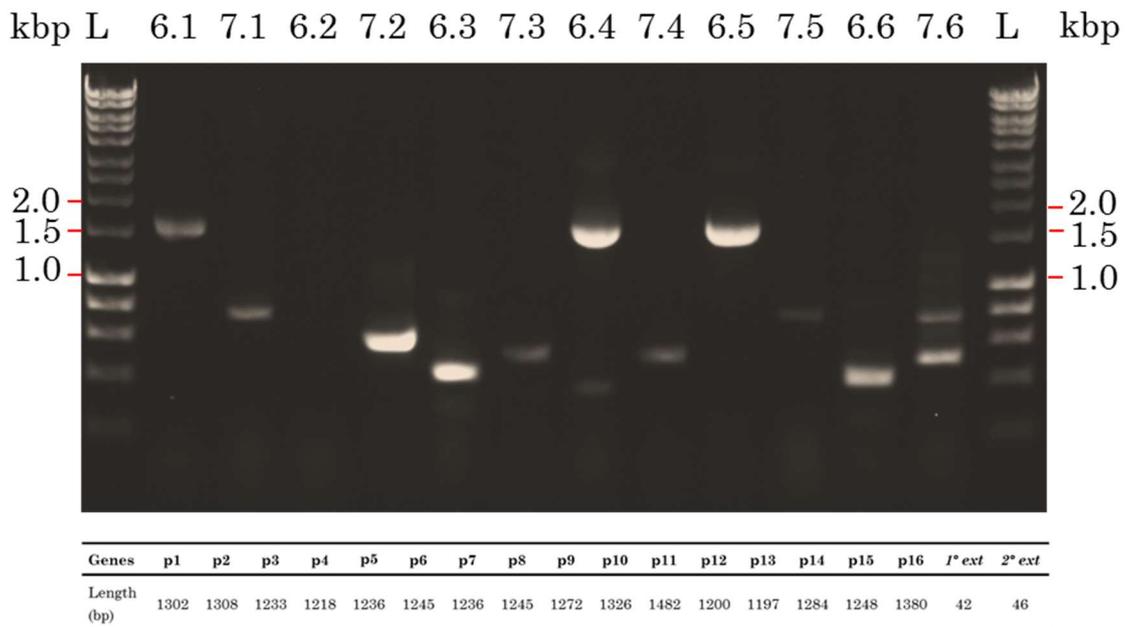
**Figure 6.1.1. Agarose gels (1%) of the PCR for the amplification of the entry clones.**

Ladder: HyperLadder 1kbp. At the top is displayed the agarose gel of first cycle of PCR, while at the bottom the agarose gel of the second PCR cycle. Genes are numbered 1-16, C<sub>N</sub> refers to the negative control PCR mix. Under each agarose gel, indications on the expected amplicon length are reported: gene of interest is coloured in black, *attB1* site in green, partial *attB2* extension in orange, *attB2* site in its whole length is coloured in magenta, the 3C-protease cleavage site in light-blue. At the bottom, a table indicating the length of each gene and the total length of the extension of 1° PCR (1° ext) and 2° PCR (2° ext) is reported. The table allows to predict the size of each expected construct. All amplicons are in-line with the expected size, though, some present non-specific bands of amplification.



**Figure 6.1.2. Transformation plates of the BP reactions between the donor vector (pDONR207) and the gene of interest (GoI).**

Transformed *E. coli* Stellar™ cells (Clontech) were plated on selective Agar + Gentamicin and incubated o/n at 37 °C. The plate, recovered the following day, are displayed in the picture above. Genes are name based on the number given to the protein they encode for: nr° 1-16, highlighted in magenta. The volume of cells injected on each plate is highlighted in black: 25 µL on the left, 75 µL on the right. Negative control plates were set up in parallel.



**Figure 6.1.3. Quality check of the cloning – screening for positive inserts.**

At the top is displayed an agarose gel (1%) of an example of colony PCR designed to amplify the inserts for the gene constructs 6 (1-6 colonies) and 7 (1-6 colonies). Ladder: HyperLadder 1kbp. Under the agarose gel, a table indicating the length of each gene and the total length of the extension of 1° PCR (1° ext) and 2° PCR (2° ext) is reported. The expected construct is equal to the sum of these three values: gene size + 1° ext + 2° ext. Amplicon of the correct size are obtained only for gene 6 – 6.1, 6.4, 6.5. Picture A) capture a PCR samples ready to be loaded onto an agarose gel 1%. Picture B) capture a 24 well plate cover by Airpore tape containing 4 mL cultures to be incubated o/n to amplify plasmids for minipreps.

## 6.2. Expression trials

Expression tests were set up for the target enzymes varying bacterial strain, media, temperature, IPTG concentration, OD of induction and length of expression (**Table 6.2.1.A**). The best conditions of expression were used in the set-up of large scale cultures (**Table 6.2.1.B**). To note that in some cases the highest yield of protein was achieved without IPTG induction, the low expression levels obtained because of a leaky promoter ensured a correct folding and accumulation of the protein of interest. For example, this was the case of p8 for which around 100 mg of protein were produced, without induction, from a 3 L culture expression incubated o/n. Six flasks of 500 mL or 750 mL cultures supplemented by ampicillin/carbenicillin at a concentration of 100 µg/mL were inoculated with 8/10 mL of an o/n pre-growth and incubated at 30 °C shaking 180 rpm until they reached OD of induction. Then, they were induced when needed (**Table 6.2.1.B**), moved to the temperature of expression until the next morning, when they were harvested at 5500 xg, 4 °C, spinning for 25 min. Pellets were resuspended in 30-50 mL of 50 mM NaH<sub>2</sub>PO<sub>4</sub> pH 7.5, 300 mM NaCl, 5 mM imidazole, 0.1% Triton, 10% glycerol and snap frozen in liquid nitrogen before storage at -80 °C.

**Table 6.2.1. Expression trials of 15 MINPPs**

The table gives a summary of the expression trials of the 15 selected enzymes: proteins, strains, media, temperatures, IPTG concentrations, OD and length of induction.

Trials	Strain	Media	Temperature	[IPTG]	OD	Length	Solubility
1 <sup>st</sup> trial p1, p6, p11 24 wells plate	BL21(DE3), Rosetta2(DE3)pLysS	LB	20 °C, 30 °C, 37 °C	0.05 mM 0.5 mM	p6: 0.5 p1, p11: 1.2-1.6	3h, o/n	Proteins are expressed but non-soluble
2 <sup>nd</sup> trial p1, p6, p9, p11, p12, p13, p14, p15 50 mL flasks	Shuffle Express T7	LB	16°C, 26°C 30°C 37 °C	0.01mM 0.05mM 0.1mM 0.5mM 1mM	0.6 1.5	o/n	Soluble proteins are expressed from each construct
3 <sup>rd</sup> trial p2, p4, p5, p7, p8, p10, p16 24 wells plate	Shuffle Express T7	LB	18°C, 24°C 30°C 37 °C	0.1mM, 1 mM	0.6	o/n	Soluble proteins are expressed from each construct

**Table 6.2.2. Best conditions of expression identified for each protein**

The table describes the best conditions of expression identified for each protein: strain, media, temperature, IPTG concentration, OD of induction and length of expression.

Protein	Strain	Media	Temperature	[IPTG]	OD	Length
p1	Shuffle Express T7	LB	24	0.5	1.0 - 1.5	o/n
p2	Shuffle Express T7	LB	24	0	-	o/n
p4	Shuffle Express T7	LB	37	1	0.6 - 0.8	o/n
p5	Shuffle Express T7	LB	30	0.1	0.6 - 0.8	o/n
p6	Shuffle Express T7	LB	30	0.5	0.6 - 0.8	o/n
p7	Shuffle Express T7	LB	37	1	0.6 - 0.8	o/n
p8	Shuffle Express T7	LB	24	0	-	o/n
p9	Shuffle Express T7	LB	37	0.5	0.6 - 0.8	o/n
p10	Shuffle Express T7	LB	18	0.1	0.6 - 0.8	o/n
p11	Shuffle Express T7	LB	24	0.5	1.0 - 1.5	o/n
p12	Shuffle Express T7	LB	37	0.5	0.6 - 0.8	o/n
p13	Shuffle Express T7	LB	37	0.5	0.6 - 0.8	o/n
p14	Shuffle Express T7	LB	37	0.5	1.0 - 1.5	o/n
p15	Shuffle Express T7	LB	18	0.5	0.6 - 0.8	o/n
p16	Shuffle Express T7	LB	24	0	-	o/n

### 6.3. Purification

The protocol used for the purification of His-tagged proteins is described in **Chapter 2.6.1**. Only p8, p10 and p15, were efficiently separated to a high degree of purity (**Figure 6.3.1**). The three proteins were expressed, purified by IMAC and dialysed o/n at 4 °C in a 3C-protease supplemented buffer to ensure His-tag removal. The tag of p8 and p15 was cleaved almost to completion, but this was not the case of p10. The solubility limit of this enzyme, close to 3 mg/mL, prevented its concentration to a small volume, which would have allowed an efficient separation of cleaved and uncleaved protein through gel filtration. Because of the low sample purity, no crystallisation trials were set up for p10. Instead, p8 and p15, whose solubility limit is 8 mg/mL and 15 mg/mL respectively, were concentrated to smaller volumes. This

allowed me to further purify the samples by gel filtration to a purity suitable for crystallisation trials.

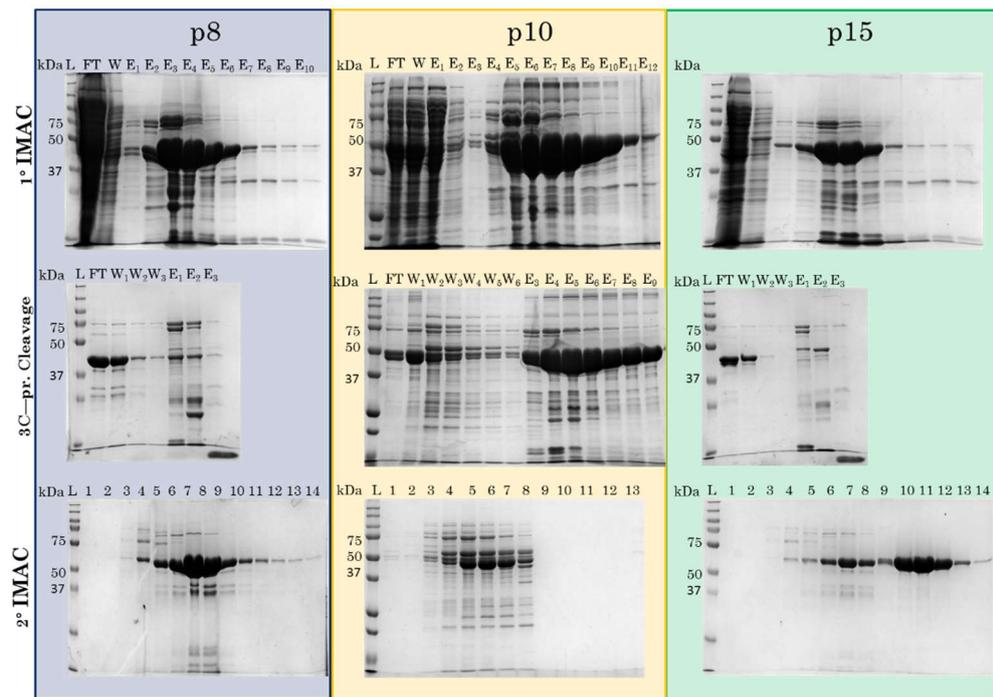
The purification of the remaining proteins was troublesome for multiple reasons:

(1) Low proteins concentration together with high level of contaminants prevented the identification of proteins p1, p4, p6, p7, p9, p11, p12, p13, p14 in the gradient elution of the first Ni-NTA/His-tag IMAC by SDS-PAGE. Protein elution was monitored instead by activity assays. The binding buffer was designed to contain low imidazole concentration (5 mM), 10 % glycerol and 0.1 % Triton, because the majority of the proteins displayed difficulties in the interaction with the resin. In fact, when a concentration of imidazole equal to 20 mM was added to the binding buffer, phytase activity was detected only in the flow through of the IMAC purification, suggesting that proteins would pass through the column without interacting with the resin. This choice of buffer produced a larger amount of contaminants, however, at least a partial enzymes purification was achieved. An example is reported in **Figure 6.3.2 – A**, which shows the SDS-PAGE of the first IMAC purification of protein p4.

(2) p1, p2, p9, p13, p16 were not able to efficiently interact with the Ni-NTA resin even using a 5 mM imidazole, 10 % glycerol and 0.1 % Triton buffer. An example is reported in **Figure 6.3.2. – B**, which shows the SDS-PAGE of the first IMAC purification of protein p16: the majority of protein is collected in the flow through (its absorbance reached saturation: 4 units) while only a small amount of protein bound the resin (the eluted fraction reached a maximum of absorbance of 0.345 units).

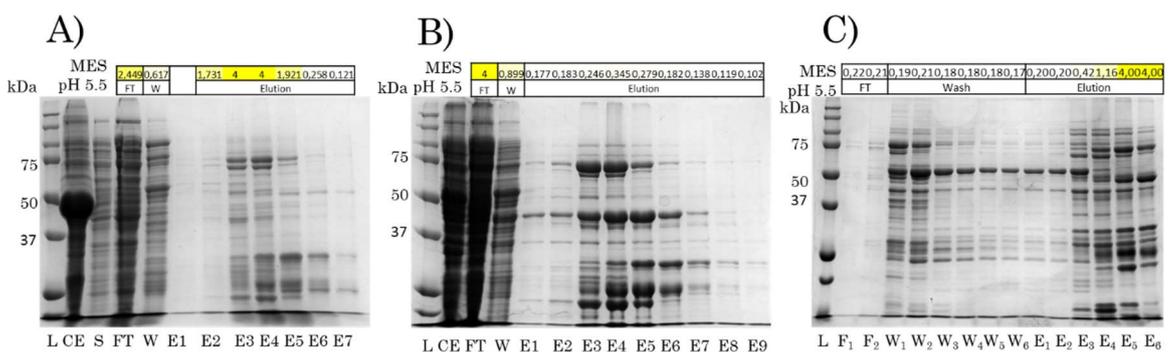
(3) The purification was prevented by an uncomplete His-tag cleavage in proteins p1, p2, p5, p11 (**Figure 6.3.2 – C**). This step would have allowed to collect the protein in the flow through of the 2<sup>nd</sup> IMAC. However, His-tag cleavage was not tested in proteins p6, p7, p9, p12, p13, p14, p16 due to time constraints.

(4) The proteins had the tendency to precipitate during dialysis and concentration, however, this was avoided just with the simple dilution of samples. An alternative to try would have been the addition of a reducing agent (e.g. DTT) to prevent the formation of intermolecular disulphide bridges in the proteins in which no disulphide bridges are predicted (**Supplemental Information 8.4.4**) or stabilising agents such as glycerol [170] or trehalose [171], in particular during spin concentration.



**Figure 6.3.1. 12% acrylamide SDS-PAGE of the purifications of proteins p8, p10, p15.**

L: Precision Plus Protein™ Dual Colour Standards, FT: flow through, W: wash, E: eluted fraction. Theoretical MWs: p8 – 50939.64 Da, p10 – 53658.26 Da, p15 – 50236.65 Da. Different background colours were chosen for each protein: p8 – light-blue, p10 – yellow, p15 – green. The first row of SDS-PAGE displays samples from the first His-tag/Ni-NTA IMAC step of purification. The second row displays a second step of IMAC aimed to isolate proteins whose His-tag was cleaved by 3C-protease o/n. Cleaved proteins are collected in the FT and His-tagged proteins (uncleaved protein of interest or His-tagged 3C-protease) in addition to contaminants should be released from the resin only during elution. The third row displays SDS-PAGE of the final separation step by gel filtration.

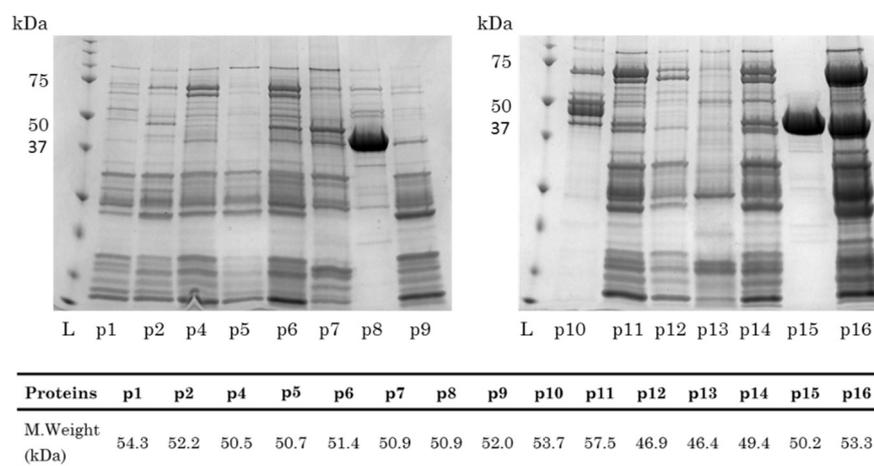


**Figure 6.3.2. 12% acrylamide SDS-PAGE of the IMAC purifications of proteins p4, p16 and p2.**

L: Precision Plus Protein™ Dual Colour Standards, CE: crude extract, S: supernatant, FT: flow through, W: wash, E: eluted fractions. At the top of the gels is reported a phosphatase activity assay (Abs unit) in 0.2M MES pH 5.5, NaCl 0.15M of 40µL samples, incubated 37°C for 30min with 10 mM pNPP. A) 1st IMAC of protein p4; B) 1st IMAC of protein p16; C) 2nd IMAC of protein p2. Theoretical MWs: p4 - 50458.17 Da, p16 - 53331.13 Da, p2 - 52214.77 Da.

Despite these problems, all the proteins were partially purified and collected in an amount that was sufficient to test their activity against pNPP and phytate, their pH profiles against pNPP and their stereospecificity. An accurate protein concentration was not estimated. Contamination from *E. coli* endogenous phytases or phosphatase was excluded because no basal activity was detected in cells during expression trials and IP<sub>5</sub> production diverged from the profile of *E. coli* AppA. However, *E. coli* AppA promoters p<sub>cyx</sub> and p<sub>appA</sub> can be induced by anaerobic conditions, phosphate starvation or by entry into stationary phase [172]. Due to time restrictions, the optimisation of the purification was not completed. Also, there has not been time to generate knock-out phytase-deficient *E. coli* strains for protein expression or to confirm protein identity by Western blots.

In Figure 6, an SDS-PAGE gel of all the protein sample stocked -80 °C displays their general high level of contamination, with the exception of p8 and p15. Nevertheless, with a few more cycle of optimisation better yield and purity could have been reached for a higher number of samples.



**Figure 6.3.3. SDS-PAGE of the final 15 MINPPs samples stocked -80°C.**

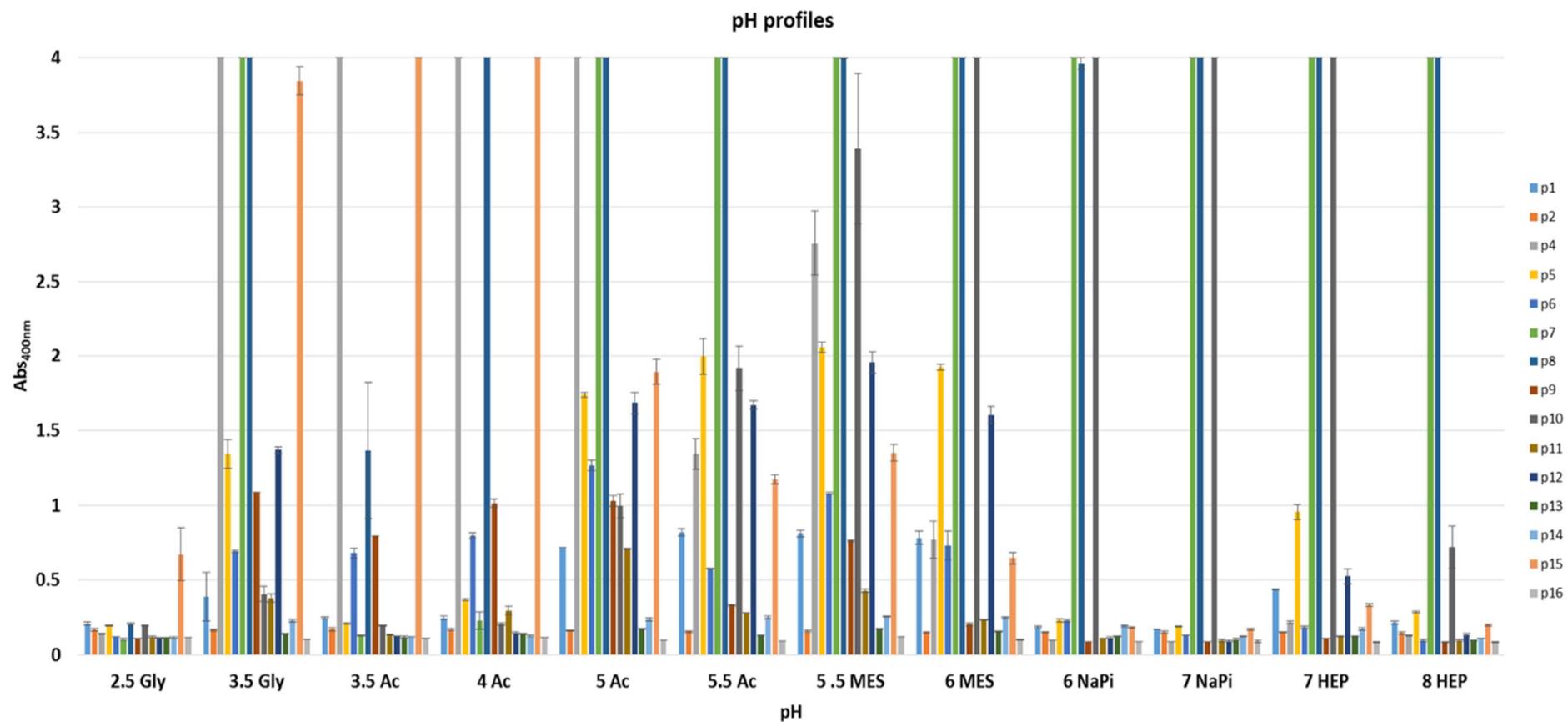
L: Precision Plus Protein™ Dual Colour Standards, p1-p16 – name of the enzyme. The table under the two SDS-PAGE gels report the molecular weight of the tagged constructs. All samples show a low degree of purity. Exceptions are proteins p8, p10 and p15.

## 6.4. pH profiles and phosphatase activity

pH profiles for each enzyme were tested after the first IMAC purification (**Figure 6.4.1**). pNPP was initially used as substrate to determine enzymes phosphatase activity. This test was chosen because it is faster than molybdenum blue assay and more reliable when dealing with partially purified phosphatases because it is not affected by the phosphate background of the sample (eventual Pi carry over from media). Phytase activity was tested instead by separating inositol polyphosphates products by HPLC and by steady-state kinetics assays on pure protein samples.

Enzymes at a concentration of approximately 1  $\mu$ M were incubated with pNPP 10 mM at 37 °C for 30 min and inactivated by the addition of 1 M NaOH. Absorbance was measure at 405 nm after 10 min of inactivation. For more details on phosphatase activity assays see **Chapter 2.7.1**. Reactions were set up in a wide range of pH: Glycine-Cl 0.2 M, NaCl 0.15 M pH 2.5 and 3.5; NaAcetate 0.2 M, NaCl 0.15 M pH 3.5, 4.0, 5.0 and 5.5; MES 0.2 M, NaCl 0.15 M pH 5.5 and 6.0; NaPhosphate 0.2 M, NaCl 0.15 M pH 6.0 and 7.0; HEPES 0.2 M, NaCl 0.15 M pH 7.0 and 8.0. Positive and negative controls were prepared in parallel as well as reads of the absorbance of each buffer, which were subtracted from the reads of the samples. Assays were carried out in duplicates.

pH profiles showed that the proteins most active in a pH range of 2.5-3.5 are p4, p7, p8 and p15, followed by p5, p12 and p9. At the pH window of 4.0-6.0, a similar response was registered, however p15 showed a lower activity at pH 5.0-6.0, which was comparable with proteins p5, p6, p9 and p12. Only a few enzymes proved to be active at a neutral/basic pH (7.0-8.0), these were p7, p8 and p10. When analysing these data, it has to be remembered that protein concentrations were not comparable because of low purity. The concentration of some enzyme could be overestimated for more than 100-fold and they could potentially be considered to be in a similar range of activity with the previously characterized *B*MINPP or *Bt*MINPP.



**Figure 6.4.1. Phosphatase activity against pNPP of the proteins p1-p16.**

The x axis reports the buffers used to maintain each pH. Absorbance is registered at 405 nm, a value of 4 represent saturation. For each protein a colour is assigned as found in the legend on the right.

## 6.5. Analysis of the product of IP<sub>6</sub> hydrolysis by HPLC

The hydrolysis of the phosphomonoester bonds of *myo*-inositol hexaphosphate by phytases generates enzyme-specific combinations of lower phosphorylated intermediates. To characterize the positional stereospecificity of IP<sub>6</sub> hydrolysis, in other words, to characterize the individual IP<sub>5</sub> species produced by each MINPP, the HPLC protocol described in **Chapter 2.7.3** was used for the separation of the products of reaction.

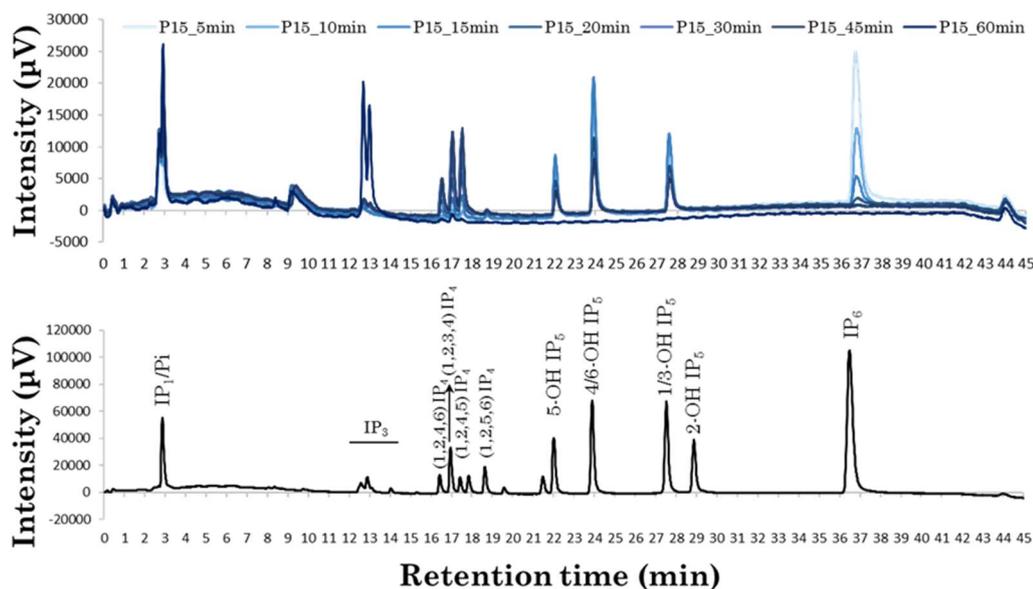
The IPs mixtures were injected on a CarboPAC PA200 column, separated in a gradient of methanesulfonic acid, mixed with Fe(NO<sub>3</sub>)<sub>2</sub> and their absorbance was measured at 290 nm. The height and the area of the produced peaks was dependent on phosphorylation level of the IPs species and on their concentration. The retention times of the IPs species in reaction was identified by comparison with a standard sample of chemically hydrolysed IP<sub>6</sub> (boiled in HCl, 120 °C, 24 h). The duration of reactions and enzyme concentrations were iteratively modified to stop IP<sub>6</sub> cleavage at comparable stages (e.g. % total IP<sub>5</sub> products was held constant (20-30 %) to compare relative species composition), **Figure 6.5.1. – A**. An example is given in **Figure 6.5.1. – B** of an experiment of IP<sub>6</sub> hydrolysis in which the production of IPs was monitored by taking time-points of reaction. The chromatogram displays reactions intermediates produced by p15 (*Amycolatopsis jejuensis* MINPP). Only IP<sub>5</sub> intermediates were considered in the evaluation of positional stereospecificity. The initial concentration of *myo*-inositol hexakisphosphate was constant and set to 1mM. The buffers chosen were the ones in which an enzymes activity maximum was detected by phosphatase assays. Reactions were incubated at 37°C and stopped by heating at 98 °C for 10 min. At least two chromatograms replicate for each protein were registered. To note, the specified molarity of the enzymes in reaction is not representative of the real enzyme concentration in the samples as only three of the enzymes have been fully purified (p8, p10, p15). The 15 MINPPs showed a variety of IP<sub>5</sub> products profiles, which are reported in **Figure 6.5.2**. They were quantified individually as percentages over the total IP<sub>5</sub> species in reaction (**Table 6.5.1**).

MINPPs from *Flammeovirga pacifica* (p1) and *Arsukibacterium* sp. MJ3 (p2), containing respectively a proton donor-less triplet “HST” and “HAQ”, were found to be active on IP<sub>6</sub>, in line with the findings of [47] for *Drosophila melanogaster* MINPPs which also bear an “HST” H motif.

A)

Enzyme	Buffer	[Enzyme]	length	IP <sub>5</sub> %
P1	0.2M MES pH 5.5, 0.15M NaCl	1000nM	15min	22.4%
P2	0.2M MES pH 5.5, 0.15M NaCl	1000nM	60min	20.3%
P4	0.2M NaAcetate pH 4.0, 0.15M NaCl	1000nM	120min	22.0%
P5	0.2M MES pH 5.5, 0.15M NaCl	1000nM	10min	21.3%
P6	0.2M NaAcetate pH 5.0, 0.15M NaCl	1000nM	240min	19.0%
P7	0.2M Glycine-Cl pH 3.5, 0.15M NaCl	1000nM	120min	25.0%
P8	0.2M MES pH 5.5, 0.15M NaCl	10nM	20min	23.0%
P9	0.2M NaAcetate pH 5.0, 0.15M NaCl	1000nM	300min	25.6%
P10	0.2M MES pH 5.5, 0.15M NaCl	100nM	45min	24.0%
P11	0.2M NaAcetate pH 5.0, 0.15M NaCl	1000nM	10min	21.1%
P12	0.2M Glycine-Cl pH 3.5, 0.15M NaCl	1000nM	120min	13.0%
P13	0.2M Glycine-Cl pH 3.5, 0.15M NaCl	1000nM	45min	26.6%
P14	0.2M MES pH 5.5, 0.15M NaCl	10000nM	300min	34.0%
P15	0.2M Glycine-Cl pH 3.0, 0.15M NaCl	100nM	60min	26.6%
P16	0.2M MES pH 5.5, 0.15M NaCl	10000nM	300min	14.2%

B)



**Figure 6.5.1. A) Reactions conditions of IP<sub>6</sub> hydrolysis by 15 MINPPs.**

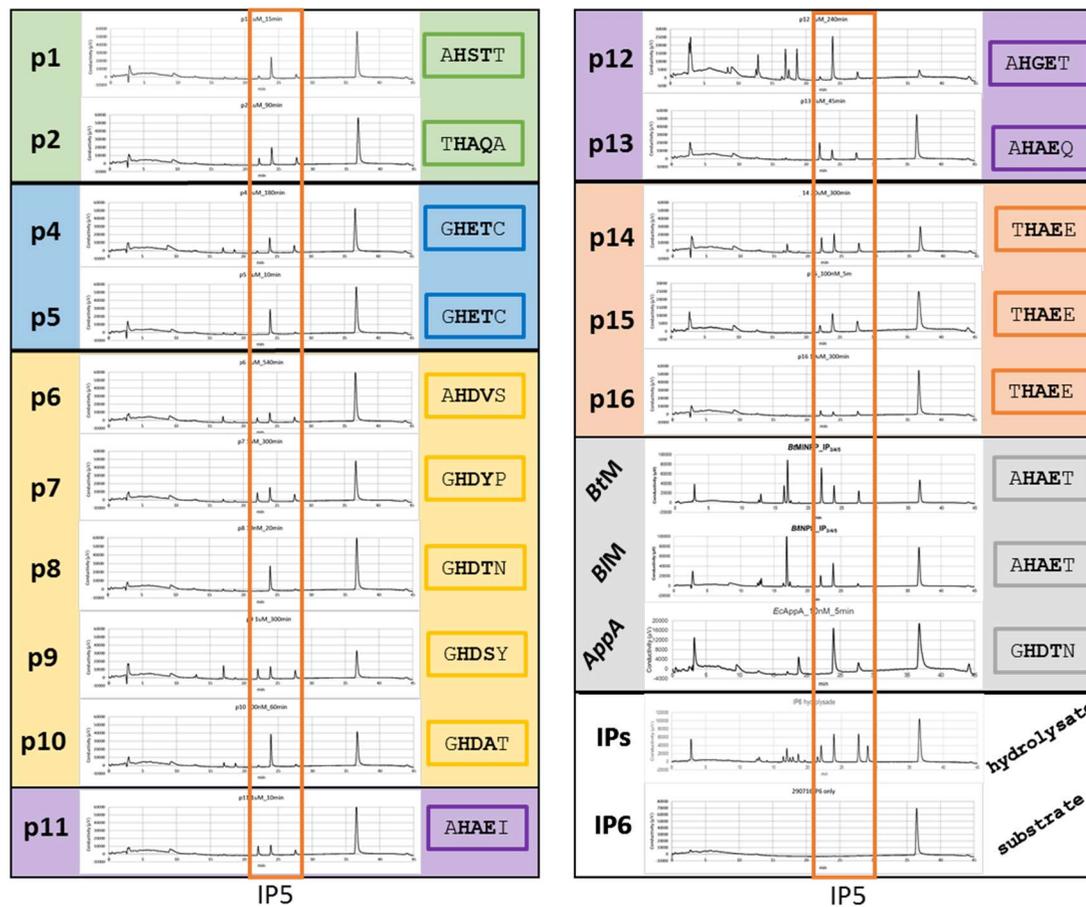
The table provides information on buffer, enzyme concentration, length of incubation and % of total IP<sub>5</sub> production. B) HPLC chromatograms of the products of IP<sub>6</sub> hydrolysis by p15. x axis: retention time (min), y axis: intensity of UV absorbance (µV). The top chromatogram shows an overlay of 7 time-points of reaction (5, 10, 15, 20, 30, 45 and 60 min incubation) coloured at increasing shades of blue. 1mM high pure IP<sub>6</sub> (Merck) is hydrolysed by 100 nM enzyme in 0.2 M Glycine-Cl pH 3.0, 0.15 M NaCl. A standard chromatograph is reported at the bottom to provide the retention time of IPs species for comparison.

The data available does not allow to accurately compare catalytic rates with other MINPPs, their concentration cannot be quantified because of low purity. However, they proved to be significantly more active than the proton donor less *EcAppA* mutant HAT. In fact, the non-pure samples of p1 and p2 showed a comparable activity to the highly purified *EcAppA* mutant using the same amount of estimated total proteins in reaction and identical assay conditions. These results suggest that proton donation in p1 and p2 may be carried out by other residues of the active site cleft or through an alternative mechanism yet to be described.

The majority of MINPPs displayed a preference for the hydrolysis of the phosphate groups on position C4 and C6 of the inositol ring of IP<sub>6</sub>, with the exception of the protein of *Aeromicrobium* sp. Root:495 which showed a preference for the cleavage of the phosphate group in position C5. MINPPs were classified according to their proton donor motif (**Figure 6.5.2**) and their level of positional stereospecificity. In regard of the latter, enzymes were grouped in 5 level of positional specificity: very high, high, medium, low, very low (**Table 6.5.1**).

MINPPs with very high positional specificity are characterized by a predominant IP<sub>5</sub> peak and individual secondary peaks accounting for less than 10% of total IP<sub>5</sub>. Proteins p5 (*Bacteroides intestinalis* MINPP), p8 (*Prevotella* sp. CAG:617 MINPP) and p10 (*Fibrobacter succinogenes* MINPP) belong to this group. High positional specific MINPPs included the proton donor-less p1 (*Flammeovirga pacifica* MINPP) and protein p12 (*Aeromicrobium* sp. Root 236 MINPP), enzymes showing a predominant peak with at least an individual secondary peak between 10% and 20% of total IP<sub>5</sub>.

MINPPs characterized by medium positional stereospecificity were p2 (*Arsukibacterium* sp. MJ3 MINPP) and p4 (*Bacteroides* sp. CAG:927 MINPP), which produced a predominant peak higher than 50% of total IP<sub>5</sub> and at least an individual secondary peak between 20% and 30% of total IP<sub>5</sub>. The larger group included six proteins with low positional stereospecificity: p6 (*Bacteroides* sp. CAG:545 MINPP), p7 (*Bacteroides* sp. CAG:770 MINPP), p11 (*Oxalobacteraceae bacterium* AB MINPP), p13 (*Aeromicrobium* sp. Root:495 MINPP), p14 (*Streptacidiphilus jeojiense* MINPP) and p15 (*Amycolatopsis jejuensis* MINPP). These proteins were characterised by a major peak accounting for less than 50% of total IP<sub>5</sub> and with a secondary peak accounting for 30% of total IP<sub>5</sub> or more. The last group included enzymes with very low positional stereospecificity. MINPPs that belong to this group, p9 (*Bacteroides bacterium* CF) and p16 (*Streptomyces* sp. AW19M42), produced a profile with all IP<sub>5</sub> peaks accounting between 30 % and 40 % of the total IP<sub>5</sub>.



**Figure 6.5.2. HPLC chromatograms of the products of IP<sub>6</sub> hydrolysis by the 15 selected MINPPs.**

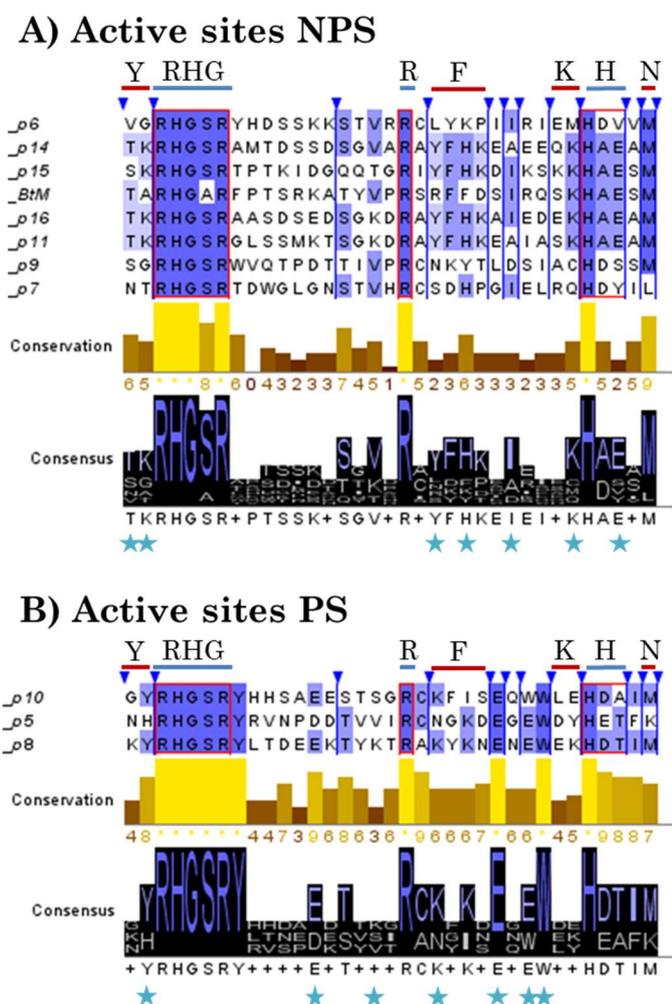
x axis: retention time (min), y axis: intensity of UV absorbance at 290 nm ( $\mu$ V). Chromatograms of protein p1-p16 are reported. They are coloured by proton donor motif (on the right of each chromatogram). *Bt*MINPP, *B*/MINPP and *Ec*AppA chromatograms are reported for comparison, as well as a chromatogram of the substrate IP<sub>6</sub> and the hydrolysate standard. IP<sub>5</sub> peaks are highlighted by an orange frame.

**Table 6.5.1. Summary table describing the distributions of individual IP<sub>5</sub> % produced by 15 MINPPs and some of their sequence features.**

IP<sub>5</sub> individual peaks % as well as details on the sequence motifs H, F and lineage are described. Peak percentages seem not to display any linear correlation with the sequence feature reported. The major IP<sub>5</sub> peak is coloured by shades of orange according to its % area on the total IP<sub>5</sub> peak area. Orange intensity: 0 – very low specific MINPPs (all peaks between 30 % and 40 % of total IP<sub>5</sub>); orange intensity: 1 – low specific MINPPs (predominant peak lower than 50 % of total IP<sub>5</sub>, secondary peaks lower than 30 %); orange intensity: 2 – medium specific MINPPs (predominant peak higher than 50 % of total IP<sub>5</sub>, secondary peaks between 20 % and 30 %); orange intensity: 3 – highly specific MINPPs (secondary peaks between 20 % and 30 %); orange intensity: 4 – very highly specific MINPPs (secondary peaks lower than 10 %).

Protein	H motif	F motif	IP <sub>5</sub> 1/3-OH (%)	IP <sub>5</sub> 4/6-OH (%)	IP <sub>5</sub> 5-OH (%)	Species	Genus
1_Fp	AHSTT	RAFK	13 ±1	77±2	10±1	<i>Flammeovirga pacifica</i>	<i>Bacteroidetes</i>
2_AMJ3	THAQA	YFNK	25±1	55±1	20±0	<i>Arsukibacterium sp. MJ3</i>	<i>γ-Proteobacteria</i>
4_BCAG927	GHETC	NHHD	23±4	68±4	8±1	<i>Bacteroides sp. CAG:927</i>	<i>Bacteroidetes</i>
5_Bi_noSP	GHETC	NGKD	7±2	92±2	1±1	<i>Bacteroides intestinalis</i>	<i>Bacteroidetes</i>
6_B. CAG545	AHDVS	YKPA	29±2	48±1	24±1	<i>Bacteroides sp. CAG:545</i>	<i>Bacteroidetes</i>
7_BCAG770	GHDYP	KYVS	29±3	44±2	27±1	<i>Bacteroides sp. 770</i>	<i>Bacteroidetes</i>
8_PCAG617	GHDTN	KYKN	6±1	93±1	1±0	<i>Prevotella sp. CAG:617</i>	<i>Bacteroidetes</i>
9_BbCF	GHDSY	NKYT	30±1	40±1	29±2	<i>Bacteroidales bacterium CF</i>	<i>Bacteroidetes</i>
10_Fs	GHDAT	KFVS	5±1	91±1	4±1	<i>Fibrobacter succinogenes</i>	<i>Fibrobacteres</i>
11_Ob	AHAEI	YFHK	20±1	42±1	37±1	<i>Oxalobacteraceae bacterium AB_14</i>	<i>β-Proteobacteria</i>
12_AR236	AHGET	RIED	10±4	84±4	6±1	<i>Aeromicrobium sp. Root236</i>	<i>Actinobacteria</i>
13_AR495	AHAEQ	RFDK	23±1	29±0	45±1	<i>Aeromicrobium sp. Root495</i>	<i>Actinobacteria</i>
14_Sj	THAEE	YFHK	25±1	43±1	31±2	<i>Streptacidiphilus jeojiense</i>	<i>Actinobacteria</i>
15_Aj	THAEE	YFHK	32±1	49±1	19±1	<i>Amycolatopsis jejuensis</i>	<i>Actinobacteria</i>
16_SAW19M42	THAEE	YFHK	32±1	30±0	37±1	<i>Streptomyces sp. AW19M42</i>	<i>Actinobacteria</i>

The active site prediction of the 15 MINPPs sequences were compared to highlight unique residues present within the groups of non-positional stereospecific (NPS, groups 4 and 5) and positional stereospecific (PS, group 1) MINPPs (**Figure 6.5.3**). The active sites were annotated to display conserved HP2 motifs (**Figure 6.5.3, blue line notations**) and conserved MINPPs-unique motifs (**Figure 6.5.3, red line notations**). Also, residues present only within each group, and therefore possibly involved in determining positional stereospecificity, were marked by a blue star (**Figure 6.5.3, star on the consensus sequence**). Some of these residues are part of the MINPPs-unique motifs Y (“TK” in NPS or “xY” in PS), F (“YxH” in NPS or “KxK” in PS) and K (“xK” in NPS). However, they are not 100% conserved within their positional stereospecificity groups. Particularly interesting are the residues “E<sub>X<sub>n</sub></sub>(E/W)W” present in PS MINPPs and absent instead in NPS enzymes.



**Figure 6.5.3. Active site alignments of non-positional stereospecific MINPPs (NPS, panel A) and positional stereospecific MINPPs (PS, panel B).**

Sequences are coloured by % identity. Conservation and consensus sequences are displayed. Also, HP2 (RHG, R, H - blue lines) and MINPPs-unique (Y, F, K, N - red-lines) motifs are annotated. Residues group-specific are highlighted by a blue star.

To collect more details on the active site configuration of NPS and PS MINPPs, the X-ray macromolecular crystal structure of a protein from each group was solved in complex with the substrate analogue inositol hexasulphate. The next paragraphs focus on the biochemical and structural characterisation of the very high positional stereospecific p8 (*Prevotella* sp. CAG:617 MINPP) and the low positional stereospecific p15 (*Amycolatopsis jejuensis* MINPP) which allowed to predict hotspots areas for stereospecificity in MINPPs (**Chapter 10**).

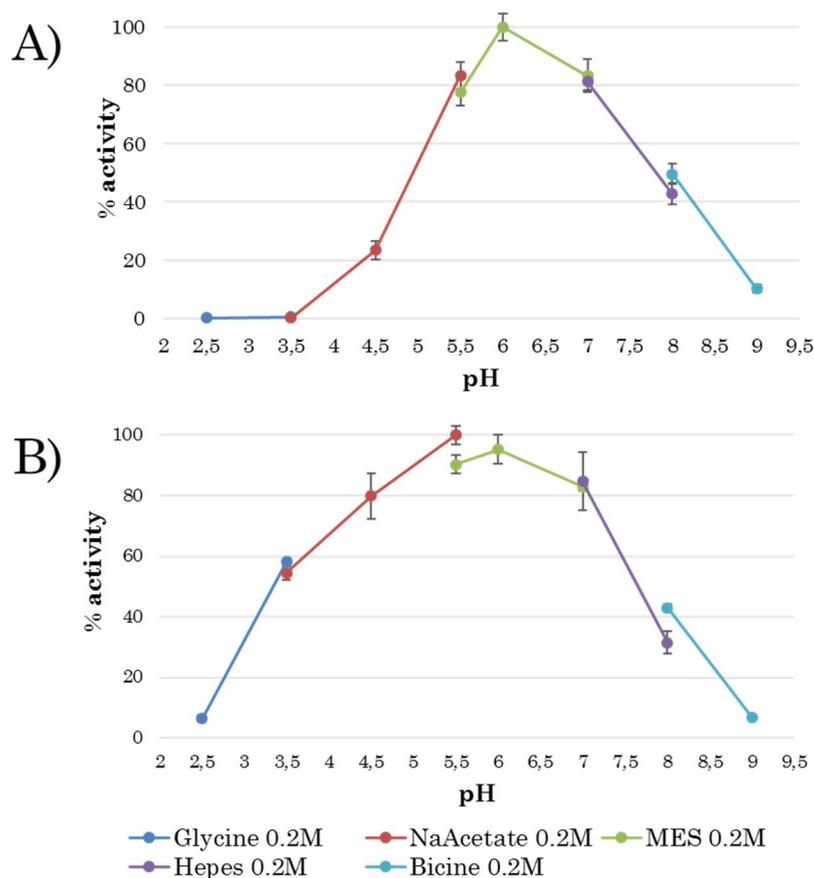
## 6.6. Biochemical and biophysical characterisation of p8 and p15

### 6.6.1. pH profile

IP<sub>6</sub> hydrolysis by p8 and p15 was tested in multiple buffers overlapping in pH range (0.2 M Glycine-Cl pH 2.5/3.5, 0.15 M NaCl; 0.2 M NaAcetate pH 3.5/4.0/5.0/5.5, 0.15 M NaCl; 0.2 M MES pH 5.5/6.0/6.5/7.0, 0.15 M NaCl; 0.2 M HEPES pH 7.0/8.0, 0.15 M NaCl; 0.2 M Bicine pH 8.0/9.0, 0.15 M NaCl). The inorganic phosphate released by p8 and p15 was quantified by monitoring the absorbance of its complex with molybdenum blue reagent at 700 nm (**Chapter 2.7.2**). Reactions of 50 µL were set up in triplicate at fixed concentrations of enzymes (100 nM) and substrate (IP<sub>6</sub> - Sigma Premium Quality: 1 mM).

Reactions were incubated at 37 °C for 15 min before inactivation by addition of molybdenum blue reagent in equal parts. Solutions were left to develop for 30 min before measuring absorbance at 700 nm. Triplicate reads of buffer-only solutions were taken as well as a phosphate calibration curve to check that the absorbance registered was in the linear range for the assay.

The pH optimum of p15 and p8 were found to be, respectively, 5.5 and 6.0, with p15 being able to process IP<sub>6</sub> at his 60 % of activity or higher in a pH range between 3.5 to 7.5, and p8 between pH 5 and 7.5 (**Figure 11**).



**Figure 6.6.1.1. pH profile of p8 and p15.**

x axis: pH, y axis: % activity. Buffers colour scheme: blue - pH 2.5/3.5, 0.2 M Glycine-Cl, 0.15 M NaCl; red - pH 3.5/4.0/5.0/5.5, 0.2 M NaAcetate, 0.15 M NaCl; green - pH 5.5/6.0/6.5/7.0, 0.2 M MES 0.15 M NaCl; violet - pH 7.0/8.0, 0.2 M HEPES, 0.15 M NaCl; light-blue - pH 8.0/9.0, 0.2 M Bicine, 0.15 M NaCl. Buffer only Abs<sub>700</sub> were subtracted from raw samples Abs<sub>700</sub>, then, the latter were averaged. The % of averaged samples Abs<sub>700</sub> to the maximum average were calculated and plotted above with the respective deviations. The top graph represents the pH profile of the enzyme p15: pH optimum 5.5. The bottom graph represents the pH profile of the enzyme p8: pH optimum 6.0.

## 6.6.2. Enzyme kinetics

Reactions of 50µL were set up in triplicate at fixed concentrations of enzymes (p8: 50 nM and p15: 20 nM), increasing concentration of IP<sub>6</sub> (25, 50, 100, 200, 400, 600, 800, 1200 µM) and incubated for 5, 10, 15, 20 min at room temperature. Buffers at pH optimum for activity were chosen to be able evaluate the best performance for each enzyme. Buffers were: 0.2 M NaAcetate pH 5.5, 0.15 M NaCl for p15 and 0.2 M MES pH 6.0, 0.15 M NaCl for p8. Reactions were inactivated by addition of molybdenum blue reagent and the Abs<sub>700</sub> was measured after 30 min incubation with the stopping reagent. Data were processed with the 'nls' function provided in R (<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nls.html>), that determines the nonlinear least-squares estimates of the parameters of a nonlinear model. In this

analysis, the non-linear model is the MM equation. The goodness of the model's fit is given by residual errors and t-test. In **Table 6.6.2.1**, the results are reported in comparison with the kinetic parameters collected for wild-type *EcAppA* (for more information see **Chapter 3**).

Different  $K_M$  values were obtained for the three proteins. p15 displayed a  $K_M$  1/3 higher than *EcAppA*, while p8 showed a higher binding affinity with a value half the one of the *E. coli* enzyme. The latter showed a comparable catalytic efficiency to p15 while p8 displayed a  $k_{cat}$  4-fold lower. Overall, the turnover number of p15 and p8 proved to be 1/3 and 1/2 lower than *EcAppA*, respectively. The comparison of the phytase units (FTU) of these three enzymes shows that, again, p8 is the protein with the lower performance: 25% of *EcAppA* FTU; instead the difference for p15 is smaller: 90% of *EcAppA* FTU.

**Table 6.6.2.1. Kinetic parameters for the hydrolysis of inositol hexasulphate at 37°C, 5min reactions.**

$V_{max}$  is expressed in  $\mu\text{M}\cdot\text{min}^{-1}$ ,  $K_M$  in  $\mu\text{M}$ ,  $k_{cat}$  in  $\text{min}^{-1}$ , the turnover number  $k_{cat}/K_M$  in  $\text{min}^{-1}\mu\text{M}^{-1}$  and phytase units FTU in  $\text{U}\cdot\text{mg}^{-1}$ .

Protein	E (nM)	Parameters	Values	St. Error	t-value	Pr(> t )
p15	50	$V_{max}$	192	6.33	30.39	8.43E-08
		$K_M$	219	22.38		
		$k_{cat}$	9613	0.13		
		$k_{cat}/K_M$	44	4.47		
		FTU	205	6.76		
p8	20	$V_{max}$	135	4.66	28.97	1.12E-07
		$K_M$	86	12.34		
		$k_{cat}$	2699	0.23		
		$k_{cat}/K_M$	32	4.54		
		FTU	57	1.96		
<i>EcAppA</i>	7.5	$V_{max}$	77	3.16	24.25	3.23E-07
		$K_M$	161	2.85		
		$k_{cat}$	10209	0.02		
		$k_{cat}/K_M$	64	9.05		
		FTU	228	9.40		

### 6.6.3. Inhibition by inositol hexasulfate (IHS)

Reactions of 50  $\mu$ L were set up in triplicate at fixed concentrations of enzymes (p8 and p15: 100 nM) and substrate (IP<sub>6</sub>: 1 mM). Serial dilutions of IHS at concentrations: 0 nM, 1 nM, 10 nM, 100 nM, 1  $\mu$ M, 10  $\mu$ M, 100  $\mu$ M, 1 mM were mixed to multiple buffered solutions centred to the buffer optimum of pH. The buffer tested for the enzyme p8 were: pH 4.5 - 0.2 M NaAcetate, 0.15 M NaCl; pH 6.0 - 0.2 M MES, 0.5 M NaCl (pH optimum) and pH 8.0; 0.2 M Bicine, 0.15 M NaCl. Instead, the buffer tested for the enzyme p15 were: pH 3.5 - 0.2 M Glycine-Cl, 0.15 M NaCl; pH 5.5 - 0.2 M NaAcetate, 0.15 M NaCl (pH optimum) and pH 7.0 - 0.2 M MES, 0.15 M NaCl. Reactions are incubated at 37 °C for 30 min before inactivation by addition of molybdenum blue reagent. Abs<sub>700</sub> was measured after 30 min incubation with the stopping solution.

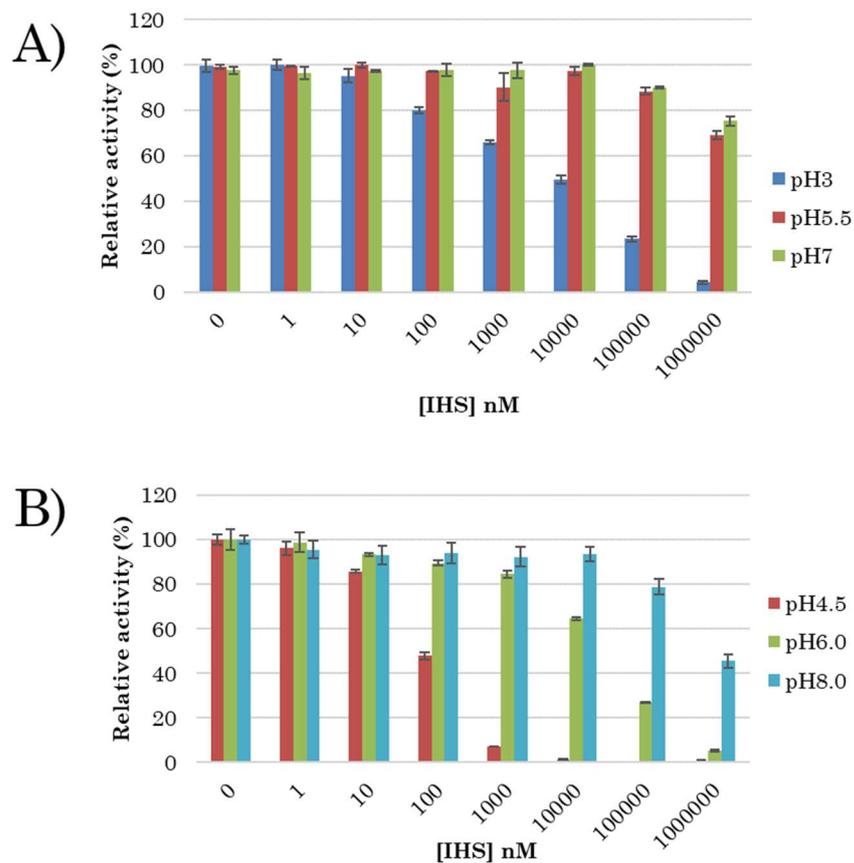
Both the enzymes showed to be strongly inhibited at low pH; specifically, the inhibitor concentration able to decrease initial velocity by 50 %, the I<sub>50</sub>, is equal to 10  $\mu$ M for p15 and 0.1  $\mu$ M for p8: the latter has an I<sub>50</sub> 100x smaller than p15 (**Figure 6.6.3.1**). The inhibition constant  $K_i$  can be calculated by using the equation [173]:

$$K_i = \frac{I}{\left(\frac{S}{K_M} + 1\right)\left(\frac{V_0}{V_i} - 1\right)}$$

where I is the inhibitor concentration, S is the substrate concentration,  $K_M$  is the Michaelis constant,  $V_0$  is the rate of non-inhibited reaction,  $V_i$  is the rate of inhibited reaction. This equation is simplified by knowing the I<sub>50</sub> because for this inhibitor concentration  $V_0$  is double  $V_i$ :

$$K_i = \frac{I_{50}}{\frac{S}{K_M} + 1}$$

Applying this equation,  $K_i$  values were calculated for p8 and p15. They are respectively:  $7.9 \pm 1.1$  nM for p8 at pH 4.5 and  $1.8 \pm 0.2$   $\mu$ M for p15 at pH 3.5. Following these results, crystals soaks were carried out at low pH for both the enzymes. A 30 mM IHS solution was dissolved at 42 °C in 0.6 M Glycine-Cl pH 3.5 to be diluted in crystallization soaks.



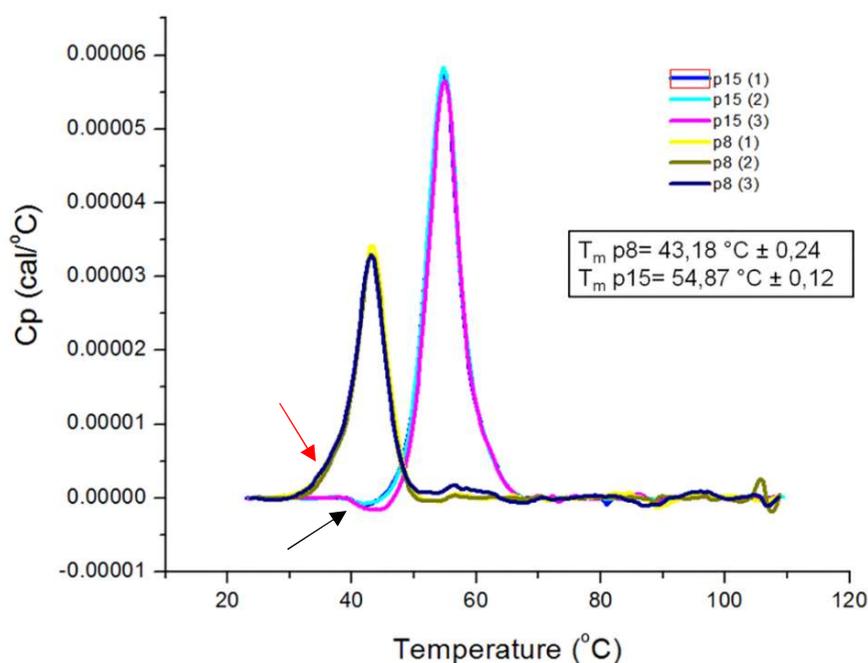
**Figure 6.6.3.1. Inhibition of the enzymes A) p15 and B) p8 by IHS.**

x axis: concentration of inhibitor, IHS in nM, y axis: % activity. Buffers colour scheme: blue - pH 3.0, 0.2 M Glycine-Cl, 0.15 M NaCl, red - pH 4.5 and 5.5, 0.2 M NaAcetate, 0.15 M NaCl, green - pH 6.0 and 7.0, 0.2 M MES 0.15 M NaCl, light-blue - pH 8.0, 0.2 M Bicine, 0.15 M NaCl. Buffer only Abs<sub>700</sub> were subtracted from raw samples Abs<sub>700</sub>, then, the latter were averaged. The % of averaged samples Abs<sub>700</sub> to the maximum average were calculated and plotted above with the respective deviations. In A) the inhibition profile of the enzyme p15: I<sub>50</sub> is equal to 10 μM at pH 3.0. In B) the inhibition profile of the enzyme p8: I<sub>50</sub> is equal to 0.1 μM at pH 3.0.

#### 6.6.4. Differential scanning calorimetry

Differential scanning calorimetry is a technique which enables understanding of folding and unfolding mechanisms of proteins in light of thermodynamics. It is particularly useful for the comparison of melting temperature between enzymes or their mutants aimed to improve stability. It provides answers to the interrogative of whether or not unfolding is a simple one-step process. If not, it may suggest which domains are the ones for which unfolding is favoured. More speculations can be drawn, but in this work, DSC is used with the sole purpose of collecting information on the melting temperature and on the shape of the peak to be able to draw hypothesis on the mechanism of unfolding. For additional details regarding the experimental procedure see **Chapter 2.7.4.**

The sources of these enzymes are mesophilic bacteria. p8 comes from metagenomic data collected on 396 samples of human stool, highly co-varying gene of unknown organisms were grouped in "metagenomic species" and *Prevotella* CAG:617, the predicted organism containing p8, was among them. The source of p15, *Amycolatopsis jejuensis*, was collected with bat guano, taken from a cave on Jeju Island, Republic of Korea, in October 2002, and isolates were able to grow at temperatures ranging from 10 to 30 °C [174]. From this information we can speculate that these enzymes may not be highly thermostable. DSC experiments estimated a  $T_m$  of 43 °C for p8 and of 55 °C for p15 (**Figure 6.6.4.1**). Re-run of the samples showed that unfolding is not reversible. The peak height for p8 is smaller because of a lower concentration of enzyme was used in the experiment (0.4 mg/mL vs 0.5 mg/mL). The peak of p8 is not symmetrical, as it would be in the case of a one-step reaction. It has a shoulder between 30° and 40° C (**Figure 6.6.4.1, red arrow**). This suggests that unfolding could be a two-step process, where a region of the protein unfolds before the rest. The peak of p15 is almost symmetrical, and unfolding can be approximated as a one-step mechanism, though it is possible to notice that the initial stage of unfolding could be exothermic, with negative values from 40 to 46 °C (**Figure 6.6.4.1, black arrow**).



**Figure 6.6.4.1. DSC experiment on p8 and p15.**

x axis: temperature (°C), y axis: heat capacity (cal/°C). On the top-left of the graph, legend describes the colour associated to each sample. Red arrow: p8 shoulder. Black arrow: possible exothermic initial stage of unfolding. Temperature ramp: 10-110 °C, scans rate: 200 °C/h. Baseline buffer was subtracted from raw data to generate the graph above. p8)  $T_m$  onset is 31.79 °C,  $T_m$  is 43.18 °C. p15)  $T_m$  onset is 46.15 °C,  $T_m$  is 54.87 °C.

## 6.7. X-ray crystal structures determination

X-ray crystallography is a powerful tool to determine a model of a protein three-dimensional structure in a spatially restricted environment when macromolecules are ordinally arranged in crystals. Macromolecules though are not static and therefore conformational changes can often occur, or, sometimes, intrinsic flexibility can turn crystallization into a difficult quest, because crystals with broken or multiple lattices won't give a constructive diffraction.

The process of crystallization of *P617*MINPP (p8) and *Aj*MINPP (p15) is a good example to show how two proteins of the same family, with low sequence identity in comparison with other MINPPs of known structure, can have different behaviours in crystallisation. p15 proved to be very easy to handle, it crystallized in the same space group in multiple conditions giving high-quality diffraction datasets that allowed structure solution through molecular replacement even with only 35% of sequence identity with *B. longum* MINPP and 26% with *B. thetaiotaomicron*. On the other hand, p8, has 28% sequence identity with *B. longum* MINPP, 25% with *B. thetaiotaomicron* MINPP, 23% with *A. jejuensis* MINPP, and even though formed very nice looking crystals, did not diffract initially beyond 3 Å. Space groups varied depending on crystallisation conditions. The structure of p8 was solved only in complex with inositol hexasulphate which may have helped in stabilising the protein conformation, making it closer to the model. p15 was solved as apo-enzyme, in complex with Pi and in complex with IHS.

The aim of this experiment was to collect data of the complexes E:IHS of this two enzymes with diverse IP<sub>6</sub> positional attack, p8 and p15, but also apo structures to test (1) the presence/absence of conformational changes upon binding of the substrate analogue, (2) the binding mode of the substrate analogue in the active site, (3) the orientations of the amino acids belonging to each phosphate-binding pockets. The solution of the p8:IHS and p5:IHS complexes allowed to model the binding of the substrate, IP<sub>6</sub>, through molecular docking studies and to test *in silico* the effects on enzymes positional stereospecificity of the active site mutation proposed in **Chapter 6.5**.

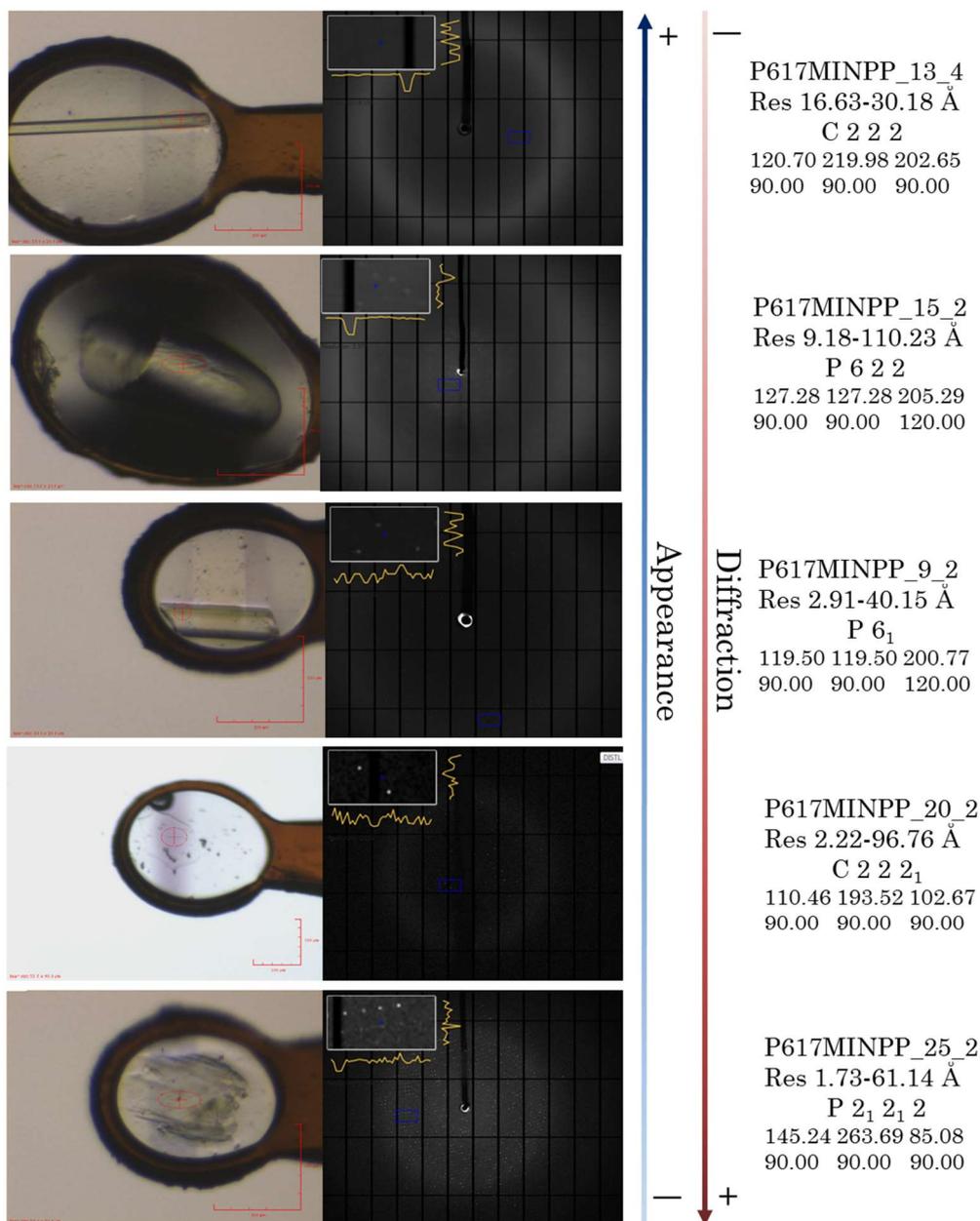
### 6.7.1. Crystallization and X-ray diffraction

Crystallization trials for both proteins were set up on gel filtered solutions >90% pure, pH buffered in 20 mM HEPES pH 7.4, 0.15 M NaCl.

Four sitting drop vapour diffusion crystallisation trials were prepared to obtain one crystal of p8 which diffracted at high resolution producing a good quality dataset. At first, multiple commercial crystallisation screenings were used: Structure Screen Eco 1&2 (Molecular Dimensions), JCSG-plus™ (Molecular Dimensions), MIDASplus™ (Molecular Dimensions) and PACT premier™ (Molecular Dimensions).

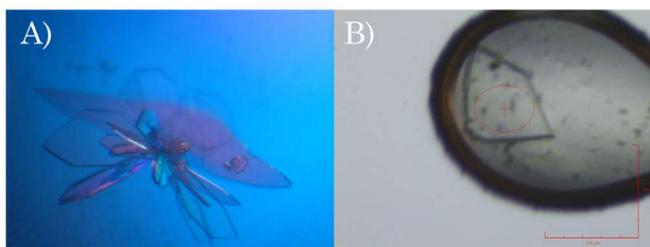
Rod-shape crystals grew in high salt conditions (1.0 M lithium sulphate, 0.5 M ammonium sulphate, 0.1 M sodium citrate pH 5.6 or 0.01 M Cobalt(II) chloride hexahydrate, 1.8 M Ammonium sulphate, MES pH 6.5), however they did not diffract beyond 3.2 Å (**Figure 6.7.1.1**, P617MINPP\_9\_2 and P617MINPP\_13\_4). Poor quality data-sets characterized by moderate to severe anisotropy and high solvent content (equal or greater than 70%) were obtained and did not allow to find a phase solution by molecular replacement. The diffraction power of the crystals was favoured by soaking with the substrate analogue IHS. This allowed to collect a data-set, P617MINPP\_9\_2, (**Figure 6.7.1.1**) at 2.9 Å, for which, a partial phase solution was found. Microcrystal of “rice” shape grew from a second trial in 25% v/v ethylene glycol or 2.0 M NaCl 10% w/v PEG 6000 and were used as seeds. Big “rice” shape crystals grew from seeded plates and were soaked with NaI or NaBr with the aim of searching for a phase solution by SIRAS, however, their resolution was very poor (e.g. **Figure 6.7.1.1**, P617MINPP\_15\_2) and these experiments were not successful. No acceptable results were obtained in absence of the inhibitor, in fact, co-crystallisation with the substrate analogue prove to be the solution. Two data-sets with improved diffraction statistics were collected for the enzyme:IHS complex: P617MINPP\_20\_2 and P617MINPP\_25\_2 (**Figure 6.7.1.1**). The latter, in particular, was the data-sets which allowed to find a phase by MR. The diffracted crystal grew from 8 mg/mL protein solution diluted 1/10 with IHS 30 mM in a precipitant solution made of 0.1 M NaCl, 30% (w/v) polypropylene glycol (PPG) 400, which also acted as cryoprotectant.

Crystallisation trials for p15 were, in contrast, very short. 96-well plates were set up using a protein solution concentrated to 10.7 mg/mL diluted in half with Structure Screen Eco 1&2 (Molecular Dimensions) or JCSG-plus™ (Molecular Dimensions) precipitant solutions. Good quality diffracting crystals were harvested from plate clusters (**Figure 6.7.1.2**) incubated 3 days at 16 °C. They grew in 0.1M NaCitrate pH 5.5, PEG 3000 (20% w/v) and were cryoprotected with 25% glycerol (v/v) in well solution.



**Figure 6.7.1.1. p8 crystal mounted on LithoLoops.**

The pictures were captured on beamline during data-collection. Five crystals and their respective X-ray diffraction sample images are presented. Resolution and unit cell parameters are reported. Funnily, crystals morphology appears inversely proportional to the quality of resolution.



**Figure 6.7.1.2. p15 protein crystals.**

**A)** Plate cluster captured in-drop prior harvesting. **B)** LithoLoop on which a plate fragment is mounted and flash-cooled. The image was captured on beamline before data-collection.

### 6.7.2. X-ray crystal structures solution of p8 and p15

A phase solution was found for p8 (from *Prevotella* sp. CAG:617) by molecular replacement with the use of Phaser-MR [81] through an iterative procedure which involved the analysis of multiple data-sets. The first preliminary results were obtained for the dataset P617MINPP\_20\_2 (**Figure 15**) of space group  $C 2 2 2_1$ , 57.7% solvent content, resolution 2.22 Å and 2 predicted monomers per ASU. A first molecule was placed using Chain A of the *Bt*MINPP X-ray crystal structure (PDB id: 4tsr, 25.31% identical), carefully edited (pruned side-chains to last common atom [71], removal of particularly flexible regions) as a model. However, Phaser-MR struggled to find the correct packing for the second monomer, probably because of the absence, in *Bt*MINPP, of an N-terminal loop region which lies at the interface between the surfaces of p8 monomers. To solve the problem a model of P617MINPP was prepared using SWISS-MODEL [175] (QMean -5.40), superposed to the partial solution, deprived only of a small region of the N-terminal loop and used again in phasing. A solution with 2 monomers per asymmetric unit was found (TOP LLG: 53.152, TOP TFZ: 9.0) in which the  $\alpha/\beta$ -domain had a good density coverage, unlike the  $\alpha$ -domain. This solution was refined using a variety of methods without any improvement (R-work 49.11%, R-free: 51.47%). However, the model that was obtained was given as an input to the DIMPLE pipeline in further data-collections. DIMPLE found a phase solution for a crystal of orthorhombic space group  $P 2_1 2_1 2$  (**Figure 6.7.1.1**, P617MINPP\_25\_2), characterised by 57.6% solvent content with 6 monomers per asymmetric unit, all complexed with the substrate analogue IHS. This solution allowed to build a complete model for the N-terminal region and to use it for subsequent phasing. The new phase solution required extensive manual rebuilding which was coupled with model morphing [176] and simulated annealing refinements using phenix.refine [90]. NCS constraints and secondary structure restraints were applied at initial stages of refinement. Data collection and refinement statistics for this structure can be found in **Table 6.7.2.1**. A ligand omit-map is presented in **Figure 6.7.2.1**, electron density coverage of the ligand is comparable in each monomer. IHS is bound to p8 directing its 6-phosphate group of the inositol ring towards the catalytic histidine, His67. A phase solution was not found for other datasets even by using the solved structure as a model.

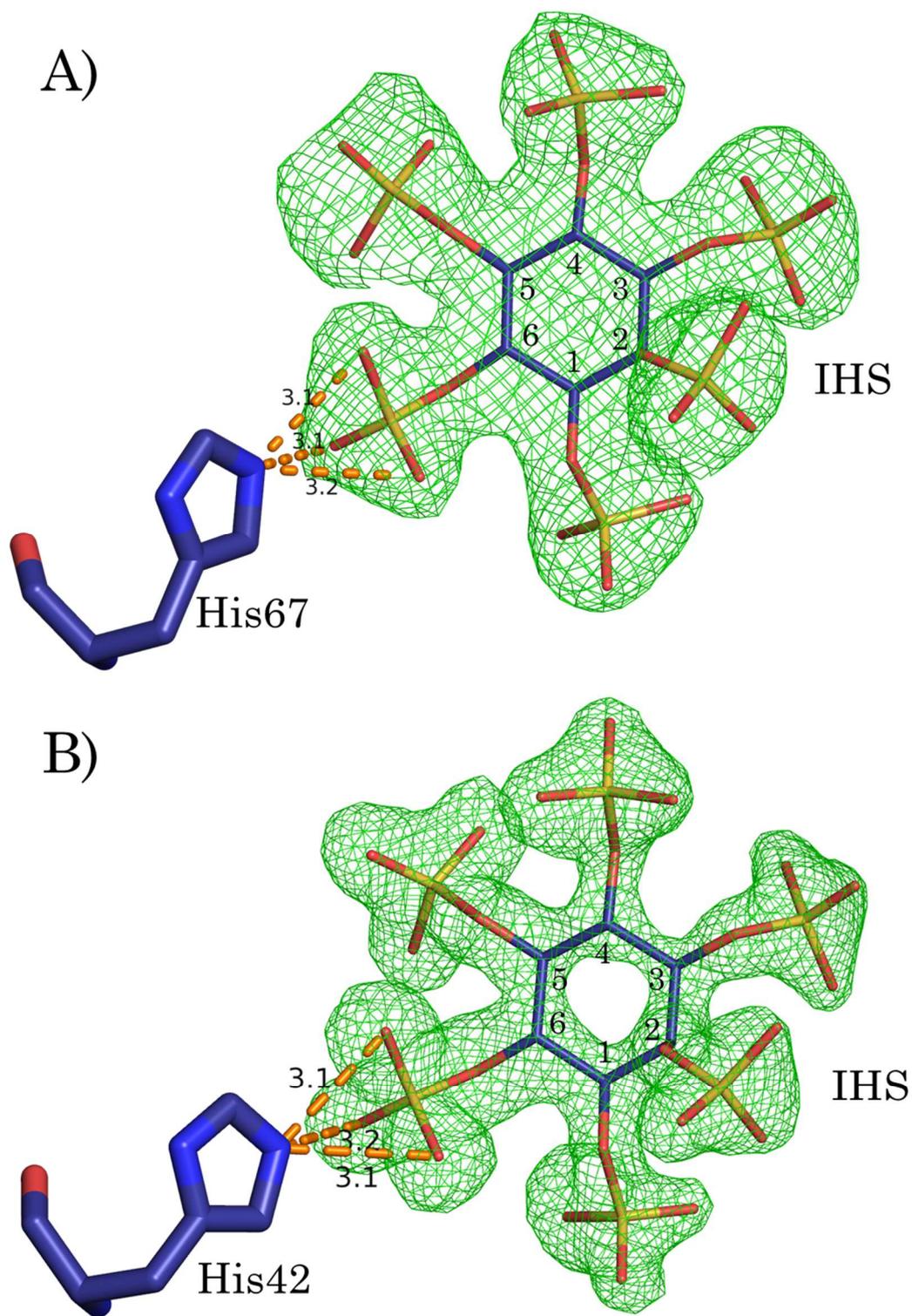
The macromolecular X-ray crystal structure solution of p15 was obtained for a crystal of monoclinic space group  $P 2_1$ , characterised by 44.9% solvent content, with 2 monomers per asymmetric unit in complex with Pi. The model used in MR was built on *Bt*MINPP (37.4% identity) sequence alignments and truncated by Chainsaw [71]

to the last common atom. Resolution was cut to 3 Å in phasing using Phaser-MR. Using these parameters the procedure provided an acceptable solution which was subsequently manually remodelled using WinCoot [87] and refined with phenix.refine [89]. Data collection refinement statistics are given in the **Table 6.7.2.1.** Inorganic phosphate was bound to the active site cleft of the two monomers present in the ASU. An apo structure of the enzyme as well as an enzyme:IHS complex were also solved. The ligand .cif file used in refinement was obtained from the WinCoot dictionary. A ligand omit-map is presented in **Figure 6.7.2.1.** The inhibitor is bound with the 6-phosphate of the inositol ring facing the catalytic histidine, His42.

**Table 6.7.2.1. Data collection and refinement statistics.**

Information describes the structures of p15 apo, in complex with Pi. Statistics of the complex p8:IHS are also reported.

Protein	p15 apo	p15:Pi	p15:IHS	p8:IHS
<b>Data Collection</b>				
Space group	P2 <sub>1</sub>	P2 <sub>1</sub>	P2 <sub>1</sub>	P 2 <sub>1</sub> 2 <sub>1</sub> 2
a , b , c (Å)	68.85, 87.14, 70.36	68.89, 86.91, 70.54	67.84, 86.42, 70.11	145.14, 263.63, 85.03
α, β, γ (°)	90.0, 97.4, 90.0	90.0, 97.4, 90.0	90.0, 98.29, 90.0	90.0, 90.0, 90.0
Resolution limit (Å)	68.3 - 2.05	69.87-1.66	32.08 - 1.41	73.37- 1.97
(I)/sd(I)	1.34	1.33	1.34	1.34
Completeness (%)	99.0	99.64	99.05	99.92
<b>Refinement Statistics</b>				
Protein monomers per asymmetric unit	2	2	2	6
Total atoms	7306	7309	6662	21869
Water molecules	740	758	826	1818
R work	22.33	21.13	16.99	19.99
R free	26.39	23.49	19.52	22.90
<u>Ramachandran Analysis (%)</u> :				
Most favoured	99.15	98.66	98.42	95.7
Favoured	0.85	1.44	1.68	0.00
Outliers	0.00	0.00	0.00	0.6
<u>RMS deviations</u> :				
Bonds (Å)	0.005	0.004	0.009	0.005
Angles (°)	0.66	0.72	1.06	0.8
Planes (Å)	0.005	0.005	0.008	0.005
<u>Mean Atomic B-value (Å<sup>2</sup>)</u>	19.87	19.20	21.46	16.99
Macromolecule	19.27	18.10	22.03	15.16
Ligands				15.85
Solvent	25.18	28.82	32.71	26.95
<u>Occupancies (%)</u>				
IHS - chain A, B, C, D, E, F			75-100, 75-100	61-100, 64-100, 60-100, 72-100, 44-100, 60-100



**Figure 6.7.2.1. Omit single-difference Fourier map of the substrate analogue IHS.**

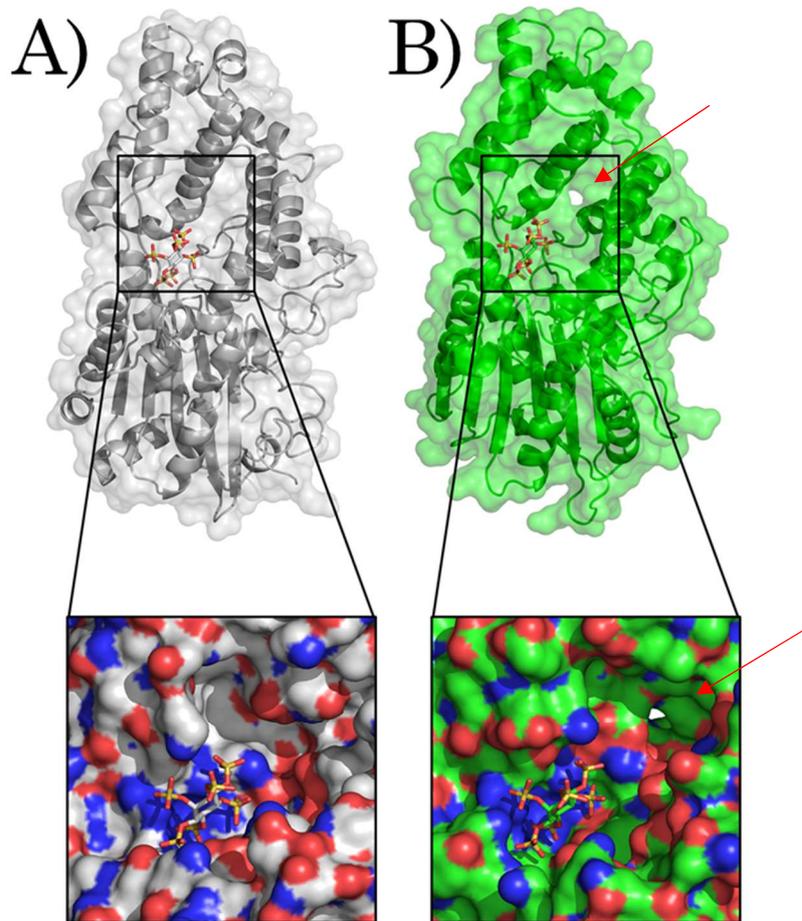
Contour level is set to  $\sigma=3.0$  r.m.s.d. **A)** Fo-Fc omit map of the density surrounding IHS in the structure of the complex p8:IHS and the corresponding models of the substrate analogue. **B)** Fo-Fc omit map of the density surrounding IHS in the structure of the complex p15:IHS and the corresponding models of the substrate analogue. The catalytic histidine makes contacts to the sulphate group bound to C6 of the inositol ring in both the structures.

## 6.8. Analysis of the crystal structures

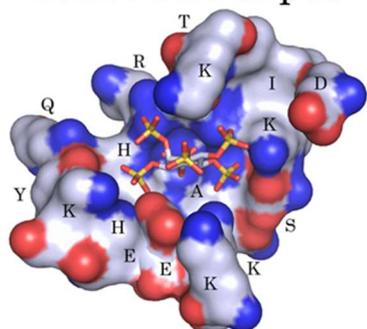
The structures of p8 and p15 are characterized by a classical HP conformation made of two domains, an  $\alpha$ -domain and  $\alpha/\beta$ -domain with the active site located in the cleft between the two (**Figure 6.8.1 – A, B**). p8, unlike other MINPPs, present a water tunnel just above the binding site (**Figure 6.8.1. – B**), which was also found in HP2 that are not phytases (**Figure 5.3.1.1.**). However, the latter enzymes have a much smaller active site in comparison with p8, and the tunnel opens just above the catalytic core residues (“RHGxR”, “R”, “HAE”). The reason why a water tunnel is present in some HP2 has not been investigated. A water tunnel may contribute to the regeneration of the proton donor, to the release of phosphate from the phosphoenzyme intermediate and the donation to an acceptor molecule while the substrate is present in the active site. The feature may be involved in product inhibition or even in the tuning of the pH optimum for enzyme activity. Particularly, it would be interesting to analyse this feature in combination with eventual conformational changes. An interesting program to use would be AQUA-DUCT [177], which trace and visualize molecular trajectories throughout MD simulations.

The architecture of the catalytic cleft of p8 is instead phytase-like. The enzyme retains the N-terminal lid found in MINPPs and fungal phytases, however the feature leaves enough space to allow the passage of water. Unfortunately, it was not possible to solve structures for apo-p8 and its complex with inorganic phosphate and so the question of whether a conformational change occurs during the catalytic cycle in this enzyme remain opens. Instead, from comparisons of the structures of apo, enzyme:Pi and enzyme:IHS complexes, no conformational change is observed in p15. Analysis of the active site cleft of the two enzymes reveals a conservation of the three-dimensional orientation of the catalytic core residues (“RHGxR”, “R”, “HAE”), while the arrangement of the remaining amino acids is protein-specific, possibly because these residues are involved in tuning the substrate binding mode.

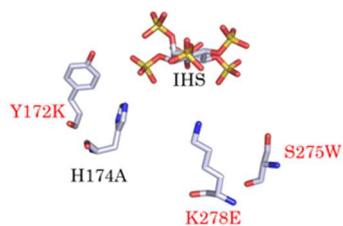
In Chapters 3 and 4, the possible importance of residues downstream of the “RHGxR” motif, namely a “PTK” triplet, in the binding of substrate was highlighted. Among the 15 MINPPs selected, enzymes in which these residues are absent were also chosen, like p8, to test the role of the lysine in the region. An analysis of the structures of p8 (“LTD”), p15 (“PTK”), *Bt*MINPP (“LSS”), *Bt*MINPP (“PTS”), *Ec*AppA (“PTK”) and *An*PhyA (“PTD”) showed that, despite the lysine is not conserved in the amino acids sequences, all the enzymes 3D structures carry a Lys in the area, within 4.0 Å from the substrate analogue and within 2/3 residues positions from the motif.



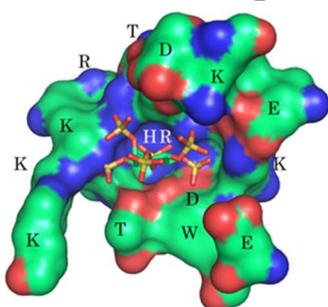
Active site of p15 (a)



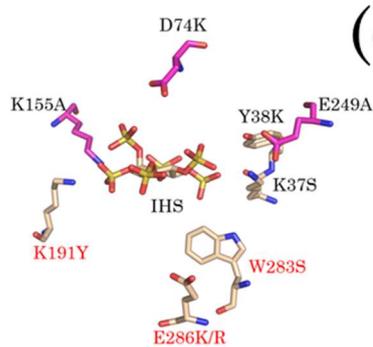
(b)



Active site of p8



(c)



(d)

Figure 6.8.1. X-ray crystal structures of p8 and p15.

**Figure 6.8.1. X-ray crystal structures of p8 and p15.**

A) Overview of the structure of the p15:IHS complex as a whole and close-up of the active site. B) Overview of the structure of the p8:IHS complex as a whole and close-up of the active site. To note the presence of a water tunnel (red arrow). (a) Surface representation of the active site of p15 in complex with IHS. The names of the residues are annotated. (c) Surface representation of the active site of p8 in complex with IHS. The names of the residues are annotated. (b, d) Residues identified as potential determinants of positional stereo-specificity are represented by sticks. *in silico* mutations tested in AutoDock Vina are annotated. Mutations coloured in red are the ones identified as affecting positional stereo-specificity by docking studies. p8 presents a crowded active site in comparison with p15. Violet coloured residues do not have a counterpart in p15.

It may be possible for this residue a role in determining substrate specificity and/or enzyme binding affinity (e.g. larger substrates could be recruited by a distal lysine or enzyme with a distal lysine could be characterized by a lower binding affinity). In p15, Lys67 coordinates the binding of the sulphate groups of IHS in the binding pockets D and E (interaction distances 3.6 Å and 3.7 Å, respectively), while in p8 Lys77 interacts with the sulphate in binding pocket C (3.4 Å). The binding pockets follow the same nomenclature as described for *B/MINPP* in **Figure 5.5.1.1**. It would be interesting to mutate this lysine residue in p8 and p15 to introduce it at different distances (in 3D space) from the conserved Thr/Ser in order to test the effect on binding affinity, substrate specificity and IP<sub>6</sub> positional stereospecificity.

A general look at the active clefts of these two enzymes reveals big differences in charge (**Figure 6.8.1 – a, c**). The active site of p15 is predominantly negatively charged, exception made for the proton donor Ser11, Tyr172 and Glu322. Five lysine residues (anticlockwise: Lys175 (F motif), Lys 278 and Lys279 (K motif), Lys12 (Y motif), Lys49 (of the “PTK” triplet)) contribute to orient the substrate. In fact, each sulphate group of IHS lies within 3.7 Å from a lysine, except for the sulphate in pocket A, which is coordinated by the catalytic histidine (His42) and the sulphate in pocket F, which is coordinated in the same manner by His320 (H motif) and His174 (F motif). Instead, in the active site of p8 is possible to recognise a negatively charged front made of Lys193, Lys191 (F motif) and Lys155 (R motif) and an electron positive front created by the proton donor motif “HDT” behind which lies a tryptophan (Trp293) which decrease the size of pocket B by steric effect (absent in p15). It is interesting to note that the benzene ring of Trp in p8 is perfectly superimposable by the benzene ring of phenylalanine (Phe254) in *EcAppA*, both positional stereospecific enzymes. In p8, like p15, the Lys residues coordinate sulphate groups on IHS. All sulphates lie within 3.4 Å from a lysine except the groups in pocket A, facing the catalytic histidine His67, and pocket C. These closer coordination of the substrate analogue may reflect

a tighter binding of IP<sub>6</sub> to p8 than to p15, in line with the kinetics data collected for the two enzymes (**Chapter 6.6.2** - p8  $K_M = 86 \mu\text{M}$ , p15  $K_M = 219 \mu\text{M}$ ).

Overall, the main differences between the two active sites seems to be correlated to: (1) changing in the size of the binding pockets (e.g. the bulky Trp reducing the size of pocket A and B in p8, addition of the “KK” in p15 (K motif)), (2) changing in the polarity of the pockets (e.g. “KK” pair of p15 or Y in p15 vs K in p8 of the F motif), (3) active site symmetry (e.g. in p15 amino acids are distributed in negatively and positively charged pairs orienting each sulphate – symmetrical active site, in p8 it is possible to identify two larger negatively and positively charged fronts plus a bulky Trp – asymmetrical active site).

## 6.9. Docking studies

From the considerations made on active site residues conservation (**Chapter 6.5**) and on active site architecture in (**Chapter 6.8**), the residues predicted to be involved in shaping positional stereospecificity in the enzyme p8 were K37, Y38, K191, W283 and E286. A potential role was not excluded also for Y38, D74, K77, K155 and E249. The reciprocal positions were identified in p15 as Y172, H174, S275 and K278 (**Figure 17.b,d**). These residues were reciprocally swapped in the two enzymes by *in-silico* mutations and the binding affinity of the substrate was studied by molecular docking in Autodock-Vina [92]. Mutations were made in Coot [87], rotamers were chosen and manually oriented on the basis of steric hindrance, trying to model the contacts with the substrate observed in the reciprocal enzyme. Docking was carried out as described in **Chapter 2.9**.

This approach may not model the real effect of a mutation in the destination enzyme because of inaccuracy of positioning which would need optimisation e.g. by MD simulations or by leaving the residues flexible during docking. In fact, it must be considered that enzymes aren't static molecules. However, this method simulates how IP<sub>6</sub> would interact to an enzyme if a residue happens to be placed in the exact position as it is found in the crystal snapshot of the reciprocal enzyme. This do not give indications on the results of a specific mutation in the enzyme at test, but this may highlight in the reciprocal enzyme links between the presence of the residues and the level of promiscuity. This choice was also made considering that often MD simulations or docking with flexible active site residues introduce large errors in modelling when parameters are not accurately tuned. Because of time constraints the simplest

approach was chosen and applied as accurately as possible with the aim of testing mutations *in vitro* at the earliest.

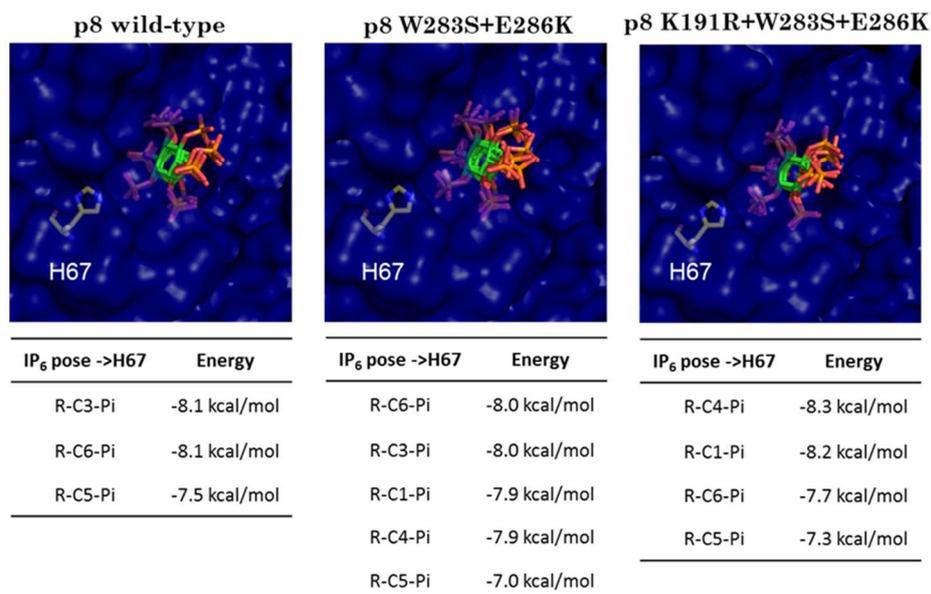
In Autodock-Vina the calculated binding affinity for each ligand pose is described by a free energy of binding (kcal/mol) which can be used to predict preferred binding modes. Also, a distance from each pose to the orientation which minimize free energy is given. Only poses within 1 kcal/mol from the binding mode representing the energy minimum were accepted. In addition, only those poses for which the phosphate group pointing towards the catalytic histidine is superimposable with the corresponding sulphate of IHS in the X-ray macromolecular crystal structure of the enzyme:IHS complex are considered. Docking trials of *in silico* mutations are reported in **Figure 6.9.1** for p8 and in **Figure 6.9.2** for p15.

Residues that could be involved in determining the positional stereospecificity in p8 are predicted to be W283 and E286. In particular, the double mutation W283S + E286K or R led to an increase of the number of the binding poses while maintaining comparable free energies of binding affinity to the wild-type enzyme. This may be a result of the absence of the steric effect produced by Trp and the formation of a new coordination of the Pi group in Pockets B and C by the introduced lysine. The E286R mutation significantly increases the number of binding poses, directing even the axial phosphate towards the catalytic histidine. However, the simulation binding affinity also increases. The mutation K37S led to a decreased binding affinity *in silico*, possibly due to the loss of phosphate group coordination. Mutations K191Y and E286K led instead to an increase in binding affinity in the simulation. *In vitro*, an increase in binding affinity could either result in an improved enzyme turnover number as consequence of an improved  $k_{cat}$  or to a loss in enzyme processivity if the stronger binding in specific pockets affect negatively substrate orientation of product release. Mutation need to be generated and tested to evaluate these eventual possibilities.

In p15 the double mutation S275W + K278E causes a loss in enzyme positional stereo-specificity. Again, this suggest the possible involvement of these two residues in determining enzyme positional stereo-specificity.

## IP<sub>6</sub> docking trials in p8

Mutation	IP <sub>6</sub> pose ->His67 (Energy)	Mutation	IP <sub>6</sub> pose ->His67 (Energy)
Wild-type	R-C3-Pi (-8.1 kcal/mol) R-C6-Pi (-8.1 kcal/mol) R-C5-Pi (-7.5 kcal/mol)	W283S+E286K	R-C6-Pi (-8.0 kcal/mol) R-C3-Pi (-8.0 kcal/mol) R-C1-Pi (-7.9 kcal/mol) R-C4-Pi (-7.9 kcal/mol)
W283S	R-C6-Pi (-8.0 kcal/mol) R-C1-Pi (-7.9 kcal/mol) R-C5-Pi (-7.5 kcal/mol)	W283S+E286R	R-C6-Pi (-8.1 kcal/mol) R-C1-Pi (-7.9 kcal/mol) R-C3-Pi (-7.8 kcal/mol) R-C2-Pi (-7.7 kcal/mol) R-C5-Pi (-7.2 kcal/mol)
K191Y	R-C3-Pi (-8.1 kcal/mol)	K37S+Y38K+K191Y	R-C6-Pi (-8.0 kcal/mol) R-C3-Pi (-7.8 kcal/mol)
K37S	R-C3-Pi (-7.6 kcal/mol) R-C6-Pi (-7.6 kcal/mol)	K37S+Y38K+W283S	-
E286K	R-C3-Pi (-8.2 kcal/mol) R-C6-Pi (-7.9 kcal/mol)	K37S+Y38K+K191Y+W238S	R-C4-Pi (-8.3 kcal/mol) R-C1-Pi (-7.9 kcal/mol) R-C5-Pi (-7.7 kcal/mol)
E286R	R-C3-Pi (-8.1 kcal/mol) R-C6-Pi (-7.9 kcal/mol)	K191Y+E286K+W283S	R-C6-Pi (-8.0 kcal/mol) R-C1-Pi (-7.9 kcal/mol)
K37S+Y38K	R-C3-Pi (-7.8 kcal/mol) R-C6-Pi (-7.7 kcal/mol) R-C4-Pi (-7.6 kcal/mol)	K191R+E286K+W283S	R-C4-Pi (-8.3 kcal/mol) R-C1-Pi (-8.2 kcal/mol) R-C6-Pi (-7.7 kcal/mol) R-C5-Pi (-7.3 kcal/mol)
K191Y+W283S	R-C3-Pi (-8.0 kcal/mol) R-C1-Pi (-8.0 kcal/mol) R-C6-Pi (-7.6 kcal/mol)	Left to try: D74K, K76A, K155A, E249A	

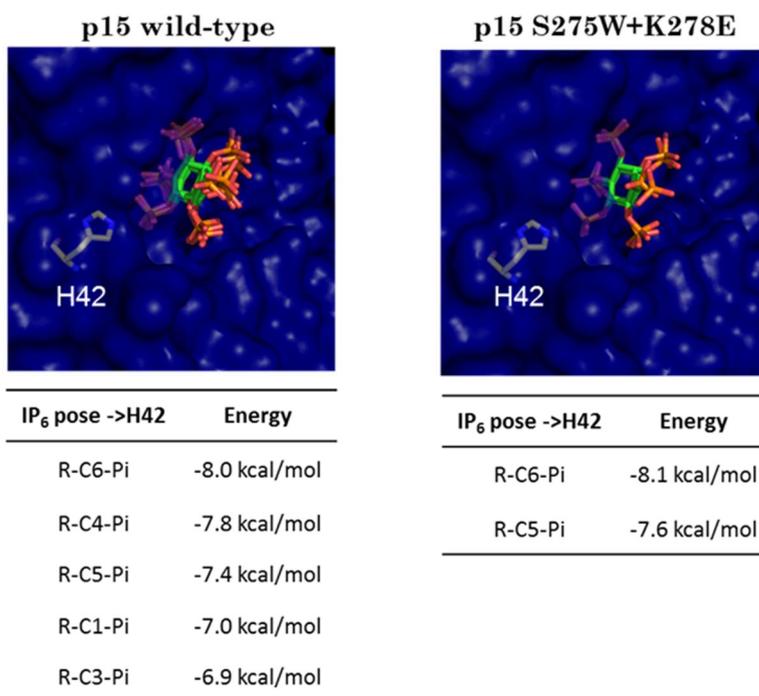


**Figure 6.9.1. Docking studies.**

Autodock-Vina [92] was the software used in the analysis. The table describes the results obtained by docking of IP<sub>6</sub> in the p8 wild type enzyme and mutants. Free energy of binding for each substrate pose in each mutant is reported. Promising mutations are highlighted in orange. Close-ups of the substrate binding modes for the wild-type and two promising mutations are shown. A box underneath the figures specify the free energies of binding calculated for the poses.

## IP<sub>6</sub> docking trials in p15

Mutation	IP <sub>6</sub> pose ->His67 (Energy)	Mutation	IP <sub>6</sub> pose ->His67 (Energy)
Wild-type	R-C6-Pi (-8.0kcal/mol) R-C4-Pi (-7.8 kcal/mol) R-C5-Pi (-7.4 kcal/mol) R-C1-Pi (-7.0 kcal/mol) R-C3-Pi (-6.9 kcal/mol)	K278E	R-C6-Pi (-7.9 kcal/mol) R-C4-Pi (-7.6 kcal/mol) R-C5-Pi (-7.4 kcal/mol)
S275W	R-C6-Pi (-8.0 kcal/mol) R-C5-Pi (-7.4 kcal/mol) R-C4-Pi (-7.2 kcal/mol)	S275W+K278E	R-C6-Pi (-8.1 kcal/mol) R-C5-Pi (-7.6 kcal/mol)
Y172K	R-C6-Pi (-7.7 kcal/mol) R-C5-Pi (-7.2 kcal/mol) R-C4-Pi (-7.0 kcal/mol)	H174A+S275W+K278E	R-C6-Pi (-7.9 kcal/mol) R-C5-Pi (-7.3 kcal/mol) R-C4-Pi (-7.2 kcal/mol)
H174A	R-C6-Pi (-7.7 kcal/mol) R-C4-Pi (-7.2 kcal/mol) R-C5-Pi (-7.2 kcal/mol) R-C1-Pi (-6.9 kcal/mol)		



**Figure 6.9.2. Docking studies.**

Autodock-Vina [92] was the software used in the analysis. The table describes the results obtained by docking of IP<sub>6</sub> in the p15 wild-type enzyme and mutants. Free energy of binding for each substrate pose in each mutant is reported. Promising mutations are highlighted in orange. Close-ups of the substrate binding modes for the wild-type and a promising mutation are shown. A box underneath the figures specify the free energies of binding calculated for the poses.

## 6.10. Discussion

Multiple inositol polyphosphate phosphatases were initially named so because of their broader substrate specificity in comparison with *A. niger* HP2 phytase [36]. However, MINPPs are subjected to enzyme-dependent levels of promiscuity. This is particularly true when evaluating the degree of positional specificity in IP<sub>6</sub> hydrolysis. For example, lily pollen phytase hydrolyse IP<sub>6</sub> to produces one IP<sub>5</sub> intermediate of reaction, 5-OH IP<sub>5</sub> [23], while *Bt*MINPP [40, 42] or *Bt*MINPP [40, 41] produce multiple IP<sub>5</sub> specie. The profile of IP<sub>5</sub> intermediates generated by 15 MINPPs revealed a palette of profiles characterized by different degrees of positional stereospecificity of IP<sub>6</sub> hydrolysis. Again, a set of highly positional specific (PS) MINPPs has been identified, mainly producing 4/6-OH IP<sub>5</sub> (p5, p8 and p10), but also non-positional stereospecific (NPS) MINPPs were found, such as p9 and p16. The evidence challenges the group nomenclature of Multiple Inositol Polyphosphate Phosphatases, which actually may highlight a peculiarity that is not spread along the whole group. MINPPs also includes proton-donor less but active enzymes like p1 and p2, as it is the case, for example, of *Drosophila melanogaster* MINPPs [47]. Posing the question of how they overcome product release. Are other active site residues able to act as proton donor? Could the solvent be involved?

To study active site differences between PS and NPS enzymes their sequences were grouped by level of promiscuity (5 groups) and compared. A number of residues were identified as potential determinants for positional stereospecificity, the majority of which lying in the conserved motifs Y, F, K or H. To explore the active site architectures of the groups the X-ray crystal structures of a PS enzyme (p8) and an NPS enzyme (p15) were solved in complex with the substrate analogue IHS. The main differences in positional stereospecificity between these enzymes appear to be related to changes in the size and charge of the amino acids of their binding pockets to produce active sites that I called “symmetrical” (in NPS enzyme) or “asymmetrical” (in PS enzyme). This is particularly relevant because of the shape and highly polar nature of the substrate. In fact, IP<sub>6</sub> is a meso compound characterized by a plane of symmetry. Moreover, its functional groups are negatively charged at the pH ranges of HP2 activity [178] and need to be solvated upon binding to a phytase. The symmetry of the active site charges and steric effect could therefore have a determining role in driving IP<sub>6</sub> to the desired position.

Despite using a different approach from the one described in this thesis, Konstrewa et al (1999) reached similar conclusion on the involvement of local

electrostatic field in determining differences in pH optimum and substrate specificity between two fungal phytases [179]. In their article, the first separation of the active site in a catalytic centre and a substrate specificity site was made. Structural and catalytic differences of the promiscuous *A. niger* pH 2.5 acid phosphatase (AnigAP) and the phytate-specific *A. niger* phytase (AnigPhyt) were discussed. It was suggested that a more neutral active site specificity site would be less selective for which phosphomonoester to bind while a highly positively charged active site would be optimised for the binding of the highly negatively charged phytate [179].

On the basis of these observations, residues identified as potential stereospecificity hotspots were mutagenized *in silico* and active sites were docked using AutoDock-Vina [92] with IP<sub>6</sub> to test their effects on binding affinity. The most promising results were obtained for the double mutants W283S + E283K/R and S275W + K278E, in p8 and p15, respectively. E283 and K278 lie in motif K, identified before as possible hotspots in determining catalytic promiscuity (**Chapter 6.5**), while the tryptophan residue is present only in highly stereoselective enzymes. These reciprocal mutations appear to alter *in silico* positional stereospecificity of IP<sub>6</sub> hydrolysis, apparently having only a minor effects on binding-affinity *in silico*. The method present limitations, as indicate in **Chapter 6.9**, however, it can be possible that attraction/repulsion forces between the introduced residues and the ligand can be strong enough to cover other effects taking similar residues conformer, in the mutated enzyme and in the reciprocal enzyme, to an energetical minimum. Site-directed mutagenesis of the residues selected in **Chapter 6.5** needs to be carried out in p8 and p15 to confirm or reject this hypothesis. It would be even more interesting to evaluate the effects of saturation mutagenesis on these positions in all the 15 enzymes, to be able to collect results on a protein sample statistically more significant. The final objective is to acquire the ability to tune promiscuity of IP<sub>6</sub> hydrolysis in HP2s of interest depending on commercial needs.

An enzyme ready to be commercialised fulfils multiple requirements: resilience to pelleting (thermostability), resistance to proteases implied in food digestion, optimal activity at the pH of the host's gut, etc. For example, altering promiscuity comes often to the expense of catalytic activity (e.g. *EcAppA* mutants). Therefore, enzymes need to be iteratively modified until an optimal aminoacidic combination is found. The process is costly and time-consuming. For this reason, a potential candidate enzyme must own properties that show it is worth the effort of the engineering and profit can actually be envisioned. MINPPs characterised in this thesis unfortunately proved to be hard to express and purify, with the exception of p8,

p10 and p15. p15 displayed catalytic parameters in line with *EcAppA* wild-type (enzyme whose engineered form is currently sold e.g. as Quantum Blue – ABVista or PhyzymeXP – Danisco), though it did not prove to be highly thermostable. Resistance to proteases was not tested. Overall, these enzymes would need to be highly engineered to be considered commercially useful. However, if this work will prove to contain correct hypothesis on positional specificity determinants, it would be interesting to use this know-how for a further engineering of commercial enzymes. If the desired level of promiscuity is introduced without affecting relevant enzyme characteristics, the mutants would be quickly ready to be sent out on the market.

In conclusion, the objective of this study to highlight residues possibly involved in determining MINPPs positional stereospecificity of IP<sub>6</sub> hydrolysis is accomplished. In the future, a confirmation of this hypothesis would come by saturation mutagenesis of the identified hotspots residues and the characterisation of the resulting mutant. Also, a role is suggested for a lysine residue downstream the RHG motif in determining the size “cut-off” for MINPPs substrate specificity. This work applies generally to HP2 phytases in establishing the basis for future positional stereospecificity engineering in this class of enzymes.

# CHAPTER 7

## 7. Final discussion

This thesis investigates the structural determinants of positional stereospecificity of IP<sub>6</sub> hydrolysis in HP2 phytases to collect information for the future engineering of enzymes capable to sustain complete degradation of IP<sub>6</sub> to inositol.

In **Chapter 3**, four mutations in *Escherichia coli* AppA (D326A, D326E, T327E and D326A/T327E), tested the role of glutamic acid as proton donor versus the wild-type aspartic acid. Positional stereospecificity in IP<sub>6</sub> hydrolysis was altered particularly for the mutants *EcAppA* HAT (proton donor-less), which produces a predominant 1/3-OH IP<sub>5</sub> peak, and *EcAppA* HET that generates 1/3-OH and 4/6-OH IP<sub>5</sub> in equal quantities. However, *EcAppA* HAT and *EcAppA* HET suffered a 1000-fold and 80-fold reduced turnover number, respectively. Inspection of the X-ray crystal structures of the mutants in complex with the substrate analogue IHS seem to suggest that non-positional specificity may be the outcome of differences in shapes and charge of the active site pockets B and F.

**Chapter 4** described the conformational change of *B*/MINPP, the first HP2 phytase seen to undergo an  $\alpha$ -domain movement during catalysis. Questions arise as to why this movement occurs and whether there are other HP2 phytases which follow this mechanism. The identification of the hinge about which motion occurs suggests that domain movement is initiated by the binding of IHS (mimicking IP<sub>6</sub>) to the highly conserved active site residues Arg48 and Arg142. Further investigations are needed to evaluate the role of the U-loop and to determine if domain motion is a common features of HP2 or it is shared only between a small number of MINPPs.

**Chapter 5** and **Chapter 6** described the selection and characterisation of 15 MINPPs representatives. Their profiles of IP<sub>5</sub> intermediates revealed that the degree of positional stereospecificity of IP<sub>6</sub> hydrolysis vary significantly from enzyme to enzyme. X-ray crystal structures of p8 (positional stereospecific (PS) enzyme) and p15 (non-positional stereospecific (NPS) enzyme) were solved in complex with the substrate analogue IHS. Studying the differences between PS and enzymes NPS active sites, a number of residues were identified as possible hotspots for the determination of those traits. The most promising results were obtained for the docking of IP<sub>6</sub> in the double mutants W283S + E283K/R and S275W + K278E, in p8 and p15, respectively. These *in silico* mutation seems to make p8 more promiscuous and p15 more selective. The bulky tryptophan reduces significantly the volume of

pocket B, while acidic or basic residues at the rear of the active site may contribute either to reduce promiscuity by repulsion or to increasing promiscuity by attraction to multiple distributed lysine residues, which may allow the substrate a higher freedom of movement. In fact, the helix possibly involved in tuning stereospecificity displays an EWxEK motif in p8 versus a KSxxKK motif in p15. Saturation mutagenesis of this portion of the active site in the two enzymes is suggested as a tool to evaluate the involvement of these residues in determining enzyme stereospecificity of IP<sub>6</sub> hydrolysis.

Information collected on *EcAppA* and MINPPs studies leads to the formulation of a theoretical hypothesis on the sequence/structure determinants behind positional specificity, which are proposed to be: (1) dimensions of the active site pockets B and F, (2) electrostatic field of the active site pockets B and F, (3) symmetry of the active site (**Chapters 6.9 and 6.10**). These studies also highlighted a possible role for a lysine downstream the RHG motif in substrate recruitment and specificity (**Chapters 3.10, 3.11, 6.8, 6.10**). The reason of the presence/absence of a water tunnel above the active site of HP is yet to be investigated (**Chapters 5.3.1, 6.8**) as well as the role of the N-terminal lid (**Chapters 5.3.2, 5.5.3, 5.6.2, 6.8**).

Results were not as expected. In fact, initial experiments were based on the simple hypothesis that the HAE proton donor triplet could be the motif able to confer promiscuity to MINPPs enzymes, versus the HDx triplet typical of positional stereospecific enzyme. However, experiments proved that (1) promiscuity of IP<sub>6</sub> hydrolysis is not a common trait of MINPPs, (2) MINPPs characterised by a HDx proton donor triplet can be promiscuous towards IP<sub>6</sub>, (3) proton donor-less MINPPs can be promiscuous towards IP<sub>6</sub>, (4) the insertion of an HAE proton donor in *EcAppA* do not widely alter positional stereo-specificity of IP<sub>6</sub> hydrolysis. Results overall showed that the proton donor may not be directly involved in determining positional stereospecificity but mutations on this triplet can affect stereospecificity when able to perturb the local charge distribution and/or the specificity pockets volume.

No previous effort has been directed to expand the stereospecificity of the family. However, my experiments are in line with the suggestions drawn by Konstrewa et al (1999), while comparing the active sites structures of *Aspergillus niger* pH 2.5 acid phosphatase (AnigAP) and *Aspergillus niger* HP phytase (AnigPhyt) [179] to understand their differences in pH optimum of activity (pH 2.5 vs pH 2.5 and 5.0, respectively) and substrate specificity (broad specificity vs specificity for phytate). Between these enzymes, again, the main difference seems to be in charge distribution. AnigAP presents 2 acid residues (D75 and E272) while AnigPhyt present two acidic

and four basic residues (E205, D239, K68, K71, K227, K228). AnigPhyt seems to be optimised to attract the highly negatively charged phytate by displaying, at pH 2.5 and 5.0, its positively charged lysines, without which, at pH 2.5, the active site would have an overall neutral electrostatic field, making the enzyme possibly less selective in the choice of the substrate to hydrolyse [179].

Instead, promiscuity has been well characterised in the alkaline phosphatase superfamily. A comprehensive review on the subject is the one by Pabis and Kamerlin (2016) [180]. In this family too, electrostatic flexibility and cooperativity of the active site appear to be determinant in promoting catalysis of multiple substrates as well as active site volumes with large polar surfaces in rigid enzymes. Computational work also led to suggest that, the enzymes of this family maintain promiscuity when the number of electrostatic interactions between enzyme and substrate exceed the minimum number of necessary interaction for catalysis [180]. From this statements, a mutation such as the bulky tryptophan in p8, the insertion of an active site loop in *B/MINPP*, or the presence/absence of a large number of positively charged lysine residues could be seen as evolutionary tools used to decrease active site volume, modify the electrostatic field of the specificity pockets and therefore tune enzyme promiscuity.

Unfortunately, site-directed mutagenesis was not carried out on p8 and p15 because of lack of time. Hypothesis on structural determinants of positional specificity of IP<sub>6</sub> hydrolysis still need to be proved. Also, *EcAppA* mutants with confirmed altered positional stereospecificity showed that this ability came at the expense of catalytic efficiency. The reason behind this behaviour of *EcAppA* mutants may be the removal of the wild-type proton donor, limiting the efficiency of the enzyme in releasing the product. It would be interesting to mutagenize a set of positions in the specificity pockets B and F maintaining unaltered the proton donor, to understand if it could be possible to modify enzyme stereospecificity without affecting catalytic parameters. Also, docking studies present limitations and on this subject I refer to **Chapter 6.9**.

To understand if any of the MINPPs other than p8 and p15 could be of use, more efforts in the optimisation of expression and purification are required. MINPPs showed the tendency to precipitate and not to bind to His-tag columns. The use of alternative bacterial hosts (e.g. *Bacillus subtilis*) would be an option to try, as well as other fusion tags, e.g. Strep tag that also avoid Ni<sup>2+</sup> contamination of the sample. p8 and p15 seems instead more promising than other MINPPs. They can be easily expressed in large quantities and purified to high purity. Both proved to hydrolyse IP<sub>6</sub> to IP<sub>1</sub> and/or phosphate without the accumulation of intermediates of reaction

(data not shown – IP<sub>1</sub> and Pi cannot be distinguished by HPLC using our method), in line with the desired enzymes profiles sought in this study. However, p8 displayed a non-reversible unfolding at temperature ( $43,18 \pm 0,24$  °C) and pH optimum of 6, while p15 has a higher melting temperature of non-reversible unfolding ( $54,87 \pm 0,12$  °C) and a pH optimum of 5.5. Both the enzyme would need to be engineered to improve thermostability and catalytic efficiency at low pH in the eventuality of a commercial use.

Overall, this work provided the theoretical ground for further rational design of positional stereospecificity of IP<sub>6</sub> hydrolysis. As alternative approach, it would be interesting to try directed evolution from comprehensive combinatorial libraries of active site mutants at residues identified in **Chapter 6.5** (or even machine-learning assisted directed protein evolution [181]), to select for enzyme turnover number and stereospecificity of IP<sub>6</sub> hydrolysis in p8, p15 and *EcAppA* [182-184]. Another possible approach would be the creation of smaller libraries using *in silico* design in place of the simple docking approach used in this thesis. Ideally, it would be even more interesting to make use of the information collected in this thesis in combination with computational methods to redesign the active site of a thermostable HP2-scaffold. The active site could then be optimised through cycles of *in silico/in vitro* experiments for the binding of IP<sub>6</sub> in desired poses (e.g using RosettaDesign [185] to define active site geometry + MD simulations to screen IP<sub>6</sub> poses).

The *B/MINPP*-characteristic U-loop is another open interrogative. Structural analysis revealed that the feature contributes to shield the active site reducing its size and further coordinating the substrate in the catalytic cleft. The loop also contributes to the stabilisation of the enzyme because mutations to alanine of the cysteines involved in a S-S bridge on the loop contribute to a significant decrease of the enzyme melting temperature [42]. This structural feature could be used as a tool in synthetic design of the wider HP2 phytases class for the development of additional enzyme functionalities. Mutations of the residues able to interact with the substrate would be required to test U-loop effect on enzyme substrate selectivity, positional stereospecificity, catalytic activity enzyme structural dynamics. Also, it would be interesting to study conformational change by NMR methods.

In conclusion, phytases and kinases seemed to have evolved to be “selectively promiscuous” towards their substrates. These enzymes appear to be finely tuned to control the diversity of the inositol phosphates population in cells, their surrounding matrices, and to regulate the multiple cellular processes in which these molecules are involved in. In the article by Raboy (2003), some of the functionalities of inositol

polyphosphate intermediates are described, as well as the metabolic pathway which leads to IP<sub>6</sub> biosynthesis [8]. It is interesting to note how phosphorylation of the axial phosphate is reported only on IP<sub>4</sub> and IP<sub>5</sub> substrates for the formation of IP<sub>5s</sub> and IP<sub>6</sub> by inositol polyphosphate 2-kinase. However, IP<sub>2s</sub> and IP<sub>3s</sub> containing a phosphate groups in position 2 are produced by phytases. The majority of HP2 are known not to be able to dephosphorylate inositol polyphosphate on the C2 carbon of the inositol ring. Therefore, it may be possible that one of the roles of these enzymes is the production in cells of IP<sub>1-3s</sub> phosphorylated on position 2. On general terms, sequential phosphorylation of *myo*-inositol, an ATP-requiring process, could be considered a stimulus for cellular growth response [47] while *myo*-inositol polyphosphate hydrolysis may activate stress response, ATP regeneration, DNA repair [38]. What differs between the pathway of *myo*-inositol and IP<sub>1-3s</sub> phosphorylation and the pathway of IP<sub>2-4s</sub> hydrolysis is the absence or presence, respectively, of a monophosphoester group in axial position 2 of the carbon ring of inositol. It would be interesting to test if maybe 2-Pi IP<sub>1-3s</sub> could act in cells as stress response activators.

# CHAPTER 8 - APPENDIX

## 8. Supplemental information – Chapter 2

### 8.1. List of protein sequences

#### 8.1.1. *Bifidobacterium longum* MINPP

>Recombinant *B. lon* MINPP1 (HIS6-3C-POI) pOPINF

MAHHHHHHSSGLEVLFFQGPMEADGRYYSSKQPYVAPNDATASSYSKAPKGYGP  
IYTESMARHGSRGLSSYKYDALLMRMAETAARDGGFKSEAIIKAEFVKNLSGITAA  
NVENGYGMLTGQGAQQHYGIGERAYQRNRSFLDQAAADGGTIAYQSSGEARAT  
ESGENFEKGFNEASGGRLIGNVSAPTPADSGNGKDFQKNPDTLYFHKVQNP  
GTSKVPGTKAYDIANNYQNFVANDATIAGAECTIGDNVDVKRASHDLLSQIFTEE  
FLAKLENGEYKWYNTTDTGTTKGGKNCAPGADASKDPDACGEVSKKIKSEYDAA  
MDLYNLYIIAADMHNENTGDHTFAFDQYFQAYADDARMFALDAEDFYEK  
GPSYAGQNETYSIAQPLDDFLNTIDARVNGGSTVATFRFAHAETMMPFAALLG  
LPGSTQQAPASTTDVYTYGNNEWRGESVTPMAANVQWDVYARKGEDPATGQRY  
TPIVRMLYNENEVPRSECTPVADGSTWYKLTTELKSCLAADHKTGQDARI

>Recombinant *B. lon* MINPP1 E401Q - (HIS6-Th-POI) pET28a

MGSSHHHHHHSSGLVPRGSHMASMEADGRYYSSKQPYVAPNDATASSYSKAPK  
GYGPIYTESMARHGSRGLSYKYDALLMRMAETAARDGGFKSEAIIKAEFVKNLSG  
ITAANVENGYGMLTGQGAQQHYGIGERAYQRNRSFLDQAAADGGTIAYQSSGEA  
RATESGENFEKGFNEASGGRLIGNVSAPTPADSGNGKDFQKNPDTLYFHKVQ  
NPDGTSKVPGTKAYDIANNYQNFVANDATIAGAECTIGDNVDVKRASHDLLSQIF  
TEEFLAKLENGEYKWYNTTDTGTTKGGKNCAPGADASKDPDACGEVSKKIKSEY  
DAAMDLYNLYIIAADMHNENTGDHTFAFDQYFQAYADDARMFALDAEDFY  
EKGPSYAGQNETYSIAQPLDDFLNTIDARVNGGSTVATFRFAHAQTMMMPFAAL  
LGLPGSTQQAPASTTDVYTYGNNEWRGESVTPMAANVQWDVYARKGEDPATGQ  
RYTPIVRMLYNENEVPRSECTPVADGSTWYKLTTELKSCLAADHKTGQDARI

### 8.1.2. *Escherichia coli* AppA

>*Escherichia coli* AppA (HIS6-3C-POI) pOPINB

MGSSHHHHHHSSGLEVLFFQGPQSEPELKLESVVIVSRHGVRAPTKATQLMQDV  
TPDAWPTWPVKLGWLTTPRGGELIAYLGHYQRQRLVADGLLAKKGCPSGQVAII  
ADVDERTRKTGEAFAAGLAPDCAITVHTQADTSSPDPLFNPLKTGVCQLDNANV  
TDAILSRAGGSIADFTGHRQTAFRELERVLNFPQSNLCLKREKQDESCSLTQALP  
SELKVSADNVSLTGAVSLASMLTEIFLLQQAQGMPEPGWGRITDSHQWNTLLSL  
HNAQFYLLQRTPEVARSRATPLLDLIMAALTPHPPQKQAYGVTLPTSVLFIAGHD  
TNLANLGGALELNWTLPGQPDNTPPGGELVFERWRRLSDNSQWIQVSLVFQTL  
QQMRDKTPLSLNTPPGEVKLTLAGCEERNAQGMCSLAGFTQIVNEARIPACSL

>*Escherichia coli* AppA (POI-KHHHHHH) pOPINA

MMQSEPELKLESVVIVSRHGVRAPTKATQLMQDVTPDAWPTWPVKLGWLTTPRG  
GELIAYLGHYQRQRLVADGLLAKKGCPSGQVAIIADVDERTRKTGEAFAAGLAP  
DCAITVHTQADTSSPDPLFNPLKTGVCQLDNANVTDAILSRAGGSIADFTGHRQT  
AFRELERVLNFPQSNLCLKREKQDESCSLTQALPSELKVSADNVSLTGAVSLASM  
LTEIFLLQQAQGMPEPGWGRITDSHQWNTLLSLHNAQFYLLQRTPEVARSRATP  
LLDLIMAALTPHPPQKQAYGVTLPTSVLFIAGHDTNLANLGGALELNWTLPGQP  
DNTPPGGELVFERWRRLSDNSQWIQVSLVFQTLQQMRDKTPLSLNTPPGEVKLT  
LAGCEERNAQGMCSLAGFTQIVNEARIPACSLKHHHHHH

>*Escherichia coli* AppA un-tagged(POI) pOPINA

MMQSEPELKLESVVIVSRHGVRAPTKATQLMQDVTPDAWPTWPVKLGWLTTPRG  
GELIAYLGHYQRQRLVADGLLAKKGCPSGQVAIIADVDERTRKTGEAFAAGLAP  
DCAITVHTQADTSSPDPLFNPLKTGVCQLDNANVTDAILSRAGGSIADFTGHRQT  
AFRELERVLNFPQSNLCLKREKQDESCSLTQALPSELKVSADNVSLTGAVSLASM  
LTEIFLLQQAQGMPEPGWGRITDSHQWNTLLSLHNAQFYLLQRTPEVARSRATP  
LLDLIMAALTPHPPQKQAYGVTLPTSVLFIAGHDTNLANLGGALELNWTLPGQP  
DNTPPGGELVFERWRRLSDNSQWIQVSLVFQTLQQMRDKTPLSLNTPPGEVKLT  
LAGCEERNAQGMCSLAGFTQIVNEARIPACSL

### 8.1.3. 15 novel MINPPs phytases

*Flammeovirga pacifica* MINPP (HIS6-3C-POI) pDEST17

>1\_Fp\_noSP\_MW54330.2Da\_ext.c.76670\_theorpI6

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMTNVNDEIDTYDEWELGKNQPY  
NFKSVTSLPEGYNLDFVGYLSRHSSRYMTKPKEDVVLYNLFENAKLNNGLKKN  
GKHLFEEIKLLLKVQRDNYGTLSSNGEEHKLLGQRMANLAPEFFNSNPKIKST  
STLISKTQDSRANFQEGISSKVNRPKFINITYDDYNDPILRAFKISPSYQSYIDSAN  
WQIYIDEYQNTAQYKELRDQILDKLFTESYIKLLEDKFKKFFYDMEHNLIIANKND  
IVNNLFCFKISHNLPEFGPNLEIFTAEDSKILSHVDNIKSYYTKGPGFKNSDIS  
YKHAITFLKYMHSNINSYIDGRSDYQGNFNFAHSTTVVPVLVLLNLDQYKEMEM  
EQWNESEMSKMATNLTWLVLEKEGEKFIQIRWNENPVQLPLKEINEHYLYSYE  
EYDAYISKILSVYGLKDSRSNYNDILLSL

*Arsukibacterium* sp. MJ3 MINPP (HIS6-3C-POI) pDEST17

>2\_AMJ3\_noSP\_MW52214.77Da\_ext.c.68885\_theorpI5.50

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMAVTSEPQAGQLAQVELYLASKT  
PYQPQQPWQDYSPPPAGFTPVMVQHVARHGSRLSSAGDDDLALQLWNKAQQL  
NGLTPLGEQLGPVLEQLYQVHQHIGYGSISGLGIAEHHQMAERLLARYAPLFSEA  
PASGQRIAVTHSGRKRAAQSADAFVQHLLTLQPALQPLIDEAKADEHTLYFNKTE  
GSEGYEHYKDNDPRLQVMQQLTEQPKTEQMAQLMLARLFSEAFIARLAQGEF  
SFTLSYDDDEVNSPTDAAMLLYSLYNITSNMPAEGDWQFQRFVLPHEAAWFAEL  
DDADSFYGRGPAFAGDDITYRLARNLVEDMLARIAEPANYVAALRFTHAQUALMP  
LAAYLGIKDASEPLAVGTAYSQSSSWRSALVSPMAANVQWDAYRNTEGEVVVR  
MLHQEREVLFGRHCCPVTTPGSYFYHFKEQLQRCLL

*Bacteroides* sp. CAG:927 MINPP (HIS6-3C-POI) pDEST17

>4\_BCAG927\_noSP\_MW50458.17Da\_ext.c.79885\_theorpI6.68

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMAELSVANRQMASNYAYPYPEL  
PLPELTDAPQGYEPFHIEHYGRHGSRWHIGEYVYSTPVSIMLKADSCHKLTPRG  
KELLAELSRIADAARGRDGELTPLGAEQHRRGIARRMVHNFPQVFADSARIDARST  
VVIRCILSMDNELQEMLAANPKLRITSDASYADMMDYMNHHDTLMRKLTDARN  
LGMPRIDSLYANTGQWIGKLFDTTKWAKDSLNTGSLFWHLFIINANSQSHKDQK  
GFYDIFTDDEITRRWTIDNADWYLSYGNSPHSSGAGQHIQRNLLNNIIHSADTAIV  
YGKPSANLRFGHETCLLPLTVLMELDHYGAPIENMEEIAGKWHNYDIFPMGGNI  
QMIFYAPKGQTPTPDNVLVKVLLNEKERHLPVATQNFPPYYKWSLLRDFYSNKLA  
KPEPKLH

*Bacteroides intestinalis* MINPP (HIS6-3C-POI) pDEST17

>5\_Bi\_noSP\_MW50.715.17Da\_ext.c.72700\_theorpI6.99

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMTAKEEIFENIYRSAANHAYPE  
PHFTMTAPPKGYKPFYLSHYARHGSRVYRNPDDYTKPLVILREAEKDGTLTELG  
EKTLCVDSLARMAKGRYGDLTPLGTRQHRGIAQRMFNNFPEVVFQGTAAVDAR  
STVVIRCILSMMAECLQLQSANPKLQIKNDASYHDMYYMNGKDDYFKIQRQKEE  
VLAVKMAFRKEHLHPERLMKSLFKSEYVVKWKVDAGKLMSYLFELAVILQNND  
TDLELYSLFTKEECYDLWLIGNLEWYIDYGPSPLTQGKMPYVEANLLENILNTA  
DTCIVKKENSATLRFGHETCVLPLACLELGDCAHYHTTDVNSLADVWRNYRIFP  
KACNIQFVVFYRKKANDDILVKVLLNEQEMKLPVESDLAPYYHWKDVEAYYRNK  
LATFRK

*Bacteroides* sp. CAG:545 MINPP (HIS6-3C-POI) pDEST17

>6\_BCAG545\_noSP\_MW51389.08Da\_ext.c.72115\_theorpI8.71

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMTREKLAADPYAGVGIYHVYK  
PGNMQDTPAPEGYNPFYISHIGRHGSRVYHDSSKKFDKAPKIRKAAETGMLTER  
GMALYRELMKVDSATVGNLGLKSELGAEHQMIAKRMYRRFPPEVFSSEERIYVD  
AASSTVRRRCQESMMAFTKRLKKERKSLKVNVDHSGDSYMAVLLYKPADYFDIVHL  
GSDVTDLSLAKALLDTTAFSSIFKDPEEGAKLVSPSWKFMREVLWIGSIAPDIRLD  
DVCVPAYYTEDMRYQLAKVNACRIYSEMCNSMESGGRRMSLTETLLSDFIAKAD  
AALTSGSQRAADLRFVHDVSVAPLSALIGIEGCDKRLPAKDVWKYWMTSEYVPM  
AMNLQMVVFYRDKDGTCKCKDVLVKFLLNEQERLVPALTPVEGPYRWDVVRKFL  
LEKVKYAHEINVRWGVAQ

*Bacteroides* sp. 770 MINPP (HIS6-3C-POI) pDEST17

>7\_BCAG770\_MW50945.94Da\_ext.c.71405\_theorpI6.27

MSYYHHHHHHLESTSLYKKAGFLEVLFGPMLSFPLLLAVLLQLQPMQMLRED  
NDRAGVNTHPYEFRTMPVTQAPKGYEPVYISHYGRHGSRTDWGLGNYTYVIEIL  
EKAKEGELLTEEGKELLNETRTVAEVHHGADGHLTRLGEWEHRELADRMFEN  
YPQVFKKGSGLIRVESSTVHRCLVSMANFTGELIRRRPGLKFEIDSDDVIMKYVS  
DHPSEHIREASGIMLEPLRKVPTDTVQVMKNLFTDPVAARKIVDNIDKFKQEKIW  
GVARIARSSGIDANVYRHLPELVYKWWDYNNRELYIRQCNSVEFGAERMKSIK  
PLVNDIVKKADEALSTGRYAADLKFGHDYPLLSLASYLHLSGVGDVVSFDEIPTR  
WNDPMNIPLASNLQIFYRSKKSQDILVKFVYNDEERTIAGLEPVSQVYKWNVDV  
KNFVNDRRD

*Prevotella* sp. CAG:617 MINPP (HIS6-3C-POI) pDEST17

>8\_PCAG617¬\_noSP\_MW50939.64Da\_ext.c.81055\_theorpI6.19

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPMTARDEIIQDPALAAGKYYAYEA  
PVSDKVKAPAGYEPFYISAFARHGSRYLTDEEKYAEPVSVLRKADREGYLTTDG  
KKALQVMERLWKEAENRYGELTAKGAAQHQGLVERMYKHYPQVFVKGAHVDA  
RSTYKTRAFLSMAAACVRLAQLNSGLLITQDASAHDAYYIKYKNKTFEQQHLLAQ  
SDSVYRIADSVYVHPARLMKQLFTRNVSAEELGVSPVVLGMGELFELDGISQSSYG  
QEGLSFLFTDDERYDMWQRNNFEWYYEKGASPLSDCCMYHLERNLLENFIMTA  
DTAIASPYRCVTLRYGHDTNLAPLAALMGMNRLQTETTDWQQIADTYRTRYIIPM  
CGNIQLIFYRRKGSDDLKPLNREVTLPVETDCAPFYHWADVRAVWQKVAD  
SIVLPDSGMQHD

*Bacteroidales bacterium* CF MINPP (HIS6-3C-POI) pDEST17

>9\_BbCF\_MW51976.43Da\_ext.c.69595\_theorpI8.27

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPMKHLYYFLTAVLLLISASLLSQT  
KEEFLKDRRHASGIYQPYFFEATKSTPAPKGYTPFYISHYGRHGSRWVQTPD  
TYTYPQEILSKAHKEGVLTPLGESVYERVDAAAKDAWNRYGDLSQLGAKEH  
KEIAERMFLSFPEVFSTKNGKRCKIYSRSTIVPRCILSMAANNEKLKELNPE  
IEFIREASDRNRYLNNKYTQAKKDSVYAIRDNFLTRNLDINMFVSKLFDKTV  
YAANNISNPLSFMRSIHLIATDIPCVDSLDFTLLDIFTDDELFTLWQGSNMS  
IYYACGPSGVNGKVV RDSTKLLVKDILDCAERAVNGGGISADLRFHDSYAI  
PLISFMDIKGMNITNDP EKIQVWSDFKVSPMGVNLQIVFYKSSKNPEILV  
KILHCEKEVEIPVKSADIAPYYK WEDFKAYYKAKLAD

*Fibrobacter succinogenes* MINPP (HIS6-3C-POI) pDEST17

>10\_Fs\_noSP\_MW53658.26Da\_ext.c.84355\_theorpI8.54

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPQVSDEELAKHPEFTSSGYLVYP  
EPVNIKYTKAPAGYKPFYISHYGRHGSRYHHSAAEYTYLFTLAKADSAQK  
LTEL GKQALVYTKVLVDKAAAPRKGDLTQVGVKQHEGIANRMSKNFGDV  
FKDWNIGG KKITPYVRSYASTSGRCIVSMAAFIGELRSLNPKIHSELISG  
KSYMKFISAFDWGK LDYSKVKTYTDESDKLWKNVNPQQFLEKLFNDYKYV  
MNNVDVTNGFYNHFFEIA TSLQGMDKPLLDEIAQAAKVPADTFVNLF  
TTEEKIARWKAQNAWWYSLEGT SPL INRPDGLNFAKPTLQNI  
LEEADAEIAVDTTINARAVQTPIAATLRFGH DATLLPLS ALMQLPIANAK  
VSDLSKLHEQW NDFRIIPMAANLQMV FYKAKNKPILVKILYNEI EQTL  
PIECKAADKVQCPAAPYYRWDDVRNFYSALLKN

*Oxalobacteraceae bacterium* AB\_14 MINPP (HIS6-3C-POI) pDEST17

>11\_Ob\_noSP\_MW57539.87Da\_ext.c.69895\_theorpI7.27

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPMEQFYQTKTPYQPQQAGASYEAP  
PQGYSAVYTTQLLARHGSRGLSSMKTDLALYKLWQLASKEKALTPLGAHLGDDLL  
QMMRANALLGYGVAGITKPGYGNETMQGVTEHKQLAERMYQRLPQLFRSAAA  
GAEPRIQLLLTSGKDRAVDSGDYFAGSLLAQQPNLQALIVRPPSLAPRAKTNHDG  
RPAGTDRFLLYFHKLSAKQDLVADDSPLRATYLASQEYQAWAKSDELREAREAA  
VLVQPQVAAAAKAVLGRLFTPAFVEGLDQGRYSAANTGTYSYTSADQQFTNKL  
GDGDTDIKSSVDAQAALYEQYAAAADMKAELKADFTRYMPAQQAAVFAATEDAI  
AFYSKGPISENGDVNYRMAQTLLNDFFGVDAIARGDLSHAAKLRFHAHEIVIP  
MAAILGLSGMSEQLPRAVSYSYNSWRGHAVAPMAANIQWDVYTNQGRTLV  
RMLYNEKEADFKRGCDRAKIAPASHYYDYAALRACYMPK

*Aeromicrobium* sp. Root236 MINPP (HIS6-3C-POI) pDEST17

>12\_AR236\_noSP\_MW46939.83Da\_ext.c.60865\_theorpI9.59

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPMDIVSNAHYANQTPYGD PATTS  
VKAPPAGYELVFLENVGRHGSRSTTADSENALAVWNAARQGKLTTPGKLF  
DDDLRKFRAAEITIGYGNLSAIGRKEWIGRRTAASYHAFLTCAAADGDDIVFRT  
TSVYRTKQSASSLLSGLRAGVPLDFQPRVDDHMRIEDGATRTGNAAIASVLR  
SDVRAAAKHVLSRLYRSSVNSLSDPVGKALDIYGMIALAPGMQDDTTVTFSTRY  
VPLADARLLGYAKDAQNFYRYGPGVKGETSSYRQARPVLTDFFSELDKRLAGGK  
NAAVFRLAHGETTMPFAALTRLPGSTKQASASSPYSYANNSWRGYVAGRMAGN  
VEWAAYRNPANGGVLVTLRYNEQPVR LAASCKPSSLDPYFYGVHMLKTCLG

*Aeromicrobium* sp. Root495 MINPP (HIS6-3C-POI) pDEST17

>13\_AR495\_noSP\_MW46397.14Da\_ext.c.42985\_theorpI6.61

MSYYHHHHHHLESTSLYKKAGFLEVLFQGPMDPSLASLSPYPASAPRTAPVPSG  
YEPVFTESLDRHGSRRTASRTDMTLTLARIAEARAAGGLRDDADELERQVRTLQ  
ADVRRIGVGELTPVGEAELRGIGARVGLRLPGLLPATRVEIWSSGVQRASDSAE  
AFRGG LAAGAPSTSIGEVEADPRLRFDKTD AEYARFLADDVAATQAIRRVAESA  
PVQAAATDVLERVFTPAYVSTLDDPAAAALSLWNLYAIVPGMGDATSADFSAFVS  
HSDAVALGTLHDADYFYRRGPSFSGQDDTYRAARVLLDDFFAAVHRRLKGGATA  
GVFRFAHAEQLIPFSALVGLPGSTQQVTPGRPYSAADNPWRGGLV SPLGGNVQW  
DVFRDDRGRVLVRLQNERQVPVAERCRPAPGTRLYYRLTELRRCLR

*Streptacidiphilus jeojiense* MINPP (HIS6-3C-POI) pDEST17

>14\_Sj\_noSP\_MW49430.15Da\_ext.c.49405\_theorpI5.20

MSYYHHHHHHLESTSLYKKAGFLEVLFGQPMQGLGTGHAAPSRQHNYTTKTPY  
APQESLSRYQSAPRGFTPVFTENVARHGSRAMTDSSDGDVAVLAVLASAQAQ GAL  
TRLGARLAPQVQSLLAGASAIGYGNLSGRGVQEQQTALRMEQRLPSLFATIVAE  
KEPIEVETSGVARAIASANAFTSGLTGGDPALAGLIQAPVTNKDLLYFHKQPQNA  
DYQAYLASDPELAAVIAEIDGEPGTARAAADVVSRLFSKDFAAAMSADDRTSFSR  
SLYELYSAAPDLAVEAPGVLDLDAFLPTADADWFAYLDDAEFEFYQKGPAPFQGR TI  
TYDMANVLLTDLFTQVEDKADGSSDKGAVLRFTHAEIEIPLAVLLGLPGSTKAA  
SATLEHPYSYGDNPWRGATVAPMAANVQWDLYRKGSRYLVRMLYNEKETAFKP  
SCKPVARGSYFYDLNELESCFDRG

*Amycolatopsis jejuensis* MINPP (HIS6-3C-POI) pDEST17

>15\_Aj\_noSP\_MW50236.65Da\_ext.c.57425\_theorpI6.33

MSYYHHHHHHLESTSLYKKAGFLEVLFGQPMETAQYSTSKTPYSPQQDIRTYQP  
PPPGFTAVFTELVSRRHGSRTPTKIDGADLLLQLWAKARDESELTSAGQDFGPTM  
ESYRAAIQKVGLGQETGRGRQELQGMADRMQRRLPELFEKIKKDATPIAVVLSQ  
QTGRIADTAKFFTARLGATDPALAPLIQQPVVDQDLLYFHKTERGKAYRDYLEN  
DQRYQETVKRIKNRDGTREAAATDILKTIFTPAFVERMEPSAVTKAAQALYDLDAI  
APDLSVEGNWHLDRFVPRHAAAWFASIDDAKSFYKKGPGFEGSDITFAMASILL  
DDFFKQAEAAARAGKLGADLRFTHAEIIPLAALMQLPGSEKQADPDEDYTYANN  
PWRGASVSPMAANLQWDIYRNGTTYLVRMLYQEKEIPFKPDCTPFTPGSHYYRL  
DELSRCFGRTAR

*Streptomyces* sp. AW19M42 (HIS6-3C-POI) pDEST17

>16\_SAW19M42\_MW53331.13Da\_Ext.c.61895\_theorpI8.91

MSYYHHHHHHLESTSLYKKAGFLEVLFGQPMQKRTVAVALTLTAAALLSTALPA  
GATSPGDYATKTPYAPQQNLRAYQQAPKGFVPVFTENVSRHGSRAASDSEDGDL  
ILALWAKAAAEGQLTHAGKRFGGDVKSLLAAMDKVGYGQLSGRGEHELVDTAG  
RLRKRLPTLFRIVRNSERIDVVNSGKDRAVDSGNLFAAALADNDPALKPLINPA  
RTDADLLYFHKSAGGEEYRDYVDNDKRLAATLEGITDQPATRТАARNVLKKIFE  
PAFVKRISAGEFSGIGTETEEAAQAVYALYGIAPTMSDEGSWNMGRIAPREARWF  
AYLSDAEDFYEKGPFGSDSDITYRMANVLLDDFFQKIDAKRAGTGNLGAELRFT  
HAEIIPLAALMGLPGSAEPASPAEPYTYANNAWRGASVAPMAANIQWDLYRKA  
GKGSKGNRGNKGSKYLVRMLYKEKQTAFKQGCKPVSKGSYFYDADELKRCFGR  
AGA

*Streptomyces himastatinicus* MINPP (HIS6-3C-POI) pDEST17

>*Streptomyces himastatinicus* MINPP

MSYYHHHHHHLESTSLYKKAGFLEVLFGGPAGAKATYSYGTKATYEPRQNARS  
YQRPPAGFAPVFTENVSRRHGSRSATDGADGELILALWDKAEAEGLLTARGREFG  
PEVRALQAAMAKVGYGNLSGRGKREMRDTAVRLAKRLPGLFQRIAKTSEKIDVV  
SSGQGRAVDSGNTFAASLGDADPALKPLIGPARTDKDLLYFHKSTGGAAYQDWL  
ENDPRLAATLKS IKDQPATHRAAGSVLKKIFKPSFVQRISDGEFASIGSDVDAAEA  
VYDLYAIAPAMSEESPGGEGRHLDRYIAPSDAAWFGYLGDTEDFYEKGPSFADS  
DITYKMADVLLDDFFQQVEAKRDGTSKLGAE LRFTHAEEIPLATLMRLPGSTKA  
VTTDGPYTYADNPWRGASVAPLGANIQWDVFRKGD TYLVRMLYNEKETAFKTS  
CRPVARGSEFYNLNELERCFGRSD

## 8.2. List of primers

### 8.2.1. In-fusion cloning

*Escherichia coli* AppA N-terminal His-tagged

>Forward primer

5' AAGTTCTGTTTT CAGGGCCCG CAG AGT GAG CCG GAG CTG AAG 3'

Overlapping  $T_m$  (°C): 63.7

>Reverse primer

5' ATGGTCTAGAAAGCT TTA CAA ACT GCA CGC CGG TAT GCG T 3'

Overlapping  $T_m$  (°C): 64.0

*Escherichia coli* AppA un-tagged

>Forward primer

5' AGGAGATATACCATG ATG AAA GCG ATC TTA ATC CCA TTT TTA TCT C 3'

Overlapping  $T_m$  (°C): 61.6

>Reverse primer

5' GTGGTGGTGG TGTTT TTA CAA ACT GCA CGC CGG TAT GC 3'

Overlapping  $T_m$  (°C): 62.4

*Escherichia coli* AppA N-terminal His-tagged

>Forward primer

5' AGGAGATATACCATG ATG AAA GCG ATC TTA ATC CCA TTT TTA TCT C 3'

Overlapping  $T_m$  (°C): 61.6

>Reverse primer

5' GTGGTGGTGG TGTTT CAA ACT GCA CGC CGG TAT GC 3'

Overlapping  $T_m$  (°C): 61.4

### 8.2.2. Gateway cloning

*Key features of primers*

3C Protease cleavage site                   CTG GAA GTT CTG TTT CAG GGC CCG  
  L   E   V   L   F   Q   G   P

attB1 recombination site

5' **GGGGACAAGTTTGTACAAAAAAGCAGGCT**TC 3'

attB2 recombination site

5' **GGGGACCACTTTGTACAAGAAAGCTGGGT**T **TTA** 3'

### First step primers

Forward primers name           **3Cprotease-gene name-F**

>1\_Fp\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG ACC AAT GTG AAC GAT GAA 3'

Overlapping T<sub>m</sub> (°C): 54.0

>2\_AMJ3\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GCA GTT ACC AGC GAA 3'

Overlapping T<sub>m</sub> (°C): 53.7

>4\_BCAG927\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GCA GAA CTG AGC GTT 3'

Overlapping T<sub>m</sub> (°C): 53.7

>5\_Bi\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG ACC GCA AAA GAA G 3'

Overlapping T<sub>m</sub> (°C): 54.5

>6\_BCAG545\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG GTT ACC CGT GAA A 3'

Overlapping T<sub>m</sub> (°C): 54.5

>7\_BCAG770

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CTG AGC TTT CCG CT CTG 3'

Overlapping T<sub>m</sub> (°C): 52.8

>8\_PCAG617\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG ACC GCA CGT G 3'

Overlapping T<sub>m</sub> (°C): 54.3

>9\_BbCF

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG AAA CAC CTG TAC TAT TTT C 3'

Overlapping T<sub>m</sub> (°C): 52.8

>10\_Fs\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG GTT AGT GAT GAA GA 3'

Overlapping T<sub>m</sub> (°C): 53.2

>11\_Ob\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GAA CAG TTT TAT CAG ACC 3'

Overlapping T<sub>m</sub> (°C): 54.0

>12\_AR236\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GAT ATT GTT AGT AAC GCA 3'

Overlapping T<sub>m</sub> (°C): 52.0

>13\_AR495\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GAT CCG AGC CTG G 3'

Overlapping T<sub>m</sub> (°C): 54.3

>14\_Sj\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG GGT CTG GGC A 3'

Overlapping T<sub>m</sub> (°C): 54.3

>15\_Aj\_noSP

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG GAA ACC GCA CAG TAT AG 3'

Overlapping T<sub>m</sub> (°C): 55.3

>16\_SAW19M42

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ATG CAG AAA CGT ACC GTT G 3'

Overlapping T<sub>m</sub> (°C): 54.5

Reverse primer name:           **attB2-gene name-R**

>1\_Fp\_noSP

5' **CAAGAAAGCTGGGT**T **TTA** CAG GCT CAG AAT AT 3'

Overlapping T<sub>m</sub>

(°C): 53.2

>2\_AMJ3\_noSP

5' **CAAGAAAGCTGGGT** **TTA** CAG GCA ACG CT 3' Overlapping  $T_m$  ( $^{\circ}$ C):  
52.8

>4\_BCAG927\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** GTG CAG TTT AGG TTC C 3' Overlapping  $T_m$   
( $^{\circ}$ C): 52.8

>5\_Bi\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** TTT GCG AAA GGT TGC CA 3'  
Overlapping  $T_m$  ( $^{\circ}$ C): 53.2

>6\_BCAG545\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** CTG TGC AAC ACC CCA 3' Overlapping  $T_m$   
( $^{\circ}$ C): 53.7

>7\_BCAG770  
5' **CAAGAAAGCTGGGT** **TTA** ATC GCG ACG ATC GTT C 3' Overlapping  $T_m$   
( $^{\circ}$ C): 54.5

>8\_PCAG617\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** ATC ATG CTG CAT ACC G 3' Overlapping  $T_m$   
( $^{\circ}$ C): 52.4

>9\_BbCF  
5' **CAAGAAAGCTGGGT** **TTA** GTC GGC CAG TTT GG 3' Overlapping  $T_m$   
( $^{\circ}$ C): 52.8

>10\_Fs\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** GTT TTT CAG TGC G 3' Overlapping  $T_m$   
( $^{\circ}$ C): 52.4

>11\_Ob\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** TTT CGG CAT ATA GCA TGC 3'  
Overlapping  $T_m$  ( $^{\circ}$ C): 54.0

>12\_AR236\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** ACC CAG GCA GGT TTT C 3' Overlapping  $T_m$   
( $^{\circ}$ C): 54.5

>13\_AR495\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** ACG CAG ACA ACG 3' Overlapping  $T_m$  ( $^{\circ}$ C):  
53.7

>14\_Sj\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** ACC ACG ATC AAA GCA G 3' Overlapping  $T_m$   
( $^{\circ}$ C): 52.4

>15\_Aj\_noSP  
5' **CAAGAAAGCTGGGT** **TTA** GCG TGC GGT ACG AC 3' Overlapping  $T_m$   
( $^{\circ}$ C): 55.2

>16\_SAW19M42  
5' **CAAGAAAGCTGGGT** **TTA** TGC ACC CGC ACG AC 3' Overlapping  $T_m$   
( $^{\circ}$ C): 55.2

### Second step primers

Forward primer attB1-3Cprotease  
5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC** **CTG GAA GTT CTG TT** 3' (45)  
Overlapping  $T_m$  ( $^{\circ}$ C): 34.4

Reverse primer attB2  
5' **GGGGACCACTTTGTACAAGAAAGCTGGGT** 3' (30)  
Overlapping  $T_m$  ( $^{\circ}$ C): 39.2

## 8.2.3. Site-directed mutagenesis of *EcAppA*

>Forward primer  
5' GGA CAC G**CG** ACT AAT **CTG GCA AAT CTC GGC G** 3'  
Overlapping  $T_m$  ( $^{\circ}$ C): 54.3

>Reverse primer  
5' **ATT AGT CGC** GTG TCC **GGC AAT AAA CAG TAC TGA AG** 3'  
Overlapping  $T_m$  ( $^{\circ}$ C): 53.2

*Escherichia coli* AppA T327E

>Forward primer

5' GGA CAC GAT GAA AAT CTG GCA AAT CTC GGC G 3'

Overlapping T<sub>m</sub> (°C): 54.3

>Reverse primer

5' ATT TTC ATC GTG TCC GGC AAT AAA CAG TAC TGA AG 3'

Overlapping T<sub>m</sub> (°C): 53.2

*Escherichia coli* AppA D326A, T327E

>Forward primer

5' GGA CAC GCG GAA AAT CTG GCA AAT CTC GGC G 3'

Overlapping T<sub>m</sub> (°C): 54.3

>Reverse primer

5' ATT TTC CGC GTG TCC GGC AAT AAA CAG TAC TGA AG 3'

Overlapping T<sub>m</sub> (°C): 53.2

*Escherichia coli* AppA D326E

>Forward primer

5' GGA CAC GAA ACT AAT CTG GCA AAT CTC GGC G 3'

Overlapping T<sub>m</sub> (°C): 54.3

>Reverse primer

5' ATT AGT TTC GTG TCC GGC AAT AAA CAG TAC TGA AG 3'

Overlapping T<sub>m</sub> (°C): 53.2

## 8.3. Supplemental information – Chapter 3

### 8.3.1. Summary table of *EcAppA* expression trials

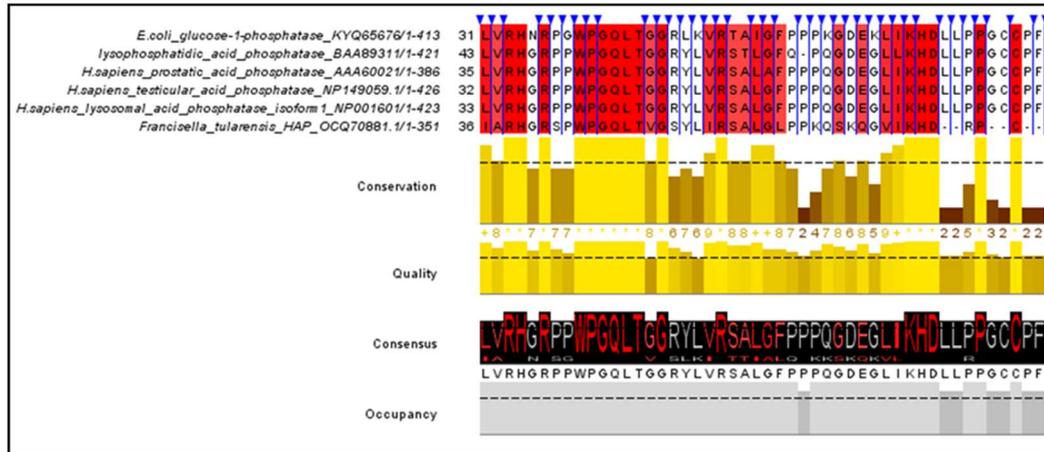
Construct	Strain	Media	Temperature	[IPTG]	OD	Length	Best solubility
pOPINB ::HIS6-3C-POI	Rosetta2(DE3)pLysS,	LB,	14 °C,	0.01 mM			Shuffle Express T7, LB, 37°C, 0.01mM IPTG, OD 0.9, o/n expression
	Rosetta-gami2(DE3)pLysS	LB+3%EtOH,	25 °C,	0.1 mM,	0.6-0.8	3h,	
	Shuffle Express T7 lysY	LB+100mM glucose	30 °C,	0.25 mM,	0.9-1.5	o/n,	
	Shuffle Express T7	TB,	37 °C	0.5 mM		o/w	
pOPINA ::POI-KHIS6	BL21(DE3),	LB	37 °C	0.05 mM,	0.6-0.8	3h,	BL21(DE3), 37°C, 0.05 mM IPTG, OD 0.8, o/n expression
	BL21(DE3)pLysS,			0.1 mM,			
	Rosetta2(DE3)pLysS			0.5 mM			
	Rosetta-gami2(DE3)pLysS						
pOPINA ::POI	BL21(DE3),	LB	37 °C	0.05 mM,	0.6-0.8	3h	BL21(DE3), 37°C, 0.05 mM IPTG, OD 0.8, o/n expression
	BL21(DE3)pLysS,			0.5 mM			
	Rosetta2(DE3)pLysS						
	Rosetta-gami2(DE3)pLysS						

**Figure 8.3.1.1. Summary of the expression trials carried out for *EcAppA*.**

HIS6-3C-POI: N-terminal cleavable His-tagged construct for cytoplasmic expression; POI-KHIS6: C-terminal His-tagged for the targeting of the protein to the periplasm; POI: untagged construct for the targeting of the protein to the periplasm. LB: Luria Bertani broth, TB: terrific broth. IPTG: isopropyl  $\beta$ -D-1-thiogalactopyranoside.

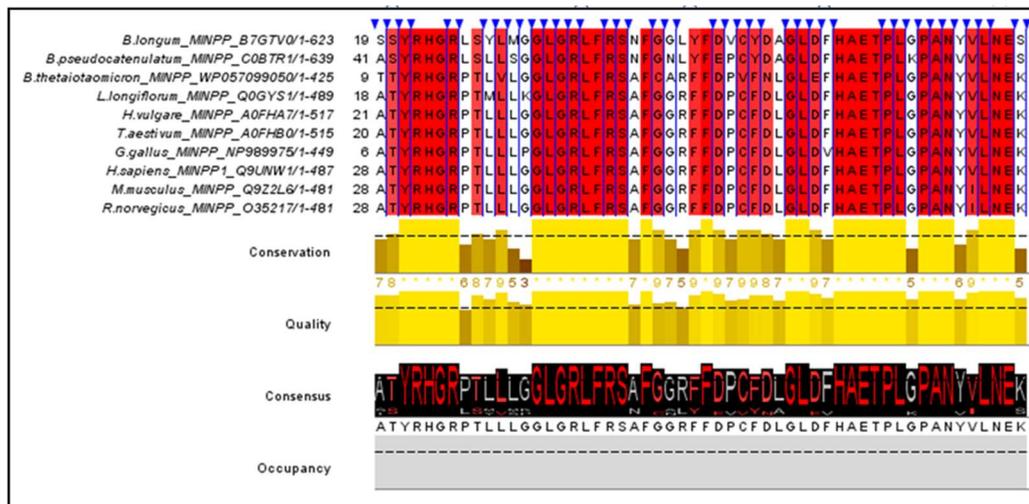
## 8.4. Supplemental information – Chapter 5

### 8.4.1. Standards multiple sequence alignments



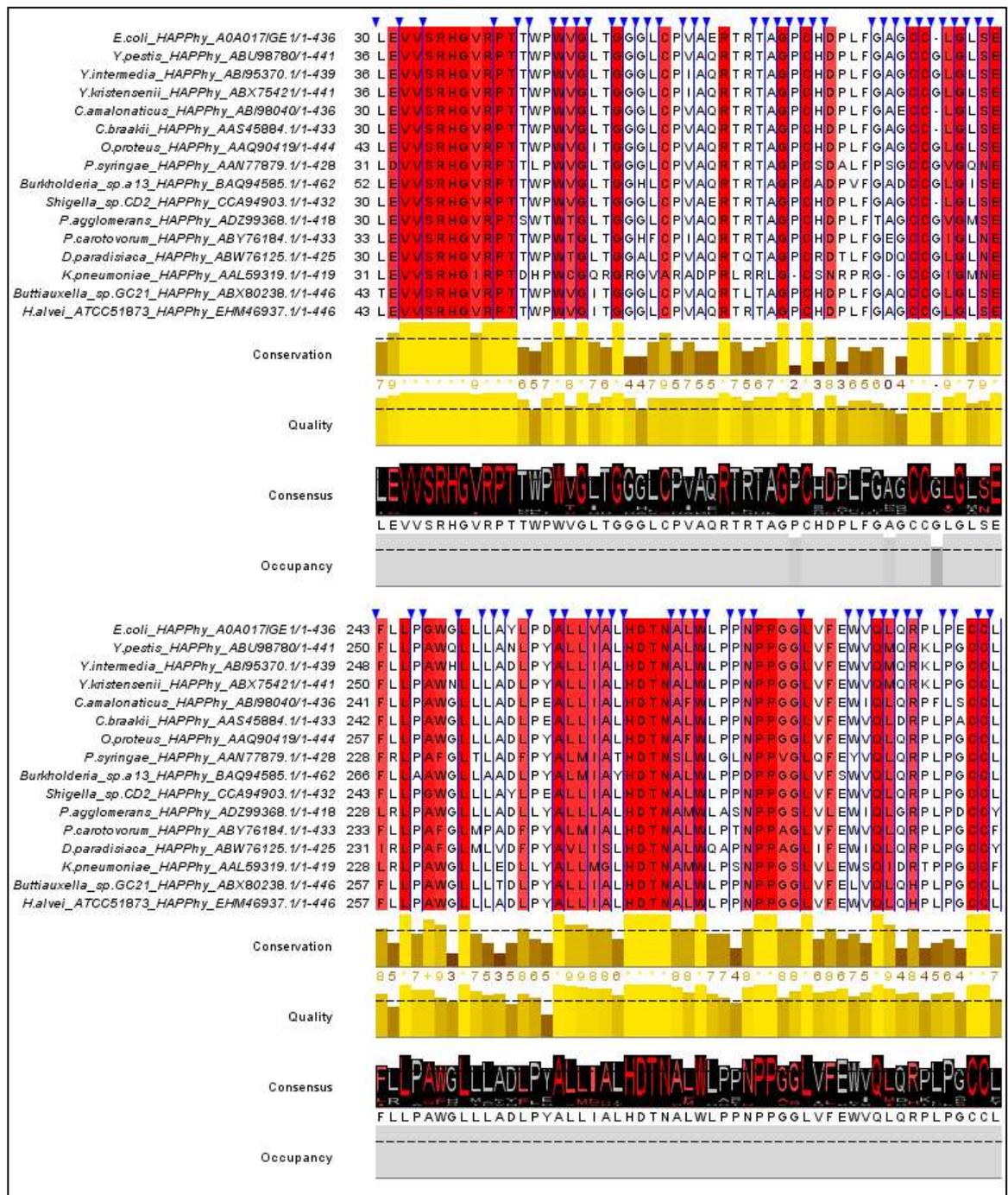
**Figure 8.4.1.1. : Multi-sequence alignment of group A – HP that are no phytases.**

Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157].



**Figure 8.4.1.2. Multi-sequence alignment of group C – MINPPs.**

Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157].



**Figure 8.4.1.3. Multi-sequence alignment of group B – bacterial phytases.**

Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157].

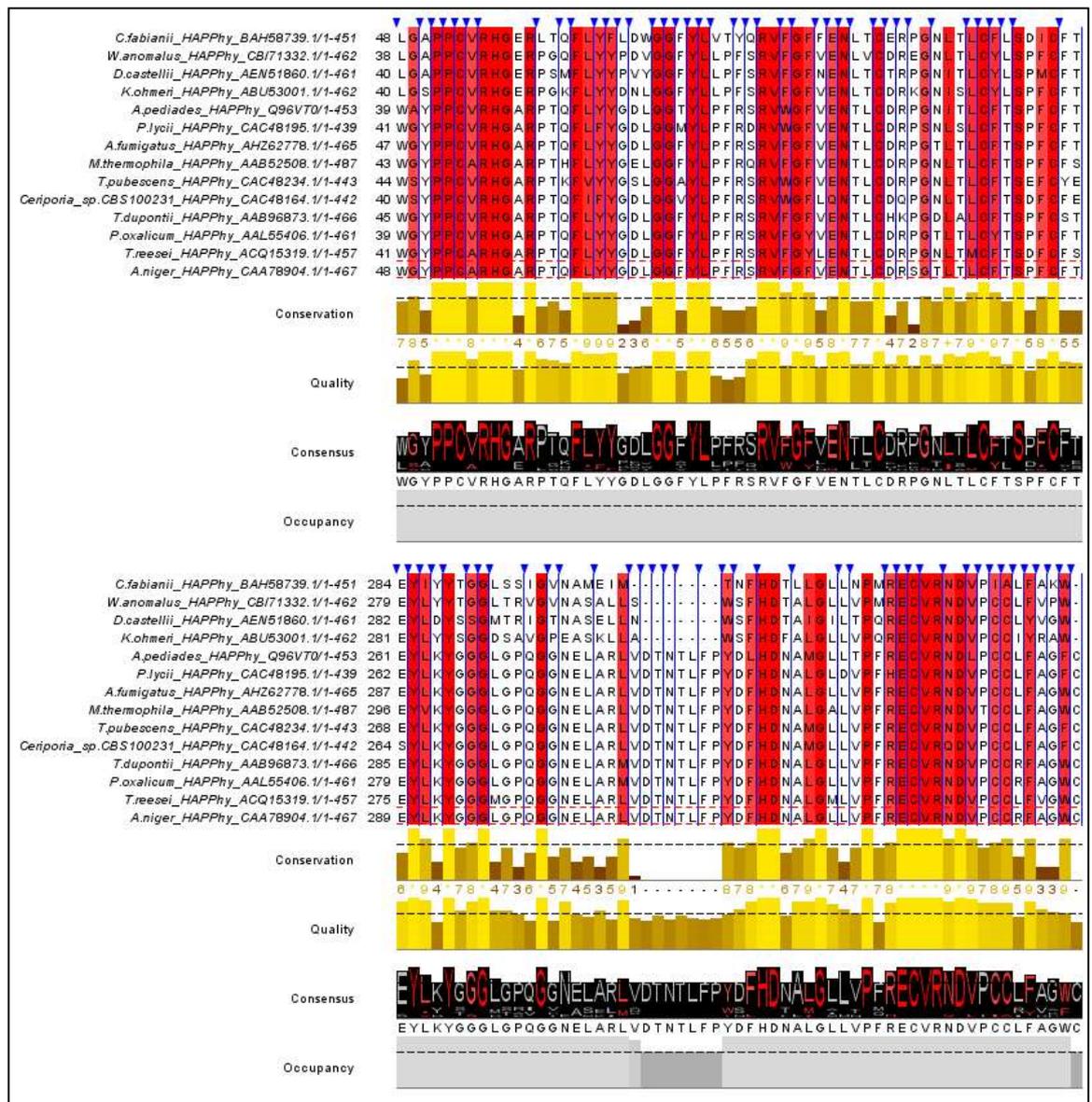
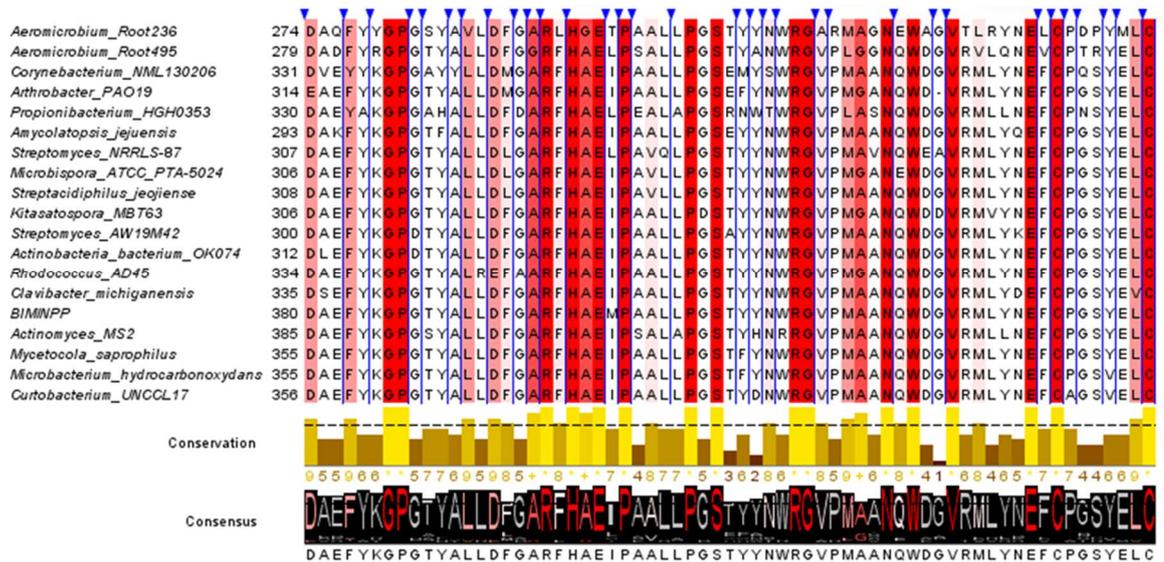
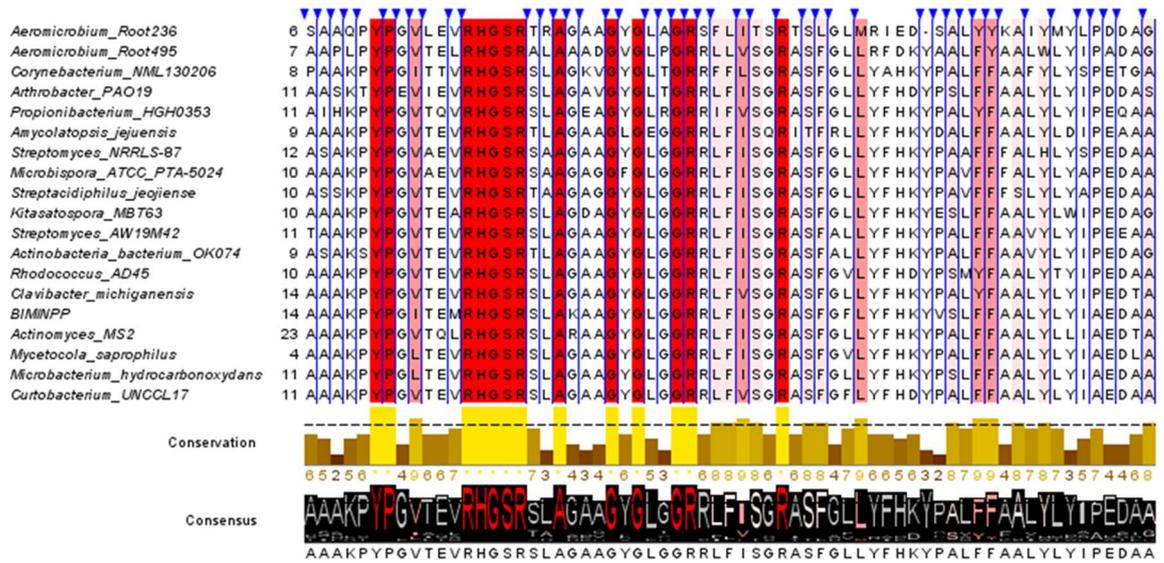


Figure 8.4.1.4. Multi-sequence alignment of group D – fungal phytases.

Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157].





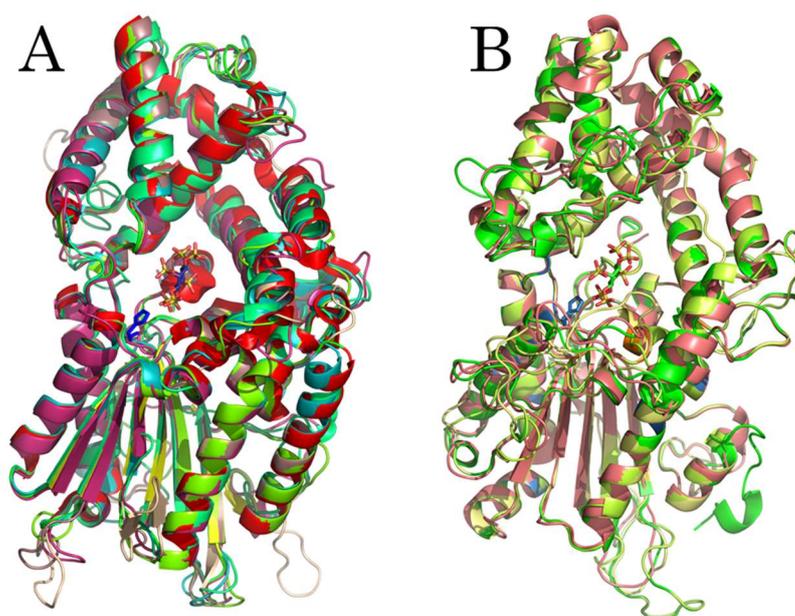
**Figure 8.4.2.2. Multiple sequence alignment of 18 *Actinobacteria* representative sequences.** Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-limits: 70% conservation, 70% consensus. The figure was created in Jalview [157].



Figure 8.4.2.3. Multiple sequence alignment of 72 *Bacteria* representative sequences.

Protein names are reported on the left of the alignment. Only regions of consensus higher than 50% are reported, blue triangles and lines mark the separation between non-adjacent amino acids. Areas of conservation from 70% to 100% are coloured in a red gradient. The bottom annotations display cut-off limits: 70% conservation, 70% consensus. The figure was created in Jalview [157]

### 8.4.3. 3D modelling of selected 16 MINPPs



Protein	template	%identity	QMean
P1 – <i>Flammeovirca pacifica</i>	BtMINPP	23.48	-9.87
P2 – <i>Arsukibacterium</i> sp. MJ3	BtMINPP	22.57	-8.52
P3 – <i>Prevotella brevis</i>	BtMINPP	24.75	-5.42
P4 – <i>Bacteroides</i> sp. CAG:927	BtMINPP	23.48	-9.87
P5 – <i>Bacteroides intestinalis</i>	BtMINPP	28.90	-4.27
P6 – <i>Bacteroides</i> sp. CAG:545	BtMINPP	27.88	-5.98
P7 – <i>Bacteroides</i> sp. CAG:770	BtMINPP	25.78	-6.51
P8 – <i>Prevotella</i> sp. CAG:617	BtMINPP	25.31	-5.40
P9 – <i>Bacteroidales bacterium</i> CF	BtMINPP	24.57	-6.7
P10 – <i>Fibrobacter succinogenes</i>	BtMINPP	26.92	-8.01
P11 – <i>Oxalobacteraceae bacterium</i> AB	BIMINPP	41.38	-8.87
P12 – <i>Aeromicrobium</i> sp. Root236	BIMINPP	40.72	-7.22
P13 – <i>Aeromicrobium</i> sp. Root495	BtMINPP	33.14	-6.29
P14 – <i>Streptacidiphilus jeojiense</i>	BIMINPP	45.52	-5.41
P15 – <i>Amycolatopsis jejuensis</i>	BtMINPP	22.54	-7.42
P16 – <i>Streptomyces</i> sp. AW19M42	BIMINPP	44.23	-6.21

**Figure 8.4.3.1. 3D modelling of 16 selected MINPPs.**

Modelling was carry out in SWISS-MODEL [175]. A) shows some of predictions modelled on *Bt*MINPP structures; B) shows some of predictions modelled on *B*/MINPP structures. The table indicate which structure was used as template for each MINPPs candidate, the degree of identity and the resulting QMEAN Z-score.

#### 8.4.4. Disulphide bridges prediction

Protein	Cys position	potentialS-S
<i>Bt</i> MINPP	C278, C291, C483, C501	C278-C291: cons. C483-C501: cons.
<i>Bt</i> MINPP	C160, C344, C396	-
P1 – <i>Flammeovirca pacifica</i>	C272	-
P2 – <i>Arsukibacterium</i> sp. MJ3	C433, C452	C433, C452: cons.
P3 – <i>Prevotella brevis</i>	C93, C223, C326, C349	-
P4 – <i>Bacteroides</i> sp. CAG:927	C87, C154, C327	-
P5 – <i>Bacteroides intestinalis</i>	C101, C156, C164, C270, C313, C330, C336, C343, C367	C330-C367: uncons.? C336-C343: cons.
P6 – <i>Bacteroides</i> sp. CAG:545	C161, C265, C284, C291, C349, C387	C291-C349: cons
P7 – <i>Bacteroides</i> sp. CAG:770	C151, C278	-
P8 – <i>Prevotella</i> sp. CAG:617	C166, C295, C296, C322, C368, C403	-
P9 – <i>Bacteroidales bacterium</i> CF	C150, C162, C256 C288, C312, C392	-
P10 – <i>Fibrobacter succinogenes</i>	C171, C436, C444	C436-C444: cons
P11 – <i>Oxalobacteraceae bacterium</i> AB	C489, C509	C489-C509: cons
P12 – <i>Aeromicrobium</i> sp. Root236	C398, C416	C398-C416: cons
P13 – <i>Aeromicrobium</i> sp. Root495	C402, C420	C402-C420: cons ?
P14 – <i>Streptacidiphilus jeojiense</i>	C433, C451	C433-C451: cons
P15 – <i>Amycolatopsis jejuensis</i>	C414, C432	C414-C432: cons ?
P16 – <i>Streptomyces</i> sp. AW19M42	C435, C453	C435-C453: cons

**Figure 8.4.4.1. S-S bridges prediction in the 16 selected MINPPs.**

The table list the Cys residues present in each protein sequence and their potential in generating S-S bridges. cons – very probably consecutive disulphide bridges, cons ? – possible consecutive disulphide bridges, uncons – unconsecutive disulphide bridges, S – Cys residues exposed at the protein surface.

## References

- [1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol* 215. 1990:403-10.
- [2] Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct*. 2012;7:12.
- [3] Suzuki U, Yoshimura K, Takaishi M. About the enzyme "phytase", which splits "anhydro-oxy-methylene diphosphoric acid". *Bulletin of the College of Agriculture, Tokyo Imperial University*. 1907;7:503-12.
- [4] Wodzinski RJ, Ullah A. Phytase. *Advances in applied microbiology*: Elsevier; 1996. p. 263-302.
- [5] Lei XG, Weaver JD, Mullaney E, Ullah AH, Azain MJ. Phytase, a new life for an "old" enzyme. *Annual review of animal biosciences*. 2013;1:283-309.
- [6] Li SA, Jiang WD, Feng L, Liu Y, Wu P, Jiang J, et al. Dietary myo-inositol deficiency decreased intestinal immune function related to NF-kappaB and TOR signaling in the intestine of young grass carp (*Ctenopharyngodon idella*). *Fish Shellfish Immunol*. 2018;76:333-46.
- [7] Dersjant-Li Y, Awati A, Schulze H, Partridge G. Phytase in non-ruminant animal nutrition: a critical review on phytase activities in the gastrointestinal tract and influencing factors. *Journal of the science of food and agriculture*. 2014.
- [8] Raboy V. myo-Inositol-1,2,3,4,5,6-hexakisphosphate. *Phytochemistry*. 2003;64:1033-43.
- [9] Haefner S, Knietsch A, Scholten E, Braun J, Lohscheidt M, Zelder O. Biotechnological production and applications of phytases. *Applied microbiology and biotechnology*. 2005;68:588-97.
- [10] Haefner S, Knietsch A, Scholten E, Braun J, Lohscheidt M, Zelder O. Biotechnological production and applications of phytases. *Applied microbiology and biotechnology*. 2005;68:588-97.
- [11] Shears SB. Inositol pyrophosphates: Why so many phosphates? *Advances in biological regulation*. 2014.
- [12] Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic acids research*. 2016:gkw952.
- [13] Boze H, Aumelas A, Moulin G. *Debaryomyces castellii* phytase. United States Patent. 2011:US 8,003,360 B2.
- [14] Viader-Salvado JM, Gallegos-Lopez JA, Carreon-Trevino JG, Castillo-Galvan M, Rojo-Dominguez A, Guerrero-Olazarán M. Design of thermostable beta-propeller phytases with activity over a broad range of pHs and their overproduction by *Pichia pastoris*. *Applied and environmental microbiology*. 2010;76:6423-30.
- [15] Shin S, Oh T-K, Ha N-C, Oh B-C, Oh B-H. Enzyme Mechanism and Catalytic Property of Beta Propeller Phytase. *Structure*. 2001;9:851-8.
- [16] Dionisio G, Madsen CK, Holm PB, Welinder KG, Jorgensen M, Stoger E, et al. Cloning and characterization of purple acid phosphatase phytases from wheat, barley, maize, and rice. *Plant physiology*. 2011;156:1087-100.
- [17] Puhl AA, Gruninger RJ, Greiner R, Janzen TW, Mosimann SC, Selinger LB. Kinetic and structural analysis of a bacterial protein tyrosine phosphatase-like myo-inositol polyphosphatase. *Protein science : a publication of the Protein Society*. 2007;16:1368-78.
- [18] Chu HM, Guo RT, Lin TW, Chou CC, Shr HL, Lai HL, et al. Structures of *Selenomonas ruminantium* phytase in complex with persulfated phytate: DSP phytase fold and mechanism for sequential substrate hydrolysis. *Structure*. 2004;12:2015-24.
- [19] Puhl AA, Greiner R, Selinger LB. Stereospecificity of myo-inositol hexakisphosphate hydrolysis by a protein tyrosine phosphatase-like inositol polyphosphatase from *Megasphaera elsdenii*. *Applied microbiology and biotechnology*. 2009;82:95-103.

- [20] Schenk G, Mitić N, Hanson GR, Comba P. Purple acid phosphatase: A journey into the function and mechanism of a colorful enzyme. *Coordination Chemistry Reviews*. 2013;257:473-82.
- [21] Tamayo-Ramos JA, Sanz-Penella JM, Yebra MJ, Monedero V, Haros M. Novel phytases from *Bifidobacterium pseudocatenulatum* ATCC 27919 and *Bifidobacterium longum* subsp. *infantis* ATCC 15697. *Applied and environmental microbiology*. 2012;78:5013-5.
- [22] Chi H, Tiller GE, Dasouki MJ, Romano PR, Wang J, O'Keefe RJ, et al. Multiple Inositol Polyphosphate Phosphatase: Evolution as a Distinct Group within the Histidine Phosphatase Family and Chromosomal Localization of the Human and Mouse Genes to Chromosomes 10q23 and 19. *Genomics* 1999;56:324-36
- [23] Mehta BD, Jog SP, Johnson SC, Murthy PP. Lily pollen alkaline phytase is a histidine phosphatase similar to mammalian multiple inositol polyphosphate phosphatase (MINPP). *Phytochemistry*. 2006;67:1874-86.
- [24] Lim D, Golovan S, Forsberg CW, Jia Z. Crystal structures of *Escherichia coli* phytase and its complex with phytate. *Nature Struct Biol*. 2000;7 108:13.
- [25] Ostanin K, Harms EH, Stevis PE, Kuciel R, Zhou M-M, Etten RLV. Overexpression, Site-directed Mutagenesis, and Mechanism of *Escherichia coli* acid phosphatase. *Journal of Biological Chemistry*. 1992;Vol. 267. No. 32:22830-6.1992.
- [26] Rigden DJ. The histidine phosphatase superfamily: structure and function. *The Biochemical journal*. 2008;409:333-48.
- [27] Etten RLV. Human Prostatic Acid Phosphatase: A Histidine Phosphatase. *Annals New York Academy of Sciences*. 1982:27-51.
- [28] Etten RLV, Davidson R, Stevis PE, MacArthur H, Moore DL. Covalent Structure, Disulfide Bonding, and Identification of Reactive Surface and Active Site Residues of Human Prostatic Acid Phosphatase. *The Journal of biological chemistry*. 1991;266:2313-9.
- [29] Ostanin; K, Saeed; A, ETTENS; RLV. Heterologous Expression of Human Prostatic Acid Phosphatase and Site-directed Mutagenesis of the Enzyme Active Site. *The Journal of biological chemistry*. 1994;Vol. 269:8971-8.
- [30] Ullah AHJ, Cummins BJ, H. Charles Dischinger J. Cyclohexanedione Modification Of Arginine At The Active Site Of *Aspergillus Ficum* Phytase. *Biochemical and biophysical research communications*. 1991;Vol. 178:45-53.
- [31] Ostanin K, Etten RLV. Asp304 of *Escherichia coli* acid phosphatase is involved in leaving group protonation. *Journal of Biological Chemistry*. 1993;269:20778-84.
- [32] Boquet PL, Manoil C, Beckwith J. Use of TnpA to detect genes for exported proteins in *Escherichia coli*. Identification of the plasmid-encoded gene for a periplasmic acid phosphatase. *Journal of Bacteriology*. 1987;Vol. 169, No. 4:1663-9.
- [33] Saini MS, Etten RLV. An Essential Carboxylic Acid Group In Human Prostate Acid Phosphatase. *Biochemica et Biophysica Acta*. 1979;568:370-6.
- [34] Liu Q, Huang Q, Lei XG, Hao Q. Crystallographic snapshots of *Aspergillus fumigatus* phytase, revealing its enzymatic dynamics. *Structure*. 2004;12:1575-83.
- [35] Greiner R, Carlsson N-G, Alminger ML. Stereospecificity of myo-inositol hexakisphosphate dephosphorylation by a phytate-degrading enzyme of *Escherichia coli*. *Journal of biotechnology*. 2000; 84 53-62.
- [36] Craxton A, Caffrey JJ, Burkhardt W, Safrany ST, Shears SB. Molecular cloning and expression of a rat hepatic multiple inositol polyphosphate phosphatase. *Biochemical Journal*. 1997;328:75-81.
- [37] Albert C, Safrany ST, Bembenek ME, Reddy KM, Reddy KK, J. R. Falck§, et al. Biological variability in the structures of diphosphoinositol polyphosphates in *Dictyostelium discoideum* and mammalian cells. *Biochemical Journal*. 1997;327:553-60.
- [38] Cho J, King JS, Qian X, Harwood AJ, Shears SB. Dephosphorylation of 2,3-bisphosphoglycerate by MIPP expands the regulatory capacity of the Rapoport-Luebering glycolytic shunt. *Proc Natl Acad Sci U S A*. 2008;105:5998-6003.

- [39] Cho J, Choi K, Darden T, Reynolds PR, Petite JN, Shears SB. Avian multiple inositol polyphosphate phosphatase is an active phytase that can be engineered to help ameliorate the planet's "phosphate crisis". *Journal of biotechnology*. 2006;126:248-59.
- [40] Li AWH. Structure-function studies of multiple inositol polyphosphate phosphatases from gut commensal bacteria. 2014.
- [41] Stentz R, Osborne S, Horn N, Li AW, Hautefort I, Bongaerts R, et al. A bacterial homolog of a eukaryotic inositol phosphate signaling enzyme mediates cross-kingdom dialog in the mammalian gut. *Cell reports*. 2014;6:646-56.
- [42] Zietek M. Characterisation of bacterial multiple inositol polyphosphate phosphatases relevant to the animal feed industry: University of East Anglia; 2017.
- [43] Romano PR, Wang J, O'Keefe RJ, Puzas JE, Rosier RN, Reynolds PR. HiPER<sub>1</sub>, a phosphatase of the endoplasmic reticulum with a role in chondrocyte maturation. *Journal of Cell Science*. 1998;111:803-13.
- [44] Kilaparty SP, Singh A, Baltosser WH, Ali N. Computational analysis reveals a successive adaptation of multiple inositol polyphosphate phosphatase 1 in higher organisms through evolution. *Evolutionary bioinformatics online*. 2014;10:239-50.
- [45] Dionisio G, Holm PB, Brinch-Pedersen H. Wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.) multiple inositol polyphosphate phosphatases (MINPPs) are phytases expressed during grain filling and germination. *Plant Biotechnol J*. 2007;5:325-38.
- [46] Mullaney EJ, Ullah AHJ. Identification of a Histidine Acid Phosphatase (phyA)-like Gene in *Arabidopsis thaliana*. *Biochemical and biophysical research communications*. 1998;251:252-5.
- [47] Cheng YL, Andrew DJ. Extracellular Mipp<sub>1</sub> Activity Confers Migratory Advantage to Epithelial Cells during Collective Migration. *Cell reports*. 2015;13:2174-88.
- [48] Evers PA, Murphy JM. The evolving world of pseudoenzymes: proteins, prejudice and zombies. *BMC Biol*. 2016;14:98.
- [49] Casali N, Preston A. *E. coli* plasmid vectors, methods and application. Humana Press. 2003.
- [50] I.Olivar F, Z.\* RLR, Greene PJ, Betlaci-I MC, Heyneker HB, Boyer HW. Construction And Characterization Of New Cloning Vehicles. *Gene*. 1977;2:95-113.
- [51] Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, et al. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic acids research*. 2007;35:e45.
- [52] Reece-Hoyes JS, Walhout AJM. Gateway Recombinational Cloning. *Cold Spring Harbor protocols*. 2018;2018:1-6.
- [53] Duong-Ly KC, Gabelli SB. Salting out of proteins using ammonium sulfate precipitation. *Methods in enzymology*. 2014;541:85-94.
- [54] Nagul EA, McKelvie ID, Worsfold P, Kolev SD. The molybdenum blue reaction for the determination of orthophosphate revisited: Opening the black box. *Analytica chimica acta*. 2015;890:60-82.
- [55] Johnson CM. Differential scanning calorimetry as a tool for protein folding and stability. *Archives of biochemistry and biophysics*. 2013;531:100-9.
- [56] Lopez MM, Makhatadze GI. Differential Scanning Calorimetry. In: Vogel HJ, editor. *Calcium-Binding Protein Protocols: Volume 2: Methods and Techniques*. Totowa, NJ: Springer New York; 2002. p. 113-9.
- [57] Weichenberger CX, Afonine PV, Kantardjieff K, Rupp B. The solvent component of macromolecular crystals. *Acta crystallographica Section D, Biological crystallography*. 2015;71:1023-38.
- [58] Berman HM, JohnWestbrook, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000;28:235-42.
- [59] Matthews BW. Solvent content of protein crystals. *J Mol Biol*. 1968;33:491-7.
- [60] Rupp B. Origin and use of crystallization phase diagrams. *Acta crystallographica Section F, Structural biology communications*. 2015;71:247-60.

- [61] Battye TG, Kontogiannis L, Johnson O, Powell HR, Leslie AG. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta crystallographica Section D, Biological crystallography*. 2011;67:271-81.
- [62] Incardona MF, Bourenkov GP, Levik K, Pieritz RA, Popov AN, Svensson O. EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis. *J Synchrotron Radiat*. 2009;16:872-9.
- [63] Winter G. xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*. 2009;43:186-90.
- [64] Powell HR. X-ray data processing. *Biosci Rep*. 2017;37.
- [65] Winter G, McAuley KE. Automated data collection for macromolecular crystallography. *Methods*. 2011;55:81-93.
- [66] Waterman DG, Winter G, Gildea RJ, Parkhurst JM, Brewster AS, Sauter NK, et al. Diffraction-geometry refinement in the DIALS framework. *Acta crystallographica Section D, Structural biology*. 2016;72:558-75.
- [67] Vonrhein C, Flensburg C, Keller P, Sharff A, Smart O, Paciorek W, et al. Data processing and analysis with the autoPROC toolbox. *Acta crystallographica Section D, Biological crystallography*. 2011;67:293-302.
- [68] Schwarzenbacher R, Godzik A, Grzechnik SK, Jaroszewsky L. The importance of alignment accuracy for molecular replacement. *Acta Crys D*. 2004;60:1229-36.
- [69] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. *Acta crystallographica Section D, Biological crystallography*. 2011;67:235-42.
- [70] Chen YW. Solution solution: using NMR models for molecular replacement. *Acta Crys D*. 2001;D57 1457-61.
- [71] Stein N. CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. *Journal of Applied Crystallography*. 2008;41:641-3.
- [72] Bunkoczi G, Read RJ. Improvement of molecular-replacement models with Sculptor. *Acta crystallographica Section D, Biological crystallography*. 2011;67:303-12.
- [73] Virtanen JJ, Zhang Y. MR-REX: molecular replacement by cooperative conformational search and occupancy optimization on low-accuracy protein models. *Acta crystallographica Section D, Structural biology*. 2018;74:606-20.
- [74] Jamrog DC, Zhang Y, Jr. GNP. SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. *Acta Crys D*. 2003;59:304-14.
- [75] Liu Q, Weaver AJ, Xiang T, Thiel DJ, Hao Q. Low-resolution molecular replacement using six-dimensional search. *Acta Crys D*. 2003;59:1016-9.
- [76] Sheriff S, Klei HE, Davies ME. Implementation of a six dimensional search using the AMoRe translation function for difficult molecular-replacement problems. *J App Crystallogr*. 1999;32:98-101.
- [77] Kissinger CR, Gehlhaar DK, Smith BA, Bouzida D. Molecular replacement by evolutionary search. *Acta Crys D*. 2001;57:1474-9.
- [78] Kissinger CR, Gehlhaar DK, Fogel DB. Rapid automated molecular replacement by evolutionary search. *acta Crys D*. 1999;55:484-91.
- [79] Glykos NM, Kokkinidis M. A stochastic approach to molecular replacement. *Acta Crys D*. 2000;56:169-74.
- [80] Vagin A, Teplyakov A. Molecular replacement with MOLREP. *Acta crystallographica Section D, Biological crystallography*. 2010;66:22-5.
- [81] McCoy AJ. Solving structures of protein complexes by molecular replacement with Phaser. *Acta crystallographica Section D, Biological crystallography*. 2007;63:32-41.
- [82] McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ. Likelihood-enhanced fast translation functions. *Acta crystallographica Section D, Biological crystallography*. 2005;61:458-64.
- [83] Storoni LC, McCoy AC, Read RJ. Likelihood-enhanced fast rotation functions. *Acta Crys*. 2004;D60:432-8.

- [84] Evans P, McCoy A. An introduction to molecular replacement. *Acta crystallographica Section D, Biological crystallography*. 2008;64:1-10.
- [85] Stubbs MT. *Protein Crystallography*. Elsevier Ltd. 2007.
- [86] McCoy A. Liking Likelihood. *Acta Crys*. 2004;D60:2169-83.
- [87] Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallographica Section D*. 2010;66:486-501.
- [88] Afonine PV, Grosse-Kunstleve RW, Adams PD, Urzhumtsev A. Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Crystallographica Section D*. 2013;69:625-34.
- [89] Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D*. 2012;68:352-67.
- [90] Headd JJ, Echols N, Afonine PV, Grosse-Kunstleve RW, Chen VB, Moriarty NW, et al. Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallographica Section D*. 2012;68:381-90.
- [91] Lengauer T, Rarey M. Computational methods for biomolecular docking. *Current opinion in structural biology*. 1996;6:402-6.
- [92] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*. 2010;31:455-61.
- [93] Wang R, Laib L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*. 2002;16:11-26.
- [94] Fei B, Cao Y, Xu H, Li X, Song T, Fei Z, et al. AppA C-terminal plays an important role in its thermostability in *Escherichia coli*. *Current microbiology*. 2013;66:374-8.
- [95] Berkmen M, Boyd D, Beckwith J. The nonconsecutive disulfide bond of *Escherichia coli* phytase (AppA) renders it dependent on the protein-disulfide isomerase, DsbC. *The Journal of biological chemistry*. 2005;280:11387-94.
- [96] Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M. SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microbial cell factories*. 2012;11:56.
- [97] Vasudevan UM, Salim SHB, Pandey A. A Comparative Analysis of Recombinant Expression and Solubility Screening of Two Phytases in *E.coli*. *Food Technol Biotechnol*. 2011;48:304-9.
- [98] Liu H, Naismith JH. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC biotechnology*. 2008;8:91.
- [99] Rodriguez E, Han Y, Lei XG. Cloning, sequencing, and expression of an *Escherichia coli* acid phosphatase-phytase gene (appA2) isolated from pig colon. *Biochemical and biophysical research communications*. 1999;257:117-23.
- [100] Golovan S, Wang G, Zhang J, Forsberg CW. Characterization and overproduction of the *Escherichia coli* appA encoded bifunctional enzyme that exhibits both phytase and acid phosphatase activities. *Can J Microbiol*. 2000; 46:59-71.
- [101] Yao MZ, Wang X, Wang W, Fu YJ, Liang AH. Improving the thermostability of *Escherichia coli* phytase, appA, by enhancement of glycosylation. *Biotechnology letters*. 2013;35:1669-76.
- [102] Fei B, Xu H, Zhang F, Li X, Ma S, Cao Y, et al. Relationship between *Escherichia coli* AppA phytase's thermostability and salt bridges. *Journal of bioscience and bioengineering*. 2013;115:623-7.
- [103] Tai HM, Yin LJ, Chen WC, Jiang ST. Overexpression of *Escherichia coli* phytase in *Pichia pastoris* and its biochemical properties. *Journal of agricultural and food chemistry*. 2013;61:6007-15.

- [104] Ariff RM, Fitrianto A, Abd Manap MY, Ideris A, Kassim A, Suhairin A, et al. Cultivation Conditions for Phytase Production from Recombinant *Escherichia coli* DH5alpha. *Microbiology insights*. 2013;6:17-28.
- [105] Klein AH, Shulla A, Reimann SA, Keating DH, Wolfe AJ. The intracellular concentration of acetyl phosphate in *Escherichia coli* is sufficient for direct phosphorylation of two-component response regulators. *J Bacteriol*. 2007;189:5574-81.
- [106] Lehmann M, Lopez-Ulibarri R, Loch C, Viarouge C, Wyss M, Loon APGMV. Exchanging the active site between phytases for altering the functional properties of the enzyme. *Protein Science* 2000;9:1866-72.
- [107] Underwood MA, German JB, Lebrilla CB, Mills DA. *Bifidobacterium longum* subspecies *infantis*: champion colonizer of the infant gut. *Pediatr Res*. 2015;77:229-35.
- [108] Li J, Santambrogio C, Brocca S, Rossetti G, Carloni P, Grandori R. Conformational effects in protein electrospray-ionization mass spectrometry. *Mass Spectrom Rev*. 2016;35:111-22.
- [109] Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica Section D, Biological crystallography*. 2010;66:12-21.
- [110] Poornam GP, Matsumoto A, Ishida H, Hayward S. A method for the analysis of domain movements in large biomolecular complexes. *Proteins*. 2009;76:201-12.
- [111] Chasles M. Note sur les propriétés générales du système de deux corps semblables entr'eux. *Bulletin des Sciences Mathématiques, Astronomiques, Physiques et Chimiques*. 1830;14:321-6.
- [112] Hayward S. Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements. *Journal of molecular biology*. 2004;339:1001-21.
- [113] Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*. 2016;44:W344-50.
- [114] Jameson GB, Anderson BF, Norris GE, Thomas DH, Baker EN. Structure of Human Apolactoferrin at 2.0 Å Resolution. Refinement and Analysis of Ligand-Induced Conformational Change. *Acta Crystallographica Section D*. 1998;54:1319-35.
- [115] Matthews LHWaBW. Structure of bacteriophage T4 lysozyme Refined at 1.7Å resolution. *J Mol Biol* 1987;193:189-99.
- [116] Dassa J, Marck C, Boquet PL. The complete nucleotide sequence of the *Escherichia coli* gene *appA* reveals significant homology between pH 2.5 acid phosphatase and glucose-1-phosphatase. *Journal Of Bacteriology*. 1990;172:5497-500.
- [117] Li J, Dong Y, Lu X, Wang L, Peng W, Zhang XC, et al. Crystal structures and biochemical studies of human lysophosphatidic acid phosphatase type 6. *Protein & cell*. 2013;4:548-61.
- [118] Fleisig H, El-Din El-Husseini A, Vincent SR. Regulation of ErbB4 phosphorylation and cleavage by a novel histidine acid phosphatase. *Neuroscience*. 2004;127:91-100.
- [119] Suter A, Everts V, Boyde A, Jones SJ, Lüllmann-Rauch R, Hartmann D, et al. Overlapping functions of lysosomal acid phosphatase (LAP) and tartrate-resistant acid phosphatase (Acp5) revealed by doubly deficient mice. *Development*. 2001;128:4899-910
- [120] Singh H, Felts RL, Schuermann JP, Reilly TJ, Tanner JJ. Crystal Structures of the histidine acid phosphatase from *Francisella tularensis* provide insight into substrate recognition. *Journal of molecular biology*. 2009;394:893-904.
- [121] Huang H, Luo H, Wang Y, Fu D, Shao N, Wang G, et al. A novel phytase from *Yersinia rohdei* with high phytate hydrolysis activity under low pH and strong pepsin conditions. *Applied microbiology and biotechnology*. 2008;80:417-26.
- [122] Luo H, Huang H, Yang P, Wang Y, Yuan T, Wu N, et al. A novel phytase *appA* from *Citrobacter amalonaticus* CGMCC 1696: gene cloning and overexpression in *Pichia pastoris*. *Current microbiology*. 2007;55:185-92.

- [123] Huang H, Luo H, Yang P, Meng K, Wang Y, Yuan T, et al. A novel phytase with preferable characteristics from *Yersinia intermedia*. *Biochemical and biophysical research communications*. 2006;350:884-9.
- [124] Kim H-W, Kim Y-O, Lee J-H, Kim K-K, Kim Y-J. Isolation and characterization of a phytase with improved properties from *Citrobacter braakii*. *Biotechnology letters*. 2003;25:1231-4.
- [125] Zinin NV, Serkina AV, Gelfand MS, Shevelev AB, Sineokoy SP. Gene cloning, expression and characterization of novel phytase from *Obesumbacterium proteus*. *FEMS microbiology letters*. 2004;236:283-90.
- [126] Fu D, Huang H, Luo H, Wang Y, Yang P, Meng K, et al. A highly pH-stable phytase from *Yersinia kristeensenii*: Cloning, expression, and characterization. *Enzyme and Microbial Technology*. 2008;42:499-505.
- [127] Shi P, Huang H, Wang Y, Luo H, Wu B, Meng K, et al. A novel phytase gene *appA* from *Buttiauxella* sp. GC21 isolated from grass carp intestine. *Aquaculture*. 2008;275:70-5.
- [128] Cho J, Lee C, Kang S, Lee J, Lee H, Bok J, et al. Molecular cloning of a phytase gene (*phy M*) from *Pseudomonas syringae* MOK1. *Current microbiology*. 2005;51:11-5.
- [129] Graminho ER, Takaya N, Nakamura A, Hoshino T. Purification, biochemical characterization, and genetic cloning of the phytase produced by *Burkholderia* sp. strain a13. *The Journal of general and applied microbiology*. 2015;61:15-23.
- [130] Pal Roy M, Mazumdar D, Dutta S, Saha SP, Ghosh S. Cloning and Expression of Phytase *appA* Gene from *Shigella* sp. CD2 in *Pichia pastoris* and Comparison of Properties with Recombinant Enzyme Expressed in *E. coli*. *PloS one*. 2016;11:e0145745.
- [131] Huang H. Novel Low-Temperature-Active Phytase from *Erwinia carotovora* var. *carotovora* ACCC 10276. *Journal of microbiology and biotechnology*. 2009;19.
- [132] Gu W, Huang H, Meng K, Yang P, Fu D, Luo H, et al. Gene cloning, expression, and characterization of a novel phytase from *Dickeya paradisiaca*. *Applied biochemistry and biotechnology*. 2009;157:113-23.
- [133] Sajidan A, Farouk A, Greiner R, Jungblut P, Muller EC, Borriss R. Molecular and physiological characterisation of a 3-phytase from soil bacterium *Klebsiella* sp. ASR1. *Applied microbiology and biotechnology*. 2004;65:110-8.
- [134] Ariza A, Moroz OV, Blagova EV, Turkenburg JP, Waterman J, Roberts SM, et al. Degradation of phytate by the 6-phytase from *Hafnia alvei*: a combined structural and solution study. *PloS one*. 2013;8:e65062.
- [135] Lassen SF, Breinholt J, Ostergaard PR, Brugger R, Bischoff A, Wyss M, et al. Expression, Gene Cloning, and Characterization of Five Novel Phytases from Four Basidiomycete Fungi: *Peniophora lycii*, *Agrocybe pediades*, a *Ceriporia* sp., and *Trametes pubescens*. *Applied and environmental microbiology*. 2001;67:4701-7.
- [136] Lee SH, Cho J, Bok J, Kang S, Choi Y, Lee PC. Characterization, gene cloning, and sequencing of a fungal phytase, *PhyA*, from *Penicillium oxalicum* PJ3. *Preparative biochemistry & biotechnology*. 2015;45:336-47.
- [137] Pasamontes L, Haiker M, Henriquez-Huecas M, Mitchell DB, Loon APGMv. Cloning of the phytases from *Emericella nidulans* and the thermophilic fungus *Talaromyces thermophilus*. *Biochimica et biophysica acta*. 1997;1353:217-23.
- [138] Lee J, Choi Y, Lee P-C, Kang S, Bok J, Cho J. Recombinant production of *Penicillium oxalicum* PJ3 phytase in *Pichia Pastoris*. *World Journal of Microbiology and Biotechnology*. 2006;23:443-6.
- [139] Dunn-Coleman N, Goedegebuur F, Ward M. *Trichoderma reisei* phytase enzymes, nucleic acids encoding such phytase enzymes, vectors and host cells incorporating same and methods of making and using same. United States Patent. 2009:US 7,510,831 B2.
- [140] Oakley AJ. The structure of *Aspergillus niger* phytase *PhyA* in complex with a phytate mimetic. *Biochemical and biophysical research communications*. 2010;397:745-9.
- [141] Phillippy BQ, Mullaney EJ. Expression of an *Aspergillus niger* Phytase (*phyA*) in *Escherichia coli*. *Journal of agricultural and food chemistry*. 1997;45:3337-42.

- [142] Ragon M, Neugnot-Roux V, Chemardin P, Moulin G, Boze H. Molecular gene cloning and overexpression of the phytase from *Debaryomyces castellii* CBS 2923. Protein expression and purification. 2008;58:275-83.
- [143] Watanabe T, Ikeda H, Masaki K, Fujii T, Iefuji H. Cloning and characterization of a novel phytase from wastewater treatment yeast *Hansenula fabianii* J640 and expression in *Pichia pastoris*. Journal of bioscience and bioengineering. 2009;108:225-30.
- [144] Joshi S, Satyanarayana T. Optimization of heterologous expression of the phytase (PPHY) of *Pichia anomala* in *P. pastoris* and its applicability in fractionating allergenic glycinin from soy protein. Journal of industrial microbiology & biotechnology. 2014;41:977-87.
- [145] Lee DC, Cottrill MA, Forsberg CW, Jia Z. Functional insights revealed by the crystal structures of *Escherichia coli* glucose-1-phosphatase. The Journal of biological chemistry. 2003;278:31412-8.
- [146] Pontoppidan K, Glitsoe V, Guggenbuhl P, Quintana AP, Nunes CS, Pettersson D, et al. In vitro and in vivo degradation of myo-inositol hexakisphosphate by a phytase from *Citrobacter braakii*. Archives of Animal Nutrition. 2012;66:431-44.
- [147] Tomschy A, Tessier M, Wyss M, Brugger R, Broger C, Schnobelen L, et al. Optimization of the catalytic properties of *Aspergillus fumigatus* phytase based on the three-dimensional structure. Protein Science. 2000;9:1304-11.
- [148] Greiner R, Konietzny U, Jany K-D. Purification and Characterization of Two Phytases from *Escherichia coli*. Archives of biochemistry and biophysics. 1993;Vol. 303, No. 1:107-13.
- [149] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. CABIOS. 1992;8:275-82.
- [150] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985;39:783-91.
- [151] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular biology and evolution. 2016;33:1870-4.
- [152] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
- [153] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics. 2004;5:113.
- [154] Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC bioinformatics. 2005;6:298.
- [155] Cedric Notredame, Higgins DG, Heringa J. T-Coffee: A Novel Method For Fast And Accurate Multiple Sequence Alignment. J Mol Biol. 2000;302:205-17.
- [156] Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic acids research. 2015;43:W7-14.
- [157] Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25:1189-91.
- [158] Robert X, Gouet P. Deciphering key features in protein structures with the new ENDscript server. Nucleic acids research. 2014;42:W320-4.
- [159] Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web server: 2015 update. Nucleic acids research. 2015;43:W30-8.
- [160] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;39:W29-37.
- [161] Xu H, Fu Y, Yang N, Ding Z, Lai Q, Zeng R. *Flammeovirga pacifica* sp. nov., isolated from deep-sea sediment. International journal of systematic and evolutionary microbiology. 2012;62:937-41.

- [162] Chan Z, Wang R, Liu S, Zhao C, Yang S, Zeng R. Draft genome sequence of an agar-degrading marine bacterium *Flammeovirga pacifica* WPAGA1. *Marine genomics*. 2015;20:23-4.
- [163] Gao C, Jin M, Yi Z, Zeng R. Characterization of a Recombinant Thermostable Arylsulfatase from Deep-Sea Bacterium *Flammeovirga pacifica*. *Journal of microbiology and biotechnology*. 2015;25:1894-901.
- [164] Zhou G, Jin M, Cai Y, Zeng R. Characterization of a thermostable and alkali-stable alpha-amylase from deep-sea bacterium *Flammeovirga pacifica*. *Int J Biol Macromol*. 2015;80:676-82.
- [165] Hou Y, Chen X, Chan Z, Zeng R. Expression and characterization of a thermostable and pH-stable  $\beta$ -agarase encoded by a new gene from *Flammeovirga pacifica* WPAGA1. *Process Biochemistry*. 2015;50:1068-75.
- [166] Lylloff JE, Hansen LB, Jepsen M, Hallin PF, Sorensen SJ, Stougaard P, et al. Draft genome sequences of two protease-producing strains of arsakibacterium, isolated from two cold and alkaline environments. *Genome Announc*. 2015;3.
- [167] Akob DM, Bohu T, Beyer A, Schaffner F, Handel M, Johnson CA, et al. Identification of Mn(II)-oxidizing bacteria from a low-pH contaminated former uranium mine. *Applied and environmental microbiology*. 2014;80:5086-97.
- [168] Fath S, Bauer AP, Liss M, Spriestersbach A, Maertens B, Hahn P, et al. Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. *PLoS one*. 2011;6:e17596.
- [169] Raab D, Graf M, Notka F, Schodl T, Wagner R. The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. *Syst Synth Biol*. 2010;4:215-25.
- [170] Vagenende V, Yap MG, Trout BL. Mechanisms of protein stabilization and prevention of protein aggregation by glycerol. *Biochemistry*. 2009;48:11084-96.
- [171] Kaushik JK, Bhat R. Why is trehalose an exceptional protein stabilizer? An analysis of the thermal stability of proteins in the presence of the compatible osmolyte trehalose. *The Journal of biological chemistry*. 2003;278:26458-65.
- [172] Atlungl T, Brondsted L. Role of the transcriptional activator AppY in regulation of the *cyx appA* operon of *Escherichia coli* by anaerobiosis, phosphate starvation, and growth phase. *Journal Of Bacteriology*. 1994;Sept:5414-22.
- [173] Brandt RB, Laux JE, Yates SW. Calculation Of Inhibitor Ki And Inhibitor Type From The Concentration Of Inhibitor For 50% Inhibition For Michaelis-Menten Enzymes. *Biochemical Medicine and Metabolic biology*. 1987;37:344-9.
- [174] Lee SD. *Amycolatopsis jejuensis* sp. nov. and *Amycolatopsis halotolerans* sp. nov., novel actinomycetes isolated from a natural cave. *International journal of systematic and evolutionary microbiology*. 2006;56:549-53.
- [175] 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 research*. 2018;46:W296-W303.
- [176] Terwilliger TC, Read RJ, Adams PD, Brunger AT, Afonine PV, Hung LW. Model morphing and sequence assignment after molecular replacement. *Acta crystallographica Section D, Biological crystallography*. 2013;69:2244-50.
- [177] Magdziarz T, Mitusinska K, Goldowska S, Pluciennik A, Stolarczyk M, Lugowska M, et al. AQUA-DUCT: a ligands tracking tool. *Bioinformatics*. 2017;33:2045-6.
- [178] Veiga N, Torres J, Macho I, Gomez K, Gonzalez G, Kremer C. Coordination, microprotonation equilibria and conformational changes of myo-inositol hexakisphosphate with pertinence to its biological function. *Dalton transactions*. 2014;43:16238-51.
- [179] Konstrewa; D, Markus Wyss, d'Arcy A, Loon APGMV. Crystal Structure of *Aspergillus niger* pH 2.5 Acid Phosphatase at 2.4 Å Resolution. *J Mol Bio*. 1999;288:965-74.
- [180] Pabis A, Kamerlin SC. Promiscuity and electrostatic flexibility in the alkaline phosphatase superfamily. *Current opinion in structural biology*. 2016;37:14-21.

- [181] Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*. 2019;116:8852-8.
- [182] Zeymer C, Hilvert D. Directed Evolution of Protein Catalysts. *Annual review of biochemistry*. 2018;87:131-57.
- [183] Porter JL, Rusli RA, Ollis DL. Directed Evolution of Enzymes for Industrial Biocatalysis. *ChemBiochem*. 2016;17:197-203.
- [184] Bunzel HA, Garrabou X, Pott M, Hilvert D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Current opinion in structural biology*. 2018;48:149-56.
- [185] Liu Y, Kuhlman B. RosettaDesign server for protein design. *Nucleic acids research*. 2006;34:W235-8.