An exploration of phosphorylases for the synthesis of carbohydrate polymers

Ellis Charles O'Neill

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

John Innes Centre

Norwich

September 2013

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

Phosphorylases are interesting enzymes with regard to both their role in metabolism and their use in the *in vitro* synthesis of carbohydrates. The disaccharide phosphorylases have attracted attention because of their strict stereo- and regiospecificity and their tractability. The polymerising phosphorylases have received less attention due to heterogeneous product formation, requiring more complex analyses. In this work three polymerising carbohydrate phosphorylases have been studied.

The plant α-1,4-glucan phosphorylase PHS2 is closely related to the well characterised mammalian glycogen phosphorylase. We present the first crystal structures of the plant enzyme which reveals a unique surface binding site. PHS2 allowed the production of novel starch like surface, both in two and three dimensions, which show some of the same properties as a native starch granule. This can now be used to study starch-active enzymes on an insoluble glucan surface which is analogous to the native starch granule.

The bacterial β-1,4-glucan phosphorylase CDP is involved in degradation of cellulose. In the reverse direction this enzyme allows the rapid synthesis of cellulose polymers in solution and also allows the synthesis of hemicellulose-like materials. The substrate specificity can in part be probed in the crystal structure presented here, which represents the first structure of a polymerising, inverting phosphorylase. Together these data provide the foundation for further work with this enzyme in the synthesis of plant cell wall related glycans.

The third enzyme studied was the β-1,3-glucan from the unsequenced alga *Euglena gracilis*, which was used for the facile enzymatic synthesis of β-glucosyl glycerols. In order to identify the sequence of this enzyme we obtained *de novo* transcriptome sequencing data from this alga, which has revealed unexpected metabolic diversity. Aside from complex carbohydrate metabolism, there are also many surprising features, including novel enzyme architectures, antioxidants only previously noted in human parasites and complex natural product synthases.

Acknowledgements

"Fortune favours the prepared mind"

- Louis Pasteur

I would like to thank all those who have helped to prepare my mind.

First and foremost: Prof Rob Field whose supervision and support have been invaluable throughout my project. Rob's all-encompassing intellect is an inspiration and role model. His willingness to allow my pursuit of new ideas and his confidence in me has encouraged my development and provided a strong foundation for my scientific pursuits.

I would like to thank the whole Field group, both past and present, who have made this such an enjoyable working environment: Martin Rejzek for his supervision; Stephan Goetz and Michael Rugen for smuggling some biology into the lab; Sergey Nepogodiev, Mike Paterson, Irina Ivanova and Mati Aguilar-Moncayo for their inspirational Chemistry; and especially Matthew Donaldson who started his PhD at the same time as me and has proved a friend and colleague throughout.

Thanks are due to all those who have helped make life at the John Innes Centre so rewarding: Clare Stevenson for indoctrinating me into the mysteries of crystallography; Dave Lawson for demystifying them; Lionel Hill and Gerhard Saalbach for mass spectrometry analysis; the BioImaging team, Kim Findlay, Elaine Barclay and Grant Calder for their assistance with microscopy; the Lab Support team who make life so much easier; and finally "The Beer Knights", with whom I have had a lot of fun.

Last but not least I would like to thank my family who have supported me through my time here and given me an occasional weekend retreat where I could relax and unwind. Without their continued support none of this would have been possible.

Table of contents

Abstract i
Acknowledgements ii
Table of contents iii
List of figures xi
List of tables xv
List of schemes xv
Abbreviations xvi

Chapter 1 – Introduction 1

1.1 The significance of carbohydrates 1
1.2 The structure of polysaccharides 2
1.2.1 Structural complexity in monosaccharides 2
1.2.2 Structural complexity in linear polysaccharides 4
1.2.3 Structural complexity provided by mixed linkages 5
1.3 Natural synthesis and degradation of carbohydrates 6
1.3.1 Glycosyltransferases (GTs) 6
1.3.2 Glycoside hydrolases 7
1.3.3 Polysaccharide lyases 7
1.3.4 Auxiliary activities 7
1.4 Carbohydrate space 8
1.4.1 Storage polysaccharides 8
1.4.2 Extra-cellular carbohydrates 8
1.5 Conclusions 9
1.6 References 10

Chapter 2 – *In vitro* synthesis of polysaccharides 13

2.1 Chemical Synthesis 13
2.1.1 Polycondensation 13
2.1.2 Ring opening polymerisation 14
2.1.3 Solid phase synthesis 14
2.2 *In vitro* enzymatic synthesis of polysaccharides 15
2.2.1 Glycoside hydrolases 16
2.2.2 Glycosyltransferases 19
2.2.3 Polysaccharide lyases 22
2.3 The natural role of phosphorylases 24
2.3.1 α-1,1-Glucan phosphorylases 26
2.3.1.1 Inverting trehalose phosphorylase 26
2.3.1.2 Trehalose-6-phosphate phosphorylase 26
2.3.1.3 Retaining trehalose phosphorylase 26
2.3.2 α-1,2-Glucan phosphporylases 26
2.3.2.1 Kojibiose phosphorylase 26
2.3.2.2 Sucrose phosphorylase 27
2.3.3 β-1,2-Glucan phosphorylases 27
2.3.4 α-1,3-Glucan phosphorylases 27
2.3.4.1 Nigerose phosphorylase 27
2.3.4.2 Glucosyl-rhamnose phosphorylase 28
2.3.5 β-1,3-Glycan phosphorylases 28
2.3.5.1 β-1,3-Glucan phosphorylases 28
2.3.5.2 Galactosyl-HexNAc phosphorylase 28
2.3.6 α-1,4-Glucan phosphorylases 29
2.3.6.1 α-1,4-Glucan phosphorylases 29
2.3.6.2 Maltose phosphorylase 29
2.3.6.3 Maltosyl phosphorylase 30

2.3.7 β-1,4-Glycan phosphorylases 30
2.3.7.1 β-1,4-Glucan phosphorylases 30
2.3.7.2 β-1,4-Mannan phosphorylases 30
2.3.7.3 Chitobiose phosphorylase 30
2.4 Conclusions on the *in vitro* synthesis of polysaccharides 31
2.5 References 32

Chapter 3 – PHS2 – an α-1,4-glucan phosphorylase 39

3.1 Introduction 39
3.1.1 α-1,4-Glucans 39
3.1.2 The *in vitro* use of α-1,4-glucan phosphorylases 44
3.1.3 Rationale 45
3.2 Cloning, expression and purification of PHS2 46
3.3 PHS2 activity in solution 47
3.3.1 Activity assays 47
3.3.1.1 Optimal conditions for PHS2 activity 47
3.3.1.2 Inhibitors of PHS2 48
3.3.1.3 Kinetic parameters of PHS2 48
3.3.2 Carbohydrate electrophoresis 50
3.3.3 Helical wrapping 53
3.3.4 Conclusions on PHS2 activity in solution 53
3.4 Crystal structure of PHS2 55
3.4.1 Protein crystallisation 55
3.4.2 Data collection and structure solution 55
3.4.3 Apo structure of PHS2 57
3.4.4 Substrate-bound structures of PHS2 61
3.4.4.1 Oligosaccharide-bound structures 61
3.4.4.2 Glc-1-P-bound structures 64
3.4.5 Conclusions on the PHS2 structures 65
3.5 Surface glucan synthesis by PHS2 67
3.5.1 SPR analysis of surface glucan synthesis by PHS2 67
3.5.1.1 Enzyme reactions on maltoheptaose surfaces 67
3.5.1.2 Fractionated glucan 68
3.5.1.3 Enzyme reactions on low FG density SPR surfaces 68
3.5.1.4 Enzyme reactions on high FG density SPR surfaces 71
3.5.1.5 Conclusions on the SPR experiments 72
3.5.2 Nanoparticles 73
3.5.2.1 Glycogen based nanoparticles 73
3.5.2.2 Synthesis of gold nanoparticles with a glycan surface 75
3.5.2.3 PHS2 catalysed synthesis on the FGNPs 75
3.5.2.4 Degradation of PHS2 synthesised glucan on FGNPs 77
3.5.2.5 Conclusion on the use of PHS2 to extend glucan nanoparticles 77
3.6 Conclusions on the use of PHS2 for the synthesis of amylose in solution and on surfaces 78
3.7 References 79

Chapter 4 – Cellodextrin phosphorylase – a β-1,4-glucan phosphorylase 85

4.1 Introduction 85
4.1.1 Cellulose 85
4.1.2 Cellobiose phosphorylases (CBP) 87
4.1.3 Cellodextrin phosphorylases (CDP) 89
4.1.4 Rationale 89
4.2 Expression and purification of CDP 90

4.3 Activity of CDP 91
4.3.1 Initial assays 91
4.3.2 Donor specificity of CDP 93
4.3.2.1 Glucan extension 93
4.3.2.2 Competitive inhibition by sugar-1-Ps 95
4.3.3 Acceptor specificity of CDP 96
4.3.3.1 β-1,4-Glucans 96
4.3.3.2 Other β-1,4-glycans 99
4.3.4 Conclusions on the activity of CDP 100
4.4 Crystal structures of CDP 102
4.4.1 Protein crystallisation 102
4.4.2 Heavy atom derivatisation 102
4.4.3 Data collection and structure solution 104
4.4.4 Apo structure of CDP 106
4.4.5 Conclusions on the structure of CDP 108
4.5 Conclusions on cellodextrin phosphorylase 108
4.6 References 109

Chapter 5 – Laminarin phosphorylase – a β-1,3-glucan phosphorylase 113

5.1 Introduction 113
5.1.1 β-1,3-Glucans 113
5.1.2 Laminarin phosphorylase 115
5.1.3 Rationale 116
5.2 Synthesis of glucosyl glycerols 117
5.2.1 Growth of *Euglena* cells 117
5.2.2 Assay for phosphorylase activity 117
5.2.3 Acceptor specificities of *Euglena* laminarin phosphorylases 118
5.2.4 Glucosyl glycerols 119
5.2.5 Biotransformation and purification of β-Glucosyl glycerols 122
5.2.6 Chemical synthesis of β-Glucosyl glycerols 124
5.2.7 Conclusions on the synthesis of glucosyl glycerols 125
5.3 Partial purification of laminarin phosphorylase from *Euglena* 126
5.3.1 Purification strategies 126
5.3.1.1 Strategy one 126
5.3.1.2 Strategy two 127
5.3.2 Proteomic analysis of partially purified laminarin phosphorylase 128
5.3.3 Conclusions on the purification of laminarin phosphorylases 128
5.4 *Euglena* transcriptome sequencing 129
5.4.1 Sequencing of the *Euglena* transcriptome 129
5.4.1.1 Growth of *Euglena* 129
5.4.1.2 Sequencing 130
5.4.2 Annotation of the *Euglena* transcriptome 132
5.4.3 Splice variants amongst the *Euglena* transcripts 134
5.5 Carbohydrate-active enzymes in the *Euglena* transcriptome 135
5.5.1 Phosphorylases 136
5.5.1.1 Trehalose phosphorylases 136
5.5.1.2 Laminarin phosphorylases 136
5.5.2 Glycosyltransferases 138
5.5.2.1 Energy storage carbohydrates 138
5.5.2.2 Glycolipids 138

5.5.2.3 Extra-cellular polysaccharides 138
5.5.3 Glycosyl hydrolases 140
5.5.3.1 Energy-storage carbohydrates 140
5.5.3.2 Mammalian GH families, absent from the *Euglena* transcriptome 140
5.5.3.3 Complex polysaccharide hydrolases 140
5.5.4 Carbohydrate binding modules 141
5.5.5 Paramylon metabolism 142
5.5.5.1 β-Glucan synthesis 143
5.5.5.1.1 GT48 143
5.5.5.1.2 GT2 143
5.5.5.1.3 GT75 144
5.5.5.2 β-Glucan degradation 144
5.5.5.2.1 GH81 145
5.5.5.2.2 GH17 145
5.5.5.2.3 GH64 145
5.5.5.2.4 GH1 145
5.5.5.2.5 GH2 145
5.5.5.2.6 GH3 145
5.5.5.2.7 GH5 and GH30 146
5.5.5.2.8 GH55 146
5.5.6 Glycolipids 148
5.5.6.1 Glycophosphatidylinositol anchor 148
5.5.6.2 Plastid glycolipids 150
5.5.7 Protein glycosylation 152
5.5.7.1 N-Glycans 152
5.5.7.2 O-GlcNAc 153
5.5.7.3 Sialic acids 153
5.5.8 Xylose containing polysaccharides 153
5.5.8.1 GT61 family 154
5.5.8.2 GT77 family 155
5.5.8.3 Other potential xylosyltransferases 155
5.5.8.4 Potential xylosyl hydrolases 156
5.5.9 Didomain CAZys 156
5.5.10 Conclusions on carbohydrate metabolism in *Euglena* 157
5.6 Conclusions on laminarin phosphorylase 158
5.7 References 159

Chapter 6 – Further details on the Euglena transcriptome 167

6.1 Introduction 167
6.1.1 *Euglena gracilis* 167
6.1.2 Rationale 170
6.2 Movement 172
6.3 Genetic control 174
6.3.1 Sexual reproduction 174
6.3.2 DNA gyrase 174
6.3.3 Gene silencing 176
6.3.4 BaseJ 177
6.4 Primary metabolism 178
6.4.1 Core metabolic enzymes 178
6.4.2 Lipids 178

6.4.2.1 Polyunsaturated fatty acids 180
6.4.2.2 Cyclopropane fatty acids 180
6.4.3 Isoprenoids 182
6.4.3.1 Isoprene biosynthesis 182
6.4.3.2 Polyprene biosynthesis 182
6.4.3.3 Carotenoids 184
6.4.3.4 Triterpenes 184
6.4.3.5 Vitamin E 185
6.4.4 Amino acids 186
6.4.4.1 Lysine biosynthesis 186
6.4.4.2 Arginine biosynthesis 188
6.4.4.3 Proline biosynthesis 188
6.4.4.4 Threonine biosynthesis 190
6.4.4.5 Branched-chain amino acid biosynthesis 190
6.4.4.6 Aromatic amino acid biosynthesis 192
6.4.4.7 Tetrapyrrole biosynthesis 193
6.4.5 Hydrogenase 193
6.5 Redox control 195
6.5.1 Vitamin C 195
6.5.2 Trypanothione 197
6.5.3 Ovothiol 198
6.5.4 Thiol analysis 199
6.6 Secondary metabolism 200
6.6.1 Indole alkaloids 200
6.6.2 Polyketides 201
6.6.3 Nonribosomal peptides 205
6.7 Conclusions of *Euglena* transcriptome sequencing 206
6.8 References 207

Chapter 7 – Future directions 217

7.1 PHS2 217
7.1.1 Further crystallography of PHS2 217
7.1.2 Engineering of PHS2 217
7.1.3 Starch-like surfaces 218
7.2 Cellodextrin phosphorylase 219
7.2.1 Further crystallography of CDP 219
7.2.2 Engineering of CDP 219
7.2.3 Cellulose-like surfaces 220
7.3 Laminarin phosphorylase 221
7.3.1 Identification of LDP 221
7.3.2 Crystallography of LDP 221
7.4 Other phosphorylases 222
7.4.1 Identification of the missing glucan phosphorylases 222
7.4.2 Identification of non-glucose phosphorylases 222
7.4.3 Polymerising phosphorylases 223
7.4.4 Relationship between hydrolases and phosphorylases 223
7.5 *Euglena gracilis* 225
7.5.1 Further sequencing of *Euglena* 225
7.5.2 Genetic control mechanisms in *Euglena* 225
7.5.3 Natural products from *Euglena* 226
7.5.4 Biotechnological uses of *Euglena* 226

7.6 Final thoughts 227
7.7 References 228

Chapter 8 – Materials and methods 230

8.1 General methods 230
8.1.1 Molecular biology methods 230
8.1.1.1 Transformation of *E. coli* 230
8.1.1.2 Agarose gel electrophoresis 230
8.1.1.3 PCR, sequencing and alignments 231
8.1.2 Purification and analysis of recombinant proteins from *E. coli* *231*
8.1.2.1 Purification of proteins 231
8.1.2.2 Denaturing polyacrylamide gel electrophoresis (SDS-PAGE) 231
8.1.2.3 Dynamic light scattering (DyLS) 232
8.1.3 Phosphorylase activity assays 232
8.1.4 Protein crystallography 232
8.1.4.1 Protein crystallisation screens 232
8.1.4.2 Preparation and transport of protein crystals to Diamond Light Source (DLS) 233
8.1.4.3 Data processing 233
8.1.5 APTS labelling of sugars 233
8.1.6 Carbohydrate electrophoresis with laser-induced fluorescence (CE-LIF) 234
8.1.7 General chemistry protocols 234
8.1.7.1 Nuclear magnetic resonance (NMR) 234
8.1.7.2 Mass spectrometry (MS) 234
8.1.7.3 Thin layer chromatography (TLC) 234
8.1.8 Atomic force microscopy (AFM) 234
8.1.9 Transmission electron microscopy (TEM) 235
8.1.9.1 Microscopy 235
8.1.9.2 Uranyl acetate (UA) staining 235
8.1.9.3 PATAg staining 235
8.1.10 Proteomics 235
8.2 Specific protocols for PHS2 236
8.2.1 Cloning of PHS2 236
8.2.2 Optimisation of PHS2 expression 238
8.2.3 Acceptor length specificity of PHS2 240
8.2.4 SPR experiments 241
8.2.4.1 Maltoheptaose surface 241
8.2.4.2 Fractionated glucan (FG) 242
8.2.4.3 Low FG density SPR surface 243
8.2.4.4 High FG density SPR surface 243
8.2.5 Fractionated glucan gold nanoparticles (FGNPs) 243
8.2.5.1 Synthesis of gold nanoparticles 243
8.2.5.2 Immobilisation of FG on AuNPs 243
8.2.5.3 Enzymatic reactions on AuNPs 244
8.3 Specific protocols for cellodextrin phosphorylase 245
8.3.1 Optimisation of CDP expression 245
8.3.1.1 Initial cloning of CDP 245
8.3.1.2 Expression of CDP 248
8.3.2 Kinetic assays of CDP 248
8.3.2.1 Inhibition of CDP 248

8.3.2.2 Donor specificity of CDP 249
8.3.2.3 Acceptor specificity of CDP 249
8.3.2.4 CE assays of CDP 250
8.3.3 X-ray crystallography of CDP 252
8.3.3.1 Crystallisation of CDP 252
8.3.3.2 Heavy metal derivatisation 252
8.3.3.3 Selenomethionine derivative of CDP 253
8.4 Specific protocols for laminarin phosphorylase 254
8.4.1 Growth of *Euglena* 254
8.4.2 *Euglena* lysis 254
8.4.3 Synthesis of β-Glc-gol 255
8.4.3.1 Enzymatic synthesis of β-Glc-gol 255
8.4.3.2 Chemical synthesis of β-Glc-gol 256
8.4.3.2.1 Chemical synthesis of O-2,3,4,6-tetra-O-acetyl-α-D-glucopyranosyl bromide (**189**) 256
8.4.3.2.2 Chemical synthesis of *O*-2,3,4,6-tetra-*O*-acetyl-β-D-glucopyranosyl-(1→2)-*O*-1,3-di-*O*-acetyl-glycerol (**161**) and *O*-2,3,4,6-tetra-*O*-acetyl-β-D-glucopyranosyl-(1→1)-*O*-2,3-di-*O*-acetyl-glycerol (**162**) 256
8.4.4 Laminarin phosphorylase purification 258
8.4.4.1 Assay for laminarin phosphorylase 258
8.4.4.2 Hydrophobic interaction chromatography (HIC) 258
8.4.4.3 Ion exchange chromatography 261
8.4.4.3.1 Cation exchange chromatography (CIEX) 261
8.4.4.3.2 Anion exchange chromatography (AIEX) 261
8.4.4.3.3 Optimisation of pH for AIEX 262
8.4.4.4 Gel filtration (GF) 263
8.4.4.5 First purification protocol 264
8.4.4.6 Second purification protocol 265
8.5 Specific protocols for *Euglena* transcriptome sequencing 266
8.5.1 RNA extraction and sequencing 266
8.5.2 Annotation of CAZys 267
8.5.3 Fatty acid analysis of *Euglena* extracts 267
8.5.4 Identification of thiols in *Euglena* 268
8.5.5 Identification of siderophore production by *Euglena* 268
8.6 References 269

Appendices 271

Appendix 1 – PHS2 sequencing 271
Appendix 2 – Synthetic gene for CDP 274
App.2.1 Nucleotide sequence of the *E. coli* codon optimised CDP 274
App.2.2. Amino acid sequence encoded in the CDP open reading frame of pET15b-CDP 274
App 2.3. Alignment of CDP and CBP from *Clostridium thermocellum* 275
Appendix 3 – Euglena proteins co-purified with laminarin phosphorylase activity 276
App 3.1 *Euglena* protein identified in Eg2.1 gel slice 276
App 3.2 *Euglena* protein identified in Eg2.2 gel slice 280
App 3.3 *Euglena* protein identified in EgC 283
App 3.4 *Euglena* protein identified in EgD 283

Appendix 4 – Carbohydrate-active enzymes in the *Euglena* transcriptome 284
Appendix 5 – Number of CAZys annotated in the database for various species 290
Appendix 6 – Alignments of didomain CAZys (lm.71174 and dm.47703) with well characterised single domain proteins 291
App.6.1 Alignment of lm.71174 with GT11s 291
App.6.2 Alignment of lm.71174 with GT15s 292
App.6.3 Alignment of dm.47703 with GT1s 293
App.6.4 Alignment of dm.47703 with GH78s 294
Appendix 7 – Rhamnosidase from *Euglena* transcript dm.44703 C-terminus (EgRase) 295
App.7.1 Synthetic gene for EgRase 295
App.7.2 Expression and purification of EgRase 296
App.7.3 Activity of EgRase 297
Appendix 8 – Euglena transcripts for core metabolic enzymes 299
App.8.1 Glycolysis/gluconeogenesis 299
App.8.2 Citrate cycle 300
App.8.3 Pentose phosphate pathway 300
App.8.4 Fatty acid biosynthesis 301
App.8.5 Calvin cycle 301
Appendix 9 – Euglena transcripts for biosynthetic enzymes 302
App 9.1 Isoprenoid biosynthesis 302
App.9.2 Vitamin E biosynthesis 304
App.9.3 Lysine biosynthesis 305
App.9.4 Arginine and proline biosynthesis 305
App.9.5 Threonine biosynthesis 306
App.9.6 Branched-chain amino acid biosynthesis 307
App.9.7 Aromatic amino acid biosynthesis 308
App.9.8 Tetrapyrrole biosynthesis 310
App.9.9 Ascorbate biosynthesis 310
App9.10 Trypanothione biosynthesis 312
App.9.11 Ovothiol biosynthesis 313
Appendix 10 – Proteomic analysis of dm.33057 314
Appendix 11 – Natural product synthases encoded in the *Euglena* transcriptome 315
App.11.1 Polyketide synthases 215
App.11.2 Nonribosomal peptide synthases 316
References 317

List of figures

Chapter 1 – Introduction 1

1.1 Comparison of D-galactose (**1**) and D-glucose (**2**) 2
1.2 Comparison of D-galactofuranose (**3**) and D-galactopyranose (**1**) 3
1.3 Comparison of α-D-glucopyranose (**4**) and β-D-glucopyranose (**2**) 3
1.4 Comparison of glucan polymers 4
1.5 Examples of mixed linkage β-glucans 5
1.6 General scheme of glycosyl transfer 6
1.7 General scheme of glycoside hydrolysis 7
1.8 General scheme of polysaccharide lyases 7
1.9 Schematic structure of pectin from plant cell wall 9

Chapter 2 – *In vitro* synthesis of polysaccharides 13

2.1 Polycondensation 13
2.2 Ring opening polymerisation 14
2.3 Mechanisms of glycoside hydrolases 16
2.4 Chemoenzymatic syntheses of cellulose 17
2.5 Chemoenzymatic synthesis of chitin 18
2.6 Glycosynthases 18
2.7 Mechanisms of glycosyltransferases 19
2.8 Enhancing the yields of glycosyltransferases 20
2.9 Human ABO blood group antigens 21
2.10 Mechanisms of polysaccharide lyases 22
2.11 Diglucose linkages for which phosphorylases are found in Nature 24
2.12 Trehalose metabolism using phosphorylases 25

Chapter 3 – PHS2 – an α-1,4-glucan phosphorylase 39

3.1 Structure of the starch granule 39
3.2 Starch granule metabolism 40
3.3 Surface changes upon enzymatic degradation of starch granules 42
3.4 Proposed products formed using α-1,4-glucan phosphorylases 43
3.5 Purification of PHS2 from *E. coli* 46
3.6 Optimum conditions for PHS2 activity 47
3.7 Inhibition assays of PHS2 48
3.8 Kinetic analysis of PHS2 49
3.9 Attachment of APTS to carbohydrates 50
3.10 Carbohydrate electrophoresis of PHS2 catalysed extension of maltooligosaccharides 51
3.11 Degradation of the PHS2 synthesised polymers 52
3.12 Formation of amylose helices using PHS2 53
3.13 PHS2 crystals 54
3.14 Structure of the PHS2 dimer 57
3.15 Pyridoxal phosphate in the active site of PHS2 58
3.16 The sites of allosteric regulation in GP are not present in PHS2 59
3.17 Details of oligosaccharides binding to α-1,4-glucan phosphorylase surface sites 63

3.18 Oligosaccharides binding to PHS2 62
3.19 Additional density when Glc-1-P was soaked into in PHS2 crystals 64
3.20 Immobilisation of glucans on SPR surfaces 66
3.21 PHS2 activity on maltoheptaose SPR surface 68
3.22 SPR analysis of PHS2 binding to the low FG density surface 69
3.23 Kinetics of low FG density surface extension using PHS2 69
3.24 PHS2-catalysed glucan synthesis on low FG density surface, as analysed by SPR and AFM 70
3.25 Typical SPR trace of enzyme reactions on the high FG density surface 72
3.26 Extension of glycogen with PHS2 73
3.27 Representative images of the glucan layer on the surface of the FGNPs 74
3.28 Extended reaction of PHS2 on FGNPs 76
3.29 Changes in the glucan layer on FGNPs 77

Chapter 4 – Cellodextrin phosphorylase – a β-1,4-glucan phosphorylase 85

4.1 Structure of β-1,4-glucans 85
4.2 Proposed products formed using cellobiose phosphorylase 86
4.3 Proposed products formed using cellodextrin phosphorylase 88
4.4 Purification of CDP from *E. coli* 90
4.5 Carbohydrate electrophoresis of CDP-synthesised polymers 91
4.6 Donor specificity of CDP, analysed using carbohydrate electrophoresis 92
4.7 Sugar-1-P binding inhibits CDP activity 94
4.8 CDP catalysed extension of APTS-labelled cellooligosaccharides 97
4.9 CDP catalysed extension of APTS-labelled β-1,4-glycans 99
4.10 Inhibition of CDP catalysed cellulose polymerisation by β-1,4-glycans 100
4.11 CDP crystals 101
4.12 X-ray energy scan at the Se *K* absorption edge of a crystal of SeMet derivatised CDP 104
4.13 Structure of the CDP dimer 106
4.14 Comparison of the structures of CDP and CBP 105
4.15 Proposed phosphate binding site of CDP 107

Chapter 5 – Laminarin phosphorylase – a β-1,3-glucan phosphorylase 113

5.1 Freeze-etch scanning electron micrograph of a paramylon granule from *Euglena gracilis* 113
5.2 Proposed products formed using laminarin phosphorylases 114
5.3 Extension of APTS-laminaritriose using *Euglena* phosphorylase 117
5.4 Acceptor specificity of phosphorylases in cleared *Euglena* cell lysate 118

5.5 Sucrose phosphorylase catalysed synthesis of 2-*O*-α-glucosyl glycerol 119
5.6 Reverse hydrolysis to form β-glucosyl glycerols 120
5.7 Proposed enzymatic synthesis of β-Glc-gol using cleared *Euglena* cell lysate 122
5.8 MALDI-ToF MS analysis of the peracetylated biotransformation of glucosyl glycerols using *Euglena* cell lysate 121
5.9 Proposed products of the glucosylation of glycerol with *Euglena* cell lysate 124
5.10 MALDI-ToF MS analysis of the products of the chemical synthesis of peracetylated glucosyl glycerols 123
5.11 ^{1}H NMR analysis of the peracetylated mono-glucosyl glycerols 125
5.12 Activity of laminarin phosphorylase partially purified from *Euglena* 127
5.13 Sequence distribution in the light and dark grown cells 130
5.14 Annotation of the *Euglena* transcriptome 131
5.15 Comparison of the HSP70 protein sequence from different strains of *Euglena* 132
5.16 Examples of splice variants amongst the *Euglena* transcripts 133
5.17 Average numbers of carbohydrate-active enzymes in different Phyla 135
5.18 Trehalose phosphorylases encoded in the *Euglena* transcriptome 136
5.19 Comparison of the glycosyltransferase family members in *Euglena* and humans 137
5.20 Comparison of the glycoside hydrolase family members in *Euglena* and humans 139
5.21 Domain architectures of the four CBM containing sequences in *Euglena* 141
5.22 Enzymes proposed to be involved in paramylon metabolism 142
5.23 Phylogeny of family GT2 proteins 144
5.24 Biosynthesis of the glycophosphatidylinositol anchor 148
5.25 Biosynthesis of plastid glycolipids 149
5.26 Biosynthesis of dolichol-linked oligosaccharide N-glycan precursor 152
5.27 Phylogeny of family GT61 proteins 154
5.28 Phylogeny of family GT77 proteins 155
5.29 Domain architecture of didomain CAZys encoded in the *Euglena* transcriptome 157

Chapter 6 – Further details on the *Euglena* transcriptome 167

6.1 Physiology of *Euglena gracilis* 167
6.2 Phylogeny of *Euglena gracilis* 168
6.3 Sources of the *Euglena* genome 169
6.4 *Euglena* movement 172
6.5 Ciprofloxacin inhibition of *Euglena* culture growth 176
6.6 BaseJ metabolism 177

6.7 Isoprenoid biosynthetic pathway 181
6.8 Polyprene synthase homology models 183
6.9 Alignment of cycloartenol synthases 184
6.10 Tocopherol biosynthesis 185
6.11 The α-aminoadipate pathway for the biosynthesis of lysine 186
6.12 Arginine and proline biosynthesis 187
6.13 Threonine and branched-chain amino acid biosynthesis 189
6.14 Aromatic amino acid biosynthesis 191
6.15 Tetrapyrrole biosynthesis 193
6.16 Ascorbic acid biosynthesis 194
6.17 Proposed nortrypanothione biosynthetic pathway 196
6.18 Ovothiol biosynthetic pathway 198
6.19 Analysis of the thiol content of *Euglena gracilis* 199
6.20 Alignment of candidate *Euglena* strictosidine synthases to known enzymes 200
6.21 Polyketide synthases encoded in the *Euglena* transcriptome 202
6.22 Insertion of a β-methyl branch on PksL in the biosynthesis of Bacillaene 203
6.23 Nonribosomal peptide synthases encoded in the *Euglena* transcriptome 204

Chapter 7 – Future directions 217

7.1 Substrate binding site of CDP and chitobiose phosphorylase 220
7.2 Family GH13 active site 224
7.3 *Euglena* natural products 226

Chapter 8 – Materials and methods 230

8.1 PCR amplification of *PHS2* 236
8.2 Plasmid map of pET151-PHS2 showing NdeI restriction sites 237
8.3 Restriction digest analysis of pET151-PHS2 plasmid from single clones 237
8.4 SDS-PAGE analysis of *E. coli* strains expressing PHS2 238
8.5 Intact mass of PHS2 239
8.6 Acceptor length specificity of PHS2 240
8.7 ^{1}H NMR of Glucidex IT12 242
8.8 Activity assays of initial CDP expression clones 245
8.9 Restriction digest of pET15b-CDP 246
8.10 Trial expression of CDP 247
8.11 Plasmid map of CDP-pET15b showing BamHI restriction sites 247
8.12 Inhibition of CDP by molybdate 248
8.13 Donor specificity of CDP 249
8.14 Acceptor length specificity of CDP 250
8.15 Insoluble product of CDP catalysed extension of APTS-(β-1,4-Glc)$_6$ 250
8.16 X-ray energy scans at the L_{III} absorption edge for heavy metal derivatives of CDP crystals 251

8.17 Purification of SeMet derivatised CDP 253
8.18 HIC purification of LDP 259
8.19 Ion exchange chromatography purification of LDP 260
8.20 AIEX purification of LDP at various pHs 262
8.21 GF purification of LDP 263
8.22 Purification of LDP by the first purification protocol 264
8.23 Purification of LDP by the second purification protocol 265

Appendices 271

App.7.1 The sequence of EgRase 295
App.7.2 Plasmid map of pET15b-dm.47703-CTerm 296
App.7.3 Purification of EgRase 297
App.7.4 Activity assay of EgRase 298
App.10.1 Proteomic analysis of dm.33057 314

List of tables

Table 2.1: Classification of phosphorylases so far characterised 23
Table 3.1: Acceptor length specificity of PHS2 50
Table 3.2: Summary of PHS2 X-ray data and model parameters 56
Table 4.1: Acceptor length specificity of CDP 96
Table 4.2: Summary of CDP X-ray data and model parameters 103
Table 5.1: Purification of LDP using strategy one 126
Table 5.2: Purification of LDP using strategy two 128
Table 5.3: Summary of *Euglena gracilis* transcriptome sequencing 130
Table 5.4: Annotation of enzymes involved in the biosynthesis of GPI anchors 147
Table 5.5: Annotation of enzymes involved in N-glycan biosynthesis 151
Table 6.1: *Euglena* transcripts encoding key genes associated with flagella (FM) or amoeboid motility (AM) 171
Table 6.2: *Euglena* transcripts encoding key genes involved in meiosis 173
Table 6.3: *Euglena* transcripts encoding key genes for RNA silencing 175
Table 6.4: *Euglena* transcripts encoding enzymes for the anaerobic fermentation of pyruvate 178
Table 6.5: Analysis of the hexane soluble extract form *Euglena* 179

List of schemes

Scheme 1: Chemical synthesis of peracetylated β-glucosyl glycerols 123
Scheme 2: The coupling of maltoheptaose to adipic acid dihydrazide 241

Abbreviations

$(\alpha\text{-}1,4\text{-}Glc)_n$	Maltodextrin (α-1,4-linked glucan) of length n
$(\beta\text{-}1,3\text{-}Glc)_n$	Laminaridextrin (β-1,3-linked glucan) of length n
$(\beta\text{-}1,4\text{-}Glc)_n$	Cellodextrin (β-1,4-linked glucan) of length n
$(\beta\text{-}1,4\text{-}Man)_n$	Mannan (β-1,4-linked mannan) of length n
$(\beta\text{-}1,4\text{-}Xyl)_n$	Xylan (β-1,4-linked xylan) of length n
Ac	Acetyl
Ac_2O	Acetic anhydride
AcOH	Acetic acid
Acr	Acarbose
ADH	Adipic acid dihydrazide
ADP	Adenosine 5'-diphosphate
AFM	Atomic force microscopy
AG	Amyloglucosidase from *Aspergillus niger*
AIEX	Anion exchange chromatography
AIM	Auto induction media
ALA	Aminolevulinic acid
AMP	Adenosine 5'-monophosphate
APTS	8-Aminopyrene-1,3,6-trisulphonic acid
AuNP	Gold nanoparticles
BBA	Barley β-amylase
BLAST	Basic local alignment search tool
BPG	β-Phenyl glucoside
CAS	Chrome azurol S
CAZy	Carbohydrate-active enzymes
CBM	Carbohydrate binding module
CBP	Cellobiose phosphorylase from *Clostridium thermocellum* unless otherwise stated
cDNA	Complementary DNA
CDP	Cellodextrin phosphorylase from *Clostridium thermocellum*
CE	Capillary electrophoresis
CIEX	Cation exchange chromatography
DCE	Dichloroethane
DGDG	Digalactosyl diacylglycerol
DLS	Diamond light source
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNJ	1-Deoxynojirimycin
DP	Degree of polymerisation
DyLS	Dynamic light scattering
E. coli	*Escherichia coli*
EDC	1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide
EDTA	Ethylenediaminetetraacetic acid

EG	*Euglena gracilis* media
ESI	Electrospray ionisation
EtOAc	Ethyl acetate
EtOH	Ethanol
FA	Fatty acid
FG	Fractionated glucan (see section 8.4.2.1)
FGNPs	Gold nanoparticles with a fractionated glucan surface (see Section 8.2.5)
FPLC	Fast protein liquid chromatography
G6P	Glucose-6-phosphate
Gal	D-Galactose
Gal-1-P	α-D-Galactose-1-P
GC	Gas chromatography
GF	Gel filtration
GGPP	Geranylgeranyl diphosphate
GH	Glycoside hydrolase
Glc	D-Glucose
Glc-1-P	α-D-Glucopyranose-1-phosphate. N.B. the β anomer is also referred to in this work
Glc-gol	Glucosyl glycerol
GlcN	D-Glucosamine
GlcN-1-P	α-D-Glucosamine-1-phosphate
GlcNAc	*N*-Acetyl-D-glucosamine
GlcNAc-1-P	α-*N*-Acetyl-D-glucosamine-1-phosphate
GO	Gene ontology
gol	Glycerol
GP	α-1,4-Glucan phosphorylase from *Oryctolagus cuniculus* or *Saccharomyces cerevisiae*
GPI	Glycophosphatidylinositol
GT	Glycosyltransferase
HEPES	4-(2-Hydroxyethyl)-1-piperazineethanesulfonic acid
Hex	Hexane
HexNAc	*N*-Acetylhexosamine
HIC	Hydrophobic interaction chromatography
HPLC	High-performance liquid chromatography
IMAC	Immobilised metal ion chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
JM	Jaworski's medium
KEGG	Kyoto encyclopedia of genes and genomes
LDP	Laminaridextrin (β-1,3-glucan) phosphorylase from *Euglena gracilis*
LB	Luria-Bertani media
MALDI	Matrix-assisted laser desorbtion/ionisation
MalP	α-1,4-Glucan phosphorylase from *Escherichia coli*
Man	D-Mannose
Man-1-P	α-D-Mannose-1-phosphate
MeOH	Methanol

MEP	Methylerythritol phosphate
MES	2-(*N*-Morpholino)ethanesulfonic acid
MGDG	Monogalactosyl diacylglycerol
MQ-H2O	Deionised water (milliQ, Millipore)
MS	Mass spectrometry
MVA	Mevalonate
MW	Molecular weight
NCBI	National Center for Biotechnology Information
NHS	*N*-Hydroxy succinimide
NMR	Nuclear magnetic resonance
NP	Nanoparticle
NRPS	Nonribosomal peptide synthase
NTA	Nitrilotriacetic acid
PAA	Porcine pancreatic α-amylase
PACE	Polyacrylamide gel carbohydrate electrophoresis
PAGE	Polyacrylamide gel electrophoresis
PATAg	Periodic acid/thiocarbhydrazide/silver
PCR	Polymerase chain reaction
PDB	Protein data bank
PEG	Polyethyleneglycol
PHS2	α-1,4-Glucan phosphorylase from *Arabidopsis thaliana*
Pi	Inorganic phosphate
PKS	Polyketide synthase
PLP	Pyridoxal phosphate
PMSF	Phenylmethanesulfonylfluoride
PNP	*p*-Nitrophenol
ProtK	ProteinaseK
RDRP	RNA-dependant RNA polymerase
Rha	L-Rhamnose
RNA	Ribonucleic acid
RU	Response units
SAM	Self-assembled monolayer - in here an alkyl thiol terminating in a short ethylene glycol moiety
SDS	Sodium dodecyl sulphate
SeMet	L-Selenomethionine
SHG	Soluble heteroglycan
SPR	Surface plasmon resonance spectroscopy
SQ	Sulfoquinovose
TBE	Tris-HCl (58 mM), Boric acid (89 mM), EDTA (3 mM)
TCA	Trichloroacetic acid
TEM	Transmission electron microscopy
TEV	Tobacco etch virus
TG	Triglycerides
TLC	Thin layer chromatography
ToF	Time of flight
Tris	Tris(hydroxymethyl)aminomethane

UA	Uranyl acetate
UDP	Uridine 5’-diphosphate
UV	Ultraviolet
V5 epitope	GKPIPNPLLGLDST
WE	Wax esters
Xyl	D-Xylose
Xyl-1-P	α-D-Xylose-1-phosphate
βCD	β-Cyclodextrin

Chapter 1 – Introduction

1.1 The significance of carbohydrates

Carbohydrates are the most abundant organic molecules on earth, and are key components in virtually every aspect of life from respiration to reproduction. Formed during photosynthesis and metabolised during respiration, carbohydrates are central to energy transduction within cells as well as acting as structural support, ligands and protein modifications.[1] All cells are coated in a surface layer of carbohydrates which can be protective, such as the peptidoglycan coat of bacteria, as well as supportive, such as the glycosaminoglycans of animal tissues. Despite the vast range of uses and the ubiquitous presence of carbohydrates in Nature, we still know very little about the structure and function of natural polysaccharides. In this age of next-generation DNA sequencing and whole cell proteomics this is principally because the structure of carbohydrates is not directly encoded in the genome. Instead they are synthesised by a set of specific transferases which can be affected by a number of different factors, including stress, disease, diet and environment.[2]

The structure of carbohydrates is extremely complex and difficult to analyse, with many sugars being isomers of each other. Great progress has been made in structural elucidation of mammalian N-glycans, based on knowledge of their biosynthesis.[3] However, for many other systems entire research programs are required to characterise a single structure, such as the core carbohydrate of the pear glycosylphosphatidylinositol anchor,[4] and well characterised synthetic analogues are often necessary to confirm the correct identification, for example in the identification of the enzymatic reactions during pectin biosynthesis.[5]

Carbohydrates also find diverse and important applications including in biomedicine,[6] materials science[7] and the growing field of nanotechnology.[8] Pure sources of well-defined carbohydrates are required for these technologies, and to some extent they can be obtained from natural sources. However, isolated polysaccharides are often produced as small quantities of impure and ill-defined material, which can have disastrous side effects. For example, over sulfation of chondroitin in clinical heparan resulted in hundreds

of cases of adverse side effects and many deaths. The impurities turned out to be a minor modification of one of the carbohydrate structures that make up the active components of the drug.[9] It is clearly desirable to synthesise defined polysaccharides to prevent contamination and to ensure precise control over the nature of the compounds.

1.2 The structures of polysaccharides

Carbohydrates are formed from the joining of diverse monosaccharide building blocks through different positions on the sugar rings, giving an extremely diverse class of compounds. These structural complications can be highlighted by comparison with short oligomers of the other biopolymers. There are four nucleotides, twenty amino acids and, considering just the most common sugars in mammals, ten monosaccharides each with two anomeric forms. This means for a simple pentanucleic acid there are 4^5 (1024) possible structures, and there are 20^5, about 3.2 million, pentapeptides; however there are over 1.5 billion possible five member polysaccharides,[10] unrestrained by simple geometric progression. There are many more monosaccharides present in Nature, vastly increasing this number to an unimaginably complex sample space, although only a few structures are actually utilised.

1.2.1 Structural complexity in monosaccharides

Monosaccharides are composed of polyhydroxylated carbon chains with multiple stereocentres and the different diastereomers can have vastly different properties. For example, the ricin toxin recognises and binds to β-galactose, but will not recognise β-glucose, with a single difference in stereochemistry at the 4 position (Fig 1.1).[11]

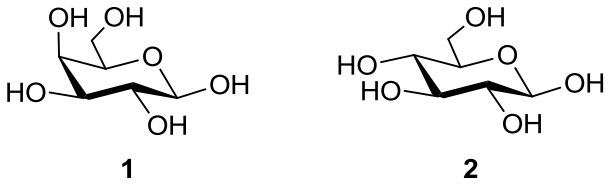
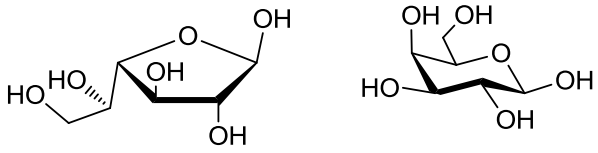


Figure 1.1: Comparison of D-galactose (**1**) and D-glucose (**2**)

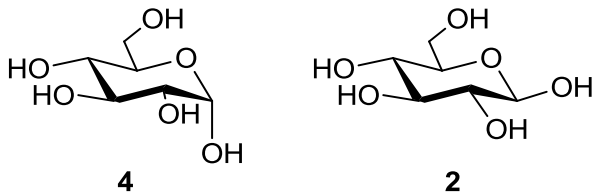
Under physiological conditions the monosaccharides undergo reversible intramolecular cyclisation, by nucleophilic attack by an hydroxyl on to the carbonyl group, resulting in ring formation. This ring can vary in size, normally five-membered (furanose) or six-membered (pyranose) rings, both of which are utilised in Nature (Fig 1.2). For instance, mammalian carbohydrates exclusively contain the more stable pyranose form of galactose,[12] whilst the furanose form is a key component of various polysaccharides and glycoconjugates in a range of bacteria, protozoa and fungi.[13]



3 **1**

Figure 1.2: Comparison of D-galactofuranose (**3**) and D-galactopyranose (**1**)

Cyclisation also introduces a new stereocentre, the anomeric centre, as the planar C=O bond is converted into an hemi-acetal (Fig 1.3). The two isomers, known as anomers, are rapidly interconverted in solution, but when a substituent is added to the anomeric oxygen, the structure becomes fixed.



4 **2**

Figure 1.3: Comparison of α-D-glucopyranose (**4**) and β-D-glucopyranose (**2**)

1.2.2 Structural complexity in linear polysaccharides

The sugars can be linked on to each other through any of the hydroxyls, giving radically different properties. Cellulose (**5**) is a long linear polymer of β-1,4-glucose and forms a ribbon with the glucose residues alternately flipping (Fig 1.4).[14] These ribbons then stack to form large bundles of hydrogen-bonded structural polymers. If residues are instead linked by α-1,4 bonds, as in amylose (**6**), the polymer twists such that it takes six residues to form a complete turn, which can assemble to form double helices.[15] Alternatively, if the residues are joined by β-1,3 linkages, as in paramylon (**7**), the polymer chains assemble to form triple helices with six residues per turn (Fig 1.4).[16]

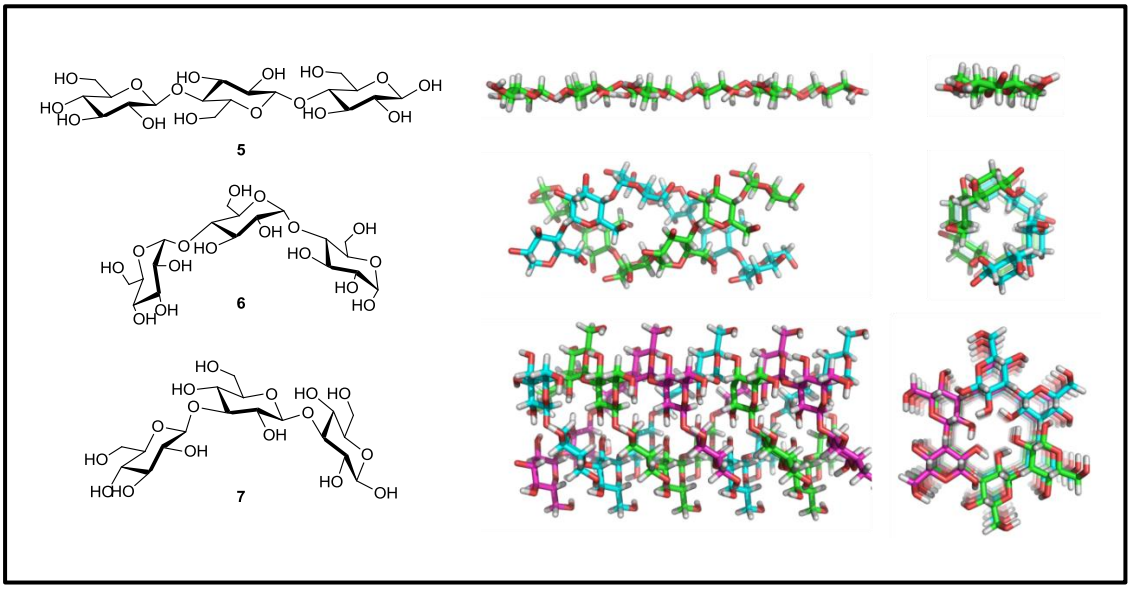


Figure 1.4: Comparison of glucan polymers

Three homo-polymers of glucose are common in nature and have remarkably different structures based on their linkage. Cellulose (**5**), a β-1,4-glucan, forms flat ribbons with the residues alternating orientation.[17] Amylose (**6**), an α-1,4-glucan, forms double helices with six residues per turn.[15] Paramylon (**7**), a β-1,3-glucan, forms triple helices with six residues per turn.[18] Structures obtained from PolySac3DB.[19]

1.2.3 Structural complexity provided by mixed linkages

Each of the different hydroxyl groups on monosaccharides can be linked to other sugars, potentially on more than one position, forming branched chains, as opposed to the exclusively linear chains of the other biopolymers, nucleic acids and proteins. β-1,3-Glucans are found throughout nature as linear chains, such as bacterial extracellular curdlan,[20] or the paramylon storage granules in *Euglena* (**7**).[21] They can also have various degrees of other linkages, which impart different molecular architectures (Fig 1.5). Kinks in the chains, such as the mixed linkage β-glucans from oat (**8**),[22] prevent the formation of the triple helices and branches, for example in laminarin (**9**), [23] ensure the polysaccharide remains soluble. One of the *Streptococcus* capsules has a β-1,2 linked glucose on every residue (**10**), giving a crowded, comb-like structure.[24]

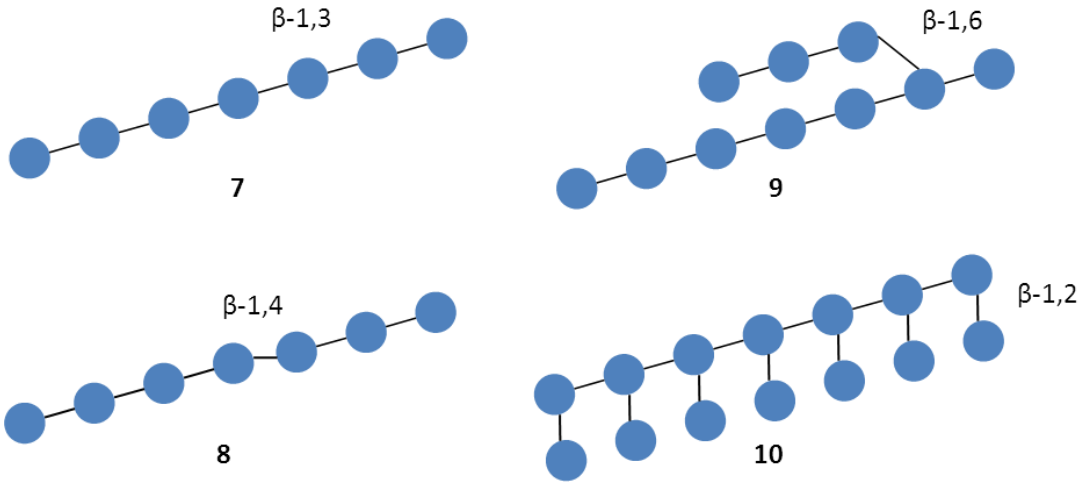


Figure 1.5: Examples of mixed linkage β-glucans

Paramylon (**7**), is a linear β-1,3-glucan.[21] Oat glucan (**8**) is a linear β-1,3-glucan which contains some β-1,4 linkages.[22] Laminarin (**9**) is a β-1,3-glucan which has β-1,6 branches.[23] *Streptococcus pneumonia* type S37 capsule (**10**) is composed of a β-1,3-backbone with β-1,2-glucoses on every residue.[24]
● Represents a glucose residue.

1.3 Natural synthesis and degradation of carbohydrates

The enzymes utilised in the natural synthesis and degradation of carbohydrates are categorised into families in the CAZy database.[25] They tend to have high specificity for their substrates, regulating the sugar and the linkage, though there are also some relatively promiscuous enzymes, such as the macrolide glycosyl transferases.[26]

1.3.1 Glycosyltransferases (GTs)

Glycosyltransferases catalyse the transfer of sugars from activated donor molecules to specific acceptor molecules (Fig 1.6). The sugars are typically activated as nucleotide phosphosugars, sugar phosphates or lipid diphosphosugars and the energy of the phospho-sugar bond is used to drive the reaction.[27] The acceptor molecule can be a carbohydrate or other molecules can accept the sugar to form glycoconjugates, including glycoproteins,[28] glycolipids,[29] and small molecules.[30]

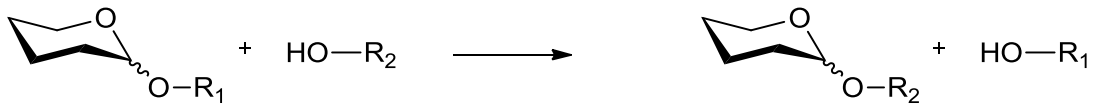


Figure 1.6: General scheme of glycosyl transfer

Glycosyltransferases catalyse the transfer of sugars from activated sugar donors, which are typically sugar phosphoesters (R_1), on to acceptor molecules (R_2), which maybe proteins, lipids, small molecules or other sugars. Often this transfer is on to oxygen, but transfer on to nitrogen, sulphur and carbon are also known.

1.3.2 Glycoside hydrolases (GHs)

Glycoside hydrolases hydrolyse the glycosidic bond between carbohydrate moieties (Fig 1.7).[31] If a molecule other than water can intercept the intermediate a transglycosidation reaction can take place. For example xyloglucan endotransferase shows no glycosidase activity, and instead transfers an oligosaccharide on to another polymer to form a new glycosidic linkage.[32]

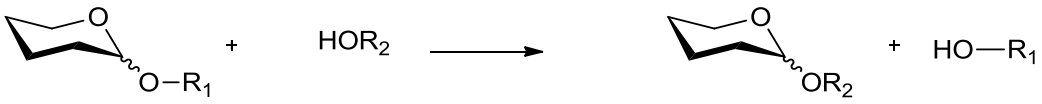


Figure 1.7: General scheme of glycoside hydrolysis

Glycoside hydrolases cleave glycosidic linkages between sugars and another molecule (R_1), transferring the sugar on to water (R_2=H) resulting in hydrolysis. If a molecule other than water ($R_2 \neq H$) can intercept the reaction then transglycosylation occurs.

1.3.3 Polysaccharide lyases

Polysaccharide lyases cleave sugars by deprotonation of glycosides, resulting in unsaturated sugars (Fig 1.8). This is particularly important in plants, since pectin, a major plant cell wall component, is composed of galacturonic acid, and is susceptible to these enzymes.[33,34]

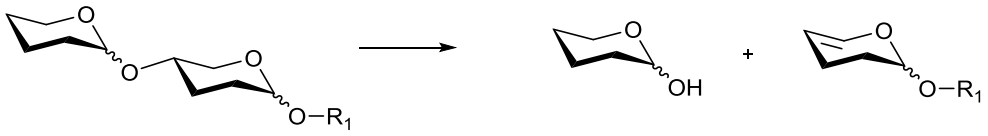


Figure 1.8: General scheme of polysaccharide lyases

Polysaccharide lyases cleave the glycosidic bond between sugars, releasing unsaturated sugars.

1.3.4 Auxiliary activities

Recently auxiliary activities, a new category of enzymes acting on carbohydrates was added to the CAZy categories.[35] It includes polysaccharide mono-oxygenases and lignin degrading enzymes.

1.4 Carbohydrate space

A huge diversity of carbohydrates is made in nature, with different properties and structures.

1.4.1 Storage polysaccharides

Energy is stored in mammalian and bacterial cells in the form of glycogen, a highly branched, soluble α-1,4/1,6-glucan.[36] The random distribution and high degree of branching ensures both energy storage capacity and accessibility to degradative enzymes. Plants also store energy in the form of α-1,4-glucans, but the highly regulated spacing of the α-1,6-branching gives an ordered, insoluble starch granule.[15]

In certain species of alga β-1,3-glucans are used as the major energy storage molecules, which can be strictly linear, as in paramylon, or branched as in laminarin.[37] Instead of polymers of glucose, some plants instead store other polysaccharides, such as polymers of fructose, including the fructans produced by *Agave*,[38] or glucomannan produced by the Voodoo Lily.[39]

1.4.2 Extra-cellular carbohydrates

Mammalian cells have flexible surfaces, which contact each other via carbohydrates, allowing communication and control of cellular shapes.[40] These surface glycans are highly specific for cell type and can be used as biomarkers for disease states such as cancer.[2] Beyond this surface layer is a region of amorphous polysaccharide, mostly composed of negatively charged glycosaminoglycans, often with specific modifications such as addition of sulphate on heparans.[41]

Plants, on the other hand, have a rigid cell wall which provides structural integrity and strength. Cellulose makes up the largest proportion of most cell walls, composed of crystalline linear β-1,4-glucan chains.[42] Insect exoskeletons, crustacean shells and some fungal cell walls are also made of a crystalline β-1,4 polymer, although this chitin is made of GlcNAc.[43] The cellulose in plant cell walls is cross linked with hemicelluloses, which are complex polysaccharides, including xylans, mannans and arabinans.[44] These can be used industrially, including Guar Gum,[45] a galactomannan, and

pectin, containing a mixture of substituted poly-galacturonic acids,[46] for example as food additives, in pharmaceutical manufacture and in the oil industry. Plant cell walls also contain charged pectins, based on a polygalacturonic acid backbone, which bind calcium, causing cross-linking and aggregation that hold the molecules in a gel.[47] When the carboxyl groups are methyl esterified, which occur during fruit ripening, they are not able to form these calcium bridges and the cell wall loses some of its structure. These pectins also have complex branching such as in Rhamnogalacturonan II, an extremely intricate, highly conserved structure, which forms borate cross links, holding the plant cell wall together (Fig 1.9).

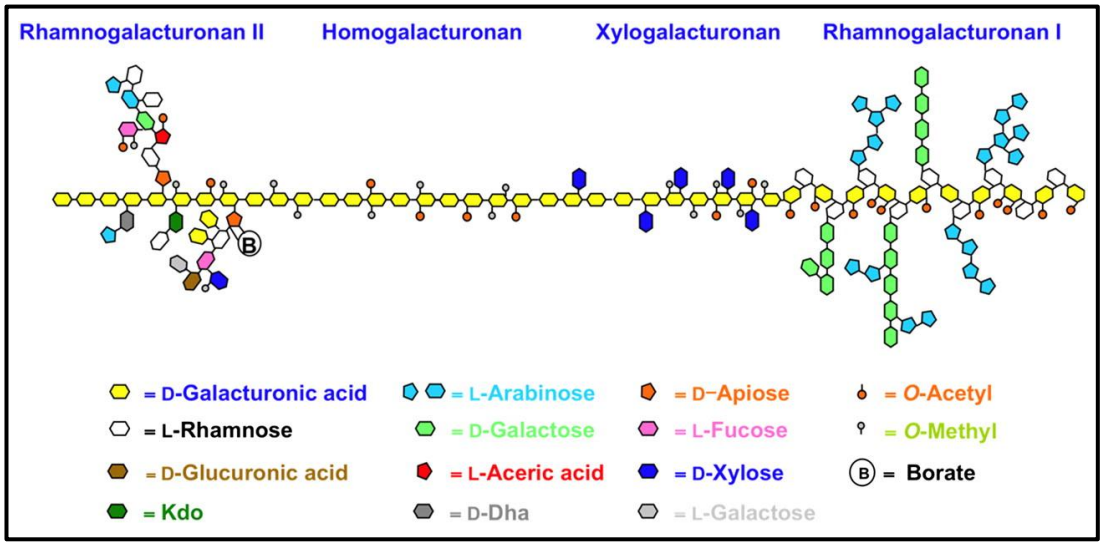


Figure 1.9: Schematic structure of pectin from plant cell wall

Pectin is composed of four different polymers which are not in equal ratios and may not actually be covalently linked. Reproduced with permission from the American Society of Plant Biologists. Copyright (2010).[46]

1.5 Conclusions

Carbohydrates are extremely complex in their structure, with various building blocks, linkages and covalent modifications. Their structure is not encoded directly in the genome, making analysis challenging. Polysaccharides are used in diverse industries from oil drilling to food manufacture.[48] In order to make the greatest use of this class of macromolecule further research is needed, both to study natural compounds and to engineer specific new materials.

1.6 References

1 Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3**, 97-130 (1993).

2 An, H. J., Kronewitter, S. R., de Leoz, M. L. A. & Lebrilla, C. B. Glycomics and disease markers. *Curr. Opin. Chem. Biol.* **13**, 601-607 (2009).

3 Antonopoulos, A., North, S. J., Haslam, S. M. & Dell, A. Glycosylation of mouse and human immune cells: insights emerging from N-glycomics analyses. *Biochem. Soc. Trans.* **39**, 1334-1340 (2011).

4 Oxley, D. & Bacic, A. Structure of the glycosylphosphatidylinositol anchor of an arabinogalactan protein from *Pyrus communis* suspension-cultured cells. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14246-14251 (1999).

5 Goetz, S. *The molecular basis of plant cell wall oligosaccharide formation* Doctor of Philosophy thesis, University of East Anglia, (2012).

6 Morelli, L., Poletti, L. & Lay, L. Carbohydrates and immunology: synthetic oligosaccharide antigens for vaccine formulation. *Eur. J. Org. Chem.* **2011**, 5723-5777 (2011).

7 Del Valle, E. M. M. Cyclodextrins and their uses: a review. *Process Biochem.* **39**, 1033-1046 (2004).

8 Eichhorn, S. J. Cellulose nanowhiskers: promising materials for advanced applications. *Soft Matter* **7**, 303-315 (2011).

9 Guerrini, M. *et al.* Oversulfated chondroitin sulfate is a contaminant in heparin associated with adverse clinical events. *Nature Biotechnol.* **26**, 669-675 (2008).

10 Werz, D. B. *et al.* Exploring the structural diversity of mammalian carbohydrates ("Glycospace") by statistical databank analysis. *ACS Chem. Biol.* **2**, 685-691 (2007).

11 Wearne, K. A., Winter, H. C., O'Shea, K. & Goldstein, I. J. Use of lectins for probing differentiated human embryonic stem cells for carbohydrates. *Glycobiology* **16**, 981-990 (2006).

12 Marlow, A. L. & Kiessling, L. L. Improved chemical synthesis of UDP-galactofuranose. *Org. Lett.* **3**, 2517-2519 (2001).

13 De Lederkremer, R. M. & Colli, W. Galactofuranose-containing glycoconjugates in Trypanosomatids. *Glycobiology* **5**, 547-552 (1995).

14 Langan, P., Nishiyama, Y. & Chanzy, H. A revised structure and hydrogen-bonding system in cellulose II from a neutron fiber diffraction analysis. *J. Am. Chem. Soc.* **121**, 9940-9946 (1999).

15 Imberty, A., Chanzy, H., Pérez, S., Bulèon, A. & Tran, V. The double-helical nature of the crystalline part of A-starch. *J. Mol. Biol.* **201**, 365-378 (1988).

16 Deslandes, Y., Marchessault, R. H. & Sarko, A. Triple-helical structure of (1->3)-β-D-glucan. *Macromolecules* **13**, 1466-1471 (1980).

17 Nishiyama, Y., Sugiyama, J., Chanzy, H. & Langan, P. Crystal structure and hydrogen bonding system in cellulose 1α, from synchrotron X-ray and neutron fiber diffraction. *J. Am. Chem. Soc.* **125**, 14300-14306 (2003).

18 Chuah, C. T., Sarko, A., Deslandes, Y. & Marchessault, R. H. Packing analysis of carbohydrates and polysaccharides. Part 14. Triple-helical crystalline structure of curdlan and paramylon hydrates. *Macromolecules* **16**, 1375-1382 (1983).

19 Anita, S. & Serge, P. PolySac3DB: an annotated data base of 3 dimensional structures of polysaccharides. *BMC Bioinformatics* **13**, 302-302 (2012).

20 Harada, T., Misaki, A. & Saito, H. Curdlan: a bacterial gel-forming β-1,3-glucan. *Arch. Biochem. Biophys.* **124**, 292-298 (1968).

21 Clarke, A. E. & Stone, B. A. Structure of the paramylon from *Euglena gracilis*. *Biochim. Biophys. Acta* **44**, 161-163 (1960).

22 Aspinall, G. O. & Carpenter, R. C. Structural investigations on the non-starchy polysaccharides of oat bran. *Carbohydr. Polym.* **4**, 271-282 (1984).

23 Anderson, F. B., Hirst, E. L., Manners, D. J. & Ross, A. G. The constitution of laminarin. Part III. The fine structure of insoluble laminarin. *J. Chem. Soc.*, 3233-3243 (1958).

24 Adeyeye, A., Jansson, P.-E., Lindberg, B. & Henrichsen, J. Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* type 37. *Carbohydr. Res.* **180**, 295-299 (1988).

25 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-D238 (2009).

26 Yang, M. *et al.* Probing the breadth of macrolide glycosyltransferases: *in vitro* remodeling of a polyketide antibiotic creates active bacterial uptake and enhances potency. *J. Am. Chem. Soc.* **127**, 9336-9337 (2005).

27 Zechel, D. L. & Withers, S. G. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11-18 (2000).

28 Kelleher, D. J. & Gilmore, R. An evolving view of the eukaryotic oligosaccharyltransferase. *Glycobiology* **16**, 47R-62R (2006).

29 Acquotti, D. *et al.* 3-Dimensional structure of the oligosaccharide chain of GM1 ganglioside revealed by a distance-mapping procedure - a rotating and laboratory frame nuclear overhauser enhancement investigation of native glycolipid in dimethyl-sulfoxide and in water dodecylphosphocholine solutions. *J. Am. Chem. Soc.* **112**, 7772-7778 (1990).

30 Bowles, D., Lim, E.-K., Poppenberger, B. & Vaistij, F. E. Glycosyltransferases of lipophilic small molecules. *Annu. Rev. Plant Biol.* **57**, 567-597 (2006).

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

32 Nishitani, K. & Tominaga, R. Endoxyloglucan transferase, a novel class of glycosyltransferase that catalyzes transfer of a segment of xyloglucan molecule to another xyloglucan molecule. *J. Biol. Chem.* **267**, 21058-21064 (1992).

33 Linhardt, R. J., Galliher, P. M. & Cooney, C. L. Polysaccharide lyases. *Appl. Biochem. Biotechnol.* **12**, 135-176 (1986).

34 Sutherland, I. W. Polysaccharide lyases. *FEMS Microbiol. Rev.* **16**, 323-347 (1995).

35 Levasseur, A., Drula, E., Lombard, V., Coutinho, P. & Henrissat, B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnology for Biofuels* **6**, 41 (2013).

36 Roach, P. J., Depaoli‐Roach, A. A., Hurley, T. D. & Tagliabracci, V. S. Glycogen and its metabolism: some new developments and old themes. *Biochem. J.* **441**, 763-787 (2012).

37 Painter, T. J. Structural evolution of glycans in algae. *Pure Appl. Chem.* **55**, 677-694 (1983).

38 Lopez, M. G., Mancilla-Margalli, N. A. & Mendoza-Diaz, G. Molecular structures of fructans from *Agave tequilana* Weber var. azul. *J. Agric. Food. Chem.* **51**, 7835-7840 (2003).

39 Katsuraya, K. *et al.* Constitution of konjac glucomannan: chemical analysis and 13C NMR spectroscopy. *Carbohydr. Polym.* **53**, 183-189 (2003).

40 Varki, A. *et al. The Essentials of Glycobiology*. (Cold Spring Harbor Laboratory Press, 2009).

41 Esko, J., Kimata, K. & Lindahl, U. in *Essentials of Glycobiology* (eds A. Varki *et al.*) Ch. 16, (Cold Spring Harboour Laboratory Press, 2009).

42 Delmer, D. P. & Amor, Y. Cellulose biosynthesis. *Plant Cell* **7**, 987-1000 (1995).

43 Saito, Y., Okano, T., Chanzy, H. & Sugiyama, J. Structural study of α chitin from the grasping spines of the arrow worm (*Sagitta* spp.). *J. Struct. Biol.* **114**, 218-228 (1995).

44 Popper, Z. A. *et al.* Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu. Rev. Plant Biol.* **62**, 567-590 (2011).

45 Srivastava, M. & Kapoor, V. P. Seed galactomannans: an overview. *Chem. Biodiversity* **2**, 295-317 (2005).

46 Harholt, J., Suttangkakul, A. & Vibe Scheller, H. Biosynthesis of pectin. *Plant Physiol.* **153**, 384-395 (2010).

47 Jarvis, M. C. Structure and properties of pectin gels in plant-cell walls. *Plant Cell Environ.* **7**, 153-164 (1984).

48 Jobling, S. Improving starch for food and industrial applications. *Curr. Opin. Plant Biol.* **7**, 210-218 (2004).

Chapter 2 – *In vitro* synthesis of polysaccharides

Polysaccharides are very complex in structure and their synthesis, for use in industry and in the understanding of Natural materials, is highly challenging. Over the years a number of chemical, enzymatic and chemoenzymatic strategies have been developed to synthesise defined oligosaccharides,[1] which will be summarised in this chapter.

2.1 Chemical synthesis

It is possible to synthesise short oligosaccharide sequences chemically but this requires a great deal of time, effort and skill. Rarely is there integration of the synthesis and application of oligosaccharides to a biological system. It is very difficult to synthesise longer oligosaccharides. Instead of stepwise syntheses, polymerisation type reactions are generally performed, to create repeating motifs.

2.1.1 Polycondensation

Polycondensation reactions, in which the alcohol promoted loss of two leaving groups unveils another alcohol, can be used to make long, polydisperse carbohydrate chains. However it is difficult to produce the desired control over stereochemistry. For example the condensation of 1,2,5-orthobenzoylacetal arabinose produced a linear arabinan with 92% 1,5- and 8% 1,2-linkages (Fig 2.1).[2]

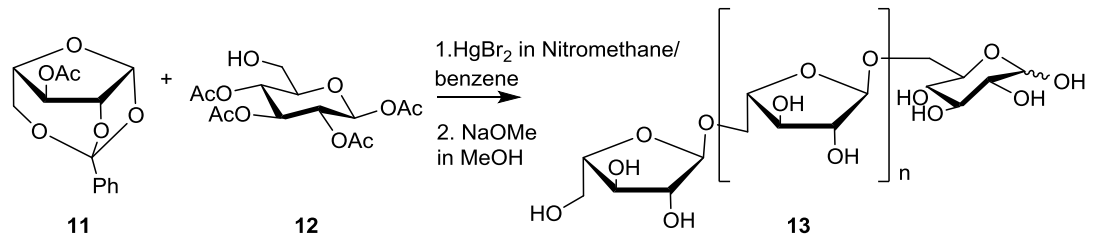
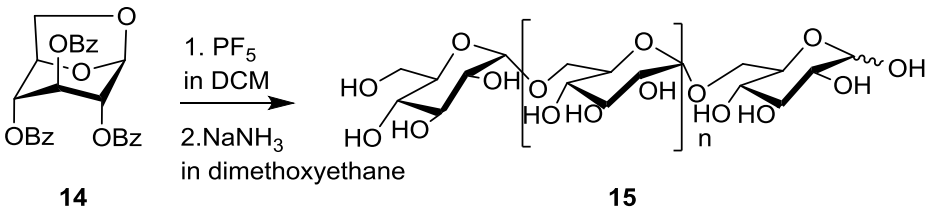


Figure 2.1: Polycondensation

The polycondensation of 1,2,5-orthobenzoylarabinofuranose (**11**) is initiated by the hydroxyl of the glucose tetraacetate (**12**) and produces predominantly 1,5-linkages (**13**).[2]

2.1.2 Ring-opening polymerisation

Ring opening polymerisation, with attack into the back of an activated ring, can be difficult to make specific for linkage and ring forms. With extensive work, it can be refined to produce both natural and unnatural polysaccharides. For example 1,6-anhydroglucose can be stereospecifically opened by phosphorous pentafluoride to form an artificial dextran, with exclusively α-linkages and an average DP of 200-225 glucose residues (Fig 2.2).[3]



14 **15**

Figure 2.2: Ring-opening polymerisation

The 1,6-anhydro ring of glucose-tribenzoate (**14**) is held in the β-position by the axial positioning of C6. The ring is then opened by attack into the back of the anomeric carbon by the anhydro ring, forming the α-1,6-glucan (**15**).[3]

2.1.3 Solid phase synthesis

Historically the synthesis of peptides and oligonucleotides faced similar problems of yield and specificity as is faced in carbohydrate synthesis; however these problems have been ameliorated by the use of automated solid-phase synthesis. An analogous approach has been developed to make oligosaccharide synthesis as routine as DNA synthesis.[4] This is achievable, despite the huge range of potential carbohydrates possible from the 10 monosaccharides present in mammalian cells, as it has been calculated that to synthesise 80% of mammalian oligosaccharides just 13 modified sugar donors are necessary.[5] However, when plant and bacterial carbohydrates are considered the number of linkages, and hence required building blocks, dramatically increases. Seeberger has constructed the first experimental oligosaccharide synthesizer able to create more complex oligosaccharides, with a larger variety of linkages, much faster than standard synthesis.[6] There has been criticism from the synthetic community as the system uses a brute force approach to the reactions, without allowing some key linkages or functional groups. Most importantly the methods fail to provide the complete anomeric control desired, one of the key aspects of carbohydrate structure.[5]

2.2. *In vitro* enzymatic synthesis of polysaccharides

In nature glycoside synthesis is carried out by glycosyltransferases and transglycosidase, and the cleavage of glycosidic bonds is performed by glycoside hydrolases and polysaccharide lyases. These have been thoroughly categorised in the CAZy database.[7] Theoretically, all the enzymes necessary to make all naturally occurring glycans are available from Nature, and could be exploited *in vitro*. However, polysaccharide synthesis is highly complicated and uses an enormous array of enzymes. Carbohydrate active enzymes make up 3% of the *Arabidopsis* genome[8] and all 1167 of these enzymes are needed to make the structures in just this one organism.

Carbohydrate active enzymes have exquisite control over the anomeric configuration of the monosaccharides and the linkages involved, providing an alternative to synthetic chemistry for precise carbohydrate synthesis. Enzymatic reactions are in principle reversible, though this is often not seen in Nature. For example, hydrolases act in such an abundance of water that no reverse reaction occurs in nature. This bi-directionality could potentially foil attempts to use enzymes, as any synthesis is offset by the degradation of the product. The equilibrium must be forced into the direction of synthesis, which can be inefficient and wasteful of starting material. Wild-type enzymes can be used under artificial conditions to force the synthetic reaction, or enzymes can be modified to prevent the undesired reaction.[9]

2.2.1 Glycoside hydrolases

Glycoside hydrolases hydrolyse the glycosidic bond between carbohydrate moieties in an acid/base attack, which can be inverting, if direct attack is by water, or retaining, if proceeding via double displacement and a covalent sugar-enzyme intermediate (Fig 2.3).[10]

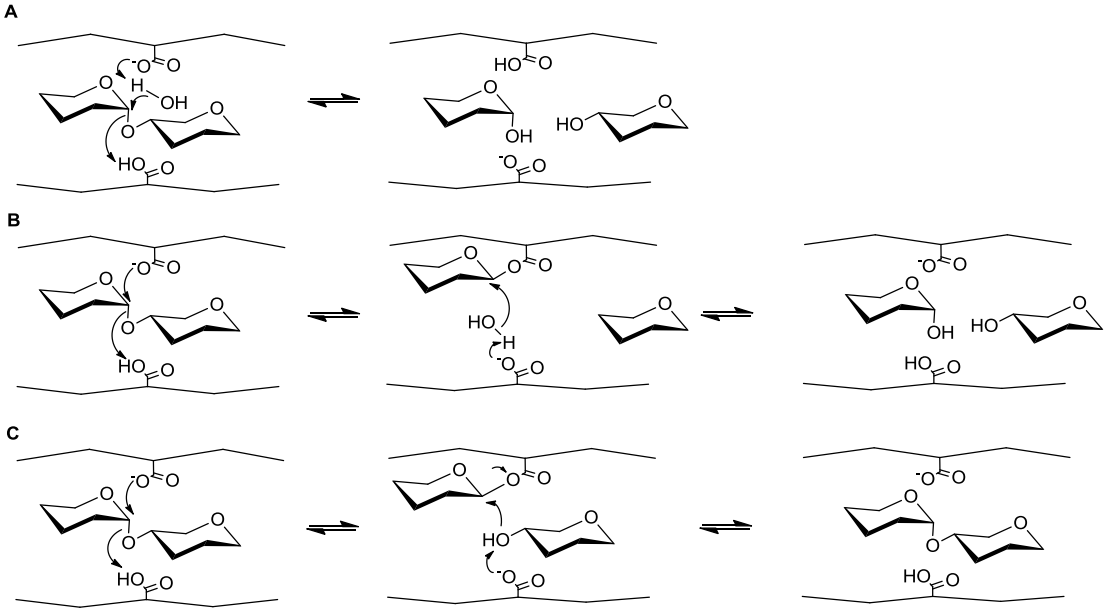


Figure 2.3: Mechanisms of glycoside hydrolases

A. The catalytic base of inverting hydrolases deprotonates a water molecule and the hydroxyl attacks the anomeric carbon. **B**. In retaining hydrolases the enzyme nucleophile attacks the anomeric carbon, forming an acyl-enzyme intermediate. The catalytic base deprotonates a water molecule, which then attacks the acyl-enzyme bond, inverting the anomeric centre back to the original conformation. **C**. In transglycosidases an enzyme-sugar intermediate is attacked by another sugar to form a glycosidic bond.

Glycosidases can be used to synthesise polysaccharides by preventing hydrolysis, either using very high concentrations of the donor saccharide, so that water is outcompeted, or by using activated sugar donors so that the glycosyl enzyme intermediate may react with the acceptor molecule faster than with water. Cellulase, for example, can be used to form artificial cellulose by condensation of β-cellobiosyl fluoride in organic solvent buffer (Fig 2.4.A).[11] The fluorine can readily leave the molecule, making the reaction proceed in the synthetic direction. The β-configuration of the fluorine is essential for activity and the linkage of the product is precisely controlled by the retaining mechanism of the enzyme. The organic solvent is used to exclude water and thus to prevent the possibility of the hydrolysis reaction. The purity of the cellulase used had a marked effect on the structure of the cellulose macromolecular structure.[12] A two enzyme strategy can be used to control the length pf the polymer formed. For example, repeated extension with β-lactosyl fluoride followed by de-galactosylation using β-galactosidase, allows only one glucose addition per cycle (Fig 2.4.B).[13]

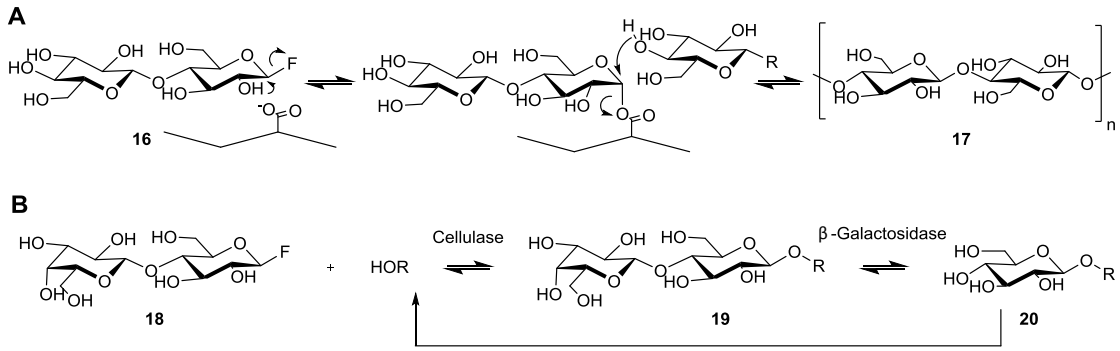


Figure 2.4: Chemoenzymatic syntheses of cellulose

A. Cellulase can readily remove the fluorine from cellobiosyl fluoride (**16**) and allow attack by the 4OH of a β-glucoside (R= $(\beta\text{-}1,4\text{-}Glc)_n$) of the enzyme-sugar intermediate to form the β-1,4-glucan product (**17**). **B**. Cellulase can also transfer lactose from lactosyl fluoride (**18**). The product (**19**) is not an acceptor for a further transfer, until it has been de-galactosylated using β-galactosidase to uncover the glucosyl residue (**20**), allowing complete control of the product length.

For the polymerisation of sugars which have an *N*-acetyl group at the 2-position, oxazoline rings can be used as the leaving group, with the opening of the ring forming the *N*-acetyl group.[14] For example, artificial chitin was synthesised using chitinase to open an oxazoline derivative of the disaccharide, controlling formation of the β-anomer (Fig 2.5).[15] Crucially the reaction proceeded well at pH 10.6, whilst the hydrolysis reaction was prevented.

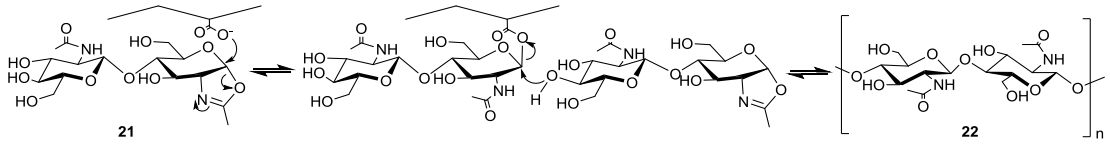


Figure 2.5: Chemoenzymatic synthesis of chitin

Chitin, a polymer of *N*-acetyl glucosamine, can be made by the opening of chitobiose oxazoline derivative (**21**) by hydrolytic enzymes, followed by attack of the polymer chain to form artificial chitin (**22**).[15]

To overcome the problems of hydrolysis using natural enzymes Withers introduced the concept of artificial "glycosynthases" (Fig 2.6).[16] A glycosynthase is a hydrolase in which the catalytic nucleophile is mutated to prevent the hydrolytic degradation of the transglycosidation products. The activated glycosyl donors are fluorinated compounds, which are readily synthesised using standard procedures. This technique uses cheaper starting materials and absolutely controls the configuration of the bond formed, whilst preventing product degradation.

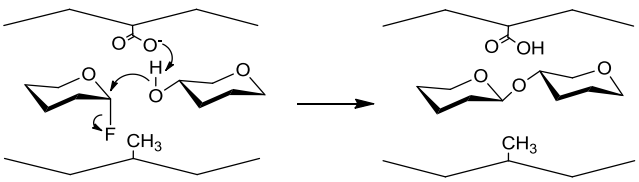


Figure 2.6: Glycosynthases

If the catalytic nucleophile of a hydrolase is mutated then the enzyme cannot attack the sugar. Instead another sugar can be deprotonated by the catalytic base and then attack the activated sugar, causing release of the stable leaving group (F^-) and formation of a new glycosidic bond. [16]

2.2.2 Glycosyltransferases

Glycosyltransferases catalyse the transfer of sugars from activated nucleotide phosphosugar, sugar phosphate or lipid diphosphosugar donor molecules to specific acceptor molecules.[17] The reactions are well characterised as either retaining or inverting (Fig 2.7).[7]

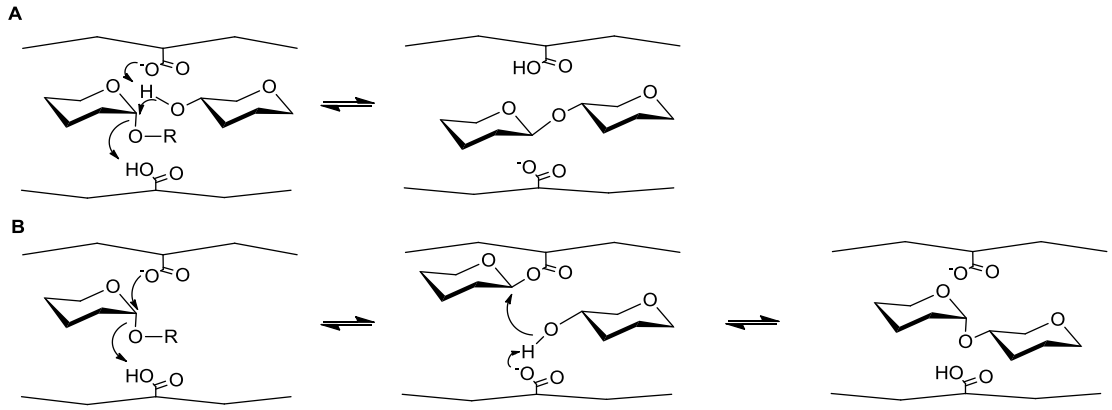


Figure 2.7: Mechanisms of glycosyltransferases

A. In inverting glycosyltransferases, direct attack by the deprotonated acceptor on the sugar nucleotide leads to inversion of the anomeric centre. **B**. Nucleophilic attack by retaining glycosyltransferases on the sugar phophoester-bond forms an acyl-enzyme intermediate. The acceptor is then activated by the catalytic base and attacks the sugar-enzyme bond forming a new glycosidic linkage. R = phosphoester such as a nucleotide diphosphate or a lipid phosphate.

Nucleotide phosphosugar cleavage, during glycosyltransferase activity, is strongly favoured, such that these enzymes show very little reverse reaction.[18] However, if a high energy glycoside is used the enzymes can be driven in reverse to efficiently synthesise sugar nucleotides.[19]

Unfortunately, glycosyltransferases are mostly integral membrane proteins (>70% in *Arabidopsis* and >85% in humans) and so are not accessible for *in vitro* reactions. Additionally, the activated glycosyl donors are often expensive and unstable, and the glycosyltransferases have a tendency towards product inhibition. To combat inhibition by the nucleotide released from glycosyl transfer, a phosphatase can be used to hydrolyse the nucleotide, which provides a large increase in yield (Fig 2.8.A).[20]

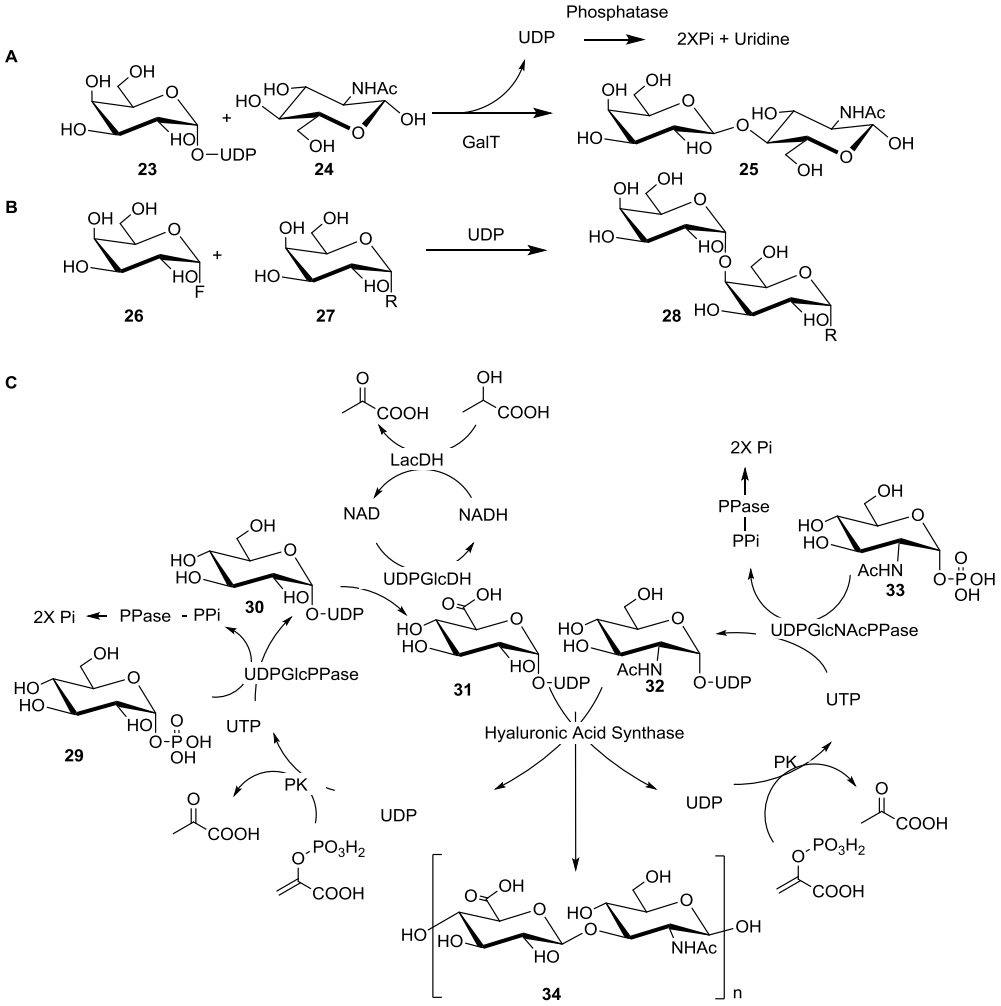


Figure 2.8: Enhancing the yields of glycosyltransferases

A. Addition of a phosphatase to the transfer of Gal from UDP-Gal (**23**) on to GlcNAc (**24**) improved the overall yield of LacNAc (**25**) from below 60% to over 90%.[20] **B**. Catalytic amounts of UDP can be used to drive the transfer of galactose from galactosyl fluoride (**26**) on to a galactose acceptor (**27**) to form the α-1,4-galactan (**28**).[21] **C**. Hyaluronic acid (**34**) was made by generating the sugar nucleotides *in situ*. The UDP-GlcNAc (**32**) was generated from UTP and GlcNAc-1-P (**33**), using GlcNAc pyrophosphatase (GlcNAcPPAse), driven by the degradation of the released pyrophosphate, by inorganic pyrophosphatase (PPAse). The UDP-GlcA (**31**) was made by first generating UDP-Glc (**30**) from Glc-1-P (**29**) and UTP, using UDP-Glc pyrophosphatase (UDPGlcPPAse), and then oxidising, using NAD and UDP-Glc dehydrogenase (UDPGlcDH). The NAD was regenerated using lactate dehydrogenase (LacDH) and pyruvate, and the UTP was regenerated by pyruvate kinase (PK), converting phosphoenolpyruvate into pyruvate. [22]

Alternatively catalytic quantities of UDP can be used to prevent the product inhibition, and also to reduce the use of expensive substrates. This technique has been used to enhance the transfer of galactose form the anomeric fluoride on to galactosides by α-galactosyltransferase, with decreased use of nucleotide and reduced inhibition (Fig 2.8.B).[21] Hyaluronic acid has been prepared using hyaluronic acid synthase and a series of other enzymes to regenerate sugar nucleotides, using energy derived from phosphoenolpyruvate (Fig 2.8.C).[22] These regenerative enzymes reduce the requirement for expensive starting materials and greatly increased the yield from 20% to 90%.

Although some glycosyltransferases can be relatively non-specific towards donor and substrate, such as the natural product glycosyltransferases,[23] many have strict donor specificity which limits the carbohydrates that can be synthesised. The human ABO blood group antigen is a carbohydrate found on the surface of many cells, notably red blood cells. The A antigen terminates in an *N*-acetyl-galactosamine, whilst the B type terminates in a galactose, both linked α-1,3 to the same core antigen (Fig 2.9). This difference in specificity is determined by four amino acids in the glycosyltransferase and exchange of these converted the antigen produced.[24] Glycosyltransferase specificity can readily be altered or relaxed using molecular biology, opening up the range of potential oligosaccharides that can be synthesised.[25] Further studies of these enzymes should enhance their use in the chemoenzymatic synthesis of carbohydrates.

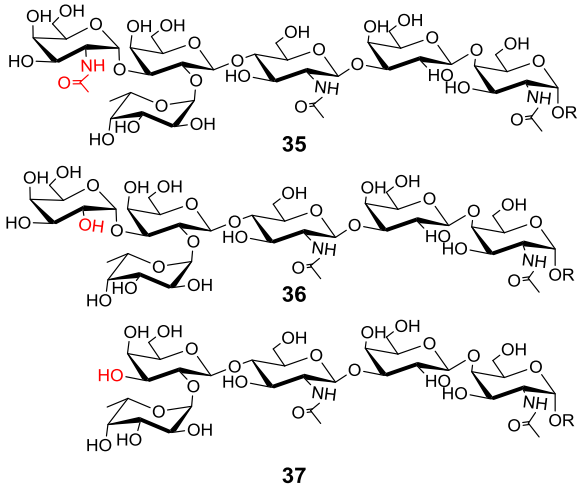


Figure 2.9: Human ABO blood group antigens

The A antigen (**35**) has GalNAc, whilst the B antigen (**36**) has Gal, and the O antigen (**37**) has no additional monosaccharide, on the 3 hydroxyl of the H antigen core.

2.2.3 Polysaccharide lyases

Polysaccharide lyases cleave sugars by deprotonation of glycosides, resulting in unsaturated sugars. The most common mode of action is the β-elimination reaction on the C5 of a six carbon carboxylate sugar, releasing a 4,5-unsaturated sugar (Fig 2.10.A). A neutral polysaccharide lyase has also been characterised, which acts on α-1,4-glucans with a similar mechanism to the inverting hydrolases (Fig 2.10.B).[26] Instead of deprotonating a water molecule to attack the sugar-enzyme intermediate, it is proposed that H2' is removed and a double bond formed which eliminates the enzyme-sugar bond, forming 1,5-anhydro-fructose.

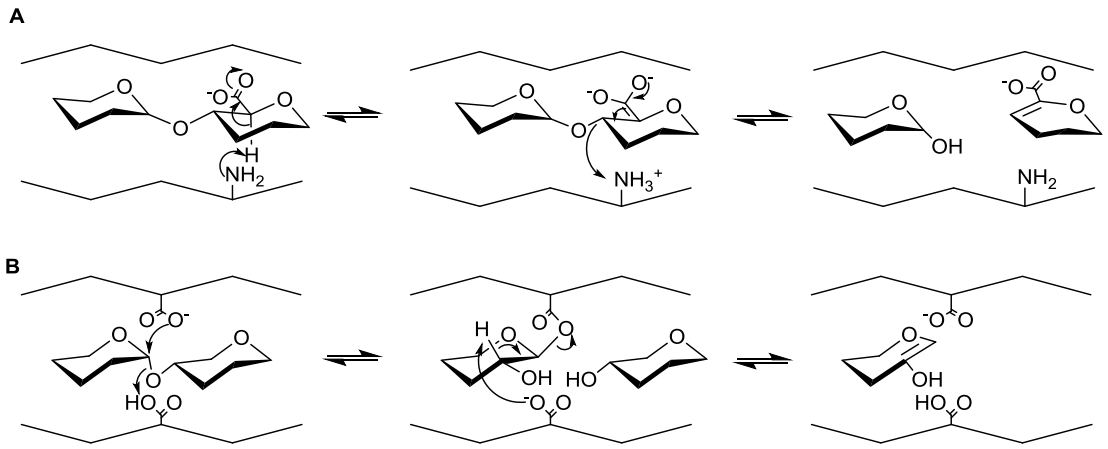


Figure 2.10: Mechanisms of polysaccharide lyases

A. In uronic acid lyases the catalytic base removes the proton alpha to the carboxyl, forming an enolate. When this reverts to the carbonyl a double bond is formed between carbons four and five, and the glycosidic bond is eliminated in an E_1cB type reaction. **B**. In α-1,4-glucan lyases a sugar-enzyme bond is formed, in a similar way to retaining hydrolases, and the bond is eliminated by the formation of a double bond, with loss of the proton on C2.

Polysaccharide lyases are used in biotechnology to remove or remodel polysaccharides and to release specific oligosaccharides.[27] There is no current use in the synthesis of polysaccharide by reversing the enzyme reaction, though in principle this is possible.

Linkage	Name	CAZy Family	Type	Polymer length	EC	Ref
α,α-1,1	Trehalose	GH65	I	2	2.4.1.64	28
α,α-1,1	Trehalose-6-P	GH65	I	2	2.4.1.216	29
α,α-1,1	Trehalose	GT4	R	2	2.4.1.231	30
α-1,2	Kojibiose	GH65	I	2+	2.4.1.230	31
α,α-1,2	Sucrose	GH13	R	2	2.4.1.7	32
β-1,2	β-1,2-Glucan	GH94	I	NA	2.4.1.-	33
α-1,3	Nigerose	GH65	I	2	2.4.1.279	34
α-1,3	Glucosyl-rhamnose	GH65	I	2	2.4.1.282	35
β-1,3	Laminaribiose	GH94	I	2	2.4.1.31	36,37
β-1,3	Laminarin	???	I	2+	2.4.1.30	38
β-1,3	β-1,3-Glucan	???	I	2+	2.4.1.97	39,40
β-1,3	Galactosyl-HexNAc	GH112	I	2	2.4.1.211	41,42
β-1,3	Galactosyl-rhamnose	GH112	I	2	2.4.1.-	43
α-1,4	Glycogen/ starch	GT35	R	4+	2.4.1.1	44,45
α-1,4	Maltose	GH65	I	2	2.4.1.8	46
α-1,4	Maltosyl	GH13	R	6+	2.4.1.169	47
β-1,4	Cellobiose	GH94	I	2	2.4.1.20	48
β-1,4	Cellodextrin	GH94	I	2+	2.4.1.49	49
β-1,4	Mannosyl-glucose	GH130	I	2	2.4.1.281	50
β-1,4	Mannan	GH130	I	2+	2.4.1.-	51
β-1,4	Chitobiose	GH94	I	2	2.4.1.280	52

Table 2.1: Classification of phosphorylases so far characterised

Type R = retaining, I = Inverting. NA = has not been defined

2.3 The natural role of phosphorylases

Of particular interest in the context of this thesis are the glycoside phosphorylases. Phosphorylases have the required stereo- and regioselectivities to allow their use in synthesising specific glycosyl linkages. Exploiting the reverse reaction of phosphorylases can be utilised to synthesise polysaccharides[53] and the sugar-phosphates are both more stable and more readily attained than nucleotide phosphosugars.

The currently characterised phosphorylases are soluble glycosyltransferases which catalyse the reversible addition of phosphate across the glycosidic bond (phosphorolysis), in the following reaction:[54]

$$(\text{Glycose})_{n+1} + \text{Orthophosphate} \rightleftharpoons (\text{Glycose})_n + \text{Glycose-1-P}$$

Phosphorylases have been found for many diverse linkages but most act on glucosyl residues (Table 2.1). Replacement of the inorganic phosphate with arsenate releases an unstable sugar-arsenate, which readily decomposes to the free monosaccharide.[55]

The mechanism can be inverting or retaining, which is determined by the enzyme's reaction mechanism and CAZy family. The configurations of the linkage and sugar phosphate produced are related by this interconversion. A range of specific phosphorylases have been identified[54] and there is now a characterised diglucose phosphorylase for every linkage except β-1,1 and α- or β-1,6 (Fig 2.11).

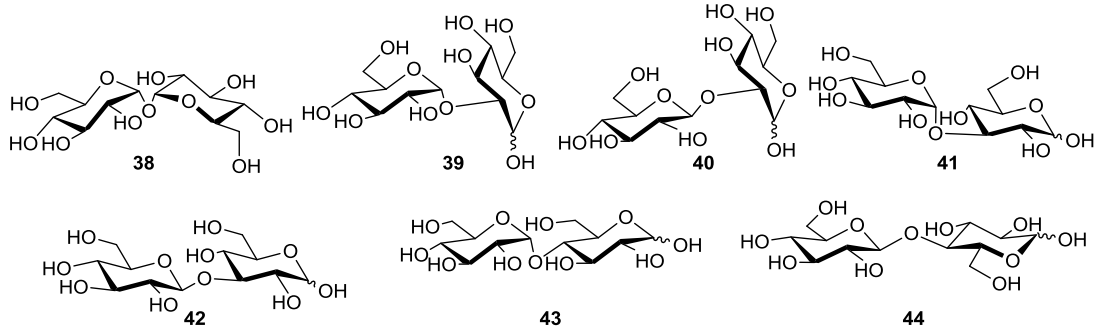


Figure 2.11: Diglucose linkages for which phosphorylases are found in Nature

Trehalose (α,α-1,1-glucosyl-glucose) (**38**). Kojibiose (α-1,2-glucosyl-glucose) (**39**). Sophorose (β-1,2-glucosyl-glucose) (**40**). Nigerose (α-1,3-glucosyl-glucose) (**41**). Laminaribiose (β-1,3-glucosyl-glucose) (**42**). Maltose (α-1,4-glucosyl-glucose) (**43**). Cellobiose (β-1,4-glucosyl-glucose) (**44**).

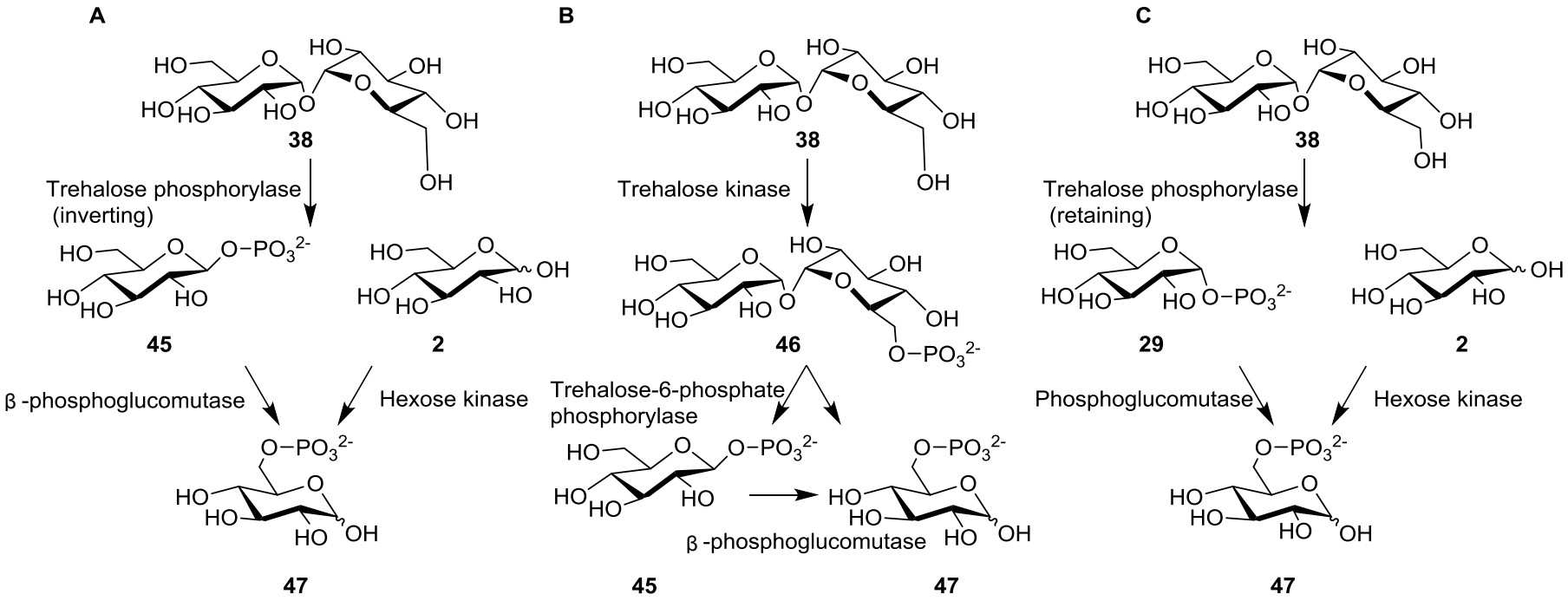


Figure 2.12: Trehalose metabolism using phosphorylases

A. Trehalose (**38**) is phosphorolysed by inverting phosphorylases from GH65 to form β-Glc-1-P (**45**) and Glc (**2**) which are converted into Glc-6-P (**47**) by β-phosphoglucomutase and hexokinase respectively.[56] **B**. Trehalose (**38**) can be phosphorylated to trehalose-6-phosphate (**46**) by trehalose kinase before being phosphorolysed by an inverting phosphorylase yielding Glc-6-P (**47**) and β-Glc-1-P (**45**), which is converted into Glc-6-P (**47**) by β-phosphoglucomutase.[29] **C**. Retaining trehalose phosphorylase from family GT4 can convert trehalose (**38**) to α-Glc-1-P (**29**) and Glc (**2**), which are converted to Glc-6-P (**47**) by phosphoglucomutase and hexokinase respectively.[57]

2.3.1 α-1,1-Glucan phosphorylases

2.3.1.1 Inverting trehalose phosphorylase

Trehalose is used in a variety of different roles including energy storage,[58] in abiotic stress responses[59] and for osmoregulation[60] and can accumulate to very high levels in a variety of organisms.[61] *Euglena gracilis* has an enzyme which shows phosphorolytic activity towards trehalose[56] and promiscuity towards the 6 position in the reverse reaction (Fig 2.12.A).[62] No sequencing data is available but the activity was also identified in bacteria, including plant symbionts[63] and thermophiles,[28] and assayed for the production of a variety of disaccharides.[28]

2.3.1.2 Trehalose-6-phosphate phosphorylase

Many acid bacteria phosphorylate the trehalose and then use trehalose-6-phosphate phosphorylase[29] to release β-Glc-1-P and Glc-6-P into the hexose phosphate pool (Fig 2.12.B).[64]

2.3.1.3 Retaining trehalose phosphorylase

An alternative trehalose phosphorylase was identified in *Flammulina velutipes* which is retaining, and thus produces α-Glc-1-P (Fig 2.12.C).[30] This enzyme has since been found in many other fungi.[57]

2.3.2 α-1,2-Glucan phosphorylases

2.3.2.1 Kojibiose phosphorylase

In nature kojibiose is found in honey[65] and in *Leuconostoc* dextran,[66] whilst the oligomer is produced by *Rhizobium*.[67] It is also part of the core antigen of *Moraxella catarrhalis*, currently under investigation as a vaccine candidate.[68]

A phosphorylase was isolated from *Thermoanaerobium brockii* which was capable of degrading kojibiose to yield β-Glc-1-P.[31] Kojibiose phosphorylase was also capable of the reverse reaction transferring more than one glucose moiety to make kojioligosaccharides,[69] and the enzyme from *Caldicellulosiruptor* can be used for the synthesis of longer oligosaccharides.[70] Mutagenesis and chimerisation with the trehalose phosphorylase from the

same organism allowed even longer polymers to be formed.[71] The enzyme strictly transferred on to glucose configured sugars but allowed either anomeric configuration and any glucoside tested.[31] This allowed the enzyme to be used in the addition of glucose to a cyclic α-glucan.[72]

2.3.2.2 Sucrose phosphorylase

Sucrose is a common plant sugar and a specific phosphorylase is found in diverse bacteria, such as *Leuconostoc mesenteroides*,[73] where it catalyses the phosphorolysis of sucrose, releasing α-Glc-1-P and fructose.[32] This enzyme is a member of GH13 and has been used for the *in vitro* transfer of glucose on to various sugars[74] and phenolics.[75]

2.3.3 β-1,2-Glucan phosphorylases

Certain bacteria including *Rhizobia*[76] and *Brucella*[77] synthesise cyclic β-1,2-glucans as osmoregulators in their cell walls, which are necessary for virulence.[78] The synthesis of these compounds proceeds at the cell membrane by initial autoglucosylation of the enzyme,[79] extension of the growing chain and cyclisation once the chain reaches the appropriate length. This length control is mediated by the C-terminal of the protein, which is an inverting phosphorylase in the GH94 family.[33] To date, this domain has not been explored for the synthesis of oligosaccharides.

2.3.4 α-1,3-Glucan phosphorylases

2.3.4.1 Nigerose phosphorylase

Recently an enzyme which degrades α-1,3-glucosyl glucose was identified in *Clostridium phytofermentans* and has been characterised as nigerose phosphorylase.[34] α-1,3-Glucan polymers are found in fungal cell walls[80] and this is presumably the source of the natural substrate for this enzyme. The enzyme shows some activity towards kojibiose and can transfer on to some other monosaccharides, though only transfers one glucose residue.

2.3.4.2 Glucosyl-rhamnose phosphorylase

α-1,3-Glucosyl rhamnose phosphorylase was[58] identified in *Clostridium phytofermentans*.[35] This structure is found in plant and bacterial cell walls and, based on the presence of this enzyme, is postulated to be in this organism's capsule, with the phosphorylase having some role in cell wall recycling.

2.3.5 β-1,3-Glycan phosphorylases

2.3.5.1 β-1,3-Glucan phosphorylases

β-1,3-Glucans are used in some algae, including *Euglena* and brown algae, as the major storage carbohydrate[81] and make up the cell walls of certain fungi, including yeasts and Basidiomycetes.[82] β-1,3-Oligoglucan phosphorylases are categorised in three sub groups based on substrate preferences. Laminaribiose phosphorylases (EC 2.4.1.31) were first discovered in *Euglena gracilis*[36] and related algae[83] and strict disaccharide phosphorylases have recently been identified in bacteria and assigned to GH94.[37,84] Laminarin phosphorylase (EC 2.4.1.30)[38] and β-1,3-glucan phosphorylase (EC 2.4.1.97) have also been identified with preferences for longer glucans.[39,40] The *Euglena* enzyme has been purified and analysed,[85] though no sequence data is available. This enzyme has proved useful in the synthesis of cell wall related oligosaccharides, including mixed linkage β-glucans,[86] and for the extension of other sugars, such as galactose and glucosamine.[38]

2.3.5.2 Galactosyl-HexNAc phosphorylase

Human mucin[87] and milk[88] contains complex oligosaccharides which are degraded by intestinal bacteria, in part by hydrolysis, but also by phosphorolysis.[41] These enzymes release Gal-1-P from Gal-β-1,3-GlcNAc or Gal-β-1,3-GalNAc[42] and belong to GH112.[89] An additional enzyme has been identified which, in the reverse reaction, transfers galactose on to rhamnose more efficiently, leading to its proposed identification as galactosyl-rhamnose phosphorylase.[43] This carbohydrate structure is found in the cell walls of plants, such as *Tanacetum vulgare*,[90] and bacteria, including certain *Klebsiella*,[91] and these may represent the natural substrates.

2.3.6 α-1,4-Glucan phosphorylases

2.3.6.1 α-1,4-Glucan phosphorylases

α-1,4-Glucan phosphorylases are present in all classes of organism from bacteria (maltodextrin phosphorylase) to animals (glycogen phosphorylase) and plants (starch phosphorylase). Structural studies of the mammalian,[92] bacterial[93] and yeast[94] glycogen phosphorylases identify them as members of GT35 family of retaining phosphorylase. Uniquely among the phosphorylases, they require a pyridoxal phosphate prosthetic group to stabilise the inorganic phosphate.[95]

Their function in nature appears to be in metabolism of glucose storage molecules, though they all have slightly different specificity for chain length and branching frequency, as expected for their different substrate specificities. The mammalian enzyme is used in the release of energy stored as glycogen and, thus, is implicated in diabetes. As such, many inhibitors have been studied.[96]

All plants have at least two forms of starch phosphorylase, localised in the chloroplast (PHS1) and the cytoplasm (PHS2). In the plastid the phosphorylase has been ascribed to starch degradation, although there is no direct *in vivo* evidence. Mutations in rice results in severely reduced starch content and altered amylopectin structure, suggesting a role in starch synthesis.[97]

2.3.6.2 Maltose phosphorylase

Maltose phosphorylase (EC 2.4.1.8) was initially identified from *Neisseria* as an enzyme which degrades maltose to β-Glc-1-P and Glc.[46] It was found to act exclusively on the disaccharide[98] and was reasonably promiscuous towards the acceptor monosaccharide.[99] The disaccharide formed during reverse phosphorolysis can be altered to kojibiose and trehalose by the engineering of a specific loop.[100]

2.3.6.3 Maltosyl phosphorylase

An enzyme was identified in *Mycobacterium* which catalysed the transfer of maltose from α-maltose-1-P on to glycogen.[47] Maltosyl phosphorylase is the only known naturally synthetic phosphorylase and mutations are lethal in *Mycobacterium*, suggesting this as a potential therapeutic target.[101] The structure of this enzyme shows a large carbohydrate binding cleft, indicating limited substrate flexibility.[102]

2.3.7 β-1,4-Glycan phosphorylases

2.3.7.1 β-1,4-Glucan phosphorylases

Cellulose represents an enormous source of energy but is exceptionally difficult to access. Many microbes have developed sophisticated capabilities for degrading this recalcitrant compound, including direct oxidation[103] and the bacterial multi-component cellulosomes.[104] The cellodextrins released by cellulose hydrolases can be imported directly into cells,[105] where they are further broken down, either by hydrolysis to glucose and cellobiose, or by phosphorolysis to α-Glc-1-P.[49] The cellobiose can be phosphorolysed by cellobiose phosphorylase, releasing Glc and α-Glc-1-P.[48] Alternatively cellobiose can be epimerised, by cellobiose 2'-epimerase, to glucosyl-mannose before hydrolysis, releasing Glc and Man.[106]

2.3.7.2 β-1,4-Mannan phosphorylases

Plant cell walls also contain large amounts of hemicellulose including β-1,4-mannan. This is structurally related to cellulose and can be hydrolysed to mannodextrins and mannobiose. These can be directly phosphorolysed by mannan phosphorylase or mannobiose phosphorylase, respectively.[51] Mannobiose can also be epimerised, by cellobiose 2-epimerase, to mannosyl-glucose which can subsequently be phosphorolysed, releasing α-Man-1-P and Glc.[50]

2.3.7.1 Chitobiose phosphorylase

Chitin has the same linkage as cellulose, composed of a homopolymer of β-1,4-GlcNAc. *Vibrio* degrades chitin extracellularly to GlcNAc and chitobiose, which is imported into the cell,[107] where it can be degraded by hydrolysis or by phosphorolysis to release GlcNAc-1-P.[52]

2.4 Conclusions on the *in vitro* synthesis of polysaccharides

Polysaccharides are incredibly complex in structure and do not yield to modern sequencing technologies. Modern mass spectrometry based techniques, including cross ring fragmentation and MS^n,[108] are facilitating structural analysis, but are restricted to highly skilled specialists working on dedicated analytical projects. This has left glycomics somewhat out in the cold in the modern genomics era. However, we now understand the basis of natural carbohydrate synthesis and the exquisite control imposed on the structures. The unexpected requirement for degradative enzymes during synthesis, for example in starch biosynthesis[109] and glycoprotein processing,[110] adds another aspect to the complexity.

The heterogeneity of polysaccharides purified from natural sources makes them unsuited to use in advanced technologies, such as pharmaceuticals and nanotechnology. With the knowledge of natural structures and biosynthesis, well defined carbohydrates have begun to be synthesised *in vitro*. Various chemical methods have been utilised to produce carbohydrates of well characterised structure and with the stereochemistry required. Natural enzymes have also been used to synthesise polysaccharides from natural substrates, although the cost of the substrates and the low yields currently limits their use. Modified enzymes and modified substrates have reduced these drawbacks and allowed the production of a wider range of oligosaccharides.

Phosphorylases can avoid some of these drawbacks and are beginning to be employed in the production of polysaccharides with defined structures and sizes. With the application of molecular biology to the enzymes involved, the potential structures available are almost unlimited. As synthetic oligosaccharides become available new drugs, medical tests and materials can be developed.

2.5 References

1 Kobayashi, S., Sakamoto, J. & Kimura, S. In vitro synthesis of cellulose and related polysaccharides. *Prog. Polym. Sci.* **26**, 1525-1560 (2001).

2 Kochetkov, N. K., Bochkov, A. F. & Yazlovetskii, I. G. Synthesis of polysaccharides. *Russ. Chem. Bull.* **17**, 1719-1724 (1968).

3 Ruckel, E. R. & Schuerch, C. Chemical synthesis of a stereoregular linear polysaccharide. *J. Am. Chem. Soc.* **88**, 2605-2606 (1966).

4 Weishaupt, M., Eller, S. & Seeberger, P. H. in *Glycomics* Vol. 478 *Methods in Enzymology* (ed M. Fukuda) Ch. 22, 463-484 (2010).

5 Werz, D. B. *et al.* Exploring the structural diversity of mammalian carbohydrates ("Glycospace") by statistical databank analysis. *ACS Chem. Biol.* **2**, 685-691 (2007).

6 Seeberger, P. H. Automated carbohydrate synthesis as platform to address fundamental aspects of glycobiology - current status and future challenges. *Carbohydr. Res.* **343**, 1889-1896 (2008).

7 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-D238 (2009).

8 Henrissat, B., Coutinho, P. M. & Davies, G. J. A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Mol. Biol.* **47**, 55-72 (2001).

9 Crout, D. H. G. & Vic, G. Glycosidases and glycosyl transferases in glycoside and oligosaccharide synthesis. *Curr. Opin. Chem. Biol.* **2**, 98-111 (1998).

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

11 Kobayashi, S., Kashiwa, K., Kawasaki, T. & Shoda, S. Novel method for polysaccharide synthesis using an enzyme - the 1st *in vitro* synthesis of cellulose via a nonbiosynthetic path utilizing cellulase as catalyst. *J. Am. Chem. Soc.* **113**, 3079-3084 (1991).

12 Kobayashi, S. *et al.* Formation and structure of artificial cellulose spherulites via enzymatic polymerization. *Biomacromol.* **1**, 168-173 (2000).

13 Kobayashi, S., Kawasaki, T., Obata, K. & Shoda, S. A novel method for synthesis of cellooligosaccharide derivatives by using enzyme catalyst. *Chem. Lett.*, 685-686 (1993).

14 Makino, A. & Kobayashi, S. Chemistry of 2-oxazolines: a crossing of cationic ring-opening polymerization and enzymatic ring-opening polyaddition. *J. Polym. Sci. Pol. Chem.* **48**, 1251-1270 (2010).

15 Kobayashi, S., Kiyosada, T. & Shoda, S. Synthesis of artificial chitin: irreversible catalytic behavior of a glycosyl hydrolase through a transition state analogue substrate. *J. Am. Chem. Soc.* **118**, 13113-13114 (1996).

16 Mackenzie, L. F., Wang, Q. P., Warren, R. A. J. & Withers, S. G. Glycosynthases: mutant glycosidases for oligosaccharide synthesis. *J. Am. Chem. Soc.* **120**, 5583-5584 (1998).

17 Zechel, D. L. & Withers, S. G. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11-18 (2000).

18 Zhang, C. *et al.* Exploiting the reversibility of natural product glycosyltransferase-catalyzed reactions. *Science* **313**, 1291-1294 (2006).

19 Gantt, R. W., Peltier-Pain, P., Cournoyer, W. J. & Thorson, J. S. Using simple donors to drive the equilibria of glycosyltransferase-catalyzed reactions. *Nat. Chem. Biol.* **7**, 685-691 (2011).

20 Unverzagt, C., Kunz, H. & Paulson, J. C. High-efficiency synthesis of sialyloligosaccharides and sialoglycopeptides. *J. Am. Chem. Soc.* **112**, 9308-9309 (1990).

21 Lougheed, B., Ly, H. D., Wakarchuk, W. W. & Withers, S. G. Glycosyl fluorides can function as substrates for nucleotide phosphosugar-dependent glycosyltransferases. *J. Biol. Chem.* **274**, 37717-37722 (1999).

22 Deluca, C. *et al.* Enzymatic-synthesis of hyaluronic-acid with regeneration of sugar nucleotides. *J. Am. Chem. Soc.* **117**, 5869-5870 (1995).

23 Yang, M. *et al.* Probing the breadth of macrolide glycosyltransferases: *in vitro* remodeling of a polyketide antibiotic creates active bacterial uptake and enhances potency. *J. Am. Chem. Soc.* **127**, 9336-9337 (2005).

24 Seto, N. O. L. *et al.* Donor substrate specificity of recombinant human blood group A, B and hybrid A/B glycosyltransferases expressed in Escherichia coli. *Eur. J. Biochem.* **259**, 770-775 (1999).

25 Williams, G. J., Zhang, C. & Thorson, J. S. Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution. *Nat. Chem. Biol.* **3**, 657-662 (2007).

26 Yu, S., Bojsen, K., Svensson, B. & Marcussen, J. α-1,4-Glucan lyases producing 1,5-anhydro-d-fructose from starch and glycogen have sequence similarity to α-glucosidases. *Biochim. et Biophys. Acta - Prot. Struc. Mol. Enzymol.* **1433**, 1-15 (1999).

27 Michaud, P., Da Costa, A., Courtois, B. & Courtois, J. Polysaccharide lyases: recent developments as biotechnological tools. *Crit. Rev. Biotechnol.* **23**, 233-266 (2003).

28 Chaen, H. *et al.* Purification and characterization of thermostable trehalose phosphorylase from *Theymoanaeyobium brockii*. *J. Applied Glycosci.* **46**, 399-405 (1999).

29 Andersson, U., Levander, F. & Rådström, P. Trehalose-6-phosphate phosphorylase is part of a novel metabolic pathway for trehalose utilization in *Lactococcus lactis*. *J. Biol. Chem.* **276**, 42707-42713 (2001).

30 Kitamoto, Y., Akashi, H., Tanaka, H. & Mori, N. α-Glucose-1-phosphate formation by a novel trehalose phosphorylase from *Flammulina velutipes*. *FEMS Microbiol. Lett.* **55**, 147-149 (1988).

31 Chaen, H. *et al.* Purification and characterization of a novel phosphorylase, kojibiose phosphorylase, from *Thermoanaerobium brockii*. *J. Applied Glycosci.* **46**, 423-429 (1999).

32 Vandamme, E. J., Van Loo, J., Machtelinckx, L. & De Laporte, A. in *Adv. Appl. Microbiol.* Vol. Volume 32 (ed I. Laskin Allen) 163-201 (Academic Press, 1987).

33 Ciocchini, A. E. *et al.* A glycosyltransferase with a length-controlling activity as a mechanism to regulate the size of polysaccharides. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 16492-16497 (2007).

34 Nihira, T., Nakai, H., Chiku, K. & Kitaoka, M. Discovery of nigerose phosphorylase from *Clostridium phytofermentans*. *Appl. Microbiol. Biotechnol.* **93**, 1513-1522 (2012).

35 Nihira, T., Nakai, H. & Kitaoka, M. 3-O-α-D-Glucopyranosyl-L-rhamnose phosphorylase from *Clostridium phytofermentans*. *Carbohydr. Res.* **350**, 94-97 (2012).

36 Marechal, L. R. & Goldemberg, S. H. Laminaribiose phosphorylase from *Euglena gracilis*. *Biochem. Biophys. Res. Commun.* **13**, 106-107 (1963).

37 Kitaoka, M., Matsuoka, Y., Mori, K., Nishimoto, M. & Hayashi, K. Characterization of a bacterial laminaribiose phosphorylase. *Biosci., Biotechnol., Biochem.* **76**, 343-348 (2012).

38 Marechal, L. R. β-1,3-Oligoglucan - orthophosphate glucosyltransferases from *Euglena gracilis*. 2. Comparative studies between laminaribiose- and β-1,3-oligoglucan phosphorylase. *Biochim. Biophys. Acta* **146**, 417-443 (1967).

39 Kauss, H. & Kriebitzsch, C. Demonstration and partial purification of a β-(1->3)-glucan phosphorylase. *Biochem. Biophys. Res. Commun.* **35**, 926-930 (1969).

40 Liénart, Y., Comtat, J. & Barnoud, F. Wall-bound 1,3-β-D-glucan: orthophosphate glucosyltransferase activity from acacia cultured cells. *Plant Sci.* **58**, 165-170 (1988).

41 Derensy-Dron, D., Krzewinski, F., Brassart, C. & Bouquelet, S. β-1,3-Galactosyl-N-acetylhexosamine phosphorylase from *Bifidobacterium bifidum* DSM 20082: characterization, partial purification and relation to mucin degradation. *Biotechnol. Appl. Biochem.* **29 (Pt 1)**, 3-10 (1999).

42 Nakajima, M., Nishimoto, M. & Kitaoka, M. Characterization of beta-1,3-galactosyl-N-acetylhexosamine phosphorylase from Propionibacterium acnes. *Appl. Microbiol. Biotechnol.* **83**, 109-115 (2009).

43 Nakajima, M., Nishimoto, M. & Kitaoka, M. Characterization of three β-galactoside phosphorylases from *Clostridium phytofermentans*: discovery of D-galactosyl-β-1->4-L-rhamnose phophorylase. *J. Biol. Chem.* **284**, 19220-19227 (2009).

44 Cori, C. F. & Cori, G. T. The activity and crystallization of phosphorylase-B. *J. Biol. Chem.* **158**, 341-345 (1945).

45 Lee, Y. P. Potato phosphorylase .1. Purification, physicochemical properties and catalytic activity. *Biochim. Biophys. Acta* **43**, 18-24 (1960).

46 Fitting, C. & Doudoroff, M. Phosphorolysis of maltose by enzyme preparations from *Neisseria meningitidis*. *J. Biol. Chem.* **199**, 153-163 (1952).

47 Elbein, A. D., Pastuszak, I., Tackett, A. J., Wilson, T. & Pan, Y. T. Last step in the conversion of trehalose to glycogen: A mycobacterial enzyme that transfers maltose from maltose-1-phosphate to glycogen. *J. Biol. Chem.* **285**, 9803-9812 (2010).

48 Sih, C. J. & McBee, R. H. A phosphorylase active on cellobiose. *Proc. Montana Acad. Sci* **15**, 21-22 (1955).

49 Zhang, Y.-H. P. & Lynd, L. R. Kinetics and relative importance of phosphorolytic and hydrolytic cleavage of cellodextrins and cellobiose in cell extracts of *Clostridium thermocellum*. *Appl. Environ. Microbiol.* **70**, 1563-1569 (2004).

50 Senoura, T. *et al.* New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase. *Biochem. Biophys. Res. Commun.* **408**, 701-706 (2011).

51 Kawahara, R. *et al.* Metabolic mechanism of mannan in a ruminal bacterium, *Ruminococcus albus*, involving two mannoside phosphorylases and cellobiose 2-epimerase: discovery of a new carbohydrate phosphorylase, β-1,4-mannooligosaccharide phosphorylase. *J. Biol. Chem.* **287**, 42389-42399 (2012).

52 Park, J. K., Keyhani, N. O. & Roseman, S. Chitin catabolism in the marine bacterium *Vibrio furniss*i: identification, molecular cloning, and characterization of an N-diacetylchitobiose phosphorylase. *J. Biol. Chem.* **275**, 33077-33083 (2000).

53 Nakai, H., Kitaoka, M., Svensson, B. & Ohtsubo, K. i. Recent development of phosphorylases possessing large potential for oligosaccharide synthesis. *Curr. Opin. Chem. Biol.* **17**, 301-309 (2013).

54 Kitaoka, M. & Hayashi, K. Carbohydrate-processing phosphorolytic enzymes. *Trends Glycosci. and Glycotechnol.* **14**, 35-50 (2002).

55 Weinhausel, A. *et al.* alpha-1,4-D-Glucan phosphorylase of gram-positive *Corynebacterium callunae*: isolation, biochemical properties and molecular shape of the enzyme from solution X-ray scattering. *Biochem. J.* **326**, 773-783 (1997).

56 Belocopitow, E. & Maréchal, L. R. Trehalose phosphorylase from *Euglena gracilis*. *Biochim. Biophys. Acta - Enzymology* **198**, 151-154 (1970).

57 Kitamoto, Y., Tanaka, H. & Osaki, N. Survey of α-glucose 1-phosphate forming trehalose phosphorylase and trehalase in various fungi including basidiomycetous mushrooms. *Mycoscience* **39**, 327-331 (1998).

58 Sussman, A. S. & Lingappa, B. T. Role of trehalose in ascospores of *Neurospora tetrasperma*. *Science* **130**, 1343 (1959).

59 López-Gómez, M. & Lluch, C. in *Abiotic Stress Responses in Plants* (eds Parvaiz Ahmad & M. N. V. Prasad) Ch. 14, 253-265 (Springer New York, 2012).

60 Madin, K. A. C. & Crowe, J. H. Anhydrobiosis in nematodes - carbohydrate and lipid-metabolism during dehydration. *J. Exp. Zool.* **193**, 335-342 (1975).

61 Elbein, A. D., Pan, Y. T., Pastuszak, I. & Carroll, D. New insights on trehalose: a multifunctional molecule. *Glycobiology* **13**, 17R-27R (2003).

62 Maréchal, L. R. & Belocopitow, E. Metabolism of trehalose in *Euglena gracilis*. *J. Biol. Chem.* **247**, 3223-3228 (1972).

63 Salminen, S. O. & Streeter, J. G. Enzymes of α,α-trehalose metabolism in soybean nodules. *Plant Physiol.* **81**, 538-541 (1986).

64 Andersson, U. & Rådström, P. β-Glucose-1-phosphate-interconverting enzymes in maltose- and trehalose-fermenting lactic acid bacteria. *Environ. Microbiol.* **4**, 81-88 (2002).

65 Watanabe, T. & Aso, K. Studies on honey II. isolation of kojibiose, nigerose, maltose and isomaltose from honey. *Tohoku J. Agric. Res.* **11**, 109 - 115 (1960).

66 Watanabe, T. *et al.* Acetolysis of *Leuconostoc mesenteroides* NRRL B-1299 dextran. Isolation and characterization of oligosaccharides containing secondary linkages from the borate-soluble fraction. *Carbohydr. Res.* **83**, 119-127 (1980).

67 Watanabe, T., Kamo, Y., Matsuda, K. & Dudman, W. F. A new, extracellular oligosaccharide, kojihexaose, from Rhizobium japonicum strain 561. *Carbohydr. Res.* **110**, 170-175 (1982).

68 Masoud, H., Perry, M. B., Brisson, J.-R., Uhrin, D. & Richards, J. C. Structural elucidation of the backbone oligosaccharide from the lipopolysaccharide of *Moraxella catarrhalis* serotype A. *Can. J. Chem.* **72**, 1466-1477 (1994).

69 Chaen, H. *et al.* Enzymatic synthesis of kojioligosaccharides using kojibiose phosphorylase. *J. Biosci. Bioeng.* **92**, 177-182 (2001).

70 Yamamoto, T. *et al.* Enzymatic properties of recombinant kojibiose phosphorylase from *Caldicellulosiruptor saccharolyticus* ATCC43494. *Biosci., Biotechnol., Biochem.* **75**, 1208-1210 (2011).

71 Yamamoto, T., Nishimoto, T., Chen, H. & Fukuda, S. Improvments of the enzymetic properties of kojibiose phosphorylase by random mutagenesis and chimerization. *J. Appl. Glycosci.*, 123-129 (2006).

72 Watanabe, H. *et al.* Enzymatic synthesis of a 2-O-α-D-glucopyranosyl cyclic tetrasaccharide by kojibiose phosphorylase. *Carbohydr. Res.* **340**, 449-454 (2005).

73 Goedl, C., Schwarz, A., Minani, A. & Nidetzky, B. Recombinant sucrose phosphorylase from *Leuconostoc mesenteroides*: Characterization, kinetic studies of transglucosylation, and application of immobilised enzyme for production of α-D-glucose 1-phosphate. *J. Biotechnol.* **129**, 77-86 (2007).

74 Kitao, S. & Sekine, H. Transglucosylation catalyzed by sucrose phosphorylase from *Leuconostoc mesenteroides* and production of glucosyl-xylitol. *Biosci., Biotechnol., Biochem.* **56**, 2011-2014 (1992).

75 Kitao, S. & Sekine, H. α-D-Glucosyl transfer to phenolic compounds by sucrose phosphorylase from *Leuconostoc mesenteroides* and production of α-arbutin. *Biosci., Biotechnol., Biochem.* **58**, 38-42 (1994).

76 Breedveld, M. W. & Miller, K. J. Cyclic β-glucans of members of the family *Rhizobiaceae*. *Microbiological Reviews* **58**, 145-161 (1994).

77 Bundle, D. R., Cherwonogrodzky, J. W. & Perry, M. B. Characterization of *Brucella* polysaccharide B. *Infection and Immunity* **56**, 1101-1106 (1988).

78 Arellano-Reynoso, B. *et al.* Cyclic β-1,2-glucan is a *Brucella* virulence factor required for intracellular survival. *Nat. Immunol.* **6**, 618-625 (2005).

79 Castro, O. A., Zorreguieta, A., Ielmini, V., Vega, G. & Ielpi, L. Cyclic β-(1,2)-glucan synthesis in Rhizobiaceae: roles of the 319-kilodalton protein intermediate. *J. Bacteriol.* **178**, 6043-6048 (1996).

80 Gorin, P. A. J. & Spencer, J. F. T. in *Advances in Carbohydrate Chemistry* Vol. Volume 23 (eds L. Wolfrom Melville & R. Stuart Tipson) 367-417 (Academic Press, 1968).

81 Painter, T. J. Structural evolution of glycans in algae. *Pure Appl. Chem.* **55**, 677-694 (1983).

82 Vandamme, E. J., De Baets, S. & Steinbüchel, A. *Polysaccharides II: polysaccharides from eukaryotes*. Vol. 6 (Wiley-VCH, 2002).

83 Manners, D. J. & Taylor, D. C. Acceptor specificity of laminaribiose phosphorylase from *Astasia ocellata*. *Biochem. J.* **94**, 17-19 (1965).

84 Nihira, T. *et al.* Characterization of a laminaribiose phosphorylase from *Acholeplasma laidlawii* PG-8A and production of 1,3-β-D-glucosyl disaccharides. *Carbohydr. Res.* **361**, 49-54 (2012).

85 Kitaoka, M., Sasaki, T. & Taniguchi, H. Purification and properties of laminaribiose phosphorylase (EC-2.4.1.31) from *Euglena-gracilis* Z. *Arch. Biochem. Biophys.* **304**, 508-514 (1993).

86 Fukamizo, T. *et al.* Enzymatic hydrolysis of 1,3-1,4-β-glucosyl oligosaccharides by 1,3-1,4-β-glucanase from *Synechocystis* PCC6803: a comparison with assays using polymer and chromophoric oligosaccharide substrates. *Arch. Biochem. Biophys.* **478**, 187-194 (2008).

87 Podolsky, D. K. Oligosaccharide structures of isolated human colonic mucin species. *J. Biol. Chem.* **260**, 5510-5515 (1985).

88 Stahl, B. *et al.* Oligosaccharides from human milk as revealed by matrix-assisted laser desorption/ionization mass spectrometry. *Anal. Biochem.* **223**, 218-226 (1994).

89 Hidaka, M. *et al.* The crystal structure of galacto-N-biose/lacto-N-biose I phosphorylase: A large deformation of a tim barrel scaffold. *JBC* **284**, 7273-7283 (2009).

90 Polle, A. Y., Ovodova, R. G., Shashkov, A. S. & Ovodov, Y. S. Some structural features of pectic polysaccharide from tansy, *Tanacetum vulgare* L. *Carbohydr. Polym.* **49**, 337-344 (2002).

91 Björndal, H., Linberg, B., Lönngren, J., Rossel, K.-G. & Nimmich, W. Structural studies of the *Klebsiella* type 47 capsular polysaccharide. *Carbohydr. Res.* **27**, 373-378 (1973).

92 Barford, D., Hu, S. H. & Johnson, L. N. Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. *J. Mol. Biol.* **218**, 233-260 (1991).

93 O'Reilly, M., Watson, K. A. & Johnson, L. N. The crystal structure of the *Escherichia coli* maltodextrin phosphorylase−acarbose complex. *Biochemistry* **38**, 5337-5345 (1999).

94 Lin, K., Rath, V. L., Dai, S. C., Fletterick, R. J. & Hwang, P. K. A protein phosphorylation switch at the conserved allosteric site in GP. *Science* **273**, 1539-1541 (1996).

95 Palm, D., Klein, H. W., Schinzel, R., Buehner, M. & Helmreich, E. J. M. The role of pyridoxal 5'-phosphate in glycogen-phosphorylase catalysis. *Biochemistry* **29**, 1099-1107 (1990).

96 Somsak, L. *et al.* New inhibitors of glycogen phosphorylase as potential antidiabetic agents. *Curr. Med. Chem.* **15**, 2933-2983 (2008).

97 Satoh, H. *et al.* Mutation of the plastidial α-glucan phosphorylase gene in rice affects the synthesis and structure of starch in the endosperm. *Plant Cell* **20**, 1833-1849 (2008).

98 Kamogawa, A., Yokobayashi, K. & Fukui, T. Properties of maltose phosphorylase from *Lactobacillus brevis*. *Agric. Biol. Chem.* **37**, 2813-2819 (1973).

99 Aisaka, K., Masuda, T. & Chikamune, T. Properties of maltose phosphorylase from *Propionibacterium freudenreichii*. *J. Ferment. Bioeng.* **82**, 171-173 (1996).

100 Nakai, H. *et al.* Rational engineering of *Lactobacillus acidophilus* NCFM maltose phosphorylase into either trehalose or kojibiose dual specificity phosphorylase. *Protein Eng., Des. Sel.* **23**, 781-787 (2010).

101 Kalscheuer, R. *et al.* Self-poisoning of *Mycobacterium tuberculosi*s by targeting GlgE in an α-glucan pathway. *Nat. Chem. Biol.* **6**, 376-384 (2010).

102 Syson, K. *et al.* Structure of *Streptomyces* maltosyltransferase GlgE, a homologue of a genetically validated anti-tuberculosis target. *J. Biol. Chem.* **286**, 38298-38310 (2011).

103 Beeson, W. T., Phillips, C. M., Cate, J. H. D. & Marletta, M. A. Oxidative cleavage of cellulose by fungal copper-dependent polysaccharide monooxygenases. *J. Am. Chem. Soc.* **134**, 890-892 (2011).

104 Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: Highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu. Rev. Biochem.* **79**, 655-681 (2010).

105 Tian, C. *et al.* Systems analysis of plant cell wall degradation by the model filamentous fungus *Neurospora crassa*. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22157-22162 (2009).

106 Tyler, T. R. & Leatherwood, J. M. Epimerization of disaccharides by enzyme preparations from *Ruminococcus albus*. *Arch. Biochem. Biophys.* **119**, 363-367 (1967).

107 Keyhani, N. O., Li, X.-B. & Roseman, S. Chitin catabolism in the marine bacterium *Vibrio furnissii*: identification and molecular cloning of a chitoporin. *J. Biol. Chem.* **275**, 33068-33076 (2000).

108 Ashline, D., Singh, S., Hanneman, A. & Reinhold, V. Congruent strategies for carbohydrate sequencing. 1. Mining structural details by MSn. *Anal. Chem.* **77**, 6250-6262 (2005).

109 Streb, S. *et al.* Starch granule biosynthesis in arabidopsis is abolished by removal of all debranching enzymes but restored by the subsequent removal of an endoamylase. *Plant Cell* **20**, 3448-3466 (2008).

110 Marth, J. D. in *Essentials of Glycobiology* (eds A. Varki *et al.*) Ch. 7, (Cold Spring Harbor Laboratory Press, 1999).

Chapter 3 – PHS2 – an α-1,4-glucan phosphorylase

3.1. Introduction

3.1.1 α-1,4-Glucans

Starch is the highest single calorie contribution to the human diet with a global requirement, for both food and industrial uses, running at several billions of tonnes. Production needs to double by 2050 when the global population is predicted to be over 9 billion. There is also increasing demand for starch as a renewable feedstock in industry, particularly for the production of biofuels.[1] The volatility in grain prices continues to cause unrest around the world and, with climate change producing unpredictable effects, this volatility is only likely to worsen. Increases in agricultural land will not be able to provide stable growth in production, so substantial improvements in crop yield are required.[2]

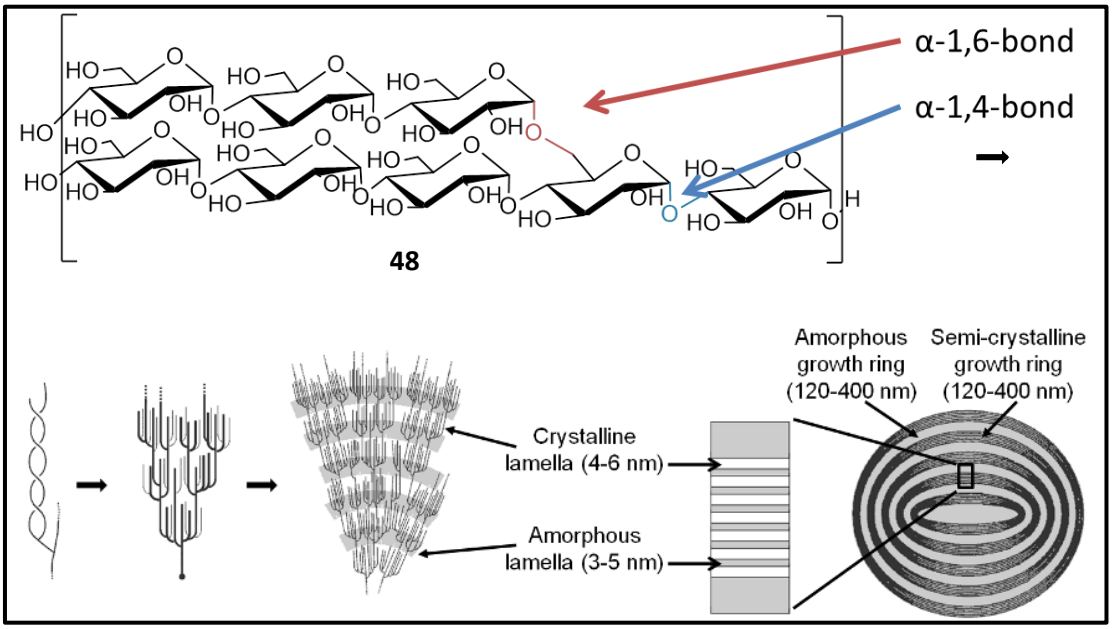


Figure 3.1: Structure of the starch granule

The α-1,4-glucans contain some α-1,6-branches (**48**) which place them parallel to each other, allowing double helix formation. The branches are specifically placed to give regions of branching, known as amorphous lamella, and regions of exclusively linear chains which form crystalline lamella, and these lamella form specific growth rings. Adapted from, Food, The Chemistry of its Components, Tom Coultate with permission from the Royal Society of Chemistry. Copyright (2002).[3]

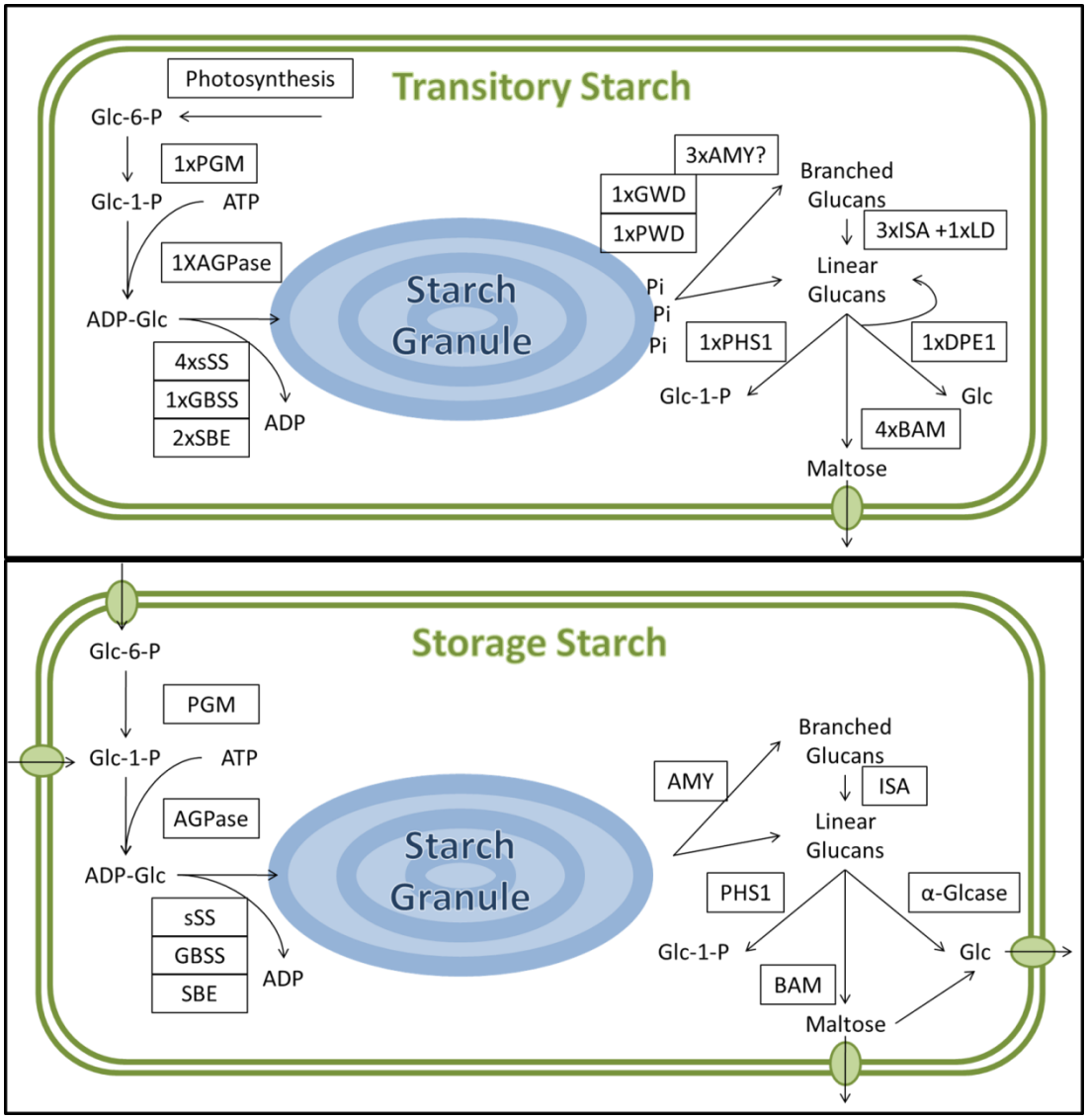


Figure 3.2: Starch granule metabolism

Glc-6-P from photosynthesis or metabolism is converted to Glc-1-P by Phosphoglucomutase (PGM) and AMP is added by ADP-Glucose Pyrophosphorylase (AGPase). The glucose is transferred on to the growing chain of the starch granules by soluble Starch Synthase (sSS) and Granule Bound Starch Synthase (GBSS) and branches are added by Starch Branching Enzyme (SBE). Degradation of the starch granule in the transitory starch granules is by initial phosphorylation, catalysed by Glucan Water Dikinase (GWD) and Phosphoglucan Water Dikinase (PWD), allowing attack of the granule, though amylases are not necessary. The storage granule is directly attacked by α-amylases (AMY) which releases branched chains, although mutants in all three isoforms have normal starch metabolism.[4] These branched glucans are debranched by Isoamylases (ISA) and Limit Dextrinase (LD). Linear glucans are degraded to release: Glc-1-P by phosphorylase (PHS1); maltose by β-amylase (BAM); and glucose, either by α-Glucosidases (α-Glcase) or by disproportionation of maltotriose by DPE1. The number of isoforms of each enzyme in *Arabidopsis* is indicated in the transitory starch diagram.[5]

Starch granules are made of α-1,4-glucans with α-1,6-branches to give long chains stacked side by side, which form alternating layers of highly ordered liquid crystalline lamella interspersed with amorphous regions (Fig 3.1).[6] This makes the surface of a starch granule highly resistant to enzymatic attack, requiring specialised enzymes, such as α-amylase, to initiate degradation.

The source of starch can have marked effects on the properties important for industrial uses. For instance, cereal starch is virtually phosphate free, whilst potato starch has high phosphate content, with increased viscosity and decreased retrogradation, factors important in the paper industry.[7] Starches may also need to be modified in a number of ways, including degradation, esterification and oxidation, although this is costly and potentially hazardous. Production of modified starches *in planta* is therefore an attractive alternative.[8] For example, the gelatinisation properties of the starch can be altered by changing the ratio of linear amylose to branched amylopectin using genetic engineering.[9,10]

In order to engineer precisely controlled starches we need to understand the biosynthetic enzymes involved, and insights are provided by continued research in the model plant *Arabidopsis*.[5] There are multiple enzymes needed to form the correct morphology of the starch granule, and the precise role of many isoforms is poorly understood (Fig 3.2).[11,12] For example, the synthesis of correct starch granules counter intuitively requires two paralogues of the degradative isoamylases, which form a complex.[13] Storage starch is stored in the amyloplasts of specialised storage organs, such as the grain or tuber, whilst transitory starch is stored in the leaf chloroplasts, as a carbon source for photosynthetic cells at night. The two plastids use different suites of enzymes to attack the starch granule.[14] In the storage organs the granule is enzymatically hydrolysed,[15] ultimately releasing glucose. In photosynthetic organs the transitory starch granule is initially phosphorylated by specific glucan kinases, before being broken down into short maltooligosaccharides.[16] There are many isoforms of each enzyme involved in starch synthesis and breakdown, but to date the differences in enzymology are interpreted based upon their gene and protein expression, rather than enzyme activity.

.

The enzymology of the starch granule actually takes place at a solid-liquid interface, rather than the solution phase used in most experiments. This can have profound effects on the reaction; for example the reaction rate of granule-bound starch synthase is enhanced on crystalline amylopectin.[17] There are few studies involving crystalline maltodextrins[18] or purified starch granules,[19] but even these measured discreet product formation, not breakdown of the starch granule. Electron microscopy clearly shows that the initial attack on the surface alters the granule (Fig 3.3),[20] undermining endpoint kinetic assessments. Direct monitoring of purified starch granules during degradation, either using electron microscopy[21] or synchrotron radiation,[22] provides snapshots of these processes during the degradation of the granule.

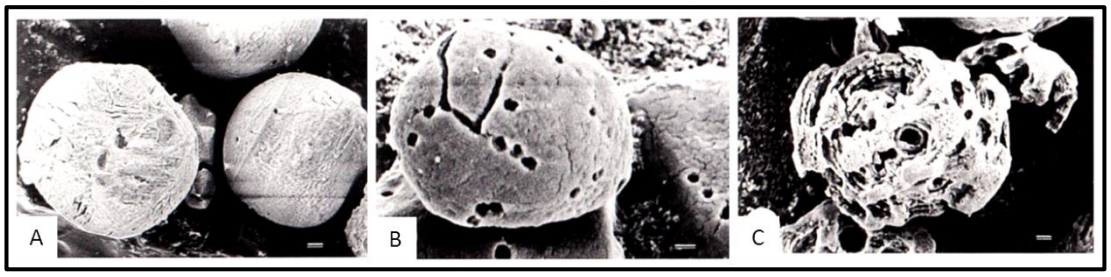


Figure 3.3: Surface changes upon enzymatic degradation of starch granules

Scanning electron micrographs of maize starch granules after hydrolysis by *Aspergillus fumigatus* alpha-amylase. **A**. High-amylose. **B**. Normal. **C**. Waxy. Bars correspond to 1 µm. Reproduced from the Journal of Cereal Science with permission from Elsevier. Copyright (1995).[21]

To study starch surface biochemistry, it is necessary to monitor what is happening to the starch substrate, rather than the released product, in real time. In order to do this a well-defined starch-like surface, with continuous measurement, is required.

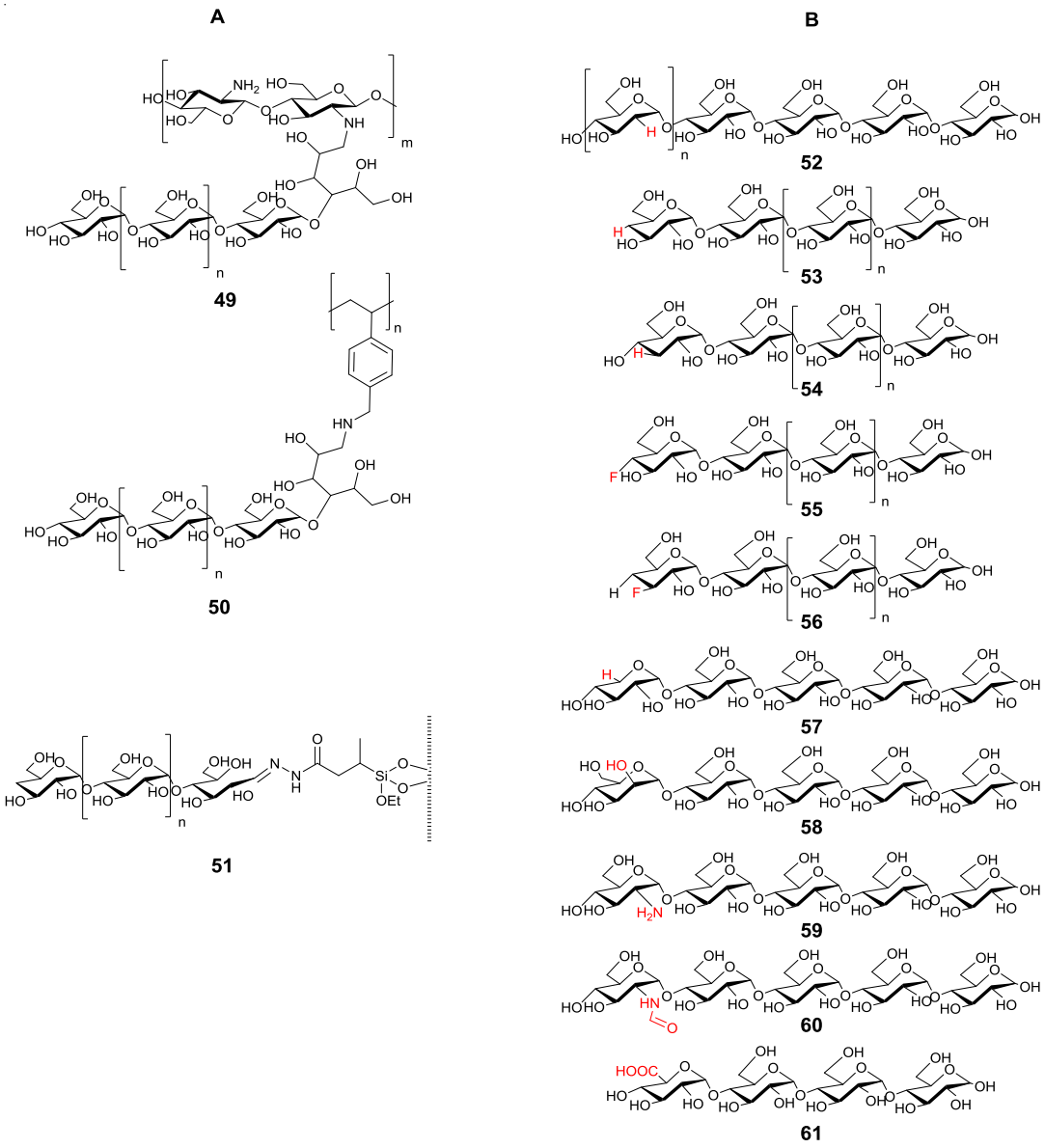


Figure 3.4: Proposed products formed using α-1,4-glucan phosphorylases

A. Phosphorylases were used to extend various maltooligosaccharide grafted compounds to form, for example, amylose grafted chitosan (**49**),[23] polystyrene (**50**)[24] and HPLC matrix (**51**).[25] **B**. Phosphorylase could transfer various sugars from their respective anomeric phosphates on to acceptors. 2-Deoxyglucans were synthesised from D-glucal (**52**),[26] which can then be phosphorolysed to synthesise 2-deoxy-α-Glc-1-P,[27] and 3- and 4-deoxy and fluoro-D-glucose were transferred on to glycogen (**53-56**) by rabbit muscle enzyme.[28] Potato phosphorylase has been used to transfer D-xylose (**57**),[29] D-mannose (**58**),[30] D-glucosamine (**59**)[31] and *N*-formyl-D-glucosamine[32] (**60**) from their respective phosphates on to maltotetraose. Thermostable phosphorylase catalysed the transfer of D-glucuronic acid on to maltotriose(**61**).[33]

3.1.2 The *in vitro* use of α-1,4-glucan phosphorylases

α-1,4-glucan phosphorylases represent a readily controllable system for the synthesis of extended amylose chains. Mammalian glycogen phosphorylases (GP) have been heavily studied, particularly with reference to their role in type II diabetes[34] and cancer[35] in humans, and a wide range of inhibitors have been tested which bind to several different sites on the enzyme, including the catalytic site, the allosteric site and the caffeine binding site.[36] They have proven less useful in biotechnology than plant phosphorylases because of the allosteric regulation, absent from the plant enzyme.[37]

α-1,4-Glucan phosphorylases have been used to synthesise many different glucan based molecules[38] both by extension of a covalent acceptor, such as maltoheptaose immobilised on chitosan[23] or polystyrene[24] (Fig 3.4.A), or by twining polysaccharides around a hydrophobic core to form a macromolecular complex, such as amylose-wrapped lipid.[39] These enzymes have allowed the production of hybrid materials such as a novel HPLC matrix, based on extension of maltoheptaose immobilised on silica,[25] and soluble single-walled carbon nano-tubes, insulated by wrapping with amylose helices.[40]

Phosphorylases also display some promiscuity towards the donor substrate. 2-Deoxy-maltooligosaccharides[26] have been synthesised from D-glucal in the presence of Pi, which can then be phosphorolysed to synthesise 2-deoxy-α-glucose-1-phosphate (Fig 3.4.B). Deoxy- and fluoro-glucose could be transferred on to glycogen by phosphorylase, but with a very low yield.[28] Alternative sugar-1-phosphates can be utilised including xylose,[29] mannose,[30] glucosamine,[31] *N*-formyl-glucosamine[32] and glucuronic acid,[33] although the products were all isolated after a single residue extension, indicating these sugars are not bound in the acceptor site of the enzyme for further extension.

3.1.3 Rationale

Many physiologically important enzymatic reactions occur on surfaces and their analysis is not trivial. A range of techniques have been developed for monitoring enzymatic reactions on surfaces, including mass spectrometry, fluorescent and radioactivity based assays.[41] Whilst it has been possible to detect glycosylation reactions on surfaces using surface plasmon resonance spectroscopy (SPR), this has typically been by alterations in lectin binding.[42] Recently direct measurement of polysaccharide synthesis on SPR surfaces has been achieved.[43,44] All the enzymes to which these techniques have been applied naturally use soluble substrates and act in solution. Any kinetic analysis, therefore, has to take into account the unnatural nature of the surface immobilised substrates. Starch-active enzymes would benefit markedly from studies using immobilised substrates which mimic the insoluble surface upon which they naturally act. It was hoped that by combining techniques for monitoring enzymatic reactions on surfaces a quantifiable starch-like surface could be generated to study starch enzymology.

Crystalline amylose is required to mimic the starch granule surface. It is known that when amylose is synthesised in solution a crystalline structure can be formed.[45] Working on surfaces, crystalline material could be seen by electron microscopy imaging of glycogen particles extended using amylosucrase[46] and measurement of the kinetics of phosphorylase-catalysed amylose synthesis was achieved using quartz crystal microbalance.[47] Gold nanoparticles can be used to simultaneously display and provide visual output of carbohydrate interactions,[48,49] using the same physical and chemical properties as SPR. These techniques were used to provide the basis for the development of approaches based on glucan-coated SPR biosensors and gold nanoparticle surfaces for the generation of enzyme-extended glucan surfaces that may be deployed for the assessment of starch-active enzymes.

3.2 Cloning, expression and purification of PHS2

PHS2, the cytosolic α-1,4-glucan phosphorylase, was cloned from *Arabidopsis thaliana* Col0 cDNA into pET151 using the TOPO cloning method. This inserts a hexahistidine tag and a V5 epitope beyond a TEV protease cleavage site on the N-terminus of the protein. The protein was successfully expressed in *E. coli* Rosetta2 cells and purified using nickel affinity chromatography and gel filtration (GF) to provide pure protein (Fig 3.5). SDS-PAGE and dynamic light scattering (DyLS) both demonstrated the monodispersity and purity of the protein and proteomic analysis of the purified protein confirmed it was indeed PHS2. The DyLS and GF gave a molecular weight in agreement for the calculated molecular weight for a dimer of PHS2 of 198 kDa.

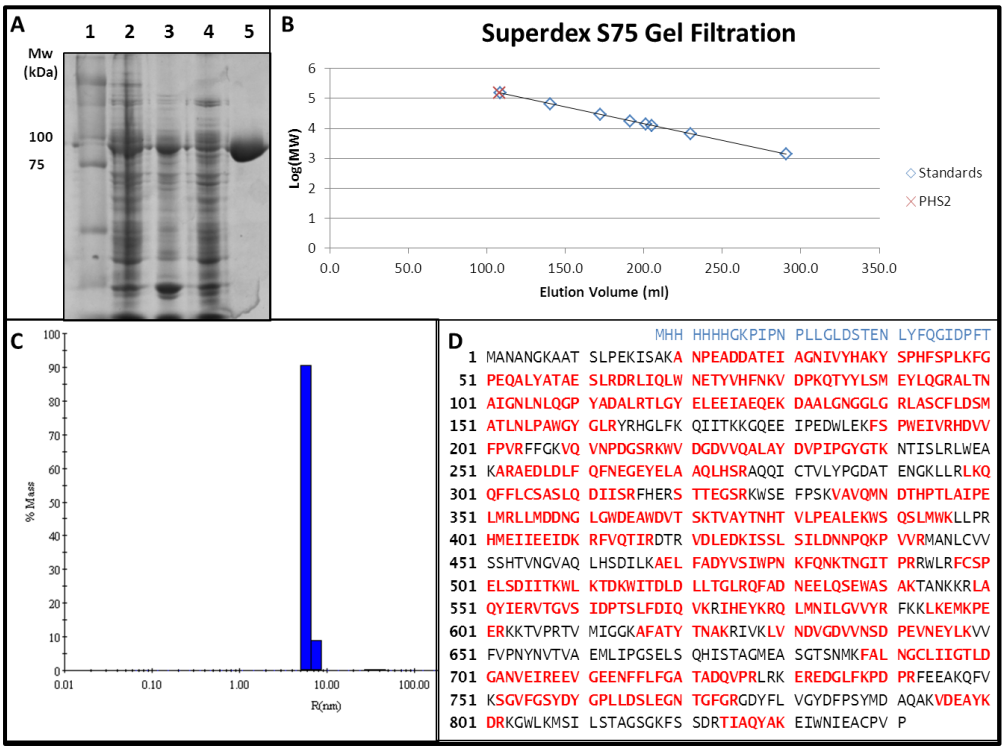


Figure 3.5: Purification of PHS2 from *E. coli*

A. SDS-PAGE of protein purification. Lane 1: Kaleidoscope protein standards (Bio-Rad), Lane 2: total cell extract, Lane 3: insoluble cell material, Lane 4: soluble cell lysate, Lane 5: purified protein, showing a single band running just under 100 kDa. **B**. Gel filtration shows a single peak eluting around 108 ml, which on a Superdex 75 gives a predicted M_w > 150 kDa when compared to standards shown (vitamin B12 – 1.355 kDa, aprotinin – 6.5 kDa, cytochrome C – 12.4 kDa, ribonuclease A – 13.7 kDa, myoglobin – 17.6 kDa, carbonic anhydrase 29 kDa, albumin 66 kDa and alcohol dehydrogenase – 150 kDa. **C**. Dynamic Light Scattering shows that protein is present as monodisperse species with a radius of 6.0 nm, correlating to a molecular weight of 221kDa. **D**. Sequence of PHS2, with artificial tag in blue, and peptides identified in MALDI-ToF spectrum after tryptic digestion in red.

3.3 PHS2 activity in solution

3.3.1 Activity assays

3.3.1.1 Optimal conditions for PHS2 activity

To assay the enzyme kinetics of PHS2 the amount of phosphate (Pi) released upon transfer of Glc from Glc-1-P on to acceptor glucans was measured. The optimum conditions for the enzyme were assayed using standard reactions performed over a range of temperatures and pHs. The temperature optimum was found to be 50 °C and the optimum pH was found to be 6.0 for the synthetic reactions of PHS2 (Fig 3.6).

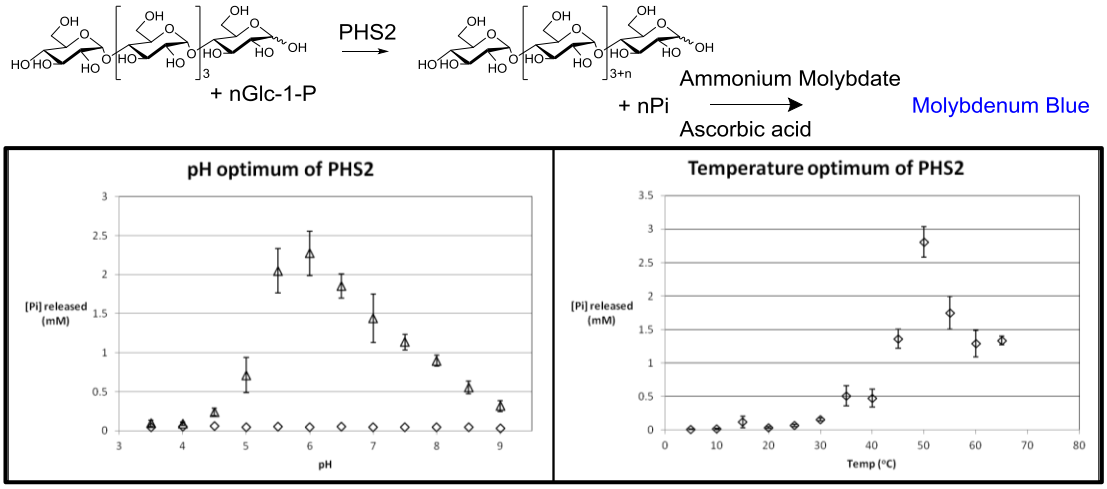


Figure 3.6: Optimum conditions for PHS2 activity

All reactions measured the release of phosphate from Glc-1-P (10 mM) in the presence of maltopentaose (1 mM) by PHS2 (8 µg/ml). Maximum activity (∆) was at pH 6.0 (100 mM Tris, MES, citrate) at 37 °C, and at 50 °C in MES (20 mM, pH 6.0). No release of phosphate is observed in the absence of enzyme (◊) across the pH range tested and for the temperature experiments a no-enzyme control was subtracted from the values obtained with the enzyme. pH variation was performed in a mixed buffer (0.1 M Tris, MES, citrate) pH determined to 0.02 units at 21 °C. Temperature variation was performed in a thermocycler (GStorm) in MES (20 mM, pH 6.0 measured at 21 °C).

3.3.1.2 Inhibitors of PHS2

Molybdate, a critical component of the activity assay, and acarbose, a known inhibitor of the *E. coli* α-1,4-glucan phosphorylase,[50] were assessed for their ability to inhibit PHS2 by measuring release of Pi from Glc-1-P by PHS2 in the presence of $(\alpha\text{-}1,4\text{-}Glc)_7$ (Fig 3.7). Molybdate was a weak inhibitor for PHS2, with an IC50 (1.4 ± 0.05 mM) well below the concentration used in the phosphorylase activity assay (200 mM), confirming this will stop the reaction. Acarbose was found to be a good inhibitor of PHS2 (IC50 = 1.3 ± 0.2 µM).

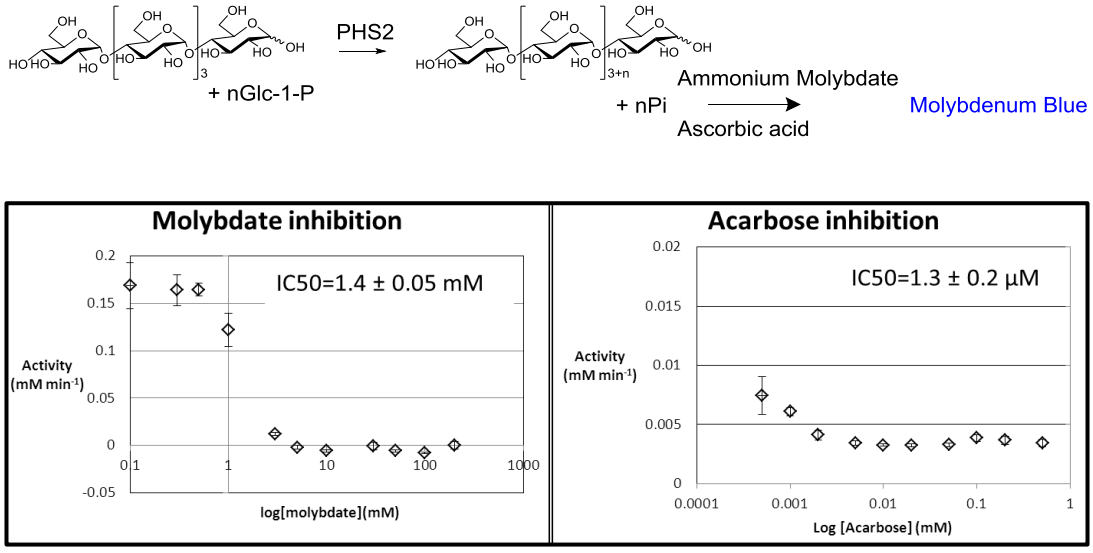
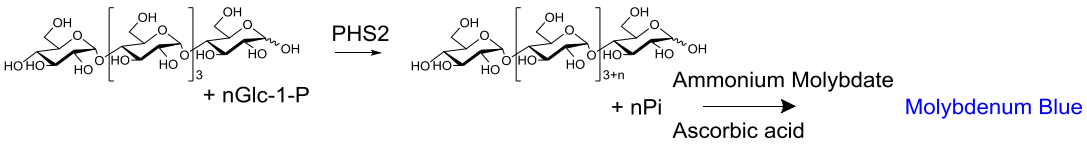


Figure 3.7: Inhibition assays of PHS2

Molybdate, a critical component of the activity assay, was first assessed for its ability to inhibit PHS2 by measuring release of Pi from Glc-1-P (10 mM) by PHS2 (2.5 µg/ml) in the presence of $(\alpha\text{-}1,4\text{-}Glc)_7$ (1 mM) in MES buffer (200 mM, pH 7.0) at 21 °C. Acarbose was then assessed as an inhibitor by measuring the release of Pi from Glc-1-P (1 mM) in the presence of $(\alpha\text{-}1,4\text{-}Glc)_7$ (1 mM) in MES buffer (20 mM, pH 6.0) by PHS2 (5 µg/ml) at 21 °C.

3.3.1.3 Kinetic parameters of PHS2

To determine apparent K_M values, reactions were performed under standard conditions, varying one substrate at a time (Fig 3.8). After 20 min the reactions were terminated by the addition of equal volumes of aqueous ammonium molybdate solution (400 mM), which had been confirmed to stop the reaction completely. Using maltoheptaose (1 mM) as the acceptor the K_M^{app} for Glc-1-P was determined to be 3.9 ± 0.5 mM.



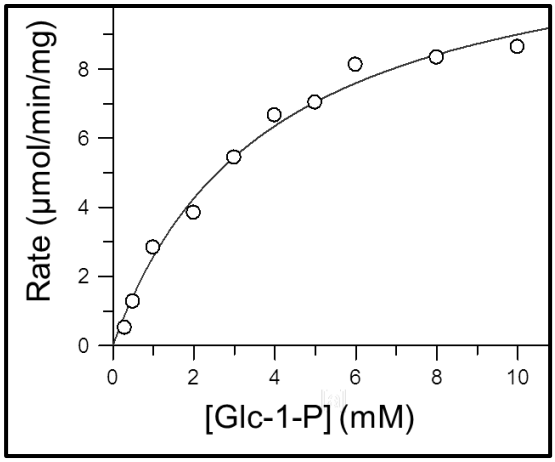
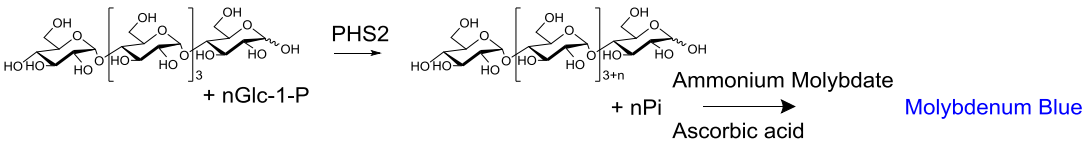


Figure 3.8: Kinetic analysis of PHS2

Experiments were conducted in 12.5 µl MES buffer (20 mM, pH 6.0) at 21 ^{o}C using 8 µg/ml PHS2. Experiments were performed in triplicate in the range 0.1-10 mM Glc-1-P, using 1 mM $(\alpha\text{-}1,4\text{-}Glc)_7$.

The specific activity for heterologously expressed PHS2 in the synthetic direction, measuring the phosphate released from Glc-1-P, under optimum conditions, was measured as 18 µmol/min/mg, giving a turnover number of 30 per second. This is similar to the rabbit muscle enzyme (25 µmol/min/mg)[51] and an order of magnitude faster than starch synthases (0.1-3 µmol/min/mg),[52,53] which have been considered for similar studies. This is also 2-3 times faster than the degradative reaction of *Arabidopsis* PHS2 measured by Lu *et al.* (6-8 µmol/min/mg).[54]

The optimum acceptor length was assayed by measuring K_M^{app} for different length maltooligosaccharides (Table 3.1). K_M^{app} decreases as acceptor length increases, leading to faster turnover, despite the decrease in the V_{max}. This indicates that for longer glucans there are more favourable binding sites along the active site cleft. The minimum acceptor length for glucosylation, as for most enzymes of this class, was found to be $(\alpha\text{-}1,4\text{-}Glc)_4$,[55] an important consideration when using this enzyme.



Acceptor	K_M^{app} (mM)	k_{cat} (1/s)	k_{cat}/K_M^{app} (1/mM/s)
maltotriose – $(\alpha\text{-}1,4\text{-}Glc)_3$	NA	NA	NA
maltotetraose – $(\alpha\text{-}1,4\text{-}Glc)_4$	2.0 ± 0.94	25 ± 6.2	12.5
maltopentaose – $(\alpha\text{-}1,4\text{-}Glc)_5$	1.1 ± 0.43	16 ± 2.8	14.5
maltohexaose – $(\alpha\text{-}1,4\text{-}Glc)_6$	1.0 ± 0.26	18 ± 2.0	18.0
maltoheptaose – $(\alpha\text{-}1,4\text{-}Glc)_7$	0.24 ± 0.05	15 ± 1.0	62.5

Table 3.1: Acceptor length specificity of PHS2
NA indicates less than 1% turnover in 20 min. See Section 8.2.3 for assay details.

3.3.2 Carbohydrate electrophoresis

Carbohydrate electrophoresis (CE) separates carbohydrates, which have been uniformly tagged with a charged and fluorescent label, on the basis of size.[56,57] Separation of amylose polymers up to 80 residues, released by the hydrolysis of starch, has been demonstrated using capillary electrophoresis.[58] In this study 8-aminopyrene-1,3,6-trisulfonic acid (APTS) was attached to carbohydrates by reductive amination (Fig 3.9)[59] and enzymatic extension was monitored using CE.

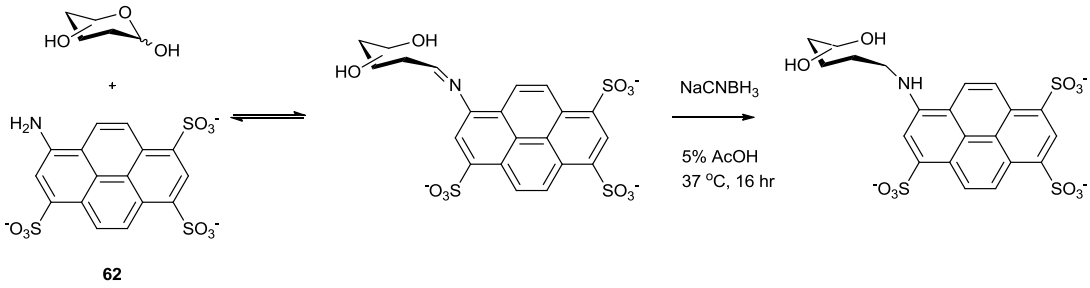


Figure 3.9: Attachment of APTS to carbohydrates

APTS (**62**) was attached to reducing carbohydrates by reductive amination of the spontaneously formed aldimine with sodium cyanoborohydride.

The synthesis reaction used APTS-(α-1,4-Glc)$_5$, which has four intact pyranose rings, as the acceptor for PHS2 and yielded a range of products with a degree of polymerisation (DP) up to 100 glucose residues (Fig 3.10). All the maltopentaose was polymerised in the reaction, although the contaminating maltotriose remained, confirming the minimum acceptor length. Varying the Glc-1-P:acceptor ratio has been established as a way to control the length of the polymer formed.[60] A coupled enzymatic reaction, to remove the inorganic phosphate released during the PHS2 reaction, can also be used to promote the synthesis of longer polymers.[61]

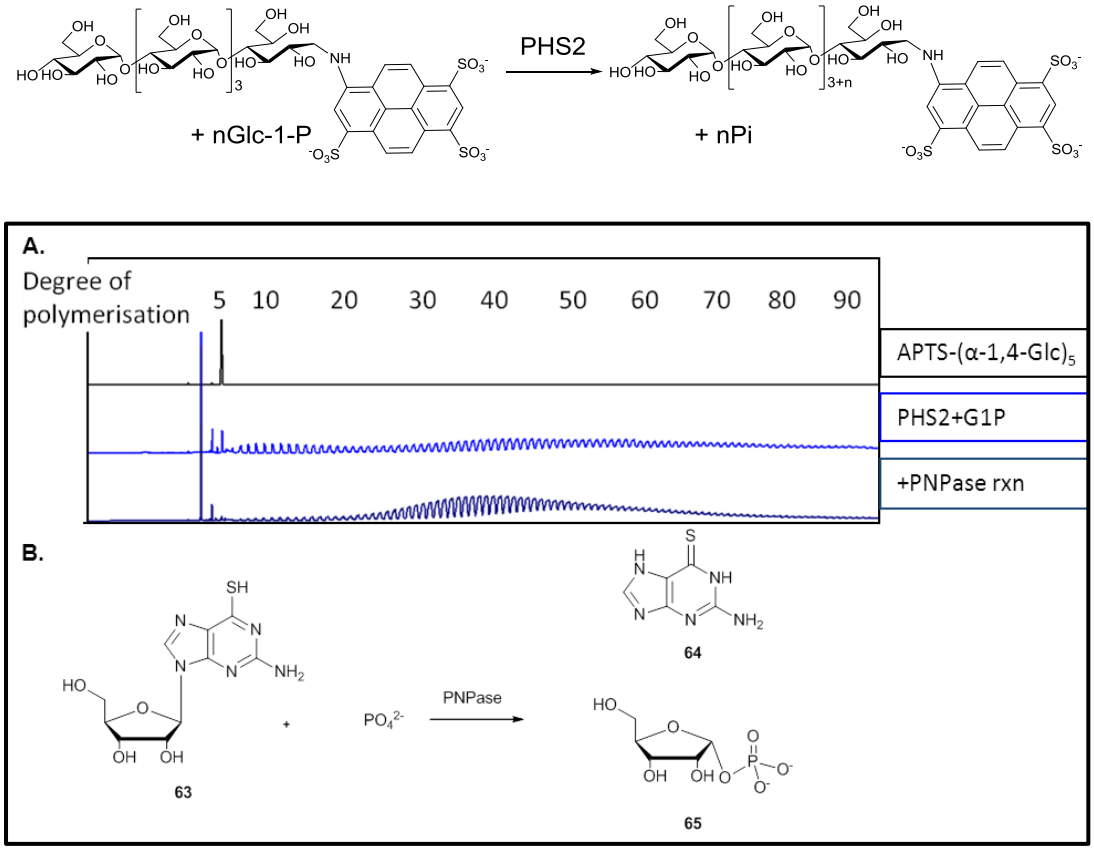


Figure 3.10: Carbohydrate electrophoresis of PHS2 catalysed extension of maltooligosaccharides

A. The reaction contained APTS-(α-1,4-Glc)$_5$ (2.2 μM), Glc-1-P (100 mM), MES (20 mM, pH 6.0) and PHS2 (5 μg/ml) in 100 μl at 30 °C for 24 hr, when it was stopped by heating to 95 °C in boiling water. PNPase rxn contained PNPase (1 U/ml, EC 2.4.2.1, Invitrogen) and amino-6-mercaptopurine riboside (**63**, 100 mM). Signals have been scaled appropriately. **B**. The PNPase catalysed removal of Pi from the reaction, by phosphorolysis of purine riboside (**63**) to purine (**64**) and ribose-1-phosphate (**65**).

In order to confirm the product formed is indeed amylose, the newly synthesised polymers were degraded. APTS-(α-1,4-Glc)$_5$ was extended by PHS2 to DP>50 (Fig 3.11). In the presence of excess inorganic phosphate PHS2 could phosphorolyse the polymer down to very short oligomers, mostly maltotetraose which is the minimum acceptor length. Amyloglucosidase (AG), a specific α-1,4-glucan hydrolase, could also degrade the newly synthesised polymer down to maltose and maltotriose, confirming the α-1,4-linkage of the PHS2-generated product.

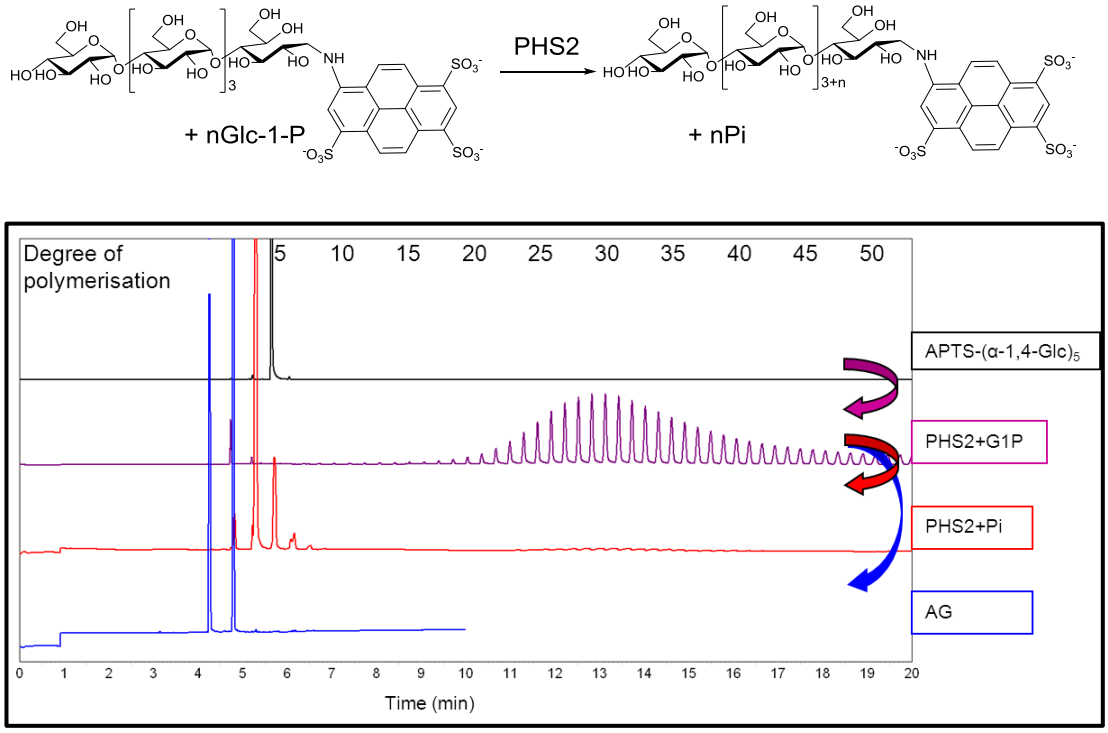


Figure 3.11: Degradation of the PHS2 synthesised polymers

The synthetic reaction (purple) contained APTS-(α-1,4-Glc)$_5$ (2.2 µM, black), Glc-1-P (10 mM), MES (20 mM, pH 6.0) and PHS2 (5 µg/ml) in 100 µl. For the phosphorolysis reaction (red) 50 µl was taken and made up to 100 µl containing extra Pi (100 mM) and fresh PHS2 (5 µg/ml) For hydrolysis 50 µl sample was made up to 100 µl containing amyloglucosidase (5 µg/ml). All reactions were carried out at 30 °C for 24 hr and terminated by heating to 95 °C in boiling water. Signals have been scaled appropriately.

3.3.3 Helical wrapping

α-1,4-Glucans naturally form helices which stain strongly with iodine. The polymers formed by PHS2 stain very strongly, indicating that helical structures are indeed formed (Fig 3.12.A). Polyaniline is a long, electrically conducting polymer with a strong blue colour and is insoluble in water.[62] The vine twining polymerisation of amylose around hydrophobic polymers has been performed with α-1,4-glucan phosphorylases to solubilise hydrophobic materials in water, including lipids[39] and single walled carbon nano-tubes.[40] When the PHS2 catalysed synthesis of amylose is performed in the presence of polyaniline the dye is solubilised (Fig 3.12.B), indicating the carbohydrate is wrapping around the hydrophobic polymer. This technique can be used to solubilise many hydrophobic polymers in aqueous solvents.

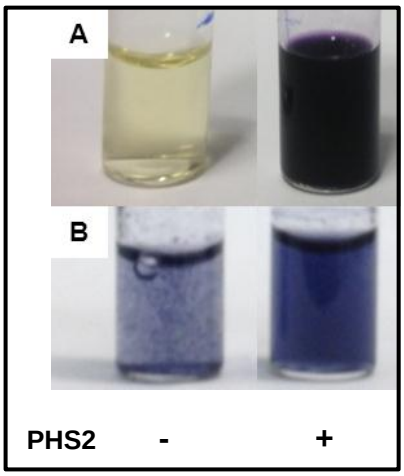


Figure 3.12: Formation of amylose helices using PHS2

PHS2 (50 μg/ml, right) polymerises amylose from $(\alpha\text{-}1,4\text{-}Glc)_5$ (1 mM) by the transfer of Glc from Glc-1-P (100 mM) at 21 °C in MES (20 mM, pH 6.5). **A**. 1% iodine saturated EtOH strongly stains the amylose helices synthesised by PHS2 in 2 hours (right). **B**. Polyaniline (1% saturated DMSO solution) can be solubilised after 16 hours by PHS2 catalysed synthesis (right).

3.3.4 Conclusions on PHS2 activity in solution

PHS2 can synthesise long polymers by transferring Glc from Glc-1-P on to maltooligosaccharide acceptors longer than four residues. This reaction is reversible and the polymers formed can be over 100 residues in length. The polymer is composed of α-1,4-glucans and forms helices that can include hydrophobic moieties, such as polyaniline, and stain with iodine.

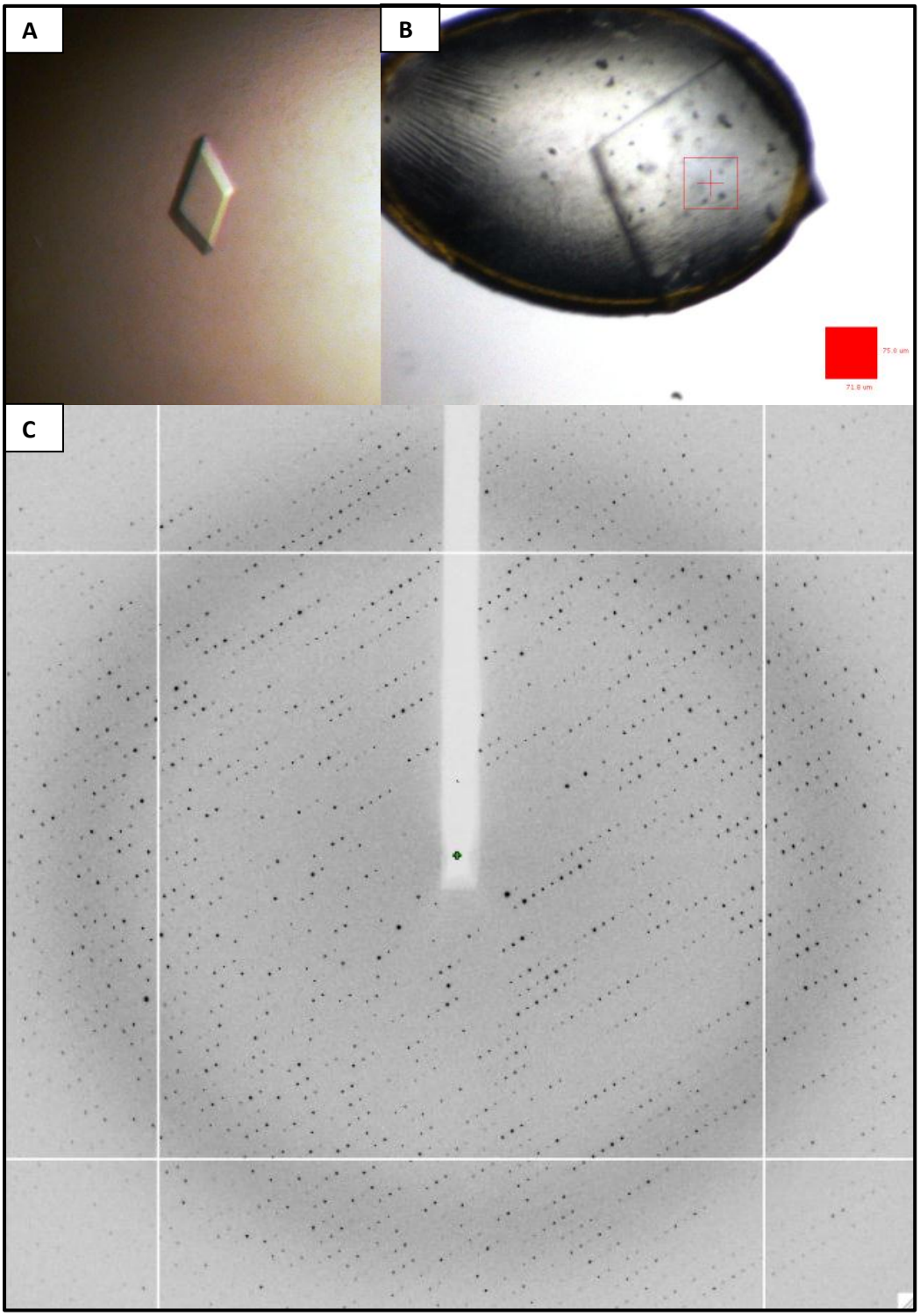


Figure 3.13: PHS2 crystals

A. A crystal of PHS2 in crystallisation drop. PHS2 crystals were approximately 100 µm rhomboids. **B**. A crystal of PHS2 in litho loop at DLS, showing beam centring on the middle of the crystal. **C**. An example diffraction pattern for PHS2.

3.4 Crystal structure of PHS2

The structure of mammalian glycogen phosphorylase (GP) has long been known, and the conformational changes associated with its allosteric and covalent control have been studied.[63,64] The crystal structure of the plant enzyme will help to rationalise the substrate promiscuity and allow the future exploitation of α-1,4-glucan phosphorylases in biotransformations.

3.4.1 Protein crystallisation

After initial crystallisation trials PEG buffered to around pH 7.5 was found to induce crystallisation. Further optimisation was performed varying pH from 6.5-8.5 in MES, HEPES, Tris, citrate and phosphate buffers, and altering the concentration and average molecular weight of PEG. Finally the counter ion for citrate was altered before producing the optimised condition of approximately 20% (w/v) PEG3350, ammonium citrate (100 mM, pH 8.25), 10% (v/v) glycerol. PHS2 (1 µl, 10 mg/ml) in 25% GF buffer was mixed with 1 µl of well solution to give the final drop. Crystals appeared after 2 weeks of hanging drop vapour diffusion and at this point 1 mM substrates in well solution were added to the drop (Fig 3.13). After a further week crystals were mounted in CryoLoops and flash cooled to -173 °C in liquid nitrogen; no additional cryoprotection was required due to the presence of glycerol in the crystallisation solution.

3.4.2 Data collection and structure solution

Data were collected at the Diamond Light Source (DLS) from an unsoaked crystal to 1.7 Å resolution (Table 3.2). A molecular replacement monomer template was prepared from the structure of rabbit muscle glycogen phosphorylase (PDB accession code 1GPB) using CHAINSAW[65] guided by an amino acid sequence alignment of GP and PHS2. A molecular replacement search with PHASER[66] using the monomer template was successful in finding two copies of the subunit, which together yielded a dimer that was consistent with the biological units of both GP and MalP. A final model with R_{work} and R_{free} values of 0.153 and 0.183 respectively, at 1.7 Å resolution, was produced to give the Apo structure.

Data set	Apo	Acarbose	Maltotriose
Data collection			
Space Group	$P2_1$	$P2_1$	$P2_1$
Cell parameters (Å)	a = 84.7, b = 117.1, c = 94.2, β = 106.7°	a = 84.3, b = 117.1, c = 95.0, β = 106.8°	a = 83.7, b = 116.1, c = 94.3, β = 107.4°
Beamline[a]	I03	I02	I02
Wavelength (Å)	0.971	0.980	0.980
Resolution range[b] (Å)	49.11 - 1.70 (1.79 - 1.70)	66.42 - 2.35 (2.48 - 2.35)	65.81 - 1.90 (2.00 - 1.90)
Unique reflections[b]	190517 (26165)	73427 (10684)	132861 (17494)
Completeness[b] (%)	98.9 (93.5)	99.9 (99.8)	98.2 (88.8)
Redundancy[b]	4.2 (3.2)	3.7 (3.7)	4.0 (3.4)
R_{merge}[b, c]	0.080 (0.433)	0.098 (0.791)	0.064 (0.558)
R_{meas}[b, d]	0.091 (0.523)	0.115 (0.926)	0.073 (0.661)
Mean I/σ(I)[e]	11.8 (2.5)	8.7 (1.6)	12.9 (2.2)
Wilson B value (Å^2)	17.8	45.6	28.1
Refinement			
Reflections: working/free[d]	180900/9587	69673/3670	126128/6643
R_{work}[f]	0.153	0.184	0.196
R_{free}[f]	0.183	0.240	0.241
Ramachandran favoured/ allowed[g] (%)	98.0/99.9	96.4/99.8	97.6/99.9
Ramachandran outliers[g]	2	3	1
rmsd bond distances (Å)	0.016	0.014	0.014
rmsd bond angles (º)	1.46	1.60	1.55
Contents of model			
Protein residues	824 (A chain) 825 (B chain)	824 (A chain) 827 (B chain)	824 (A chain) 824 (B chain)
Pyridoxal phosphate	2	2	2
Ligands[h]	0	2 x ACR; 3 x GLC	2 x M3
Precipitant/cryoprotectant molecules[h]	2 x GOL; 4 x PEG	1 x GOL	1 x GOL; 1 x PEG
Water molecules	1844	344	919
Average atomic displacement parameters (Å^2)			
Main chain atoms	17.2	45.4	31.9
Side chain atoms	18.8	47.3	33.9
Pyridoxal phosphate	15.9	44.1	31.0
Ligands[h]	-	ACR: 72.5; GLC: 77.1	M3: 39.4
Precipitant/cryoprotectant molecules[h]	GOL: 30.8; PEG: 34.6	GOL: 48.3	GOL: 35.0; PEG: 35.0
Water molecules	27.8	39.7	37.5
Overall	19.2	46.4	33.2
PDB accession code	4BQE	4BQF	4BQI

Table 3.2: Summary of PHS2 X-ray data and model parameters

[a] I02, I03 = beamlines at the Diamond Light Source (Oxfordshire, UK).
[b] The figures in brackets indicate the values for outer resolution shell.
[c] $R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl and $\langle I(hkl)\rangle$ is the weighted average intensity for all observations i of reflection hkl.
[d] $R_{meas} = \sum_{hkl} [N/(N-1)]^{½} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where N is the number of observations of reflection hkl.
[e] The data sets were split into "working" and "free" sets comprising 95% and 5% of the data, respectively. The free set was not used for refinement.
[f] The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum(|F_{obs} - F_{calc}|)/\sum|F_{obs}| \times 100$, where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively.
[g] As calculated using MOLPROBITY.[67]
[h] Abbreviations: ACR = acarbose; GOL = glycerol; GLC = glucose; PEG = polyethylene glycol; M3 = maltotriose.

3.4.3 Apo structure of PHS2

In the asymmetric unit of the crystal two copies of the protein were found, which together comprise the biologically relevant dimer (Fig 3.14). The dimer interface of PHS2 has about 2600 Å^2 of solvent accessible surface buried per monomer, which is comparable in extent to that of the GP dimer (2500 Å^2), with 24 H-bonds giving a calculated solvation free energy gain of 23.7 kcal/mol upon dimerisation.[68]

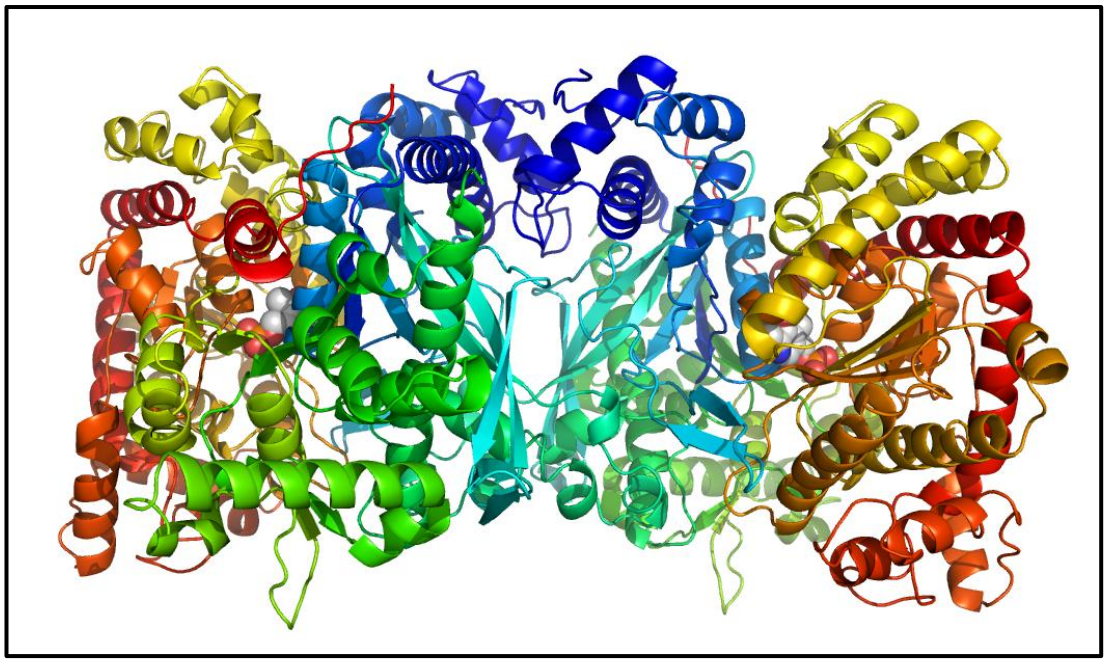


Figure 3.14: Structure of the PHS2 dimer

PHS2 forms a dimer, which has an interface of around 2600 Å^2. The prosthetic group PLP is shown as spheres, indicating the position of the active site.

As expected there is a pyridoxal phosphate (PLP) covalently bound, about 20 Å from the surface, by formation of a Schiff base with Lys687 (Fig 3.15). The phosphate group acts as a catalytic acid/base in the reaction, and also has a role in coordinating the phosphate in the active site.[69]

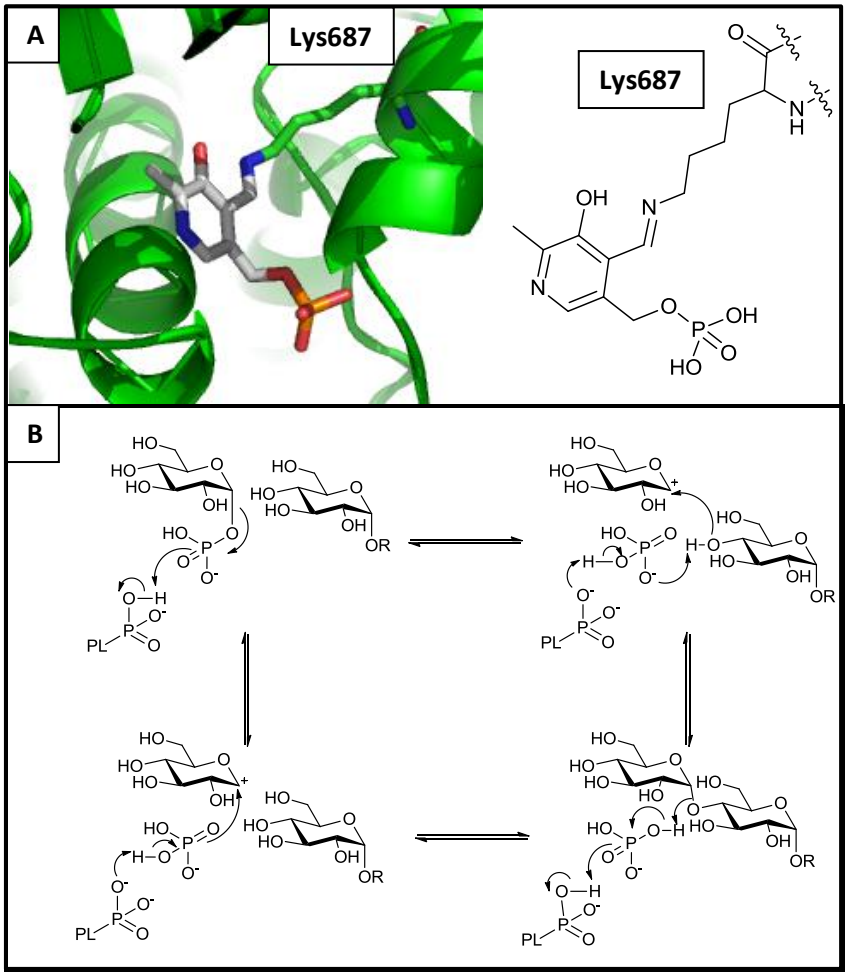


Figure 3.15: Pyridoxal phosphate in the active site of PHS2

A. Pyridoxal phosphate is bound in the active site of PHS2 by formation of a Schiff base with the Lys687. **B**. The catalytic mechanism of α-1,4-glucan phosphorylases, showing the reversibility of these enzymes and the role of PLP in the reaction.[69] PL = Enzyme bound pyridoxal. R = $(\alpha\text{-}1,4\text{-}Glc)_n$.

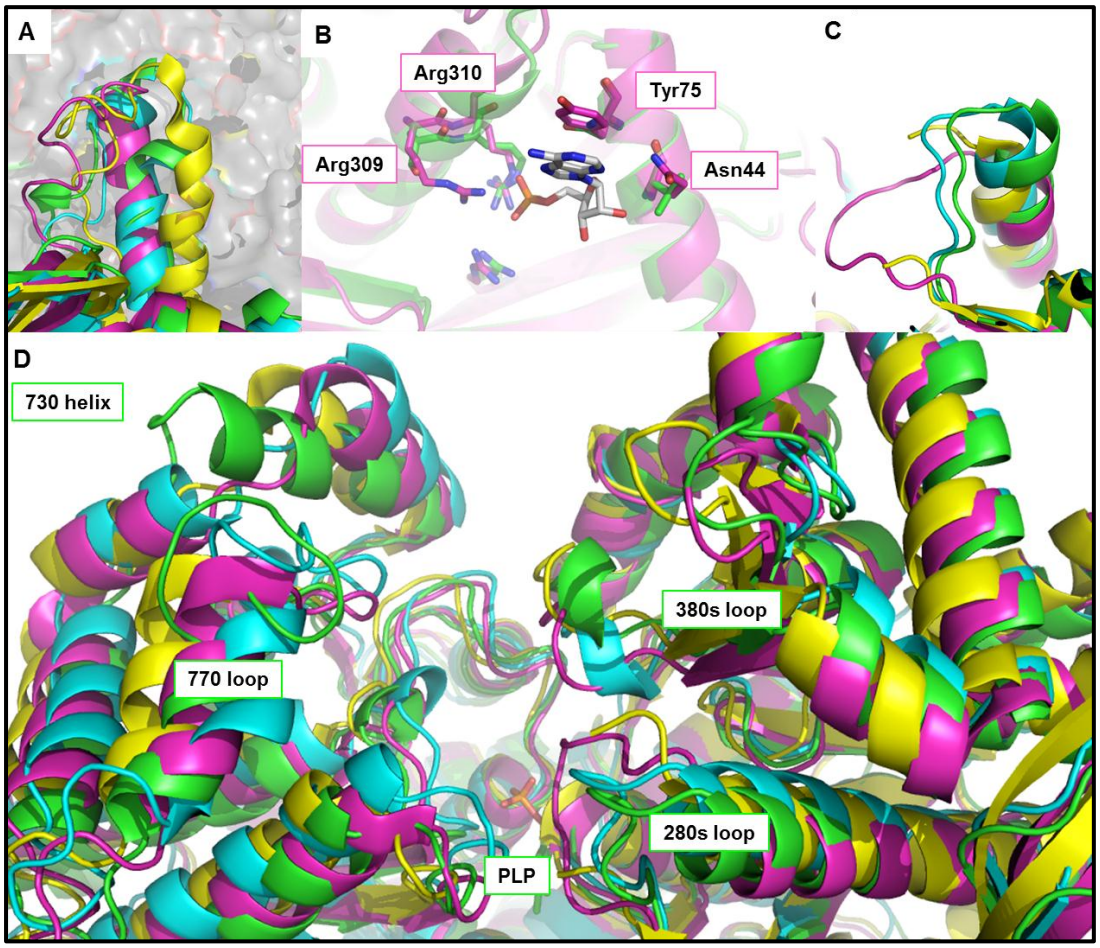


Figure 3.16: The sites of allosteric regulation in GP are not present in PHS2

A. The tower helices, involved in inter-subunit interactions, in PHS2 and MalP adopt a similar conformation whereas the GP structure adopts a variable conformation, depending on state. Chain B surface in grey. **B**. PHS2 does not have the residues involved in AMP binding with the exception of two arginines involved in phosphate binding.[63] AMP shown from GP structure. **C**. The 280s loop, which controls access to the active site and thus allosteric regulation in GP, adopts a more open conformation in PHS2 and MalP. **D**. The active site in PHS2 and MalP is substantially more open than in either conformation of GP. PHS2 in green. GP R state in magenta, T state in yellow. MalP in blue.

There are many close structural homologues of PHS2 in the PDB, with an RMSD less than 2.0,[70] which come from just seven non-redundant proteins: mammalian[63] and yeast GP structures[71] and the *E. coli* maltodextrin phosphorylase, MalP.[72] At the subunit level, PHS2 most closely resembles the A form mammalian enzymes and the yeast enzyme, whilst at the dimer level, PHS2 most closely resembles the B form mammalian enzymes. The latter suggests that PHS2 has a quaternary structure that is closest to the non-phosphorylated, less active, state of the mammalian enzymes. However, the N-terminus of PHS2, the site of phosphorylation in the mammalian and yeast enzymes, takes up a unique conformation which does not impinge on the active site.

In the inactive state of the mammalian enzyme, the so-called tower helices are almost anti-parallel, whilst they are almost perpendicular in the active form.[63] By contrast, in PHS2 the tower helices cross each other at an angle of roughly 55°, which is similar to the situation in yeast GP and MalP (Fig 3.16.A). In addition, there is no recognisable binding site for the allosteric activator AMP at the dimer interface (Fig 3.16.B), where binding of AMP changes the quaternary structure of the dimer such that the tower helices re-orient with respect to one another.[63] This has the effect of drawing the 280's loops away from the active sites. Thus PHS2 would be expected to be insensitive to inhibitors that target this site in the mammalian enzymes.

The active site is more accessible in PHS2 than in the rabbit muscle enzyme, partly due to the 380's loop adopting a more open conformation, and the shorter 280's loop (Fig 3.16.C), which would appear to be incapable of gating access to the active site or of binding allosteric inhibitors. The 730 helix and 770 loop are also further from the active site (Fig 3.16.D). Overall the structures indicate PHS2 is structurally more closely related to the less elaborately regulated mammalian liver and yeast GPs and the unregulated MalP, and it is not possible for PHS2 to be regulated in the same way as the rabbit muscle enzyme.

3.4.4 Substrate-bound structures of PHS2

Crystals which had been soaked with acarbose (**66**, Acr), Glc-1-P, $(\alpha\text{-}1,4\text{-}Glc)_3$, 1,4-dideoxy-1,4-imino-D-arabinitol, β-cyclodextrin (βCD), $(\alpha\text{-}1,4\text{-}Glc)_4$+Glc-1-P, $(\alpha\text{-}1,4\text{-}Glc)_4$+Pi, and $(\alpha\text{-}1,4\text{-}Glc)_7$+Pi all diffracted successfully and were solved from the Apo structure, with which they were isomorphous. Only the $(\alpha\text{-}1,4\text{-}Glc)_3$ and Acr showed sufficient unique additional density to merit full refinement.

3.4.4.1 Oligosaccharide-bound structures

Upon soaking with oligosaccharides the PHS2 protein crystals tended to crack, although this was likely attributable to ligands binding near crystal contacts, rather than to conformational changes. Nevertheless, useable X-ray data were collected for soaks with $(\alpha\text{-}1,4\text{-}Glc)_3$, $(\alpha\text{-}1,4\text{-}Glc)_4$ and β-cyclodextrin (βCD). All three oligosaccharides bound to the surface of the enzyme, although only three glucosyl moieties could be resolved in the electron density for the longer polymers. This is the first time this surface site has been seen to date, although it is adjacent, and almost contiguous with, the $(\alpha\text{-}1,4\text{-}Glc)_5$ site previously observed in rabbit muscle GP.[73] The PHS2 site is roughly 40 Å from the PLP in a direct line, whilst the GP site borders the mouth of the active site channel. The GP surface binding site is almost identical in MalP, but all the interacting amino acids are different in PHS2 (Fig 3.17.A). The presence of Asn435 across the binding site would appear to exclude the probability of binding in this site, and so the failure to find bound substrate is hardly surprising. The unique PHS2 surface binding pocket is composed of amino acids identical to those in GP and MalP, with the exception of Leu361Met (Fig 3.17.B). The fold is substantially the same although the loop 210-216 is truncated in MalP. Indeed, the juxtaposition of the reducing end of the $(\alpha\text{-}1,4\text{-}Glc)_3$ in PHS2, and the non-reducing end of the $(\alpha\text{-}1,4\text{-}Glc)_5$ in GP, suggests that the latter could theoretically bind longer oligosaccharides that span both sites. It is therefore surprising that no carbohydrates have been observed in this site in GP or MalP.

The enzymes may associate with their soluble substrates via this surface site, rather than remaining free in the cytosol. Theses alternative surface sites may determine the binding to the different substrates, SHG for PHS2 and glycogen for GP. GP also prefers branched acceptors whilst PHS2 prefers longer, unbranched acceptors,[37] which may be determined by the proximity of these surface sites to the active sites.

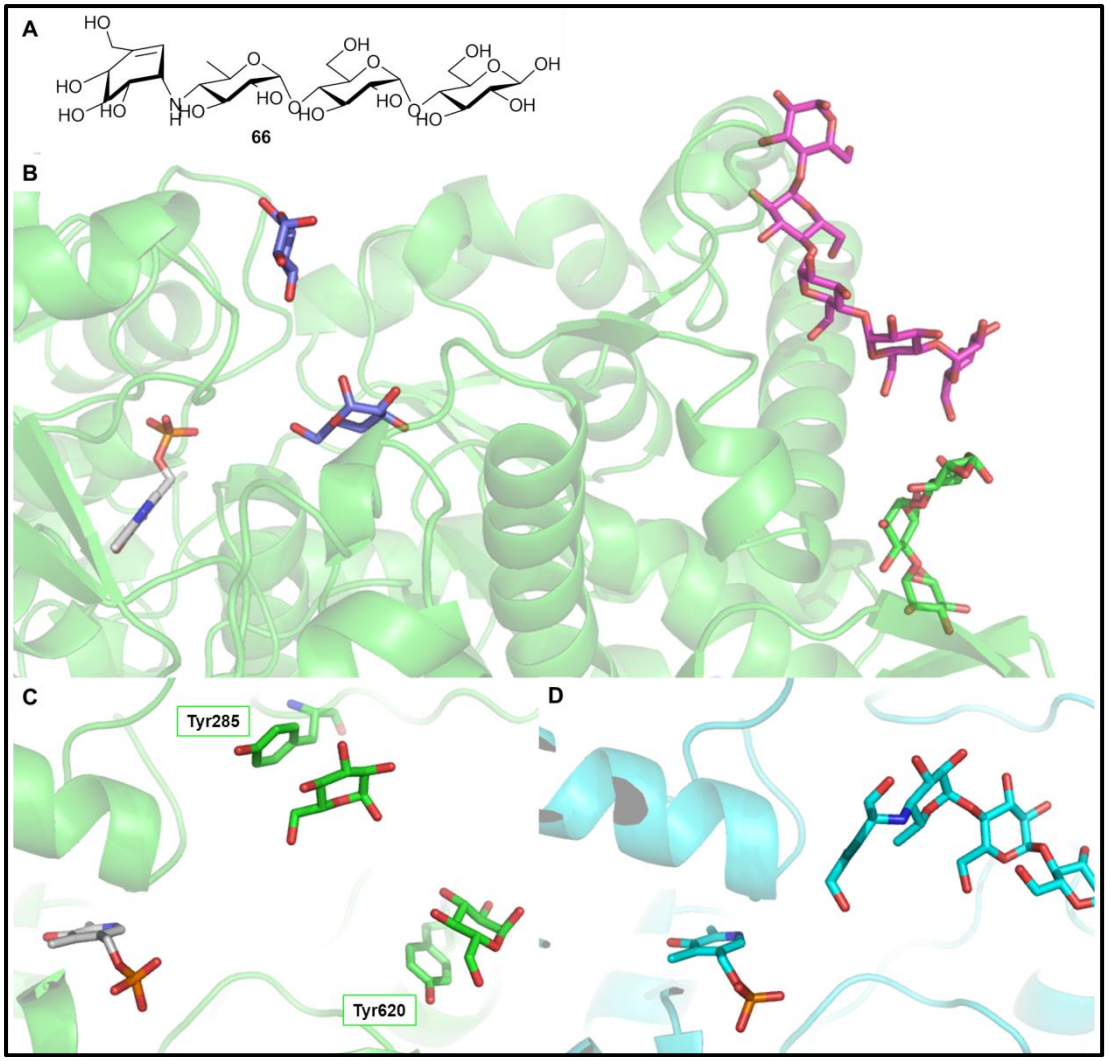


Figure 3.18: Oligosaccharides binding to PHS2

A. The structure of acarbose (**66**). **B**. Oligosaccharides (green) bind to the surface of PHS2 in a unique position compared to GP (magenta). In the active site only two glucosyl units from acarbose can be resolved (purple), neither of which is close enough to the active site pyridoxal phosphate (white) to act as an acceptor. **C**. Acarbose binds to the active site of PHS2 but can only be resolved as two glucosyl residues, adjacent to Tyr285 and 620, roughly corresponding to the 4-amino-1,6-dideoxy glucose and reducing terminal glucoses of acarbose bound in MalP (**D**).[74]

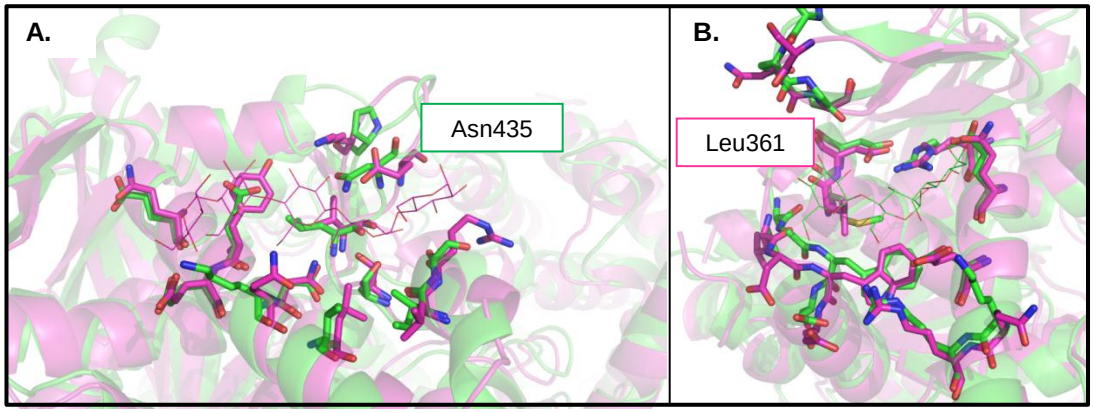


Figure 3.17: Details of oligosaccharides binding to α-1,4-glucan phosphorylase surface sites

A. The GP (magenta) surface binding site is not present in PHS2 (green) and the conformation of Asn435 would not allow substrates in this site. **B**. The PHS2 surface binding site is conserved in GP except for the presence of Leu361 instead of Met. Maltooligosaccharides are shown as lines in the colour of the binding protein.

Since no oligosaccharides were successfully bound in the active site, attempts were made to soak the pseudo-tetrasaccharide inhibitor acarbose into crystals of PHS2. Acarbose is a potent inhibitor of α-amylases, binding in the substrate binding site across the active site.[75] In this study a structure was obtained with acarbose bound to the same surface site of PHS2 as the maltotriose, making very similar contacts to the protein (Fig 3.18). In one subunit, the whole of the inhibitor was resolved, whereas only the two central rings could be seen in the other subunit. This was, at least in part, due to the proximity of a symmetry related subunit, which effectively blocks the site for the cyclohexitol ring. There was, in addition, discontinuous residual density in the active site channel (Fig 3.18.C). In one subunit, it was possible to model two separate glucosyl moieties, although it was not clear if they were parts of the same or different acarbose molecules. In the second subunit of PHS2, only the inner site had sufficient density to model a glucosyl residue. A comparison with the acarbose complex of MalP[74] showed that the ring closest to the PLP (some 9 Å distant) is between the 4-amino-4,6-dideoxy-D-glucopyranose and cyclohexitol of the acarbose in the MalP complex, whilst the ring further from the PLP (some 14 Å distant) roughly corresponded to the glucosyl moiety at the reducing end of the acarbose in the MalP structure. However the substrates do not directly overlay indicating there are differences in the substrate binding in the Active sites of these enzymes.

3.4.4.2 Glc-1-P-bound structures

Addition of 1 µl of Glc-1-P (10 mM) to crystallisation drops resulted in precipitation of the PHS2, although when a 1 mM solution was added crystals could be grown which diffracted successfully. As this concentration is below the K_M for Glc-1-P of this enzyme (see Section 3.3.1) the structure does not show complete occupancy and not enough extra density can be resolved to be reported as bound substrate. This extra density might equate to one Glc-1-P π-stacking to Tyr620 in one subunit and Tyr285 in the other subunit, making the same contacts as the glucose moieties in the acarbose bound structures (Fig 3.19). When Glc-1-P and $(\alpha\text{-}1,4\text{-}Glc)_4$ were soaked into the PHS2 crystals one subunit had additional density, adjacent to the PLP, which might equate to Glc-1-P and matches the binding site of Glc-1-P in the active site of MalP.[76]

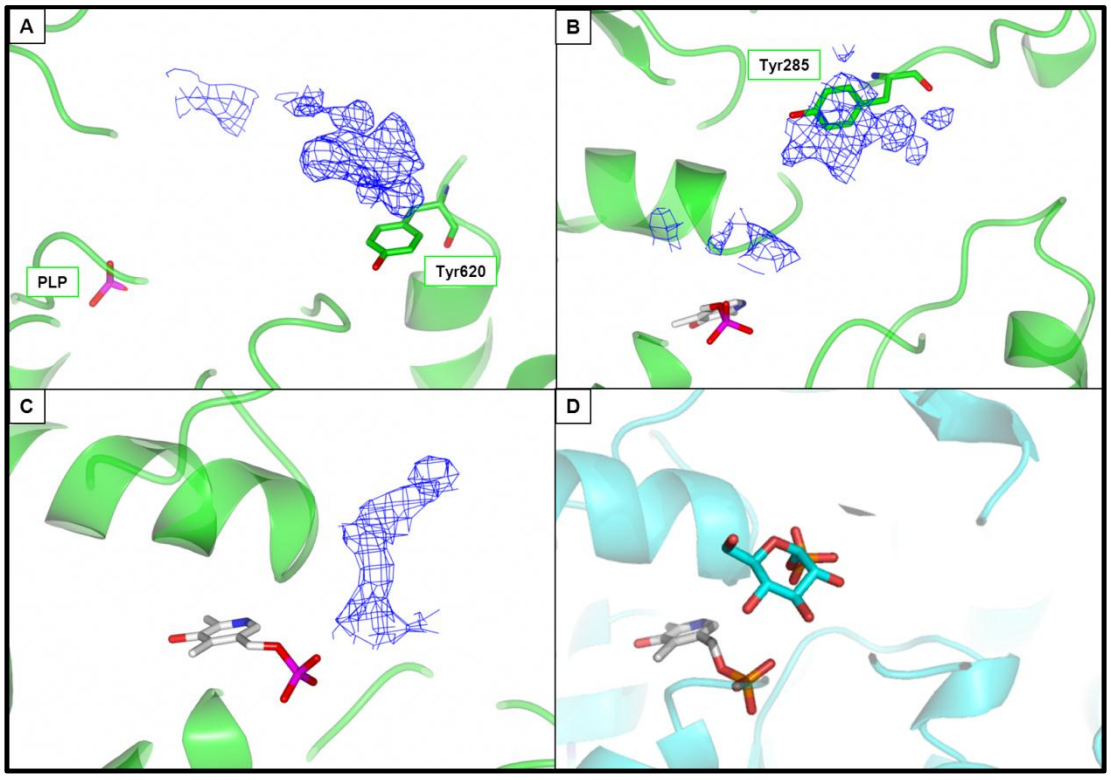


Figure 3.19: Additional density when Glc-1-P was soaked into in PHS2 crystals

A. Density adjacent to Tyr620. **B**. Density adjacent to Tyr285. **C**. Density adjacent to PLP in active site when $(\alpha\text{-}1,4\text{-}Glc)_4$ is also soaked in to the crystal. **D**. Glc-1-P in MalP active site from 1L5V.[76]

3.4.5 Conclusions on the PHS2 structures

The active site of PHS2 is substantially similar to that of GP and MalP. Whilst it has not been possible to resolve oligosaccharides bound in the active site, the structures show the importance of π-stacking for oligosaccharide binding. The fact that acarbose doesn't directly overlay with related structures suggests that the PHS2 active site contains some important differences. The unprecedented surface binding site is consistent with the unusual substrate upon which PHS2 acts. With this range of structures we can see how the plant α-1,4-glucan phosphorylase has adapted for its substrate and lost the controls present in glycogen phosphorylase.

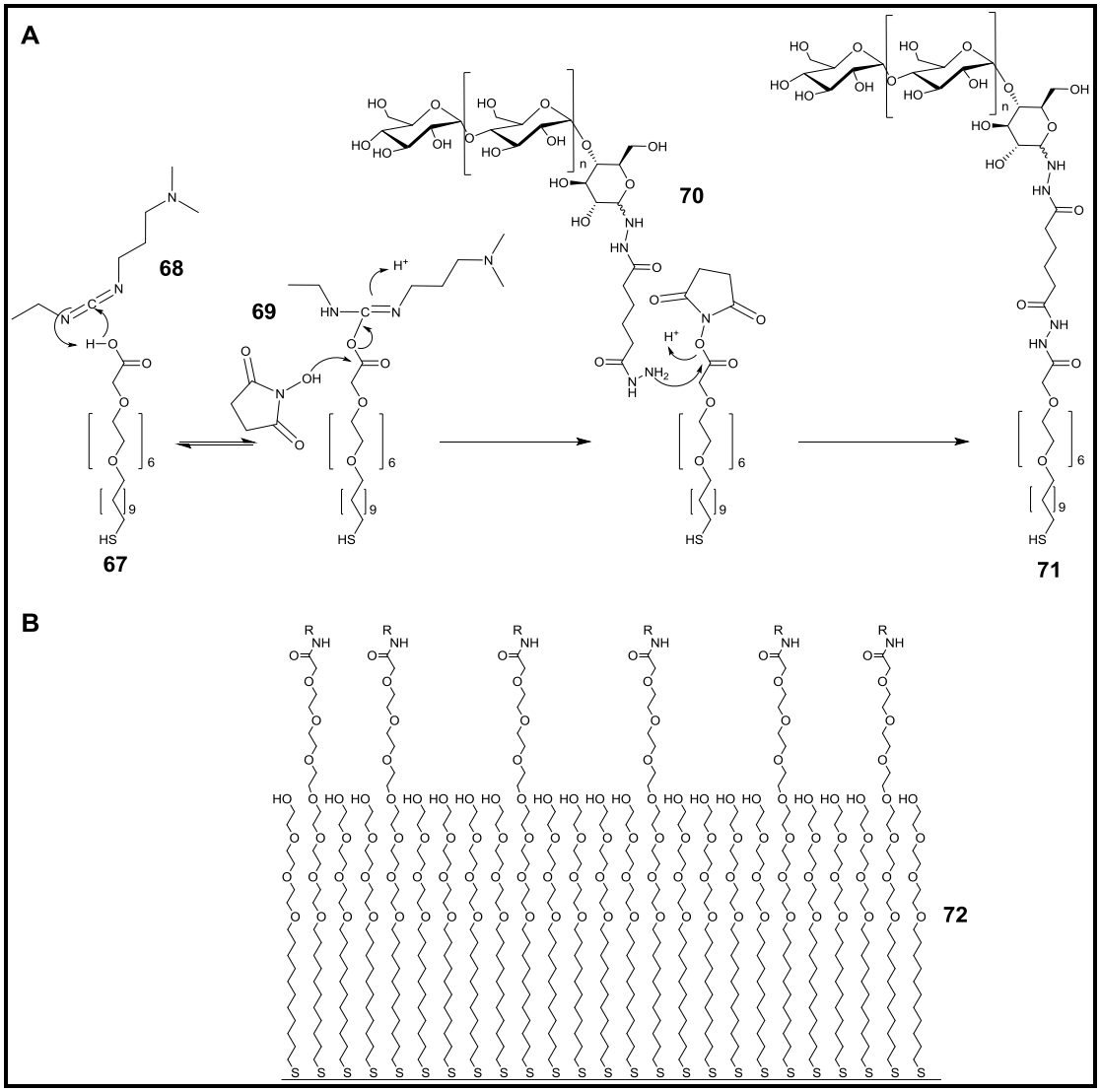


Figure 3.20: Immobilisation of glucans on SPR surfaces

A. The EDC (**68**)/NHS (**69**) promoted coupling of hydrazide glycosides (**70**) to the SAM linker (**67**). **B**. A schematic of the SAM composed of mixed spacer (**72**) and linker molecules (**71**). R = $(\alpha\text{-}1,4\text{-}Glc)_n$.

3.5 Surface glucan synthesis by PHS2

In order to assay starch active enzymes quantifiable starch-like surfaces are required. PHS2 can rapidly synthesise α-1,4-glucans in solution and, by immobilising acceptors, insoluble glucan surfaces were developed.

3.5.1 SPR analysis of surface glucan synthesis by PHS2

Glucans for surface immobilisation were first functionalised with adipic acid dihydrazide (ADH) through their reducing termini.[77] The SPR surfaces were made by peptide coupling of ADH-glucoside to alkane-thiol-carboxylates and the formation of self-assembled monolayers (SAM) on the gold surface of the SPR chips (Fig 3.20).[78,79] This technique allowed the rapid assembly of different glucans at various concentrations, insulated from the SPR surface.

3.5.1.1 Enzyme reactions on maltoheptaose surfaces

In order to analyse PHS2 activity on a surface analogous to a starch granule, $(\alpha\text{-}1,4\text{-}Glc)_7$ was immobilised on an SPR chip. PHS2 binds well to the surface and the binding is enhanced by Glc-1-P, but no synthesis was observed upon co-injection with Glc-1-P on this surface (Fig 3.21). The crystallographic evidence would suggest the surface binding site might account for the donor free binding, whilst in the presence of Glc-1-P the active site binds more tightly to the surface. The lack of synthesis would suggest either the glucan structure on the surface cannot reach into the active site or the conformation of the surface prevents PHS2 binding in an active mode, possibly mediated by the surface binding site elucidated in the crystal structures (see Section 3.4.4.1).

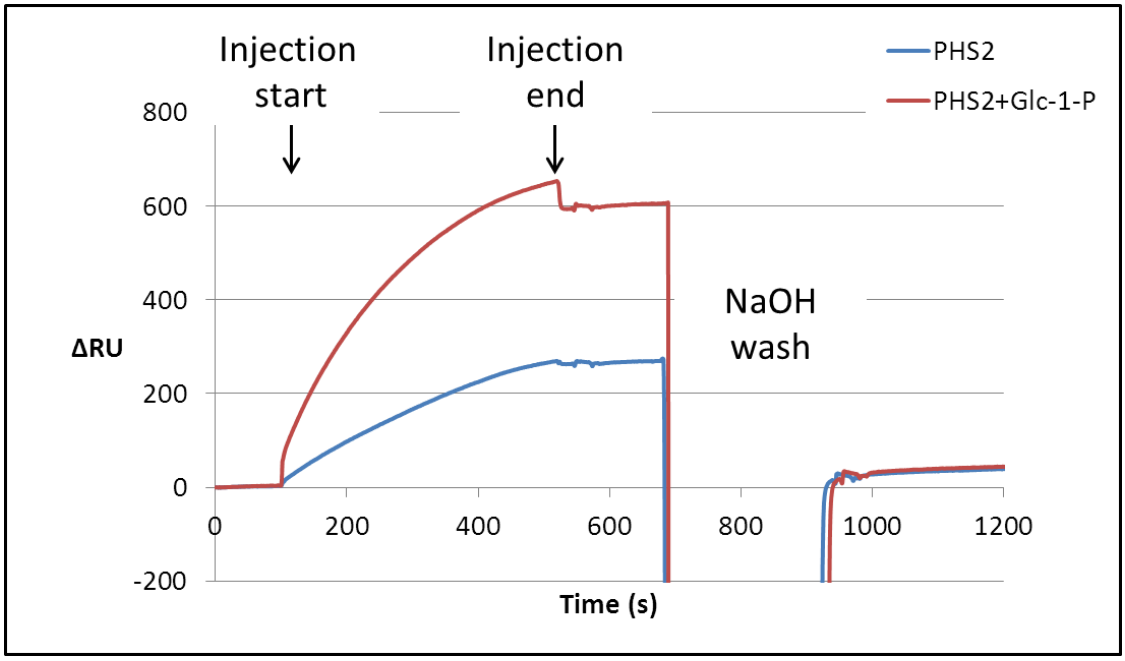


Figure 3.21: PHS2 activity on maltoheptaose SPR surface

PHS2 (15 µg/ml) binds to the maltoheptaose surface and is almost completely removed by NaOH (10 mM) wash. When co-injected with Glc-1-P (4 mM) a larger response is seen over the course of the injection, but this is also removed by the NaOH wash.

3.5.1.2 Fractionated glucan

As an alternative to $(\alpha\text{-}1,4\text{-Glc})_7$, fractionated glucan (FG), a partially degraded starch fragment, was immobilised on the SAM as this is likely to provide a more open and accessible surface. FG was obtained by removing low DP contaminants from the Glucidex™, which would not act as acceptors for PHS2. This material has 3% branching, as judged by ^{1}H NMR and an average molecular weight of 19 kDa (DP ~ 110) (Section 8.2.3.2). [80]

3.5.1.3 Enzyme reactions on low FG density SPR surfaces

A SAM surface, containing 2% linker, was immobilised on the SPR surface and FG was coupled, using the same chemistry as for the $(\alpha\text{-}1,4\text{-Glc})_7$ surface, and interrogated using SPR. PHS2 was able to bind rapidly to this surface, with the response showing typical surface binding kinetics and slow dissociation (Fig 3.2). It was not possible to extract surface binding kinetics from the data collected, due to the high K_a of this protein for the surface.

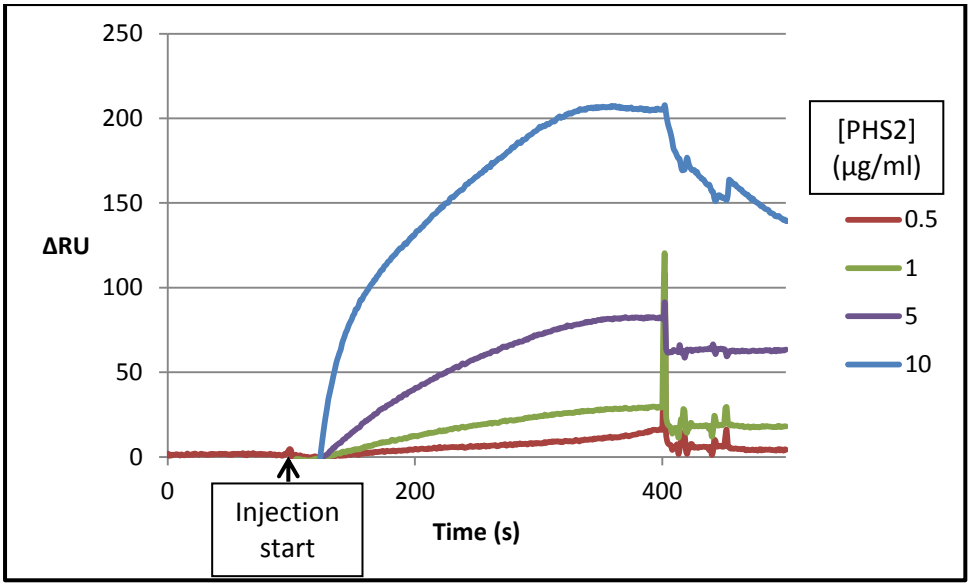


Figure 3.22: SPR analysis of PHS2 binding to the low FG density surface

The low sugar density surface was interrogated by injecting 100 µl PHS2 (0.5-10 µg/ml in running buffer) on to the surface, normalised 10 s after injection.

When PHS2 (5 µg/ml) was co-injected with Glc-1-P a larger, stable response was observed indicating surface extension. After extension with PHS2 the surface could be repeatedly probed with enzymes showing consistent results. PHS2 further extended the surface by a similar amount and the response positively correlates with Glc-1-P concentration (Fig 3.23), confirming the response is due to synthesis, not protein-surface interactions.

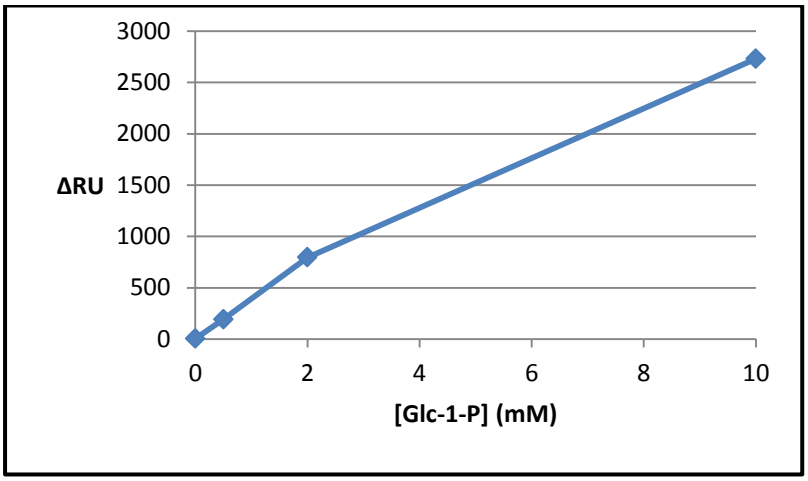


Figure 3.23: Kinetics of low FG density surface extension using PHS2

As the concentration of Glc-1-P increases, co-injected with a constant concentration of PHS2 (5 µg/ml), the response increases.

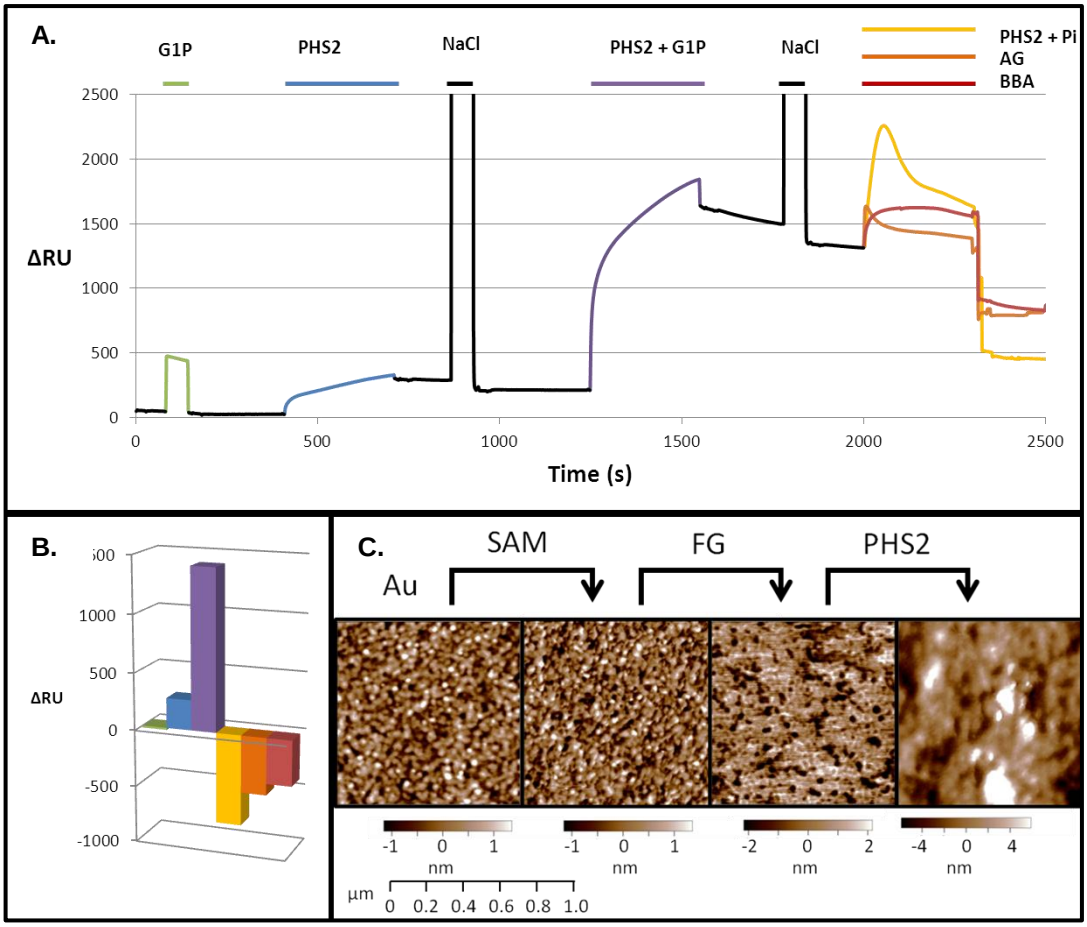


Figure 3.24: PHS2-catalysed glucan synthesis on low FG density surface, as analysed by SPR and AFM

A. Composite SPR sensogram and **B**. quantification of the final change in response, using the low sugar density surface. Green: When Glc-1-P (10 mM) was injected over the surface (20 µl injection) there was a bulk change that returns to baseline when the injection ends. Blue: When PHS2 (5 µg/ml) was injected alone there was around 250 RU signal increase, which did not decrease significantly when the injection ended, or with NaCl wash (100 mM, 20 µl). Purple: When PHS2 (5 µg/ml) and Glc-1-P (10 mM) were co-injected, a large response (1200 RU) was seen, giving rise to an elevated and stable baseline after NaCl wash (100 mM, 20 µl). Yellow: When PHS2 (5 µg/ml) and Pi (10 mM) were co-injected a large decrease in signal (800 RU) was observed. Orange: Amyloglucosidase (1 mg/ml, *Aspergillus niger*, Sigma) caused a significant decrease in signal (500 RU) when injected over the surface. Red: Barley β-amylase (3 mg/ml, Megazyme) also gave a decrease (400 RU) in signal when injected over the extended surface. All enzyme reactions used 100 µl injections. **C**. AFM of SPR chip showing a gold sputtered surface, with a standard deviation, corresponding to surface roughness, of 536 pm. After addition of the 2% SAM layer the surface remains smooth (SD 563 pm). Coupling of FG to the surface caused an increase in roughness (SD 1.069 nm); when PHS2 (5 µg/m) and Glc-1-P (10 mM) were co-injected, a large change in surface morphology was observed, with an increase in SD from the mean of the surface thickness to 2.8 nm.

After this initial extension the surface can then be enzymatically degraded: PHS2 can phosphorolyse the surface when co-injected with Pi, showing a decrease of over 1000 RU, after initial protein binding to the surface; specific α-1,4-glucan hydrolases, amyloglucosidase (AG) and barley β-amylase (BBA), showing a decrease of approximately 450 and 400 RU respectively (Fig 3.24). These results confirm the increase in response is indeed caused by the PHS2 mediated synthesis of amylose.

Previously it has been reported that atomic force microscopy (AFM) can be used to analyse soluble polysaccharides synthesised on SPR surfaces[43] and this technique was used to measure the insoluble starch like surface in this study. The initial gold surface and the surface after immobilisation of the SAM were relatively smooth, with a standard deviation from the mean of 0.5-0.6 nm (Fig 3.24.C). Immobilisation of FG caused the surface to become rougher (SD = 1.1 nm) and after PHS2 injection the surface became much more uneven (SD = 2.8 nm). These results confirm that PHS2 catalyses the addition of substantial amounts of material to the glucan surface which can be removed by α-1,4-glucan hydrolases.

3.5.1.3 Enzyme reactions on high FG density SPR surfaces

A higher sugar density surface was prepared by coupling FG to the SAM linker in solution and then applying this to the surface, without dilution with the SAM spacer. Figure 3.25 shows a typical SPR trace of enzyme injections over this high FG density surface. When PHS2 alone is injected over the surface a much greater response is seen, compared to the low FG density surface, as there is more ligand immobilised on the surface. The co-injection of PHS2 and Glc-1-P gave an increase of around 5000 RU of polysaccharide synthesis. Hydrolases were then used to test this surface: amyloglucosidase, which cannot degrade granular starch, is unable to degrade this surface; α-amylase, which can readily degrade starch granules, is able to degrade the surface, removing nearly all the newly synthesised glucan. α-Amylase is able to hydrolyse both α-1,4 and α-1,6 glucosidic linkages, so can remove some of the branch points already present in the FG surface, rendering the surface unusable.

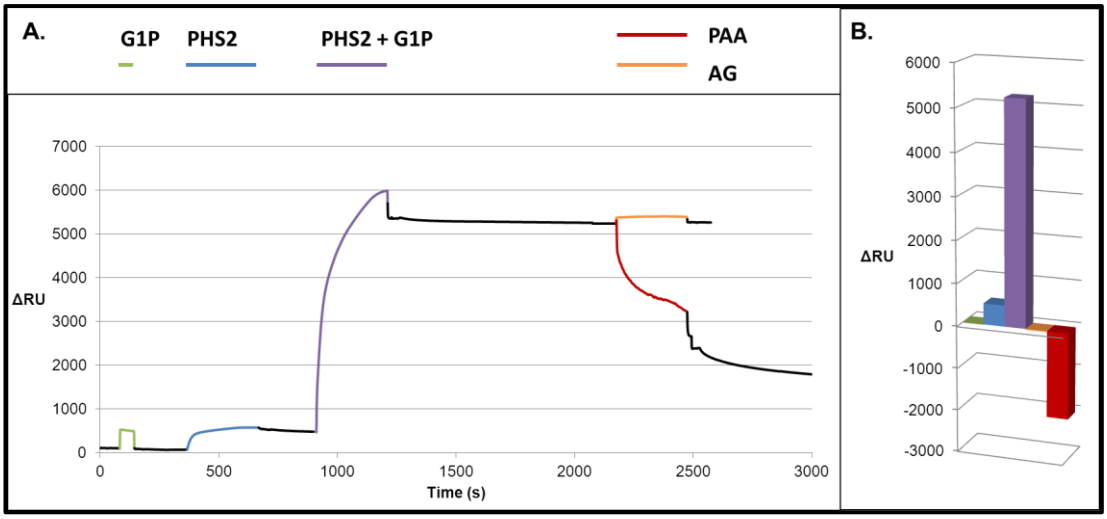


Figure 3.25: Typical SPR trace of enzyme reactions on the high FG density surface

A. Composite SPR sensogram and **B**. quantification of the final change in response, using the high sugar density surface. Green: When Glc-1-P (10 mM) was injected over the surface (20 µl injection) there was a bulk change that returns to baseline when the injection ends. Blue: When PHS2 (5 µg/ml) was injected alone there was around 500 RU signal increase, which did not decrease significantly when the injection ended. Purple: When PHS2 (5 µg/ml) and Glc-1-P (10 mM) were co-injected, a large response (5000 RU) was seen, giving rise to an elevated and stable baseline. Orange: Amyloglucosidase (AG, 1 mg/ml) caused no decrease in signal when injected over the surface. Maroon: Porcine pancreatic α-amylase (PAA, 3 mg/ml, Megazyme) also gave a large decrease (>2000 RU) in signal when injected over the extended surface, and continued to degrade the surface after the end of the injection by a further 1500 RU, as the enzyme slowly dissociates and continues to degrade the glucans. All enzyme reactions used 100 µl injections.

3.5.1.5 Conclusions on the SPR experiments

The SPR surfaces behave differently depending on the density of the immobilised glucan. When relatively dilute the surface behaves in the same way as in solution, with all the tested hydrolases able to degrade the surface. This confirms that PHS2 is synthesising large amounts of α-1,4-glucan on the surfaces. When a denser surface is used the material takes on the pattern of enzyme resistance seen in starch granules. These results indicate that the amylo-polymer created by PHS2 on the high density surface, forms a macromolecular complex, which may be crystalline, preventing access by most enzymes to the surface.

3.5.2 Nanoparticles

Whilst a quantifiable 2D system is relevant for the analysis of starch active enzymes a third spacial dimension is needed to represent a starch granule. As can be seen in Figure 3.3, enzymatic degradation of the starch granule remodels the surface in three dimensions. The physical and chemical properties and uniform shape and size[81] of gold nanoparticles make them excellent materials for the study of biological interactions.[82] Uses include the rapid detection of DNA,[83] as anticancer vaccines[84] and for the detection of carbohydrate binding proteins[85,86] based upon the interactions between surface plasmons, altering the colour, upon aggregation. Citrate stabilised gold nanoparticles (AuNP) of approximately 15 nm provide an excellent carbohydrate display system[49] and are simple to make according to the procedure of Turkevich.[81]

3.5.2.1 Glycogen based nanoparticles

Extension of glycogen particles using amylosucrase showed an increase in the particle size, when imaged using transmission electron microscopy (TEM), followed by a contraction of the material upon ageing of the material, leading to the proposal that the amylose chains are forming double helices and crystallites.[45] In this study the extension of glycogen by PHS2 gives a similar pattern of behaviour. The material stained dark blue with iodine only after extension, indicating the formation of ordered amylose double helices (Fig 3.26). Particle size had also increased, when imaged using TEM.

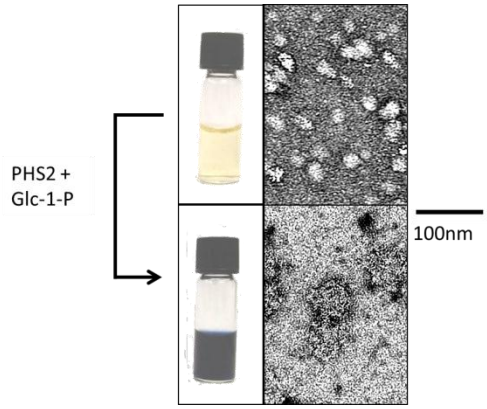


Figure 3.26: Extension of glycogen with PHS2

Glycogen (1 mg/ml) does not stain with iodine (0.5% EtOH saturated with I_2) but after extension with PHS2 (10 µg/ml, 10 mM Glc-1-P, 20 mM MES pH 6.5) the material stains very strongly, and is seen under TEM (50000x, 2% UA stained) as larger formations.

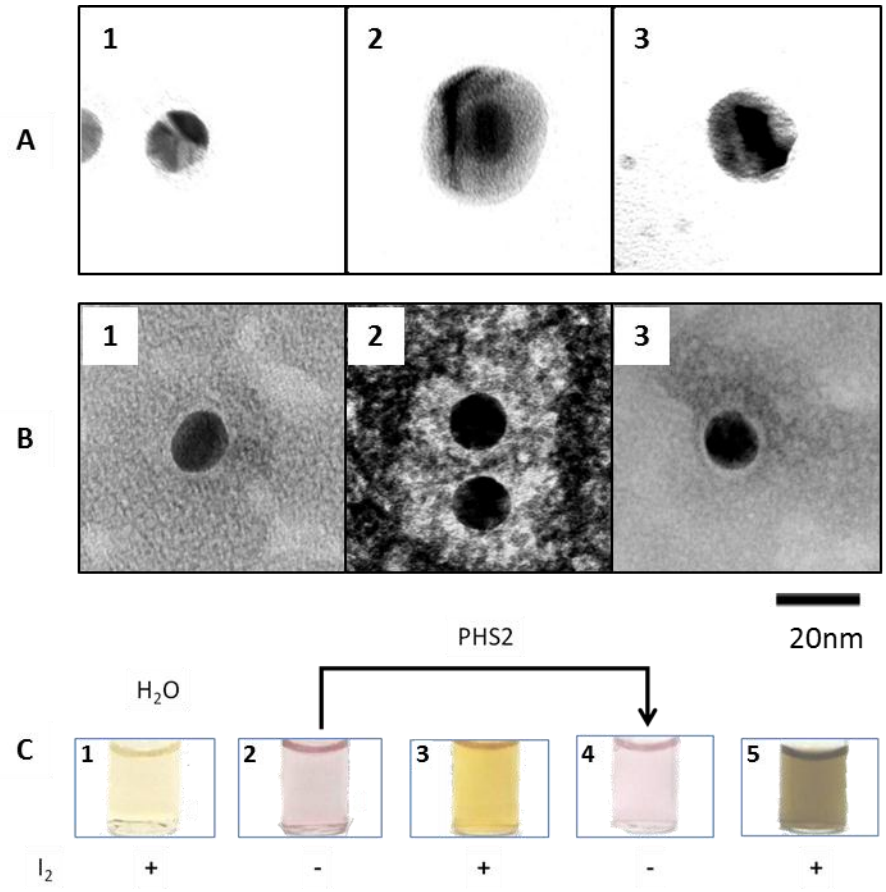


Figure 3.27: Representative images of the glucan layer on the surface of the FGNPs

Gold nanoparticles with FG immobilised on the surface were imaged with TEM, either stained with PATAg (**A**) or 2% uranyl acetate (**B**). Both staining methods show the carbohydrate on the nanoparticles increases from around 1 ± 0.09 nm (**1**) to around 4.0 ± 0.13 nm after treatment with PHS2 (10 µg/ml) and Glc-1-P (10 mM) (**2**). 24 hours after the reaction had been stopped a thinner layer of about 2.3 ± 0.09 nm can be observed (**3**). See Fig 3.29 for quantification. **C**. **1**. Five µl of ethanol saturated with iodine diluted into 1 ml water has a pale yellow colour. **2**. Gold nanoparticles with FG. **3**. Iodine only weakly stains the FG nanoparticles. **4**. The extension of the immobilised carbohydrate on the surface of the nanoparticles by PHS2 does not change the colour of the gold nanoparticles. **5**. Iodine stains nanoparticles brown after treatment with PHS2.

3.5.2.2 Synthesis of gold nanoparticles with a glycan surface

Gold nanoparticles were prepared by the citrate reduction of gold chloride[81] and had a regular round shape and an average diameter of 14.7 ± 0.09 nm, according to TEM. The high FG density surface was immobilised on these AuNPs, using the same technique as for the SPR surface, to form FGNPs, which were probed with various enzymes.

3.5.2.3 PHS2 catalysed synthesis on the FGNPs

When imaged by TEM after negative staining with the carbohydrate-specific periodic acid/thiocarbhydrazide/silver (PATAg)[87] or 2% uranyl acetate (UA), a narrow unstained zone was visible around the FG-NPs (0.9 ± 0.05 nm), indicating the thickness of the SAM and FG layer on the particles (Fig 3.27.A/B). When imaged immediately after a one hour incubation with PHS2 and Glc-1-P, a much larger unstained zone (4.0 ± 0.13 nm) was visible around the particles. After 24 hours of aging at 4 °C, this clear zone had collapsed to a much smaller size (2.3 ± 0.09 nm), suggesting that some reordering and packing of the newly synthesised glucan chains had occurred.

The PHS2-mediated extension of FGNPs had no effect on the red colour of the AuNPs themselves (Fig 3.27.C.2 to 4), indicating that the particles had not been affected by the glucan extension. However the PHS2-extended FGNPs stained strongly with iodine (Fig 3.27.C.5), in contrast to the starting FGNPs (Fig 3.27.C.3) indicating that the material synthesised on the particles had arranged to form ordered amylose helices rather than a random unstructured arrangement of chains.[88]

With high enzyme concentrations and at extended time periods, a precipitate of nanoparticles in an entangled micron scale amylose matrix was formed, without causing a change in the colour (Fig 3.28). This indicates there is no aggregation, which would cause interaction between the gold plasmons. While some way short of the 0.5-100 µm length scale of the typical starch granules,[89] these preliminary data show potential for the elaboration of micron scale carbohydrate architectures by PHS2-mediated synthesis.

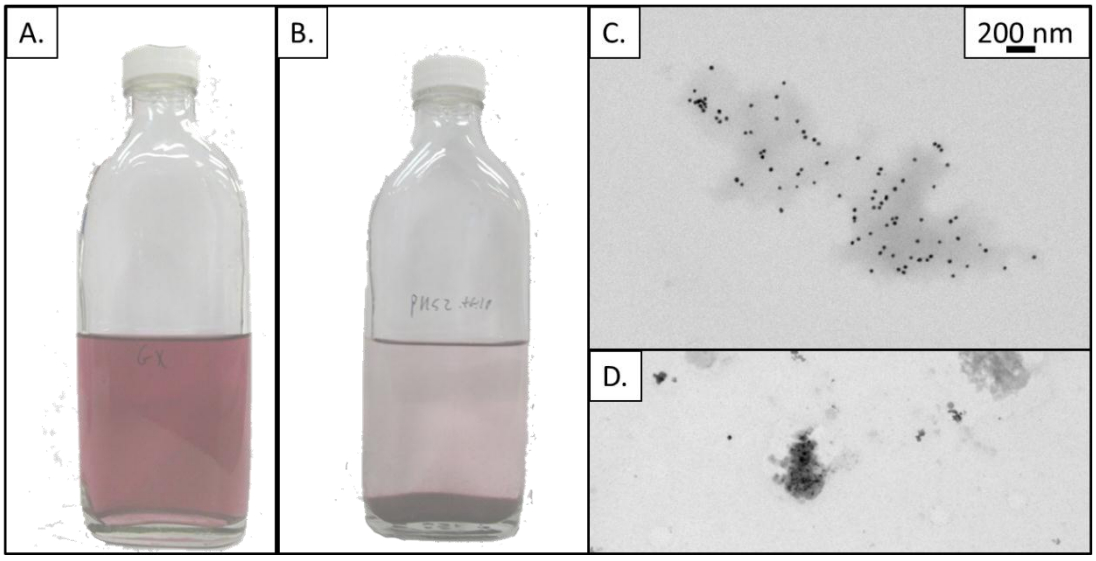


Figure 3.28: Extended reaction of PHS2 on FGNPs

A. FGNPs show characteristic red colour. **B**. PHS2 (5 µg/ml) and Glc-1-P (10 mM) were added to 75 ml of the FG-NPS for 16 hours at 21 °C. The red colour of the precipitant indicates that the nanoparticles had precipitated, but not aggregated. **C**. TEM (11500x magnification) shows extensive material has been synthesised when stained with UA. **D**. Positive staining with PATAg shows the material synthesised is carbohydrate.

3.5.2.4 Degradation of PHS2 synthesised glucan on FGNPs

FGNPs that had been extended by PHS2 and aged for 24 hours were probed with α-1,4-glucan hydrolases. Amyloglucosidase, which cannot degrade starch granules, caused no significant change in the zone around the particles which did not stain with UA. Treatment with PPA, which can degrade starch granules, caused a modest reduction (0.7 ± 0.09 nm) in the extended surface.

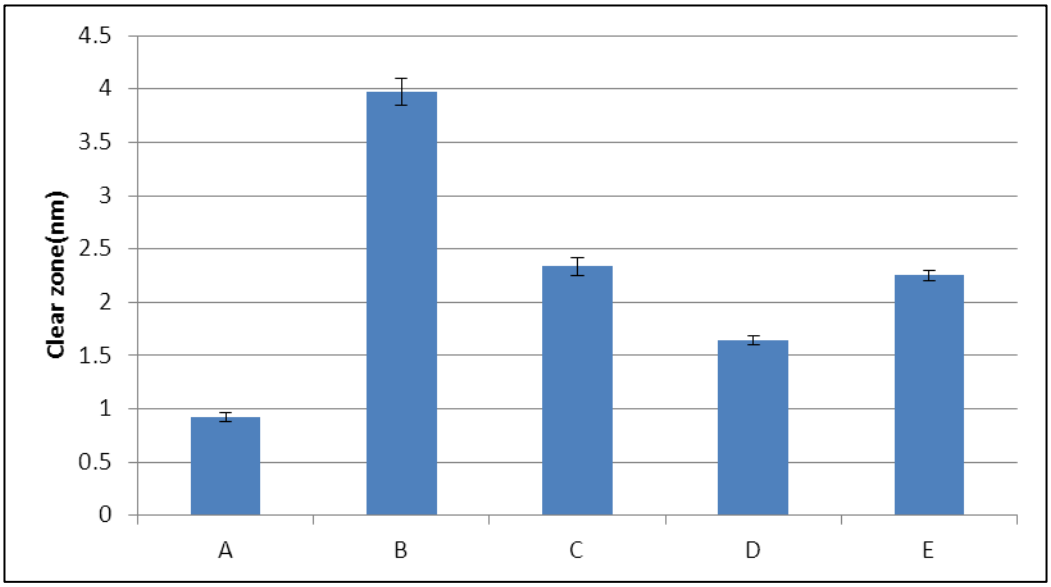


Figure 3.29: Changes in the glucan layer on FGNPs

The glucan layer on the FG-nanoparticles (**A**) increased after treatment with PHS2 (10 µg/ml, 10 mM Glc-1-P) (**B**). 24 hours after PHS2 was inactivated the thickness of the clear zone around the FG-NPs had decreased (**C**). After treatment with PAA (0.3 mg/ml) a further decrease was apparent (**D**), whilst AG (1 mg/ml) caused no decrease in the clear zone around the particles (**E**). Clear zones were measured using TEM with UA negative staining. Means and SE were determined with 31-65 particles for each treatment and all conclusions are at the 5% confidence level (t-test).

3.5.2.5 Conclusions on the use of PHS2 to extend glucan nanoparticles

PHS2 is able to synthesise polymeric α-1,4-glucans from a fractionated glucan primer immobilised on gold nanoparticles. The resultant glycopolymer stains with iodine in a manner similar to amylose and has the same resistance to amyloglucosidase, and susceptibility to α-amylase, as native starch granules. These particles can be used to analyse starch active enzymes and as models for the synthesis and modification of biocompatible starch layers on inorganic surfaces.

3.6 Conclusions on the use of PHS2 for the synthesis of amylose in solution and on surfaces

In order to study the surface enzymology of the starch granule we need new ways to investigate enzymatic reactions at solid-liquid interfaces. Due to its catalytic efficiency, we have explored the plant phosphorylase PHS2 for the rapid synthesis of linear amylose chains.

In solution assays, linear amylose-like structures of up to 100 sugar units could be synthesised. Similar syntheses could also be achieved on gold sensor chip surfaces: on its own, the phosphorylase enzyme bound tightly to the glucan-coated sensor surface, presumably using the surface binding site elucidated in these crystal structures of PHS2. When co-injected with the donor substrate PHS2 was able to synthesise glucans, which could be monitored conveniently in real time by SPR. The resulting extended surfaces were responsive to starch-active enzymes in a manner reminiscent of native starch granules, suggesting the potential for this approach to generate physiologically relevant starch-like surfaces.

PHS2-catalysed glucan polymerisation could also be monitored on gold nanoparticles using TEM. These nanoparticle-based glucan surfaces were subject to aging, with the glucan layer appearing to reorganise and assemble into a much more tightly packed structure over time. This data is interpreted as glucan chains wrapping back on themselves to produce inter- or intra-chain hydrogen bonds, as proposed for crystalline material generated by the enzymatic debranching of starch[90] or the amylosucrase-mediated extension of glycogen nanoparticles.[45]

The 2D and 3D surfaces developed in this study represent novel tools for the study of surface enzymology, which will be explored further to investigate enzymes associated with starch granule metabolism and the industrial processing of this bulk commodity.

3.7 References

1 Smith, A. M. Prospects for increasing starch and sucrose yields for bioethanol production. *Plant J.* **54**, 546-558 (2008).

2 Diouf, J. FAO's Director-General on how to feed the world in 2050. *Population and Development Review* **35**, 837-839 (2009).

3 Coultate, T. P. *Food: the chemistry of its components*. (Royal Society of Chemistry, 2002).

4 Yu, T.-S. *et al.* α-Amylase is not required for breakdown of transitory starch in *Arabidopsis* leaves. *J. Biol. Chem.* **280**, 9773-9779 (2005).

5 Santelia, D. & Zeeman, S. C. Progress in *Arabidopsis* starch research and potential biotechnological applications. *Curr. Opin. Biotechnol.* **22**, 271-280 (2011).

6 Waigh, T. A., Perry, P., Riekel, C., Gidley, M. J. & Donald, A. M. Chiral side-chain liquid-crystalline polymeric properties of starch. *Macromolecules* **31**, 7980-7984 (1998).

7 Blennow, A., Bay-Smidt, A. M., Leonhardt, P., Bandsholm, O. & Madsen, M. H. Starch paste stickiness is a relevant native starch selection criterion for wet-end paper manufacturing. *Starch - Stärke* **55**, 381-389 (2003).

8 Jobling, S. Improving starch for food and industrial applications. *Curr. Opin. Plant Biol.* **7**, 210-218 (2004).

9 Jobling, S. A., Westcott, R. J., Tayal, A., Jeffcoat, R. & Schwall, G. P. Production of a freeze-thaw-stable potato starch by antisense inhibition of three starch synthase genes. *Nat. Biotechnol.* **20**, 295-299 (2002).

10 Visser, R. G. F. *et al.* Inhibition of the expression of the gene for granule-bound starch synthase in potato by antisense constructs. *Mol. Gen. Genet.* **225**, 289-296 (1991).

11 Tetlow, I. J. Starch biosynthesis in developing seeds. *Seed Sci. Res.* **21**, 5-32 (2011).

12 Ball, S. G. & Morell, M. K. From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule. *Annu. Rev. Plant Biol.* **54**, 207-233 (2003).

13 Bustos, R. *et al.* Starch granule initiation is controlled by a heteromultimeric isoamylase in potato tubers. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2215-2220 (2004).

14 Smith, A. M., Zeeman, S. C. & Smith, S. M. Starch degradation. *Annu. Rev. Plant Biol.* **56**, 73-98 (2005).

15 Radchuk, V. V. *et al.* Spatiotemporal profiling of starch biosynthesis and degradation in the developing barley grain. *Plant Physiol.* **150**, 190-204 (2009).

16 Fettke, J. *et al.* Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. *J. Exp. Bot.* **60**, 2907-2922 (2009).

17 Edwards, A. *et al.* Specificity of starch synthase isoforms from potato. *Eur. J. Biochem.* **266**, 724-736 (1999).

18 Hejazi, M., Fettke, J., Paris, O. & Steup, M. The two plastidial starch-related dikinases sequentially phosphorylate glucosyl residues at the surface of both the A- and B-Type allomorphs of crystallized maltodextrins but the mode of action differs. *Plant Physiol.* **150**, 962-976 (2009).

19 Edner, C. *et al.* Glucan, water dikinase activity stimulates breakdown of starch granules by plastidial β-amylases. *Plant Physiol.* **145**, 17-28 (2007).

20 Sun, Z. & Henson, C. A. Degradation of native starch granules by barley α-glucosidases. *Plant Physiol.* **94**, 320-327 (1990).

21 Planchot, V., Colonna, P., Gallant, D. J. & Bouchet, B. Extensive degradation of native starch granules by alpha-amylase from *Aspergillus fumigatus*. *J. Cereal Sci.* **21**, 163-171 (1995).

22 Tawil, G. *et al. In situ* tracking of enzymatic breakdown of starch granules by synchrotron UV fluorescence microscopy. *Anal. Chem.* **83**, 989-993 (2011).

23 Kaneko, Y., Matsuda, S. & Kadokawa, J. Chemoenzymatic syntheses of amylose-grafted chitin and chitosan. *Biomacromol.* **8**, 3959-3964 (2007).

24 Loos, K. & Müller, A. H. E. New routes to the synthesis of amylose-block-polystyrene rod−coil block copolymers. *Biomacromol.* **3**, 368-373 (2002).

25 Breitinger, H. G. Synthesis of silica-bound amylose by phosphorolytic elongation of immobilised maltoheptaosyl hydrazides. *Tetrahedron Lett.* **43**, 6127-6131 (2002).

26 Klein, H. W., Palm, D. & Helmreich, E. J. M. General acid-base catalysis of α-glucan phosphorylases - stereospecific glucosyl transfer from D-glucal is a pyridoxal 5'-phosphate and ortho-phosphate (arsenate) dependent reaction. *Biochemistry* **21**, 6675-6684 (1982).

27 Percival, M. D. & Withers, S. G. Applications of enzymes in the synthesis and hydrolytic study of 2-deoxy-α-D-glucopyranosyl phosphate. *Can. J. Chem.* **66**, 1970-1972 (1988).

28 Withers, S. G. The enzymic synthesis and N.M.R. characterisation of specifically deoxygenated and fluorinated glycogens. *Carbohydr. Res.* **197**, 61-73 (1990).

29 Nawaji, M., Izawa, H., Kaneko, Y. & Kadokawa, J. I. Enzymatic synthesis of α-D-xylosylated malto-oligosaccharides by phosphorylase-catalyzed xylosylation. *J. Carbohydr. Chem.* **27**, 214-222 (2008).

30 Evers, B. & Thiem, J. Further syntheses employing phosphorylase. *Bioorg. Med. Chem.* **5**, 857-863 (1997).

31 Nawaji, M., Izawa, H., Kaneko, Y. & Kadokawa, J.-i. Enzymatic α-glucosaminylation of maltooligosaccharides catalyzed by phosphorylase. *Carbohydr. Res.* **343**, 2692-2696 (2008).

32 Kawazoe, S., Izawa, H., Nawaji, M., Kaneko, Y. & Kadokawa, J.-i. Phosphorylase-catalyzed *N*-formyl-α-glucosaminylation of maltooligosaccharides. *Carbohydr. Res.* **345**, 631-636 (2010).

33 Umegatani, Y. *et al.* Enzymatic α-glucuronylation of maltooligosaccharides using α-glucuronic acid 1-phosphate as glycosyl donor catalyzed by a thermostable phosphorylase from *Aquifex aeolicus* VF5. *Carbohydr. Res.* **350**, 81-85 (2012).

34 Treadway, J. L., Mendys, P. & Hoover, D. J. Glycogen phosphorylase inhibitors for treatment of type 2 diabetes mellitus. *Expert Opin. Investig. Drugs* **10**, 439-454 (2001).

35 Lee, W. N. P. *et al.* Metabolic sensitivity of pancreatic tumour cell apoptosis to glycogen phosphorylase inhibitor treatment. *Br. J. Cancer* **91**, 2094-2100 (2004).

36 Somsak, L. *et al.* New inhibitors of glycogen phosphorylase as potential antidiabetic agents. *Curr. Med. Chem.* **15**, 2933-2983 (2008).

37 Fukui, T., Shimomura, S. & Nakano, K. Potato and rabbit muscle phosphorylases - comparative studies on the structure, function and regulation of regulatory and non-regulatory enzymes. *Mol. Cell. Biochem.* **42**, 129-144 (1982).

38 Kadokawa, J.-i. & Kobayashi, S. Polymer synthesis by enzymatic catalysis. *Curr. Opin. Chem. Biol.* **14**, 145-153 (2010).

39 Gelders, G. G., Goesaert, H. & Delcour, J. A. Potato phosphorylase catalyzed synthesis of amylose-lipid complexes. *Biomacromol.* **6**, 2622-2629 (2005).

40 Yang, L. Q. *et al. In situ* synthesis of amylose/single-walled carbon nanotubes supramolecular assembly. *Carbohydr. Res.* **343**, 2463-2467 (2008).

41 Gray, C. J., Weissenborn, M. J., Eyers, C. E. & Flitsch, S. L. Enzymatic reactions on immobilised substrates. *Chem. Soc. Rev.* (2013).

42 Plath, C., Weimar, T., Peters, H. & Peters, T. Assaying sialyltransferase activity with surface plasmon resonance. *ChemBioChem* **7**, 1226-1230 (2006).

43 Cle, C. *et al.* Detection of transglucosidase-catalyzed polysaccharide synthesis on a surface in real time using surface plasmon resonance spectroscopy. *J. Am. Chem. Soc.* **130**, 15234-15235 (2008).

44 Cle, C., Martin, C., Field, R. A., Kuzmic, P. & Bornemann, S. Detection of enzyme-catalyzed polysaccharide synthesis on surfaces. *Biocat. Biotrans.* **28**, 64-71 (2010).

45 Buleon, A., Veronese, G. & Putaux, J. L. Self-association and crystallization of amylose. *Aust. J. Chem.* **60**, 706-718 (2007).

46 Putaux, J. L., Potocki-Veronese, G., Remaud-Simeon, M. & Buleon, A. α-D-Gucan-based dendritic nanoparticles prepared by in vitro enzymatic chain extension of glycogen. *Biomacromol.* **7**, 1720-1728 (2006).

47 Nishino, H., Murakawa, A., Mori, T. & Okahata, Y. Kinetic studies of AMP-dependent phosphorolysis of amylopectin catalyzed by phosphorylase b on a 27 MHz quartz-crystal microbalance. *J. Am. Chem. Soc.* **126**, 14752-14757 (2004).

48 Schofield, C. L., Field, R. A. & Russell, D. A. Glyconanoparticles for the colorimetric detection of cholera toxin. *Anal. Chem.* **79**, 1356-1361 (2007).

49 Schofield, C. L. *et al.* Colorimetric detection of *Ricinus communis* agglutinin 120 using optimally presented carbohydrate-stabilised gold nanoparticles. *Analyst* **133**, 626-634 (2008).

50 Watson, K. A. *et al.* Phosphorylase recognition and phosphorolysis of its oligosaccharide substrate: answers to a long outstanding question. *EMBO J.* **18**, 4619-4632 (1999).

51 Vereb, G., Fodor, A. & Bot, G. Kinetic characterization of rabbit skeletal muscle phosphorylase ab hybrid. *Biochim. et Biophys. Acta - Prot. Struc. Mol. Enzymol.* **915**, 19-27 (1987).

52 Ozbun, J. L., Hawker, J. S. & Preiss, J. Soluble adenosine diphosphate glucose-α-1,4-glucan α-4-glucosyltransferases from spinach leaves. *Biochem. J.* **126**, 953-963 (1972).

53 Pollock, C. & Preiss, J. The citrate-stimulated starch synthase of starchy maize kernels: purification and properties. *Arch. Biochem. Biophys.* **204**, 578-588 (1980).

54 Lu, Y., Steichen, J. M., Yao, J. & Sharkey, T. D. The role of cytosolic α-glucan phosphorylase in maltose metabolism and the comparison of amylomaltase in *Arabidopsis* and *Escherichia coli*. *Plant Physiol.* **142**, 878-889 (2006).

55 Suganuma, T., Kitazono, J. I., Yoshinaga, K., Fujimoto, S. & Nagahama, T. Determination of kinetic-parameters for maltotriose and higher maltooligosaccharides in the reactions catalyzed by α-D-glucan phosphorylase from potato. *Carbohydr. Res.* **217**, 213-220 (1991).

56 Goubet, F., Jackson, P., Deery, M. J. & Dupree, P. Polysaccharide analysis using carbohydrate gel electrophoresis: a method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Anal. Biochem.* **300**, 53-68 (2002).

57 Prifti, E., Goetz, S., Nepogodiev, S. A. & Field, R. A. Synthesis of fluorescently labelled rhamnosides: probes for the evaluation of rhamnogalacturonan II biosynthetic enzymes. *Carbohydr. Res.* **346**, 1617-1621 (2011).

58 Morell, M. K., Samuel, M. S. & O'Shea, M. G. Analysis of starch structure using fluorophore-assisted carbohydrate electrophoresis. *Electrophoresis* **19**, 2603-2611 (1998).

59 Chiesa, C. & Horvath, C. Capillary zone electrophoresis of malto-oligosaccharides derivatized with 8-aminonaphthalene-1,3,6-trisulfonic acid. *J. Chromatogr.* **645**, 337-352 (1993).

60 van der Vlist, J. & Loos, K. Amylose and amylopectin hybrid materials via enzymatic pathways. *Macromol. Symp.* **254**, 54-61 (2007).

61 Webb, M. R. A continuous spectrophotometric assay for inorganic phosphate and for measuring phosphate release kinetics in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 4884-4887 (1992).

62 Lux, F. Properties of electronically conductive polyaniline: a comparison between well-known literature data and some recent experimental findings. *Polymer* **35**, 2915-2936 (1994).

63 Barford, D., Hu, S. H. & Johnson, L. N. Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. *J. Mol. Biol.* **218**, 233-260 (1991).

64 Buchbinder, J. L., Rath, V. L. & Fletterick, R. J. Structural relationships among regulated and unregulated phosphorylases. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 191-209 (2001).

65 Stein, N. CHAINSAW: a program for mutating pdb files used as templates in molecular replacement. *J. Appl. Crystallogr.* **41**, 641-643 (2008).

66 McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674 (2007).

67 Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375-383 (2007).

68 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774-797 (2007).

69 Palm, D., Klein, H. W., Schinzel, R., Buehner, M. & Helmreich, E. J. M. The role of pyridoxal 5'-phosphate in glycogen-phosphorylase catalysis. *Biochemistry* **29**, 1099-1107 (1990).

70 Holm, L. & Sander, C. DALI: a network tool for protein structure comparison. *Trends Biochem. Sci* **20**, 478-480 (1995).

71 Lin, K., Rath, V. L., Dai, S. C., Fletterick, R. J. & Hwang, P. K. A protein phosphorylation switch at the conserved allosteric site in GP. *Science* **273**, 1539-1541 (1996).

72 O'Reilly, M., Watson, K. A., Schinzel, R., Palm, D. & Johnson, L. N. Oligosaccharide substrate binding in *Escherichia coli* maltodextrin phosphorylase. *Nat. Struct. Biol.* **4**, 405-412 (1997).

73 Pinotsis, N., Leonidas, D. D., Chrysina, E. D., Oikonomakos, N. G. & Mavridis, I. M. The binding of beta- and gamma-cyclodextrins to glycogen phosphorylase b: kinetic and crystallographic studies. *Protein Sci.* **12**, 1914-1924 (2003).

74 O'Reilly, M., Watson, K. A. & Johnson, L. N. The crystal structure of the *Escherichia coli* maltodextrin phosphorylase−acarbose complex. *Biochemistry* **38**, 5337-5345 (1999).

75 Qian, M. X., Haser, R., Buisson, G., Duee, E. & Payan, F. The active center of a mammalian α-amylase. Structure of the complex of a pancreatic α-amylase with a carbohydrate inhibitor refined to 2.2-A resolution. *Biochemistry* **33**, 6284-6294 (1994).

76 Geremia, S., Campagnolo, M., Schinzel, R. & Johnson, L. N. Enzymatic catalysis in crystals of *Escherichia coli* maltodextrin phosphorylase. *J. Mol. Biol.* **322**, 413-423 (2002).

77 Flinn, N. S., Quibell, M., Monk, T. P., Ramjee, M. K. & Urch, C. J. A single-step method for the production of sugar hydrazides: intermediates for the chemoselective preparation of glycoconjugates. *Bioconjugate Chem.* **16**, 722-728 (2005).

78 Lahiri, J., Isaacs, L., Tien, J. & Whitesides, G. M. A strategy for the generation of surfaces presenting ligands for studies of binding based on an active ester as a common reactive intermediate: a surface plasmon resonance study. *Anal. Chem.* **71**, 777-790 (1999).

79 Zhi, Z. L. *et al.* A versatile gold surface approach for fabrication and interrogation of glycoarrays. *ChemBioChem* **9**, 1568-1575 (2008).

80 Kilburn, D., Claude, J., Schweizer, T., Alam, A. & Ubbink, J. Carbohydrate polymers in amorphous states: an integrated thermodynamic and nanostructural investigation. *Biomacromol.* **6**, 864-879 (2005).

81 Enustun, B. V. & Turkevich, J. Coagulation of colloidal gold. *J. Am. Chem. Soc.* **85**, 3317-3328 (1963).

82 Saha, K., Agasti, S. S., Kim, C., Li, X. & Rotello, V. M. Gold nanoparticles in chemical and biological sensing. *Chem. Rev.* **112**, 2739-2779 (2012).

83 Nam, J. M., Stoeva, S. I. & Mirkin, C. A. Bio-bar-code-based DNA detection with PCR-like sensitivity. *J. Am. Chem. Soc.* **126**, 5932-5933 (2004).

84 Ojeda, R., de Paz, J. L., Barrientos, A. G., Martín-Lomas, M. & Penadés, S. Preparation of multifunctional glyconanoparticles as a platform for potential carbohydrate-based anticancer vaccines. *Carbohydr. Res.* **342**, 448-459 (2007).

85 Otsuka, H., Akiyama, Y., Nagasaki, Y. & Kataoka, K. Quantitative and reversible lectin-induced association of gold nanoparticles modified with α-Lactosyl-ω-mercapto-poly(ethylene glycol). *J. Am. Chem. Soc.* **123**, 8226-8230 (2001).

86 Hone, D. C., Haines, A. H. & Russell, D. A. Rapid, quantitative colorimetric detection of a lectin using mannose-stabilized gold nanoparticles. *Langmuir* **19**, 7141-7144 (2003).

87 Thiery, J. P. Demonstration of polysaccharides in thin sections by electron microscopy. *J. Microsc.-Oxf.* **6**, 987-1018 (1967).

88 Bates, F. L., French, D. & Rundle, R. E. Amylose and amylopectin content of starches determined by their iodine complex formation. *J. Am. Chem. Soc.* **65**, 142-148 (1943).

89 Jane, J.-L., Kasemsuwan, T., Leas, S., Zobel, H. & Robyt, J. F. Anthology of starch granule morphology by scanning electron microscopy. *Starch - Stärke* **46**, 121-129 (1994).

90 Pohu, A., Planchot, V., Putaux, J. L., Colonna, P. & Buléon, A. Split crystallization during debranching of maltodextrins at high concentration by isoamylase. *Biomacromol.* **5**, 1792-1798 (2004).

Chapter 4 – Cellodextrin phosphorylase – a β-1,4-glucan phosphorylase

4.1 Introduction

4.1.1 Cellulose

Cellulose is the most abundant organic compound on Earth[1] and is extremely strong and resistant to degradation. Composed of β-1,4-linked glucose, the residues alternate orientation and the structure should possibly be considered, at least in terms of molecular shape, as a polymer of cellobiose (Fig 4.1). Together with its structural homologue, chitin, this isoform is the main structural support in plants, insects, fungi and to a certain extent in bacteria[2] and oomycetes.[3] Cellulose and chitin are synthesised as parallel chains which form crystalline fibres.[4-6]

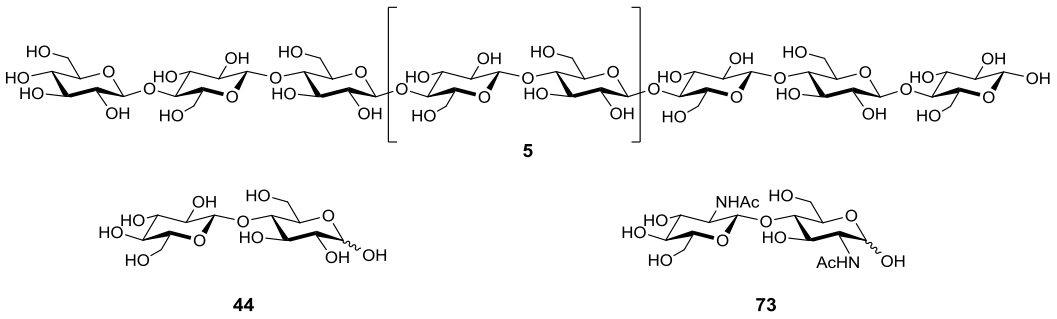


Figure 4.1: Structure of β-1,4-glucans

β-1,4-Glucans, such as cellulose (**5**), are composed of β-1,4-linked glucose residues, alternating in orientation, which gives a flat ribbon shape. The repeating motif in is the disaccharide cellobiose (**44**), or in the case of chitin is chitobiose (**73**).

The global biofuel industry is beginning to utilise the structural sugars in plant biomass. Commercial usage of this lignocellulosic biomass is hampered by the relatively slow rate of sugar release from plant material,[7] due to the fact that the cellulose is rarely found in a pure form, but is instead combined with hemicellulose and lignin. The future direction of work in this field is likely to involve modifying the cellulose chains, such that they are more easily attacked, and altering the hemicellulose and lignin content *in planta*, thus making the release of sugars more efficient.[7] Bacterial cellulose is synthesised in a much purer form, requiring less processing before being incorporated into a range of products as diverse as wound dressings,[8] food thickeners[9] and composite materials.[10]

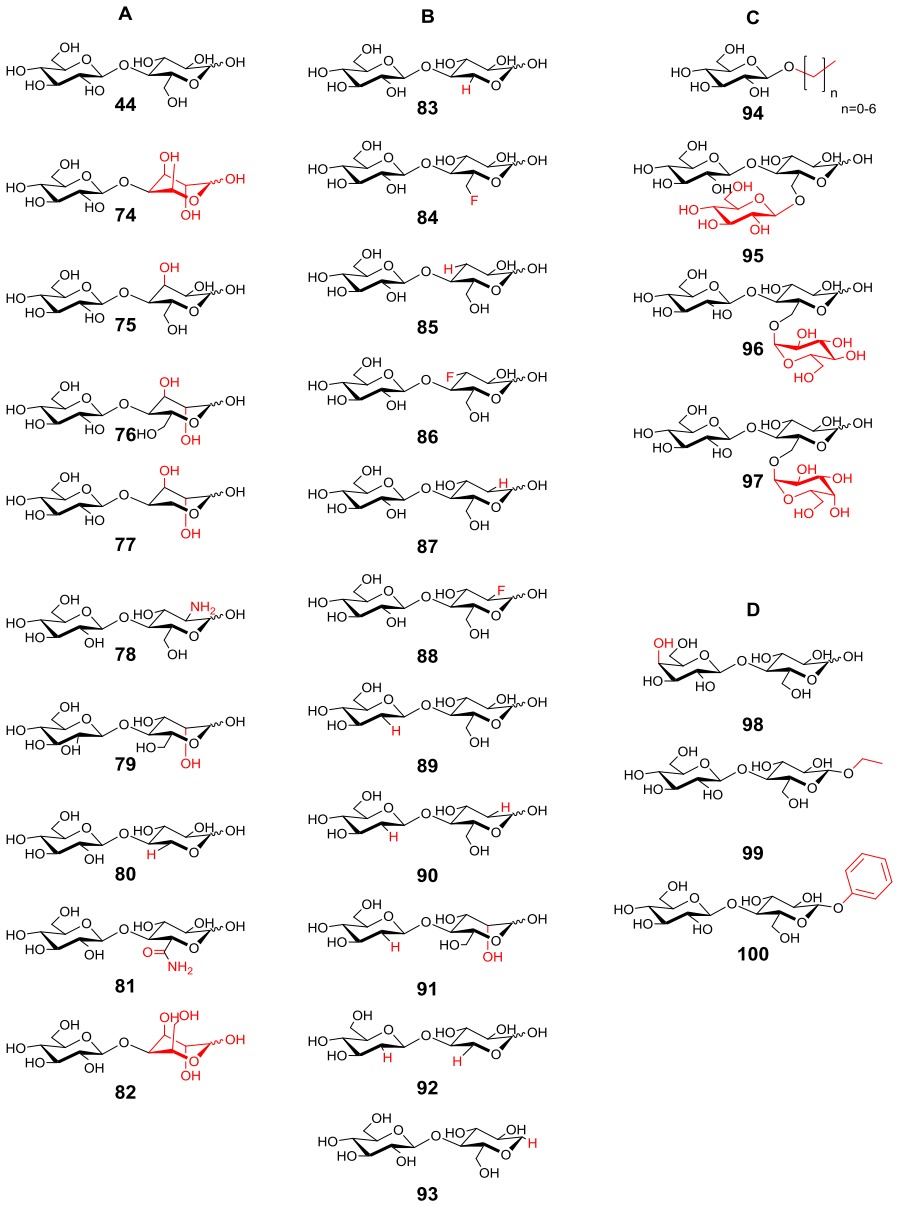


Figure 4.2: Proposed products formed using cellobiose phosphorylases

A. Glc can be transferred from Glc-1-P on to various monosaccharides including glucose (**44**), L-fucose (**74**), D-allose (**75**), D-altrose (**76**),[11] D-arabinose (**77**),[12] D-glucosamine (**78**), D-mannose (**79**), D-xylose (**80**),[13] D-glucuronamide (**81**)[14] and L-galactose (**82**).[15] **B**. Deoxy and fluoro-cellobioses could also be made using 6-deoxy- (**83**), 6-fluoro- (**84**), 3-deoxy- (**85**), 3-fluoro- (**86**), 2-deoxy- (**87**) and 2-fluoro- (**88**) D-glucose as acceptors.[16] When glucal was used as the donor 2'-deoxy-cellobiose (**89**), 2,2'-dideoxy-cellobiose (**90**), and 2'-deoxy-glucosyl-xylose (**91**) and mannose (**92**) were formed using appropriate acceptors.[17] 1-Deoxy-D-glucose, known as 1,5-anhydroglucitol, also accepted glucose to form 1,5-anhydrocellobiitol (**93**).[18] **C**. Alkyl chains (**94**) were acceptors as were the 6-glycsoylated glucoses gentiobiose (**95**), isomaltose (**96**) and melibiose (**97**).[12] **D**. Mutagenesis allowed formation of lactose phosphorylase which tolerated galactose at the non-reducing position (**98**)[19] and allow addition of glucose to the reducing end of ethyl (**99**) or phenyl (**100**) glucosides.[20]

β-1,4-Linked polysaccharides, other than cellulose, are also used in many structural polymers. For example, β-1,4-linked mannan and xylan are used in plant secondary cell walls, while chitin (**73**) is used in the exoskeleton of insects and crustaceans. Research into the degradation of these materials, for fine chemical production[21] and to enhance release of sugars from biomass,[22] is on-going.

A substantial amount of research effort has been spent studying the microbial degradation of cellulose for the identification of enzymes useful in the biotechnological conversion of biomass.[23] For studying cellulases various substrates are used: delignified wood which contains some hemicellulose; microcrystalline cellulose, from which hemicellulose and amorphous cellulose have been removed by acid treatment; bacterial cellulose, which does not have the same crystalline form;[24] carboxymethyl cellulose, which is soluble and can be degraded by non-cellulolytic enzymes. Thus a system allowing direct measurement of pure cellulose degradation is highly desirable.

4.1.2 Cellobiose phosphorylases (CBP)

Disaccharide β-1,4-glucan phosphorylases, which have been more thoroughly studied than oligosaccharide phosphorylases, include galactosyl-rhamnose,[25] mannosyl-glucose,[26] mannobiose,[27] chitobiose[28] and cellobiose[29] phosphorylases. The last of these has been particularly heavily studied with respect to substrate specificity (Fig 4.2). Whilst many assays utilise only phosphate release to indicate turnover,[15] purification of the products is required to confirm product formation rather than hydrolysis of the donor. 3-Deoxy-, 3-deoxy-3-fluoro- and 3-epimers of glucose are acceptors for CBP, but 3-*O*-methyl glucose was not, suggesting steric hindrance at the 3OH.[11] Detailed substrate interactions were studied with deoxy- and fluoro-sugars, all showing reasonable kinetics except the 3-deoxy-3-fluoro-glucose.[16] Glucosamine, mannose and xylose were also shown to be acceptors[13] and glucosyl-xylose (**80**) was produced, with a 60% yield, using the cell extract of *Cellvibrio gilvus*.[30] A 28% yield of disaccharide (**77**) was achieved from the transfer from Glc-1-P on to arabinose.[12] Glucal was also a substrate, in the presence of Pi, allowing formation of 2'-deoxyglucosyl disaccharides.[17] In addition, trisaccharides could be synthesised using the 1,6-disaccharides gentiobiose, melibiose and isomaltose as acceptors for glucose[12] and, whilst glucuronic acid was not an acceptor, glucuronamide was.[14] Linear alcohols as small as methanol were utilised as acceptors,[31] as was 1,5-anhydro-glucitol.[18]

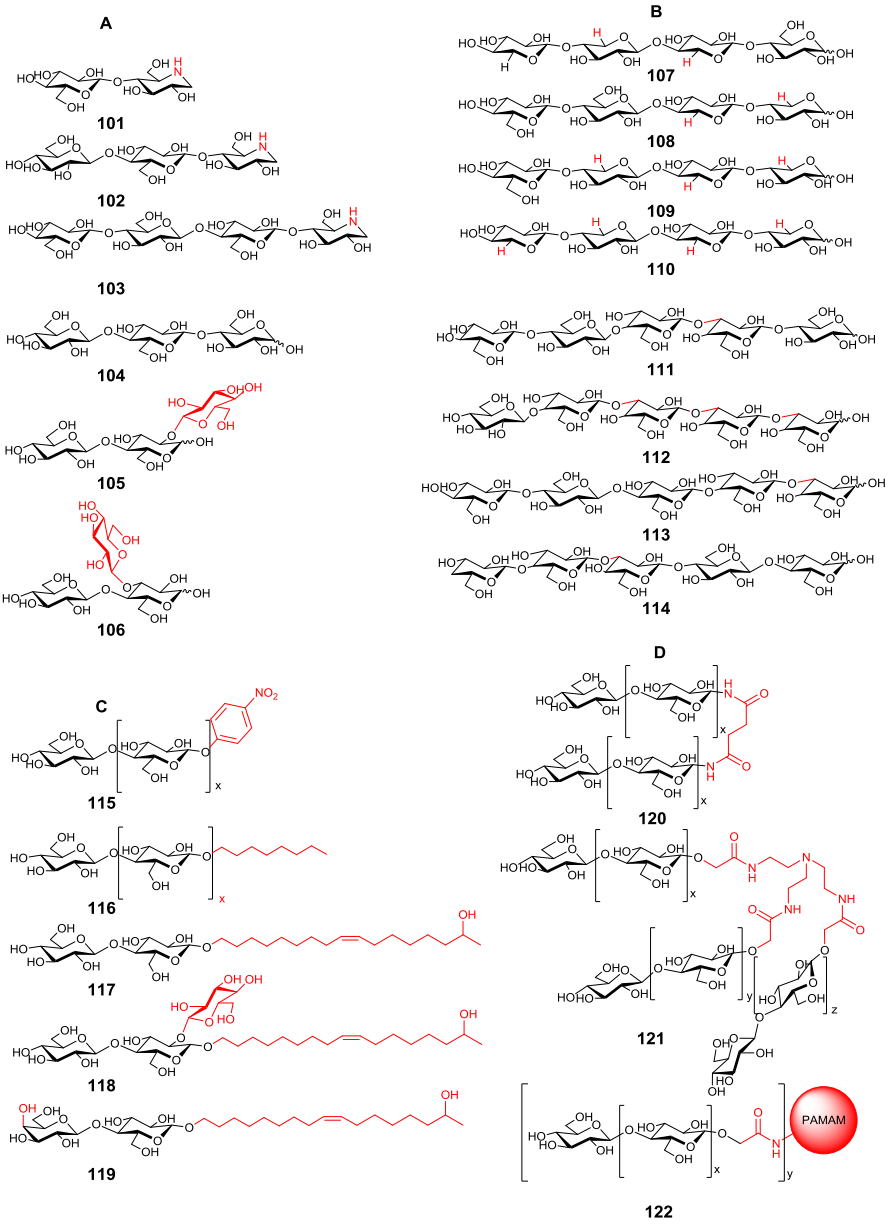


Figure 4.3: Proposed products formed using cellodextrin phosphorylase

A. Cellodextrin phosphorylase can be used to extend deoxynojirimycin (DNJ) to form pseudo cellooligosaccharides for use as cellulase inhibitors (**101-103**).[32] Trisaccharides could be purified from the transfer of glucose on to the 4-hydroxyl of cellobiose (**104**), sophorose (**105**) and laminaribiose (**106**). **B**. Specific xyloglucan oligosaccharides can be made by using either Xyl-1-P or Glc-1-P as the sugar donors and the synthesised xylo- and cellooligosaccharides as acceptors. (**107-110**)[33] and β-1,3/1,4 glucans can be made using laminarin phosphorylase and CDP (**111-114**).[34] **C**. The enzymes were also capable of extending glucosides including *p*-nitophenol glucoside (**115**) and β-octyl glucoside (**116**).[35] Novel glycolipids could be made by adding glucose to glucosyl (**117**) and sophorsyl (**118**) lipids and a single transfer from galactose-1-P on to the glucosyl lipid was achieved, yielding lactosyl lipid (**119**).[36] **D**. Dendrimers were also made by extending cellobiose derivatised succinamide (**120**), tris(2-aminoethyl)amine (**121**) and polyamidoamine (**122**) with CDP.[37]

4.1.3 Cellodextrin phosphorylases (CDP)

The most studied polymerising β-1,4-glucan phosphorylase is cellodextrin phosphorylase,[38] closely related to the cellobiose phosphorylases. This enzyme has been used to synthesise cellooligosaccharide-DNJ (Fig 4.3.A),[32] as cellulase inhibitors. β-Linked disaccharides acted as acceptors,[12] showing that the enzyme is permissive for anomeric substitution of β-glucosyl acceptors. CDP was capable of transferring xylose from Xyl-1-P, and could also transfer Glc and Xyl on to β-linked xylose, facilitating the synthesis of a series of xylans related to hemicellulose (Fig 4.3.B).[33] By alternating between CDP and laminarin phosphorylase, both of which will transfer on to any β-glucan, specific β-1,3/1,4-glucans related to plant cell walls were synthesised.[34] Alkyl and PNP glycosides were also assessed as acceptors for CDP[35] and novel glycolipids were produced (Fig 4.3.C).[36] Dendrimers based on cellulose could readily be synthesised by the extension of cellobiose dendrimers (Fig 4.3.D).[37]

4.1.4 Rationale

β-1,4-Linked oligosaccharides have interesting properties, including structural strength and resistance to degradation. In order to better utilise these carbohydrates in new materials and in biomedicine, pure, well-defined material is required. In addition, well-characterised standards are needed for the study of the natural role of these compounds in Nature.

In this study, cellodextrin phosphorylase was investigated for the production of cellulose oligomers and the possibility of synthesising β-1,4-oligosacchaides based on alternative sugars explored. This enzyme is flexible towards certain acceptors, but more inflexible with regard to the donor. Inhibition experiments were used to probe the interaction between the substrates and the enzyme. In order to study this further X-ray crystallography was used to characterise the structure of cellodextrin phosphorylase. The comparison between cellobiose phosphorylase and cellodextrin phosphorylase from the same organism shows the structural changes necessary for accommodation of longer polymers in the phosphorylase active site and rationalise the engineering of polymerising phosphorylases.

4.2 Expression and purification of CDP

The *E. coli* codon optimised CDP gene from *Clostridium thermocellum* was synthesised and subcloned into the BamHI site of pET15b, which inserts a hexahistidine tag behind a thrombin cleavage site at the N-terminus of the protein. After optimisation the protein was successfully expressed in *E.coli* BL21 and purified using Ni-IMAC and GF to obtain >98% pure protein, as judged by SDS-PAGE (Fig 4.4). The elution from the GF column at 120 ml indicates the protein is approximately 111 kDa, comparable to the calculated M_W of 114.3 for the monomer. Dynamic light scattering shows the protein is present as a monodisperse species with a molecular weight of 277 kDa, which is comparable to the M_W of the dimer. A total yield of 10 mg of purified CDP per litre culture was obtained and concentrated to 40 mg/ml.

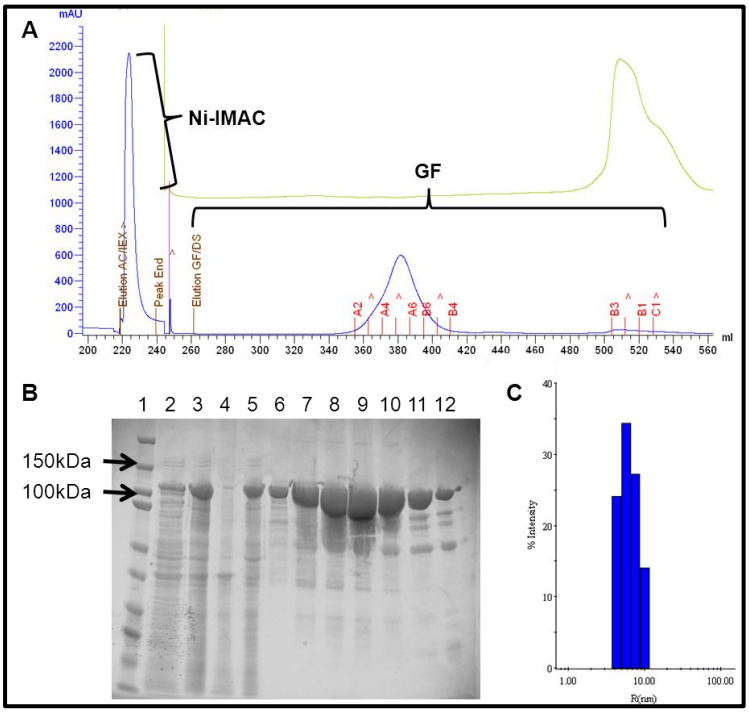


Figure 4.4: Purification of CDP from *E. coli*

A. Purification of CDP was by the standard protein purification protocol (Section 8.1.2). CDP was eluted from the Ni-IMAC by single step increase of imidazole at 220 ml. The whole peak was injected on the gel filtration column at 260ml to give a single species (A2-B4) at 120 ml elution volume, as judged by A_{280} (blue), with some low molecular weight contaminants eluting in the column volume, as indicated by the conductivity trace (green). **B**. SDS-PAGE of the fractions from GF show a major band, consistent with the calculated MW of 114 kDa. 1: Kaleidoscope protein standards (Bio-Rad). 2: total cell extract of uninduced cells. 3: total cell extract of induced cells. 4: insoluble induced cell material. 5: soluble induced cell lysate. 6-12: fractions A2-B5 from GF. **C**. DyLS of CDP (0.1 mg/ml) shows the protein is present as a monodisperse species with a radius of 6.6 nm.

4.3 Activity of CDP

4.3.1 Initial assays

The activity of CDP was confirmed by synthesis of cellooligosaccharides, monitored using carbohydrate electrophoresis (Fig 4.5). CDP was able to extend APTS-labelled cellohexaose (APTS-(β-1,4-Glc)$_6$) to long polymers by transferring glucose from Glc-1-P on to the acceptor, although most of the synthesised material was insoluble and removed during the sample preparation (Fig 8.14). CDP can also clearly phosphorolyse both the APTS-(β-1,4-Glc)$_6$ and the CDP-synthesised polymers to cellobiose and cellotriose. Cellulase can degrade the CDP-synthesised polymers, confirming the material is indeed cellulose-based.

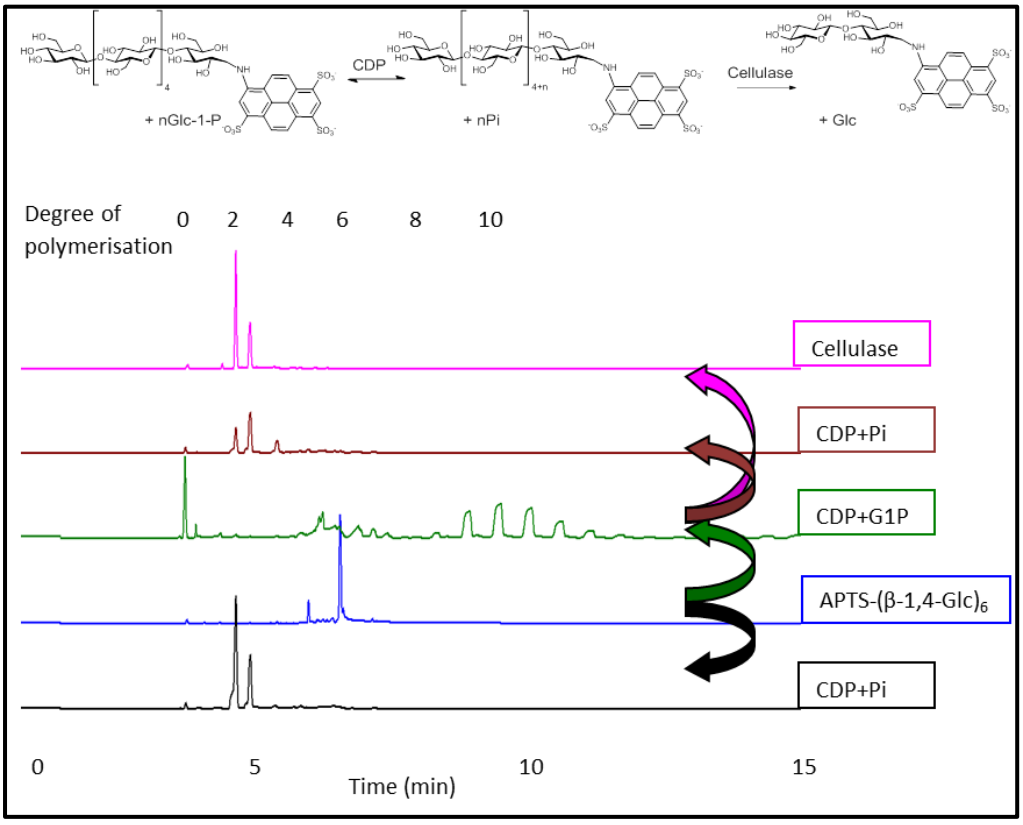


Figure 4.5: Carbohydrate electrophoresis of CDP-synthesised polymers

The activity of the phosphorylase was confirmed by assaying the ability of CDP (0.5 µg) to transfer Glc from Glc-1-P (10 mM) on to APTS-(β-1,4-Glc)$_6$ (0.72 µM) in 100 µl MES (20 mM, pH 7.5). After 1 hour at 37 °C the reactions were terminated by heating to 95 °C in a boiling water bath for 15 min and centrifuging at 16,000 g for 5 min (CDP+G1P). 30 µl of this synthetic reaction was further probed by degradation with either cellulase (0.94 U) from *T. longibrachum* (Megazyme) in NaAc (100 mM, pH 4.5), or CDP (0.5 µg) in Pi (100 mM, pH 7.0), for 1 hour at 37 °C (CDP+Pi), before again heating to 95 °C and centrifuging as before.

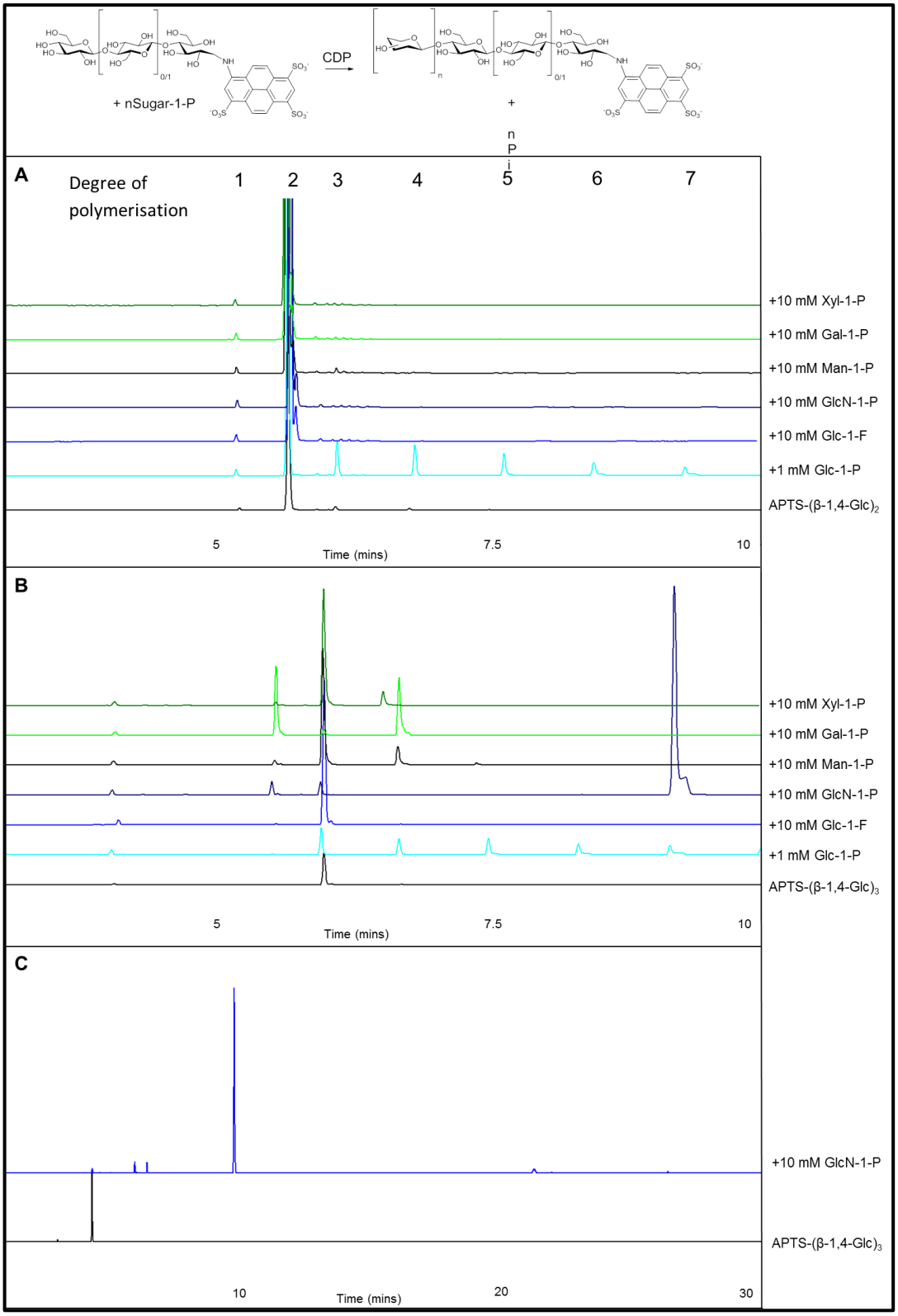


Figure 4.6: Donor specificity of CDP, analysed using carbohydrate electrophoresis

A. Extension of APTS-cellobiose with various sugar-1-P donors. **B**. Extension of APTS-cellotriose with various sugar-1-P donors. **C**. Extension of APTS-cellotriose with GlcN-1-P. Assays were carried out using CDP (5 µg/ml) at 40 °C with APTS-labelled acceptor (2 µM) in HEPES (50 mM, pH 7.5) followed by heating to 95 °C and centrifuging at 16,000 g.

4.3.2 Donor specificity of CDP

4.3.2.1 Glucan extension

In order to determine the capability of CDP for the synthesis of other β-1,4-glycan polymers the transfer of other sugars on to cellooligosaccharides was monitored using CE. APTS-(β-1,4-Glc)$_2$ was initially used, even though this is a poor acceptor (Section 4.3.3), as this would reduce the chance of phosphorolysis of the acceptor, releasing the natural donor Glc-1-P (Fig 4.6). Glc-1-P is an excellent donor, rapidly extending the polymer. The other sugar-1-Ps showed no significant new peaks, which does not agree with the literature.[33]

APTS-(β-1,4-Glc)$_3$ was then used as a more efficent acceptor, which was confirmed by the faster extension when using Glc-1-P. There is very little material visible in the Glc-1-P reactions, as the long polymers formed are insoluble and so removed during sample preparation. In the presence of Gal-1-P disproportionation is the only reaction occurring, caused by hydrolytically released Pi, allowing phosphorolysis of one cellooligosaccharide and transfer of Glc on to another acceptor molecule. This is apparent from the equal amount of APTS-(β-1,4-Glc)$_2$ and APTS-(β-1,4-Glc)$_4$ formed. In the reactions which has contained Man-1-P or Xyl-1-P there is a small additional peak which may correspond to the addition of one of these sugars to the acceptor.

Most strikingly the reaction with GlcN-1-P shows a large new peak appearing, greatly retarded on the electrophoretogram. This is proposed to be caused by a single turnover adding one GlcN residue which, at the pH of the running buffer (pH 4.75), is protonated. This alters the charge on the molecule causing it to elute much later. With a much longer electrophoresis another peak is visible suggesting another addition, either of a further GlcN or of a Glc.

We have been unable to confirm the identity of these or any other APTS-labelled products by mass-spectrometry. This has been a consistent drawback of this technique across numerous projects.

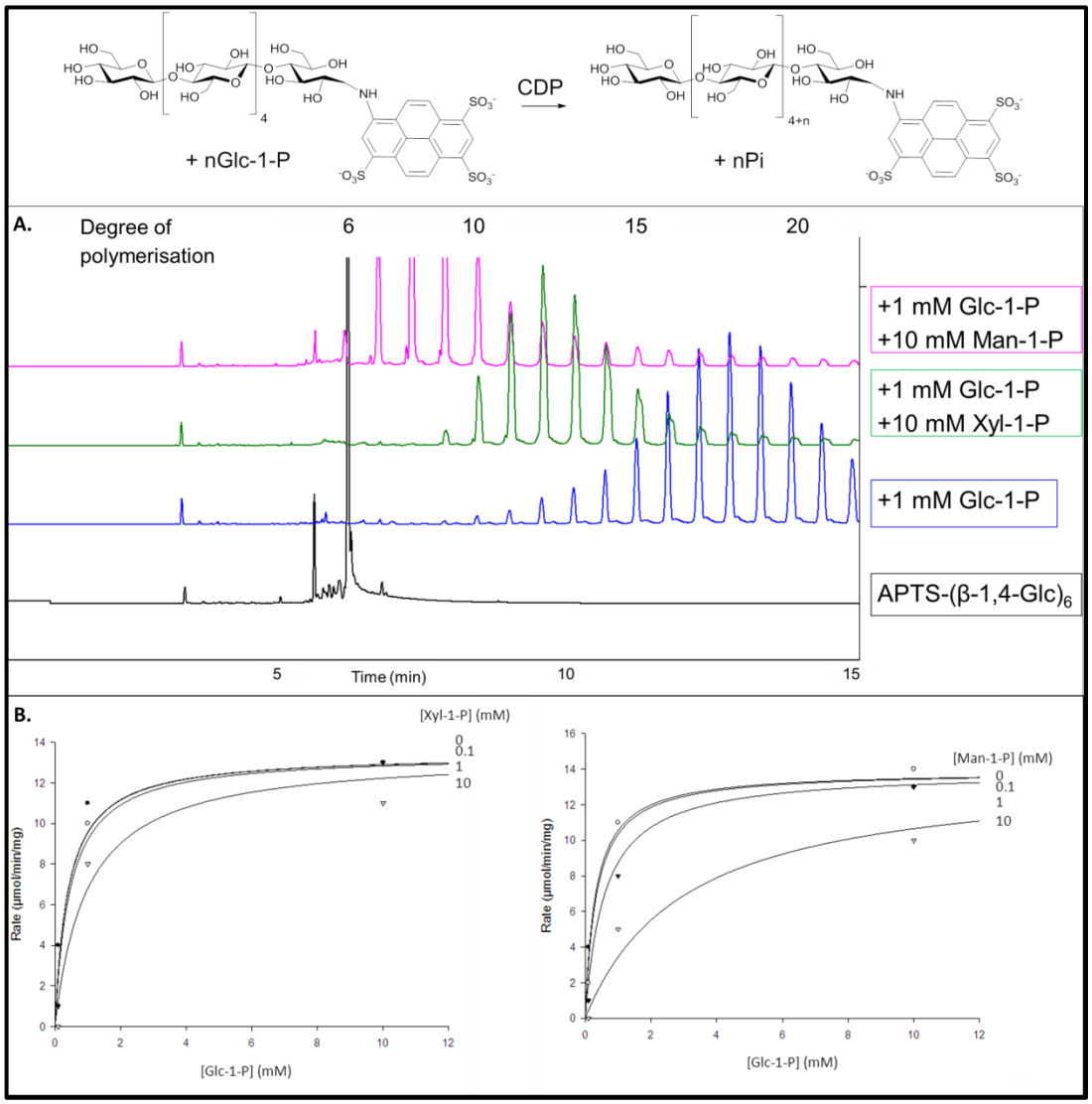


Figure 4.7: Sugar-1-P binding inhibits CDP activity

A. CE was performed with standard conditions varying Glc-1-P concentration with a range of Xyl-1-P or Man-1-P concentrations. **B**. The DP of the peak with the largest area was calculated and thus the average turnover of the enzyme, allowing a calculation of the approximate inhibition constants. Assays were carried out using CDP (5 µg/ml) at 40 °C with APTS-(β-1,4-Glc)$_6$ (2 µM) in HEPES (50 mM, pH 7.5) followed by heating to 95 °C and centrifuging at 16 000 g.

4.3.2.3 Competitive inhibition by sugar-1-Ps

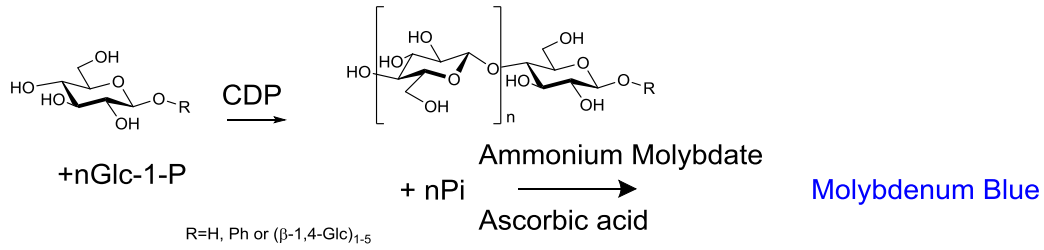
Competition assays were performed in order to understand whether the lack of efficient activity is caused by failure to bind the substrates (K_M) or inefficient catalysis (k_{cat}) (Fig 4.7). Reaction mixtures, containing different concentrations of Glc-1-P, had competing substrates added, to assess the impact on polymerisation.

Xyl-1-P decreased the average polymer length indicating binding to the active site, consistent with the ability for Xyl-1-P to act as a donor for CDP.[33] Man-1-P causes a much greater decrease in polymer length, indicating tighter binding, but is a much poorer donor than Glc-1-P. This suggests that Man-1-P has a lower turnover, possibly because of the deactivation of the phosphate ester by the axial hydroxyl on C2. Repeating these experiments over a range of substrate and inhibitor concentrations it was possible to calculate the K_i of Xyl-1-P (6.2 ± 2.6 mM) and Man-1-P (1.2 ± 0.6 mM). As these are substrates this reduces the reliability of the values. These calculations are based on competitive inhibition as the V_{max} is unchanged (13.4 ± 0.5 μmol/min/mg) and the K_M^{app} changes with varying inhibitor concentration, to 0.4 mM ± 0.06 mM with no inhibitor. However the acceptor concentration (2 μM) is significantly below K_M so these results do not match literature values (K_M^{app} = 4.7 mM,[39] V_{max} = 64.2.[40]).

4.3.3 Acceptor specificity of CDP

4.3.3.1 β-1,4-Glucans

The length specificity of CDP for unlabelled acceptors was determined by measurement of the phosphate released from 10 mM Glc-1-P over a range of acceptor concentrations. Using this assay the capability of CDP to use cellooligosaccharides as acceptors was assayed (Table 4.1). Glucose was found to be a poor acceptor, whilst the monosaccharide analogue phenyl β-D-glucopyranoside (BPG) was a better acceptor, probably because of its fixed anomeric configuration. The cellooligosaccharides had a decreasing K_M up to DP5, indicating there are stabilising interactions with the sugars further from the catalytic site. Cellohexaose has a higher K_M, which may be an effect of its limited solubility. The k_{cat} also follows a similar pattern, with turnover decreasing as acceptor length increases, which may be an effect of the longer oligomers binding more tightly the active site.



Acceptor	K_M^{app} (mM)	k_{cat} (1/s)	k_{cat}/K_M^{app} (1/mM/s)
Glucose – (Glc)	NA	NA	NA
BPG	(24 ± 13)	(15 ± 6.3)	0.63
Cellobiose – (β-1,4-Glc)$_2$	2.6 ± 0.18	17 ± 0.50	6.5
Cellotriose – (β-1,4-Glc)$_3$	0.68 ± 0.076	9.5 ± 0.35	14
Cellotetraose – (β-1,4-Glc)$_4$	0.54 ± 0.13	5.0 ± 0.25	9.3
Cellopentaose – (β-1,4-Glc)$_5$	0.36 ± 0.076	4.3 ± 0.47	12
Cellohexaose – (β-1,4-Glc)$_6$	1.9 ± 0.76	7.6 ± 1.8	4.0

Table 4.1: Acceptor length specificity for CDP.

Glucose is an acceptor, but the activity is too low for kinetic analysis. BPG is a poor acceptor and the values calculated are outside the range of the assay.

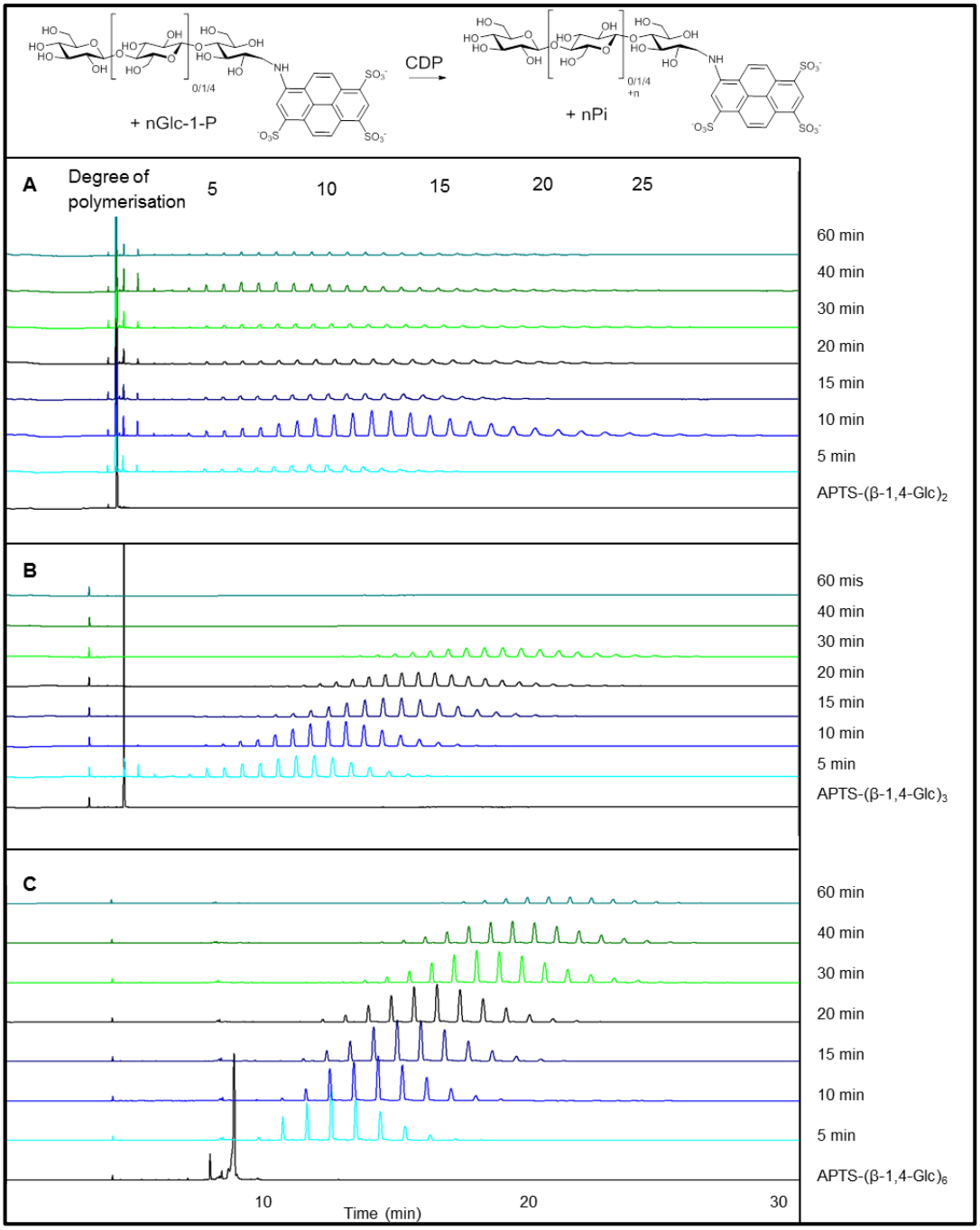


Figure 4.8: CDP catalysed extension of APTS-labelled cellooligosaccharides

A. Extension of APTS-cellobiose. **B**. Extension of APTS-cellotriose. **C**. Extension of APTS-cellohexaose. Assays were carried out using CDP (5 µg/ml) at 40 °C with Glc-1-P (10 mM) and APTS-labelled acceptor (2 µM) in HEPES (50 mM, pH 7.5) followed by heating to 95 °C and centrifuging at 16 000 g.

The ability of CDP to use APTS labelled (β-1,4-Glc)$_2$, (β-1,4-Glc)$_3$ and (β-1,4-Glc)$_6$ was monitored over time using CE (Fig 4.8).

APTS-(β-1,4-Glc)$_2$ is a poor acceptor and when a single glucose residue is added the formed APTS-(β-1,4-Glc)$_3$ is a much better acceptor and is rapidly polymerised to insoluble cellooligosaccharides. BPG is a poor acceptor in solution and the reductively-aminated cellobiose will have a similar Glc configuration. It is unclear whether the APTS itself prevents access of the acceptor to the active site, or whether this low activity is purely caused by the lack of a second pyranose ring.

APTS-(β-1,4-Glc)$_3$ is a much better acceptor, being completely consumed within 10 min. Interestingly, after 5 min there is still some DP 3, 4 and a very little amount of 5 material, though there is apparently no DP 6 polymer. This suggests either, that DP 6 and 7 are much better acceptors than DP 3, 4 and 5, or that for short oligomers the reaction is processive and for longer oligomers is non-processive.

APTS-(β-1,4-Glc)$_6$ is a good acceptor for CDP forming long polymers rapidly. This can be seen in the CE as a non-processive reaction, with all oligomers extending continuously until they become insoluble at DP>25.

4.3.3.2 Other β-1,4-glycans

Xylotriose ((β-1,4-Xyl)$_3$) and mannotriose ((β-1,4-Man)$_3$) were labelled with APTS and assessed as acceptors for CDP (Fig 4.9). CDP is not able to transfer glucose, or any other tested sugar, from the anomeric phosphates on to (β-1,4-Man)$_3$-APTS and there is no apparent phosphorolysis. This indicates the -1 binding site of CDP cannot tolerate an axial 2OH. (β-1,4-Xyl)$_3$-APTS on the other hand is a reasonable acceptor of Glc, forming long polymers. There may be some slight transfer of Xyl or Man from their respective phosphates (marked with arrows in Fig 4.9.B). This indicates the 6 hydroxyl is not required in the -1 binding site for CDP.

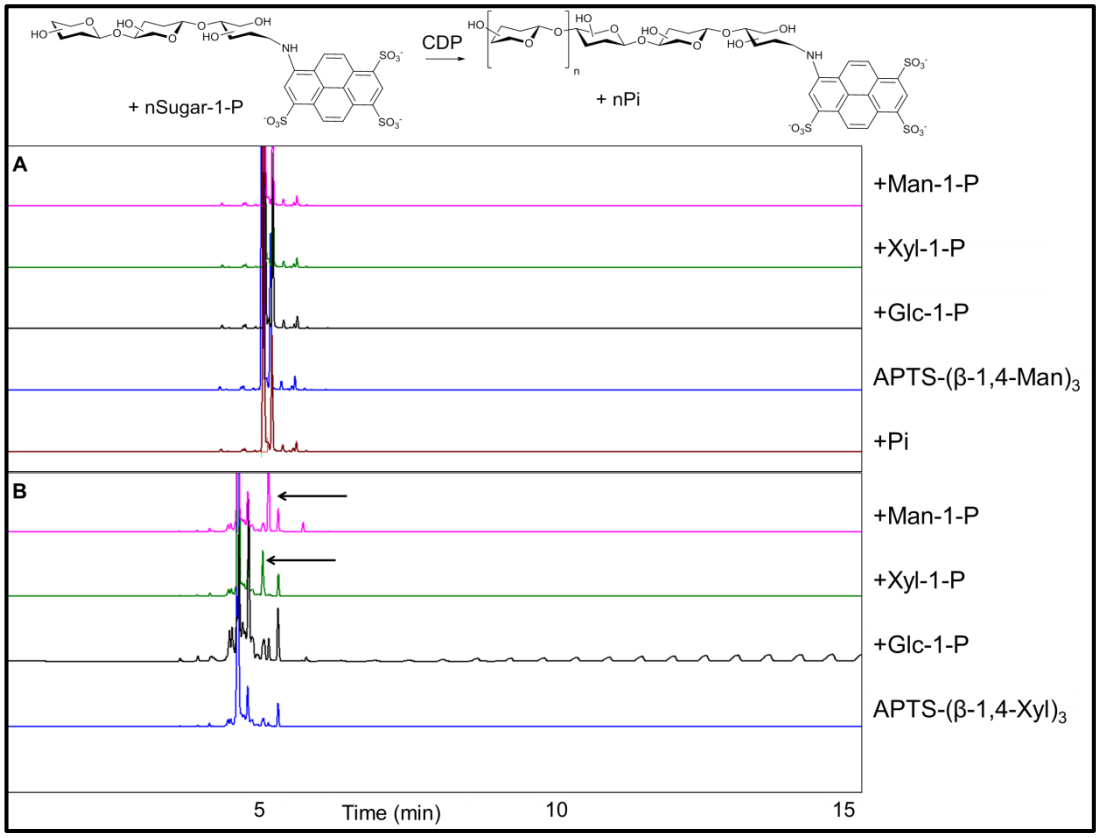


Figure 4.9: CDP catalysed extension of APTS-labelled β-1,4-glycans

A. Extension of APTS-(β-1,4-Man)$_3$, which is not an acceptor for CDP. **B**. Extension of APTS-(β-1,4-Xyl)$_3$, which is an acceptor for CDP catalysed transfer of glucose, and to a much lower extent xylose and mannose (arrows indicate new peaks formed). Assays were carried out using CDP (5 µg/ml) at 40 °C with Glc-1-P (10 mM) and APTS-labelled acceptor (13.5 µM) in HEPES (50 mM, pH 7.5) followed by heating to 95 °C and centrifuging at 16 000 g.

In order to probe this further, competition experiments were carried out using standard assay conditions (Fig 4.10). When a large excess of $(\beta\text{-}1,4\text{-}Xyl)_3$ was added, there was a significant decrease in the average extension of the APTS-$(\beta\text{-}1,4\text{-}Glc)_6$ acceptor indicating binding to the active site of the enzyme. Small additional shoulders on the peaks may indicate some Xyl-1-P is formed by phosphorolysis and this is then transferred on to the acceptor. This is consistent with xylans acting as acceptors for CDP.[33] Addition of $(\beta\text{-}1,4\text{-}Man)_3$ has no impact on the average extension indicating there is no binding to the active site.

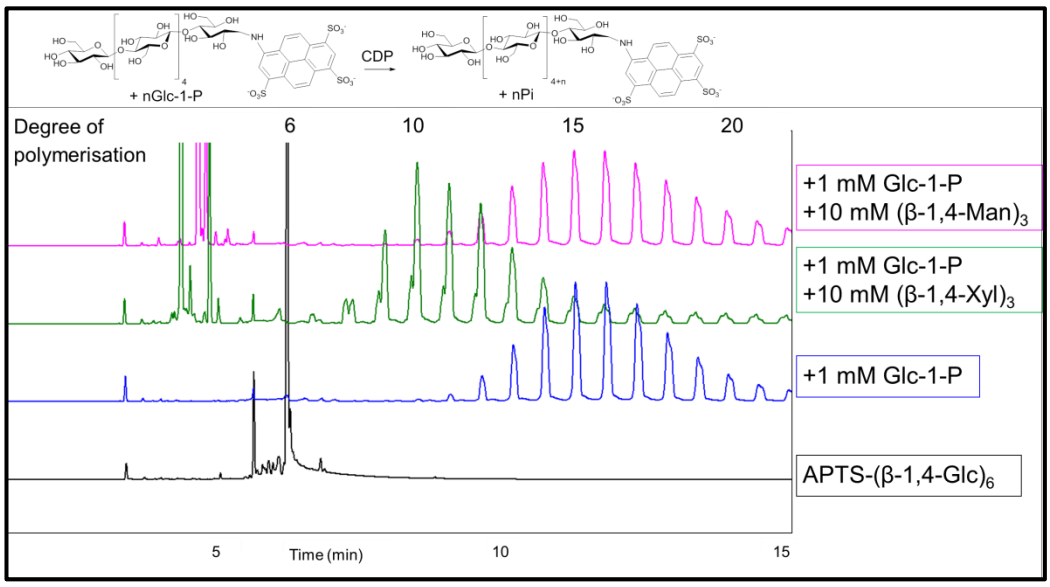


Figure 4.10: Inhibition of CDP catalysed cellulose polymerisation by β-1,4-glycans

Standard reactions had unlabelled trisaccharides added to assess the impact on the CDP catalysed reaction. The average DP of glucose polymer was less when $(\beta\text{-}1,4\text{-}Xyl)_3$ was added but not with $(\beta\text{-}1,4\text{-}Man)_3$. Assays were carried out using CDP (5 µg/ml) at 40 °C with Glc-1-P (10 mM) and APTS-$(\beta\text{-}1,4\text{-}Glc)_6$ (13.5 µM) in HEPES (50 mM, pH 7.5) followed by heating to 95 °C and centrifuging at 16 000 g.

4.3.4 Conclusions on the activity of CDP

Extensive work has previously been carried out on cellobiose phosphorylases to assess the substrate promiscuity and to synthesise disaccharides. Monitoring polymerisation reactions is less straightforward, but by utilising the combination of resolution and dynamic range provided by capillary electrophoresis the long polymers produced by this enzyme could be resolved. This technique shows that different polymers can be produced with mixed substrates and can resolve the relative impacts of binding and catalysis on the turnover of non-natural substrates, giving clues to the nature of enzyme-substrate interactions.

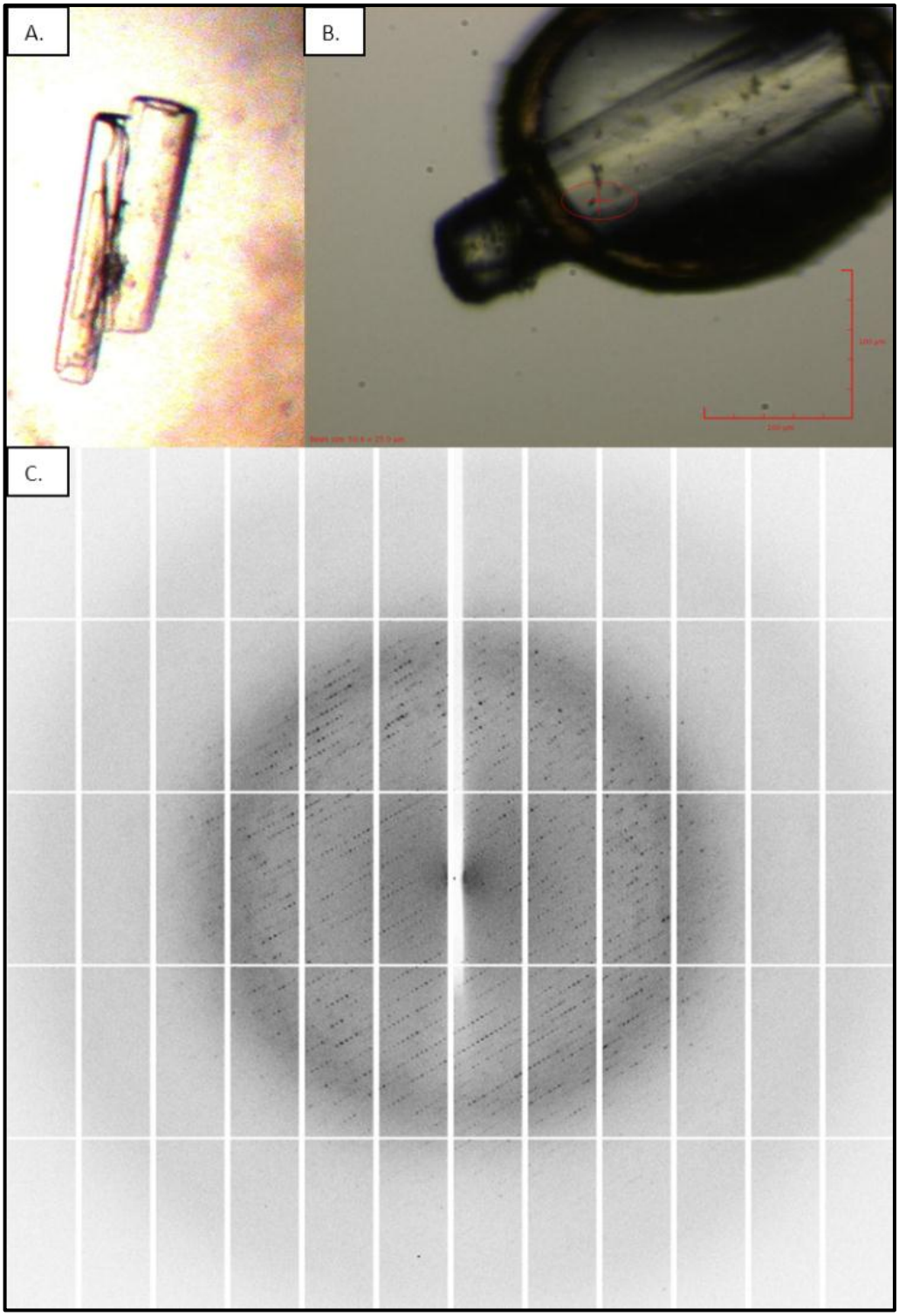


Figure 4.11: CDP crystals

A. A Crystal of CDP in the crystallisation drop. Crystals were clusters of approximately 50 x 200 µm rods which were separated before mounting in the loop. **B**. Crystal of CDP in litho loop at DLS, showing elliptical beam centred on the middle of the crystal. **C**. Example diffraction pattern of CDP crystal, showing diffraction to 3 Å.

4.4 Crystallisation of CDP

In order to better understand substrate recognition by CDP and inform engineering strategies, structural studies were carried out.

4.4.1 Protein crystallisation

After screening crystallisation conditions CDP was crystallised from approximately 20% (w/v) PEG3350, 100 mM HEPES (pH 7.5), KCl (300 mM) using the hanging drop vapour diffusion method at 20 °C. Crystals appeared after 2 days and were soaked in crystallisation solution containing 20% (v/v) glycerol, as cryoprotectant, and flash cooled to -173 °C in liquid nitrogen (Fig 4.11).

4.4.2 Heavy atom derivatisation

Attempts were made to soak the crystals with heavy metals to facilitate phasing by single-wavelength anomalous diffraction. Whilst mercury, gold and platinum derivatised crystals gave good diffraction and strong fluorescence (see Section 8.3.3.2). However, the anomalous signals were too weak for structure solution and the selenomethionine (SeMet) strategy was adopted instead. Protein was expressed from the same cells as above using the metabolic inhibition strategy in the presence of exogenously supplied SeMet.[41] The protein was purified from one litre of culture to give a final yield of 12 mg of purified protein, which was crystallised in the same conditions as the native enzyme. Crystals grown from the SeMet containing CDP showed a clear selenium *K*-edge in the X-ray energy scan (Fig 4.12) and diffracted to around 3.5 Å resolution.

Data set	SeMet peak	Native 1	Native 2
Data collection			
Space Group	$P2_1$	$P2_1$	$P2_1$
Cell parameters (Å)	*a* = 84.68 *b* = 151.80, *c* = 91.97, β = 114.6	*a* = 84.75 *b* = 151.50, *c* = 91.47, β = 114.7	*a* = 84.64 *b* = 151.80, *c* = 91.97, β = 114.6
Beamline[a]	I02	I04	I02
Wavelength (Å)	0.9795	0.9760	0.9795
Resolution range[b] (Å)	83.62 - 3.50 (3.59 - 3.50)	73.93 - 2.55 (2.62 - 2.55)	50.60 - 2.30 (2.36 - 2.30)
Unique reflections[b]	26456 (1937)	68073 (4901)	93120 (6720)
Completeness[b] (%)	99.0 (99.1)	99.7 (96.8)	99.6 (97.3)
Redundancy[b]	3.5 (3.4)	5.2 (4.9)	5.6 (3.4)
R_{merge}[b, c]	0.043 (0.142)	0.097 (0.880)	0.084 (0.759)
R_{meas}[b, d]	0.078 (0.174)	0.120 (1.092)	0.102 (1.035)
Mean I/σ(I)[e]	18.9 (7.1)	12.9 (2.0)	14.3 (2.0)
Wilson B value (Å^2)	72.9	63.9	57.9
Refinement			
Reflections: working/free[d]	-	-	88414/4668
R_{work}[f]	-	-	0.191 (0.317)
R_{free}[f]	-	-	0.223 (0.334)
Ramachandran favoured/ allowed[g] (%)	-	-	95.8/3.9
Ramachandran outliers[g]	-	-	6
rmsd bond distances (Å)	-	-	0.008
rmsd bond angles (º)	-	-	1.18
Contents of model			
Protein residues	-	-	984 (1-984)/978 (1-694; 653-795; 799-984)
Ligands[h]	-	-	6xCl
Precipitant/cryoprotectant molecules[h]	-	-	0
Water molecules			233
Average atomic displacement parameters (Å^2)			
Main chain atoms	-	-	66.7
Ligands[h]	-	-	59.6
Water molecules	-	-	52.1
Overall	-	-	66.5
PDB accession code	-	-	

Table 4.2: Summary of CDP X-ray data and model parameters

[a] I02, I04 = beamlines at the Diamond Light Source (Oxfordshire, UK).
[b] The figures in brackets indicate the values for outer resolution shell.
[c] $R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th observation of reflection *hkl* and $\langle I(hkl)\rangle$ is the weighted average intensity for all observations *i* of reflection *hkl*.
[d] $R_{meas} = \sum_{hkl} [N/(N-1)]^{½} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where *N* is the number of observations of reflection *hkl*.
[e] The data sets were split into "working" and "free" sets comprising 95% and 5% of the data, respectively. The free set was not used for refinement.
[f] The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum(|F_{obs} - F_{calc}|)/\sum|F_{obs}| \times 100$, where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively.
[g] As calculated using MOLPROBITY.[42]

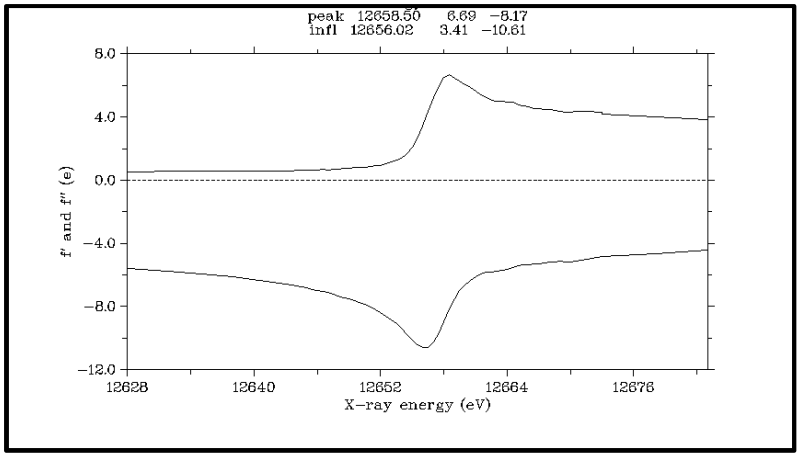


Figure 4.12: X-ray energy scan at the Se *K* absorption edge of a crystal of SeMet derivatised CDP

Crystals of SeMet derived CDP showed a good absorption edge near the expected 12666 eV.

4.4.3 Data collection and structure solution

The structure was solved by Single Isomorphous Replacement with Anomalous Scattering (SIRAS) using Phenix[43] by combining this data set with the native data to 2.55 Å resolution. A homology model of CDP, based on the structure of cellobiose phosphorylase from *Cellvibrio gilvus* (2CQT), the closest homologue (16.3% identity) for which the structure has been solved,[44] was manually docked into the resultant experimentally phased electron density map. A higher resolution native data set (2.3 Å resolution) was subsequently collected, allowing extension of the resolution of the model (Table 4.2). The protein backbone was automatically fitted into the electron density maps using the Buccaneer package,[45] followed by several rounds of manual model building and refinement in the CCP4 software suite[46] to give a final model with R_{work} and R_{free} values of 0.191 and 0.222 respectively.

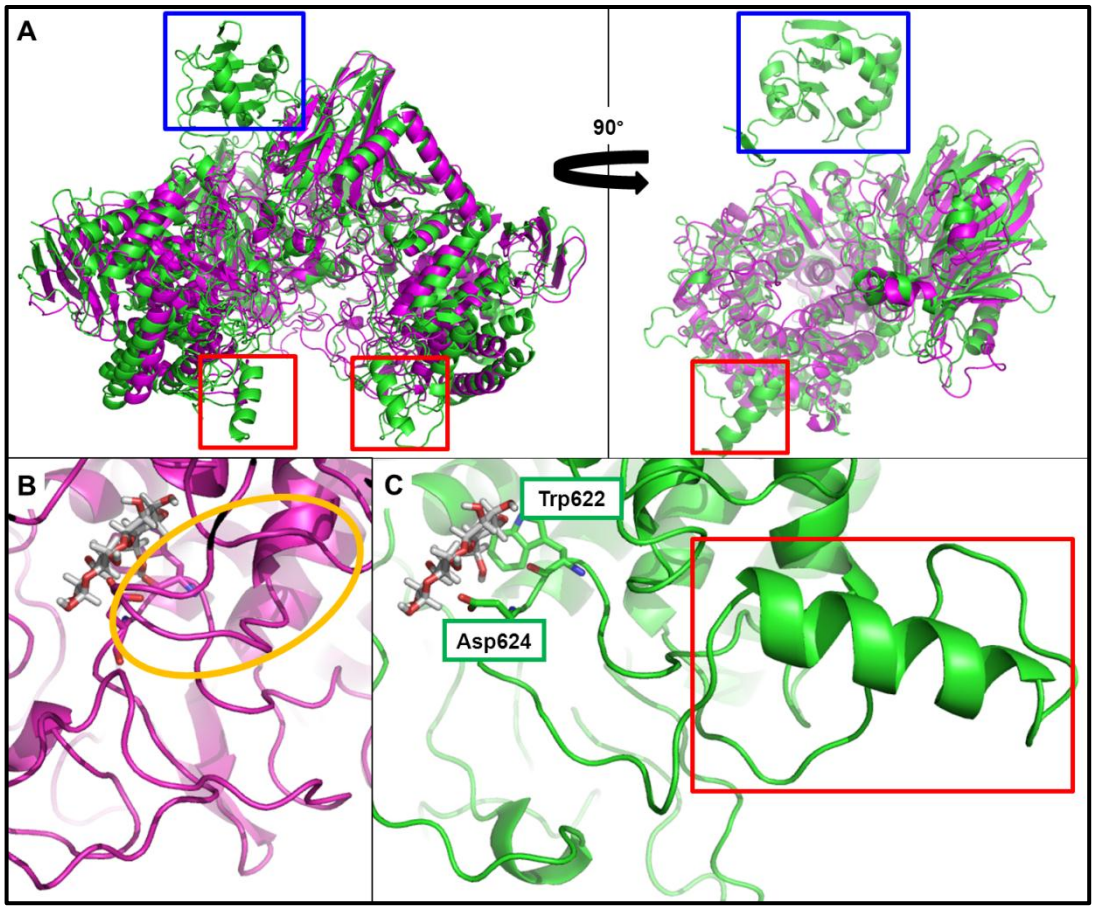


Figure 4.14: Comparison of the structures of CDP and CBP

A. CDP (green) has an N-terminal extension (blue box) over CBP (3QDE,[47] magenta), with clearly defined structure. Helix 630-660 (red box) has moved out of the active site compared to CBP. This structure was rotated through 90° and one subunit removed. **B**. CBP active site with cellobiose modelled form *C. uda* CBP (3S4A). Loop 490-510 (yellow) caps the substrate binding site. **C**. Helix 630-660 in CDP does not cap the active site compared to loop 490-510 in CBP. The hydrophobic platform (Trp622) and the catalytic base (Asp424) are both apparent in the structure. For alignment of primary protein sequences see Appendix 2.3.

4.4.4 Apo structure of CDP

The structure of the CDP is a dimer, which is consistent with the DyLS, but not GF. The dimer interface of CDP has about 4800 Å^2 of solvent accessible surface buried per monomer, giving an overall solvation free energy gain ΔG = -37.0 kcal/mol upon dimerization, as well as 59 H-bonds and 26 salt bridges (Fig 4.13).[48]

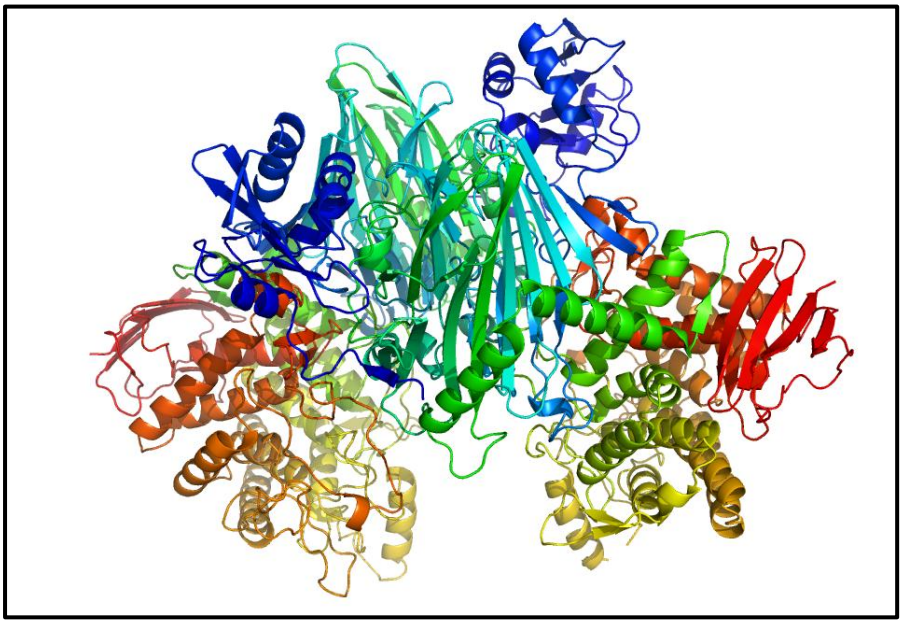


Figure 4.13: Structure of the CDP dimer

CDP forms a dimer, coloured from the N-terminus (blue) to the C-terminus (red) of each monomer.

There are few close structural homologues of CDP in the PDB: 17 entries give a DALI Z score of greater than 30,[49] although these come from just four non-redundant proteins: cellobiose phosphorylases from *Cellulomonas uda*, *Cellulomonas gilvus*[44] and *Clostridium thermocellum* (CBP)[47] and chitobiose phosphorylase from *Vibrio proteolyticus*.[50] Compared to these enzymes, CDP has an N-terminal extension of over 120 amino acids, which forms a well ordered globular domain with no known structural or sequence homologues (Fig 4.14.A) and might be involved in protein-protein interactions. The disaccharide phosphorylases have a relatively closed active site, with Loop 490-510 capping the substrate binding site (Fig 4.14.B). In CDP there is a 12 amino acid extension in the equivalent of this loop, Helix 630-660, which has moved away from the substrate binding site, with substantial remodelling of this side of the active site (Fig 4.14.C). This opens up the active site to accommodate the extended glucan substrate of this enzyme.

The residues identified in the CBP structure as the hydrophobic platform (Trp622) and the catalytic aspartate (Asp624)[47] are both apparent in the active site of CDP (Fig 4.14.C). Previously, it has been noted that the lactose phosphorylase capability of cellobiose phosphorylase can be enhanced by two mutations,[19] one of which (T→I) is in the highly altered region of the active site and the other (N→A) is already present in CDP.

Three regions of extra density could be resolved in the active site of each monomer in the structure of CDP. These were all modelled as chloride ions, based on their local environments and the high concentration in the crystallisation solution. One of these sites directly overlays with the active site phosphate in the structure of CBP from the same organism (Fig 4.15), suggesting this site probably binds phosphate during catalysis.

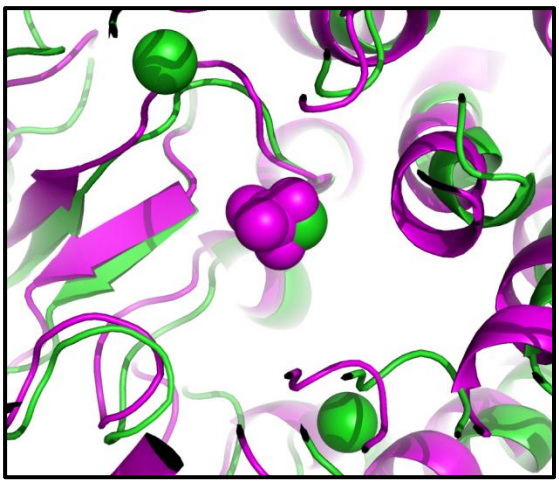


Figure 4.15: Proposed phosphate binding site of CDP

Phosphate has been resolved in the active site of CBP (3QDE, magenta).[47] Extra density can be resolved in the same place in the CDP structure (green), which is modelled as a Cl^-. Two other Cl^- atoms are also resolved in the active site.

4.4.5 Conclusions on the structure of CDP

The crystal structure of CDP is the first solved for an inverting polymerising phosphorylase. There is a large and unique N-terminal extension, which may be involved in making inter-protein contacts. The key difference between the relatively closed structure of the CBP and this structure rationalises the ability of CDP to accept the longer polymers. Work is on-going to achieve structures of CDP with oligosaccharides bound, which will help elucidate the substrate specificity of this enzyme.

4.5 Conclusions on cellodextrin phosphorylase

CDP can be used to synthesis long β-1,4-glucans which rapidly precipitate out of solution. There is a very slow transfer of other sugars, including mannose, xylose and glucosamine, and competition assays have helped to define the substrate recognition by CDP. Using CE xylotriose was also found to be a reasonable acceptor and to outcompete the APTS-labelled β-1,4-glucan, indicating it binds to the active site.

The crystal structure shows substantial rearrangement of the active site compared to cellobiose phosphorylase, which allows activity on longer polymers. There is also some evidence that phosphate binds in the same position as in the cellobiose phosphorylase. Together with the substrate binding assays the crystal structures will help to inform the synthesis of plant cell wall related β-1,4-oligomers.

4.6 References

1 Brown, R. M. Cellulose structure and biosynthesis: what is in store for the 21st century? *J. Polym. Sci., Part A: Polym. Chem.* **42**, 487-495 (2004).

2 Römling, U. Molecular biology of cellulose production in bacteria. *Res. Microbiol.* **153**, 205-212 (2002).

3 Lin, C. C. & Aronson, J. Chitin and cellulose in the cell walls of the oomycete, *Apodachlya* sp. *Archiv. Mikrobiol.* **72**, 111-114 (1970).

4 Brown, R. M. & Saxena, I. M. Cellulose biosynthesis: a model for understanding the assembly of biopolymers. *Plant Physiol. Biochem.* **38**, 57-67 (2000).

5 Delmer, D. P. & Amor, Y. Cellulose biosynthesis. *Plant Cell* **7**, 987-1000 (1995).

6 Saito, Y., Okano, T., Chanzy, H. & Sugiyama, J. Structural study of α chitin from the grasping spines of the arrow worm (*Sagitta* spp.). *J. Struct. Biol.* **114**, 218-228 (1995).

7 Himmel, M. E. *et al.* Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* **315**, 804-807 (2007).

8 Fontana, J. D. *et al. Acetobacter* cellulose pellicle as a temporary skin substitute. *Appl. Biochem. Biotechnol.* **24-25**, 253-264 (1990).

9 Okiyama, A., Motoki, M. & Yamanaka, S. Bacterial cellulose III. Development of a new form of cellulose. *Food Hydrocolloid.* **6**, 493-501 (1993).

10 Yoshinaga, F., Tonouchi, N. & Watanabe, K. Research progress in production of bacterial cellulose by aeration and agitation culture and its application as a new industrial material. *Biosci., Biotechnol., Biochem.* **61**, 219-224 (1997).

11 Percy, A., Ono, H., Watt, D. & Hayashi, K. Synthesis of β-D-glucopyranosyl-(1->4)-D-arabinose, β-D-glucopyranosyl-(1->4)-L-fucose and β-D-glucopyranosyl-(1->4)-D-altrose catalysed by cellobiose phosphorylase from *Cellvibrio gilvus*. *Carbohydr. Res.* **305**, 543-548 (1997).

12 Nakai, H. *et al.* Efficient chemoenzymatic oligosaccharide synthesis by reverse phosphorolysis using cellobiose phosphorylase and cellodextrin phosphorylase from *Clostridium thermocellum*. *Biochimie* **92**, 1818-1826 (2010).

13 Kitaoka, M., Taniguchi, H. & Hayashi, K. Characterization of cellobiose phosphorylase and cellodextrin phosphorylase. *J. Applied Glycosci.* **49**, 221-227 (2002).

14 Percy, A., Ono, H. & Hayashi, K. Acceptor specificity of cellobiose phosphorylase from *Cellvibrio gilvus*: synthesis of three branched trisaccharides. *Carbohydr. Res.* **308**, 423-429 (1998).

15 Alexander, J. K. Purification and specificity of cellobiose phosphorylase from *Clostridium thermocellum*. *J. Biol. Chem.* **243**, 2899-2904 (1968).

16 Nidetzky, B., Eis, C. & Albert, M. Role of non-covalent enzyme-substrate interactions in the reaction catalysed by cellobiose phosphorylase from *Cellulomonas uda*. *Biochem. J.* **351**, 649-659 (2000).

17 Kitaoka, M., Nomura, S., Yoshida, M. & Hayashi, K. Reaction on D-glucal by an inverting phosphorylase to synthesize derivatives of 2-deoxy-β-D-arabino-hexopyranosyl-(1->4)-D-glucose (2''-deoxycellobiose). *Carbohydr. Res.* **341**, 545-549 (2006).

18 Kim, Y.-K., Kitaoka, M., Krishnareddy, M., Mori, Y. & Hayashi, K. Kinetic studies of a recombinant cellobiose phosphorylase (CBP) of the *Clostridium thermocellum* YM4 strain expressed in *Escherichia coli*. *J. Biochem.* **132**, 197-203 (2002).

19 De Groeve, M. R. M. *et al.* Creating lactose phosphorylase enzymes by directed evolution of cellobiose phosphorylase. *Protein Eng., Des. Sel.* **22**, 393-399 (2009).

20 De Groeve, M. R. M. *et al.* Construction of cellobiose phosphorylase variants with broadened acceptor specificity towards anomerically substituted glucosides. *Biotechnol. Bioeng.* **107**, 413-420 (2010).

21 Kuk, J. H. *et al.* Production of N-acetyl-β-D-glucosamine from chitin by *Aeromonas* sp. GJ-18 crude enzyme. *Appl. Microbiol. Biotechnol.* **68**, 384-389 (2005).

22 Bronnenmeier, K., Kern, A., Liebl, W. & Staudenbauer, W. L. Purification of *Thermotoga maritima* enzymes for the degradation of cellulosic materials. *Appl. Environ. Microbiol.* **61**, 1399-1407 (1995).

23 Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506-577 (2002).

24 Atalla, R. H. & Vanderhart, D. L. Native cellulose: a composite of two distinct crystalline forms. *Science* **223**, 283-285 (1984).

25 Nakajima, M., Nishimoto, M. & Kitaoka, M. Characterization of three β-galactoside phosphorylases from *Clostridium phytofermentans*: discovery of D-galactosyl-β-1->4-L-rhamnose phophorylase. *J. Biol. Chem.* **284**, 19220-19227 (2009).

26 Senoura, T. *et al.* New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase. *Biochem. Biophys. Res. Commun.* **408**, 701-706 (2011).

27 Kawahara, R. *et al.* Metabolic mechanism of mannan in a ruminal bacterium, *Ruminococcus albus*, involving two mannoside phosphorylases and cellobiose 2-epimerase: discovery of a new carbohydrate phosphorylase, β-1,4-mannooligosaccharide phosphorylase. *J. Biol. Chem.* **287**, 42389-42399 (2012).

28 Park, J. K., Keyhani, N. O. & Roseman, S. Chitin catabolism in the marine bacterium *Vibrio furnissi*: identification, molecular cloning, and characterization of an N-diacetylchitobiose phosphorylase. *J. Biol. Chem.* **275**, 33077-33083 (2000).

29 Sih, C. J. & McBee, R. H. A phosphorylase active on cellobiose. *Proc. Montana Acad. Sci* **15**, 21-22 (1955).

30 Kitaoka, M., Taniguchi, H. & Sasaki, T. Production of glucosyl-xylose using *Cellvibrio gilvus* cells and its properties. *Appl. Microbiol. Biotechnol.* **34**, 178-182 (1990).

31 Kino, K. *et al.* A new method of synthesis of alkyl β-glycosides using sucrose as sugar donor. *Biosci., Biotechnol., Biochem.* **72**, 2415-2417 (2008).

32 Kawaguchi, T. *et al.* Cloning, nucleotide sequence, and expression of the *Clostridium thermocellum* cellodextrin phosphorylase gene and its application to synthesis of cellulase inhibitors. *J. Ferment. Bioeng.* **85**, 144-149 (1998).

33 Shintate, K., Kitaoka, M., Kim, Y. K. & Hayashi, K. Enzymatic synthesis of a library of β-(1->4) hetero-D-glucose and D-xylose-based oligosaccharides employing cellodextrin phosphorylase. *Carbohydr. Res.* **338**, 1981-1990 (2003).

34 Fukamizo, T. *et al.* Enzymatic hydrolysis of 1,3-1,4-β-glucosyl oligosaccharides by 1,3-1,4-β-glucanase from *Synechocystis* PCC6803: a comparison with assays using polymer and chromophoric oligosaccharide substrates. *Arch. Biochem. Biophys.* **478**, 187-194 (2008).

35 Hai Tran, G., Desmet, T., De Groeve, M. R. M. & Soetaert, W. Probing the active site of cellodextrin phosphorylase from *Clostridium stercorarium*: kinetic characterization, ligand docking, and site-directed mutagenesis. *Biotechnol. Progr.* **27**, 326-332 (2011).

36 Tran, H. G. *et al.* Biocatalytic production of novel glycolipids with cellodextrin phosphorylase. *Bioresour. Technol.* **115**, 84-87 (2012).

37 Choudhury, Ambar K., Kitaoka, M. & Hayashi, K. Synthesis of a cellobiosylated dimer and trimer and of cellobiose-coated polyamidoamine (PAMAM) dendrimers to study accessibility of an enzyme, cellodextrin phosphorylase. *Eur. J. Org. Chem.* **2003**, 2462-2470 (2003).

38 Sheth, K. & Alexande.Jk. Cellodextrin phosphorylase from *Clostridium thermocellum*. *Biochim. Biophys. Acta* **148**, 808-810 (1967).

39 Sheth, K. & Alexander, J. K. Purification and Properties of β-1,4-Oligoglucan:Orthophosphate Glucosyltransferase from Clostridium thermocellum. *J. Biol. Chem.* **244**, 457-464 (1969).

40 Arai, M., Tanaka, K. & Kawaguchi, T. Purification and properties of cellodextrin phosphorylase from *Clostridium thermocellum*. *J. Ferment. Bioeng.* **77**, 239-242 (1994).

41 Doublié, S. in *Methods Enzymol.* Vol. Volume 276 (ed Charles W. Carter, Jr.) 523-530 (Academic Press, 1997).

42 Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375-383 (2007).

43 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 213-221 (2010).

44 Hidaka, M. *et al.* Structural dissection of the reaction mechanism of cellobiose phosphorylase. *Biochem. J.* **398**, 37-43 (2006).

45 Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **62**, 1002-1011 (2006).

46 Bailey, S. The CCP4 suite - programs for protein crystallography. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **50**, 760-763 (1994).

47 Bianchetti, C. M., Elsen, N. L., Fox, B. G. & Phillips, G. N., Jr. Structure of cellobiose phosphorylase from *Clostridium thermocellum* in complex with phosphate. *Acta Crystallogr. F. Struct.* **67**, 1345-1349 (2011).

48 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774-797 (2007).

49 Holm, L. & Sander, C. DALI: a network tool for protein structure comparison. *Trends Biochem. Sci* **20**, 478-480 (1995).

50 Hidaka, M. *et al.* Chitobiose phosphorylase from *Vibrio proteolyticus*, a member of glycosyl transferase family 36, has a clan GH-L-like $(\alpha/\alpha)_6$ barrel fold. *Structure* **12**, 937-947 (2004).

Chapter 5 – Laminarin phosphorylase – a β-1,3-glucan phosphorylase

5.1 Introduction

5.1.1 β-1,3-Glucans

The chemistry, biochemistry and biology of β-1,3-glucans has been the focus of study for many years and has been extensively reviewed by Bacic *et al*.[1] This class of glycans is found throughout nature as exclusively β-1,3-linked chains, such as paramylon, the algal storage glucan,[2] or the bacterial extracellular curdlan.[3] They also contain various degrees of other linkages including β-1,4, as in oat glucan,[4] β-1,6 as in laminarin[5] or, in the case of a *Streptococcus* capsule, β-1,2.[6] β-Glucans are used structurally in the cell walls of some fungi, such *Aspergillus*,[7] and as cell wall reinforcing material in plants.[8] They are also used in some algae, including brown algae and diatoms, as the major storage carbohydrate,[9] in preference to the α-1,4-glucans used in plants and animals. For example, the β-1,3-glucan storage polysaccharide in the Euglenoid algae, known as paramylon, is a cytosolic crystalline granule, with characteristic shape dependent upon the species of *Euglena* (Fig 5.1).[10]

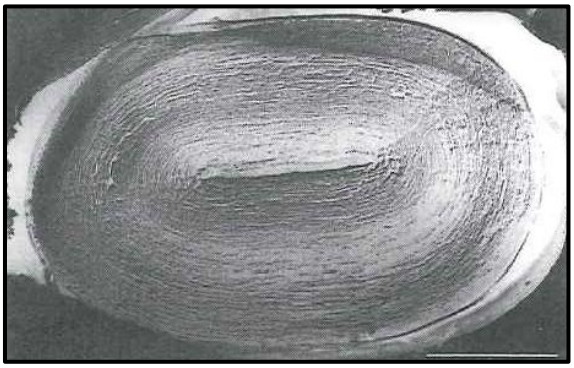


Figure 5.1: Freeze-etch scanning electron micrograph of a paramylon granule from *Euglena gracilis*

The β-1,3-glucans form 4 nm ridges in the paramylon granule, which is the width of the glucan triple-helix (scale bar = 500 nm). Reprinted from *Protoplasma,* with permission from Springer-Verlag. Copyright (1988).[11]

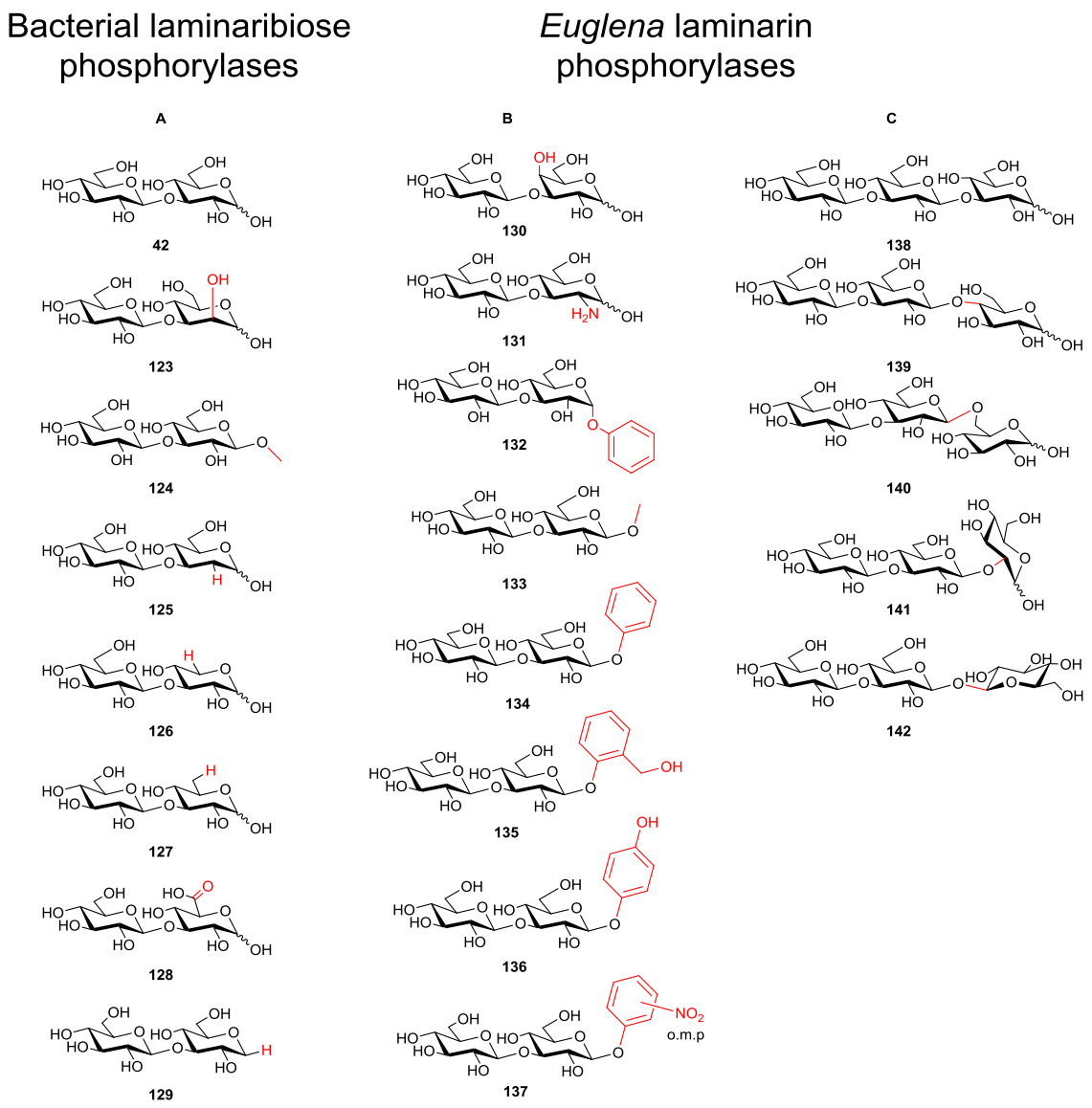


Figure 5.2: Proposed products formed using laminarin phosphorylases

A. Proposed products formed when glucose was transferred by bacterial laminaribiose phosphorylases on to D-glucose (**42**), D-mannose (**123**), β-methyl-D-glucoside (**124**), 2-deoxy-D-glucose (**125**), D-xylose (**126**), 6-deoxy-D-glucose (**127**),[12] D-glucuronic acid (**128**) and 1,5-anhydroglucitol (**129**).[13] **B**. Phosphorylases from *Euglena* are capable of transferring glucose on to the same sugars as the bacterial phosphorylases and in addition D-galactose (**130**), D-glucosamine (**131**), α-phenyl (**132**), β-methyl (**133**) and β-phenyl (**134**) D-glucosides, salicin (**135**), arbutin (**136**)[14] and *o*/*m*/*p*-nitrophenyl β-glucosides (**137**). **C**. Glucose can also be transferred on to laminaribiose (**138**), cellobiose (**139**), gentiobiose (**140**), sophorose (**141**)[15] and β,β-trehalose (**142**).[16] These are only identified by the Pi released from Glc-1-P by the phosphorylase in the presence of these acceptors and their stereochemistry has not been confirmed.

β-1,3-Glucans are plant immune elicitors, acting as pathogen associated molecular patterns (PAMPs), indicating to the plant the likely presence of a fungal parasite.[17] By applying these to crops, and activating the plant immune response, the yield of defence compounds, such as capsaicin in pepper,[18] increased and mildew could be controlled on strawberries.[19]

The β-1,3-glucan curdlan has been extensively investigated in the food industry,[20] where it is used for its excellent thermal stability and resistance to freeze-thaw.[21] This has led to research into improvements in yield and purification of this material from the producing bacteria.[22] There is a growing interest in biomedical applications of β-1,3-glucans, particularly with respect to immunological activity.[23] β-1,3-Glucans from diverse sources, including curdlan, fungal cell walls[24] and paramylon,[25] have been shown to targeting vaccines to immune cells, enhancing their immunogenicity. They also trigger immune cells, including natural killers and monocytes, which protect against cancer.[26] These glucans can lower cholesterol,[27] protect against HIV[28] and sulphated β-1,3-glucans can be used to prevent blood clots.[29] In addition, the triple helical structure of β-1,3-glucans[30] allows them to be used in diverse biotechnological applications, including as part of a chemosensor for acarbose[31] and for the co-complexation of DNA strands.[32,33]

5.1.2 Laminarin phosphorylase

Two laminaribiose phosphorylases have been cloned from β-1,3-glucan metabolising bacteria, *Paenibacillus sp. YM-1* and *Acholeoplasma laidlawii*.[12,13] These have been well characterised and belong to the GH94 family, alongside CDP (see Chapter 4). These two enzymes show reasonable promiscuity towards substitution at the 2 and 6 positions of the acceptor and are strict disaccharide phosphorylases (Fig 5.2.A). The *Euglena* laminarin phosphorylases were partially purified and analysed in 1993, but no sequence data is available.[15] They are able to form polymers, as opposed to the exclusively disaccharide phosphorylases found in bacteria, and allow large substituents on the reducing terminal of the acceptor, including other sugars (Fig 5.2.B+C). This has proved useful in the synthesis of cell wall related oligosaccharides, in combination with CDP, to form specific mixed β-1,3/1,4-glucan oligosaccharides.[34]

5.1.3 Rationale

In order to exploit β-1,3-glucans, a source of defined and tractable material would be highly desirable. The ability to specifically modify parts of the molecule, by inserting or appending other chemical moieties, including sugars, would allow further understanding of the mode of action, and give advances in the biotechnological utilisation of this carbohydrate.

In order to utilise the laminarin phosphorylase from *Euglena* (LDP) the enzyme was purified from cells and tested for activity. The utility of this enzyme for the glycosylation of small compounds was investigated. To obtain high purity enzymes heterologous expression is desirable, requiring a protein sequence. Sequencing of the transcriptome of *Euglena gracilis* var. *saccharophila*, which avoided the complications of genome sequencing,[35] revealed the range of carbohydrate active enzymes used by this alga. It was hoped that combining this transcriptome with proteomic analysis of the purified enzyme would allow identification of the sequence of the LDP protein.

5.2 Synthesis of glucosyl glycerols

5.2.1 Growth of *Euglena* cells

Euglena gracilis var. saccharophila was obtained from the culture collection of algae and protozoa (www.ccap.ac.uk) and grown in the recommended EG:JM media with added glucose. Cells were grown for 7 days in the dark and collected by centrifugation at 800 g. These cells were then lysed using a cell disruptor and cleared by centrifugation at 30,000 g.

5.2.2 Assay for phosphorylase activity

Two assays were employed to confirm the activity of the phosphorylase. Initially the release of phosphate from Glc-1-P in the presence of Glc was measured using the method of De Groeve *et al*.[36] In order to confirm the activity measured was indeed that of a laminarin phosphorylase, samples were then analysed for their ability to synthesise β-1,3-glucans using CE. Activity was confirmed by the extension of APTS-labelled laminaritriose to over 20 residues, in the presence of Glc-1-P, as analysed using capillary electrophoresis (Fig 5.3).

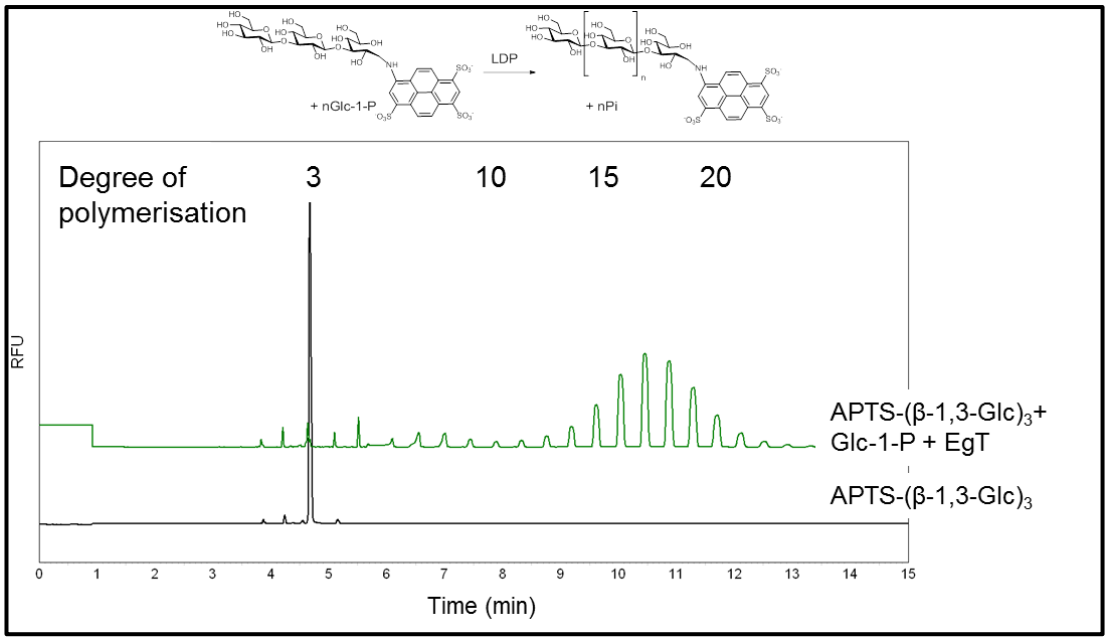
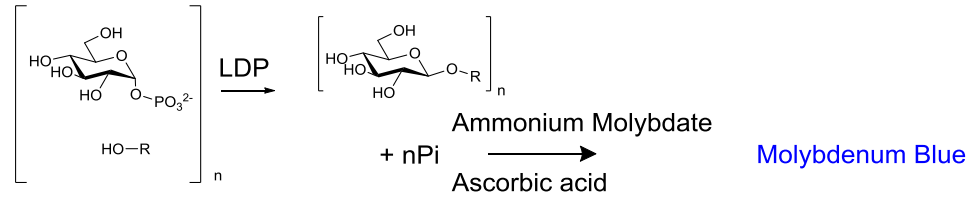


Figure 5.3: Extension of APTS-laminaritriose using *Euglena* phosphorylase

Cleared cell lysate (EgT, 5 µl) was used to extend APTS-(β-1,3-Glc)$_3$ (9.2 µM) by transferring Glc from Glc-1-P (20 mM) in buffer (20 mM MES pH 7.0) at 30 °C in 20 min.

5.2.3 Acceptor specificities of *Euglena* laminarin phosphorylases

The acceptor specificity of the *Euglena* phosphorylases was assayed by quantifying the amount of phosphate released from Glc-1-P in the presence of various acceptors (Fig 5.4). *Euglena* cell lysate was used with no further purification. Maltose was found to be a reasonable substrate, although the non-reducing terminal glucosyl moiety is held in the α-anomeric configuration, which is known to be non-permissive for the laminarin phosphorylases.[14] This led to the hypothesis that there is some hydrolysis of maltose to glucose, which then acts as the acceptor. Glucosamine and xylose are reasonably good acceptors, presumably because of their close similarity in structure to glucose. There is some enhancement of phosphate release from Glc-1-P in the presence of all the tested sugars and sugar alcohols, but the precise structure of the products is unclear.



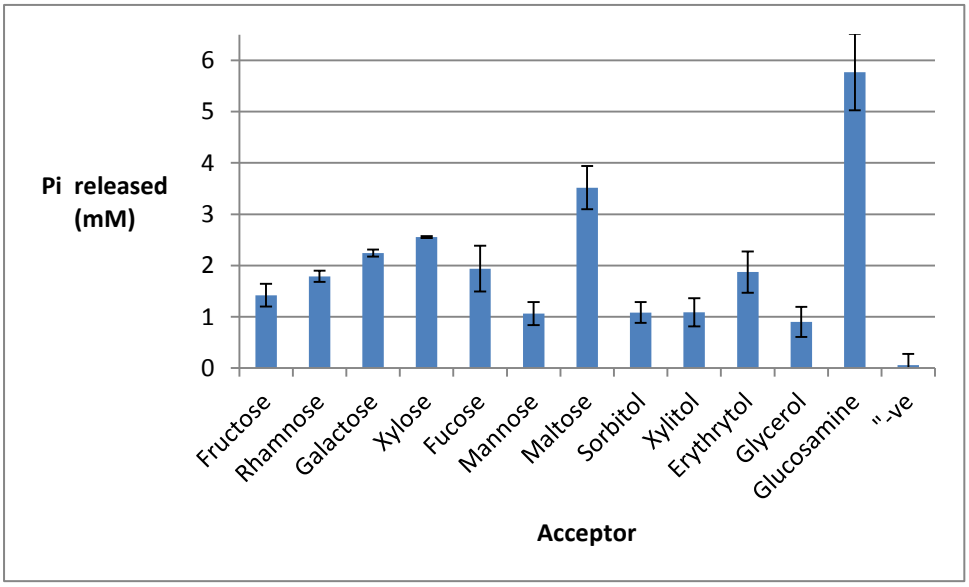


Figure 5.4: Acceptor specificity of phosphorylases in cleared *Euglena* cell lysate

Cleared *Euglena* cell lysate (1 mg/ml total protein) was used to transfer Glc from Glc-1-P (10 mM) on to various acceptors (10 mM) in reaction buffer (20 mM MES, pH 6.5) for 72 hr at 37 °C. The amount of phosphate released was measured using the phosphate release assay.

5.2.4 Glucosyl glycerols

Glucosyl glycerols (Glc-gol) are found in many organisms including plants[37] and bacteria.[38] The α-anomers are produced in many species in response to stress,[39] where they act as potent osmolytes with high water solubility and no net charge. Mixtures of mono-glucosyl-α-D-glycerol are stable at 160 °C and have low hygroscopicity yet a high water holding capacity,[40] making them attractive compounds for industry. Maltooligosaccharyl-α-D-glycerols have also been found to be inhibitors of digestive enzymes including pancreatic α-amylase[41] and an intestinal enzyme mixture.[42] The less studied β-glucosyl glycerols (β-Glc-gol), and derivatives, are also naturally occurring, for example found in the leaves of *Lillium longiflorum*.[43]

The natural biosynthetic pathway for the formation of Glc-gol, which involves transfer of glucose from unstable ADP-Glucose on to glycerol-3-phosphate followed by dephosphorylation,[44] is too complicated to be used industrially. Chemical syntheses are unsuitable for commercial production as they show low yields and a lack of stereo and regiospecificity.[45] 2-*O*-α-Glucosyl glycerol (**146**) has been prepared by Nidetzky's group using an *in vitro* phosphorylase catalysed biotransformation (Fig 5.5).[46] Sucrose phosphorylase (E.C 2.4.1.7) was used to transfer glucose from sucrose (**143**) on to glycerol (**144**), and the product was purified using simple column chromatography on a mixture of activated charcoal and calcined Celite. This method is currently used by bitop AG (www.bitop.de) to produce α-Glc-gol commercially, to be used as a cosmetic moisturiser, with further potential including use as a novel sweetener, as a biotechnological stabiliser and in cancer therapy.[47]

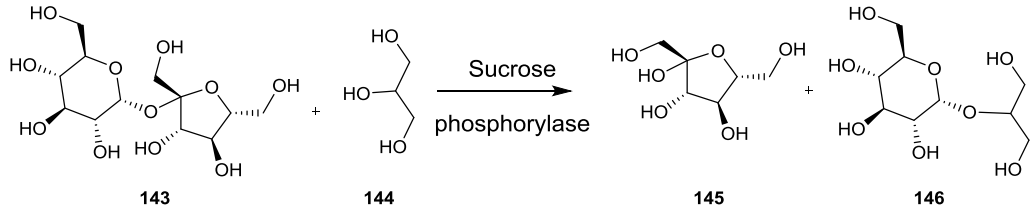


Figure 5.5: Sucrose phosphorylase catalysed synthesis of 2-*O*-α-glucosyl glycerol

Sucrose phosphorylase catalyses the transfer of glucose from sucrose (**143**) on to various acceptors, including glycerol (**144**), releasing fructose (**145**).[46] This is used in the commercial synthesis of α-Glc-gol (**146**).

β-Glucosyl glycerol is also reported to have useful properties, such as in the suppression of the tumour promoting activity of Epstein-Barr Virus.[48] The isomers of β-Glc-gol have been enzymatically synthesised using catalytically unfavourable transglycosylation or reverse hydrolysis. However these tended to yield a mixture of products which are difficult to separate.[49] For example the almond β-glucosidase-catalysed reverse hydrolysis can achieve yields of up to 54%, but the mixture of isomers produced (**147**,**148** and **149** in a 2:15:15 ratio) require separation (Fig 5.7).[50]

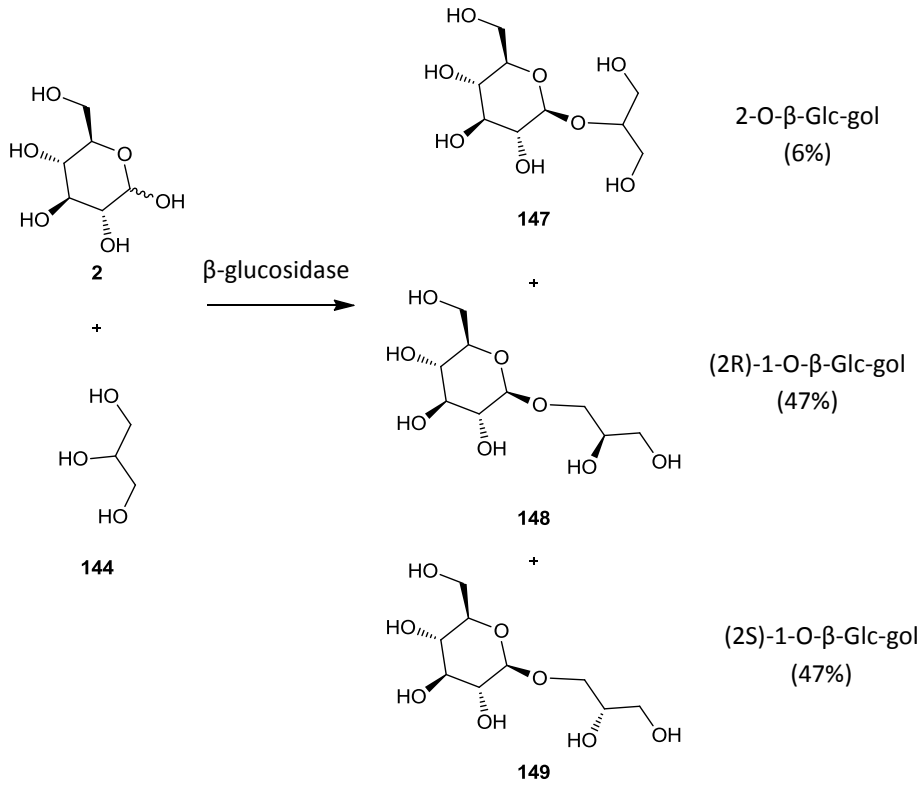


Figure 5.6: Reverse hydrolysis to form β-glucosyl glycerols

Almond β-glucosidase can be used to transfer glucose (**2**) on to glycerol (**144**), forming a mixture of 2-*O*-β-glucosyl glycerol (**147**), (2R)-1-*O*-β-glucosyl glycerol (**148**), (2S)-1-*O*-β-glucosyl glycerol (**149**).[50]

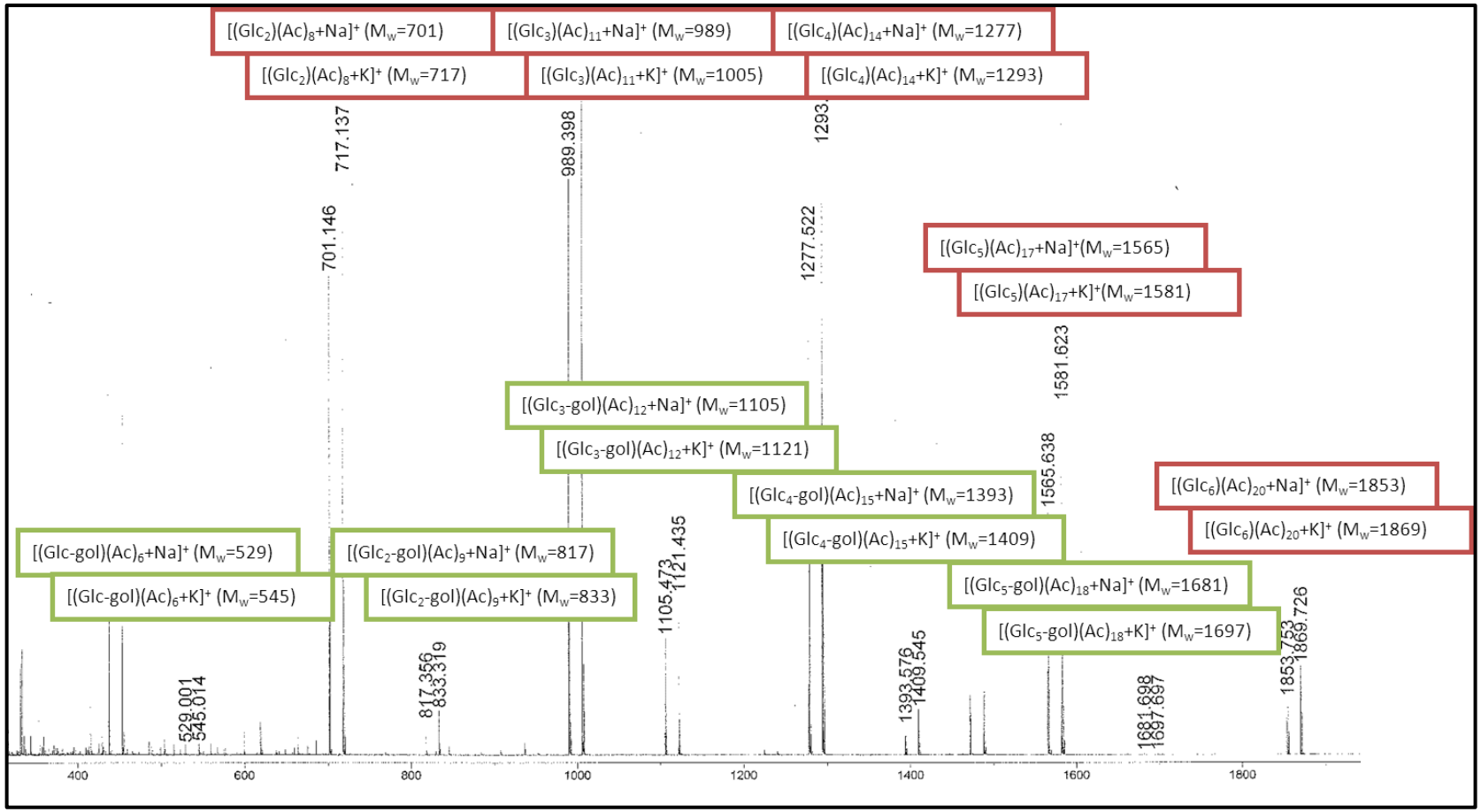
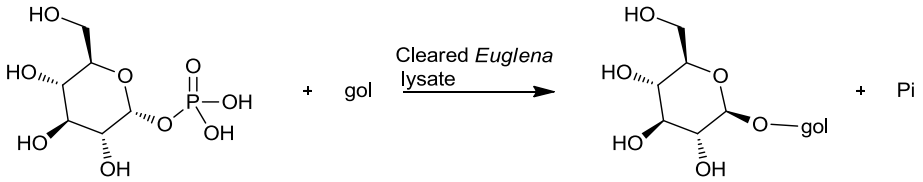


Figure 5.8: MALDI-ToF MS analysis of the peracetylated biotransformation of glucosyl glycerols using *Euglena* cell lysate

Highlighted in green are the glucosylated glycerols, presumed to be β-1,3-glucans, terminating in glycerol. Highlighted in red are the oligoglucosides, presumed to be β-1,3-glucans, synthesised by the repeated transfer of Glc from Glc-1-P on to Glc.

5.2.5 Biotransformation and purification of β-Glucosyl glycerols

When cleared *Euglena* lysate was incubated with Glc-1-P and glycerol, inorganic phosphate was released (see Section 5.2.3). To confirm that the product formed was the potentially valuable β-Glc-gol the reaction was repeated using 100 mg of Glc-1-P in a 10 ml reaction, with a yield of approximately 1.5% based on phosphate release, although the protocol has not been optimised (Fig 5.7).

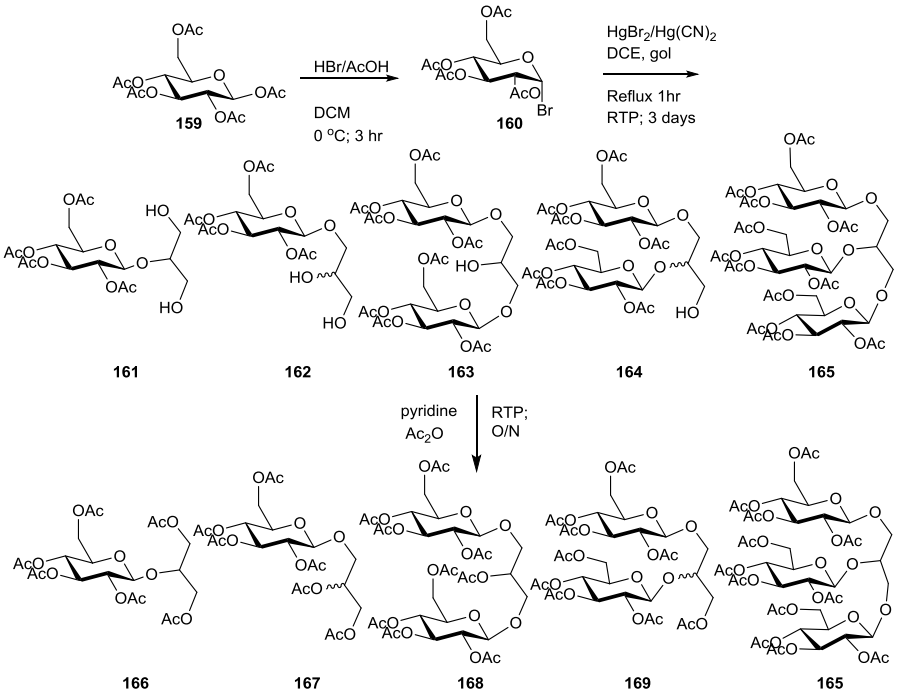


29 **144** **150**

Figure 5.7: Proposed enzymatic synthesis of β-Glc-gol using cleared *Euglena* cell lysate

Cleared *Euglena* cell lysate was used to transfer glucose from Glc-1-P (**29**) on to glycerol (**144**) to form glucosyl glycerol with undefined regio- and stereochemistry (**150**).

After removing the protein, the reaction mixture was subject to anion exchange chromatography to remove the buffer, unreacted Glc-1-P and any released Pi. The unbound material was acetylated using acetic anhydride (Ac_2O) in pyridine. MALDI-ToF MS indicated that poly-glucosylated glycerol had been produced (Fig 5.8, green). These are more likely to be β-1,3-glucan monosubstituted glycerol, rather than multisubstituted glycerol, although insufficient quantities of these products were produced for isolation and further characterisation (Fig 5.9). There was also a high proportion of glucose polymers produced (Fig 5.8, red), assumed to be β-1,3-glucans, formed by the reverse phosphorolysis, catalysed by LDP, on to Glc, which had been released by hydrolysis of Glc-1-P or the β-Glc-gol product. The acetylated mono-glucosylated products (**150**) could readily be isolated using silica gel chromatography, with a hexane:ethyl acetate gradient, to provide TLC homogenous material. The product identity was confirmed by MS and NMR analysis was consistent with literature data for peracetylated 2-*O*-β-glucosyl glycerol (Fig 5.11).



Scheme 1: Chemical synthesis of peracetylated β-glucosyl glycerols

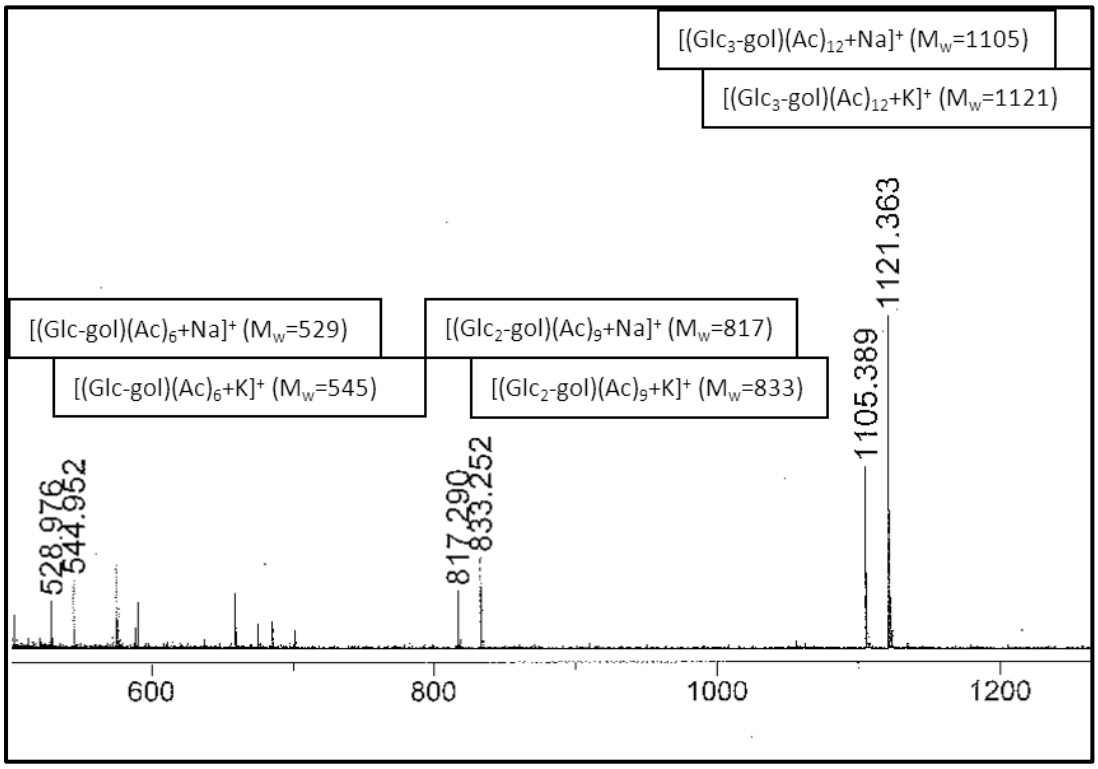


Figure 5.10: MALDI-ToF MS analysis of the products of the chemical synthesis of peracetylated glucosyl glycerols

The crude reaction mixture shows peaks consistent with the sodium and potassium adducts of peracetylated Glc-gol (**166-167**), Glc_2-gol (**168-169**) and Glc_3-gol (**165**).

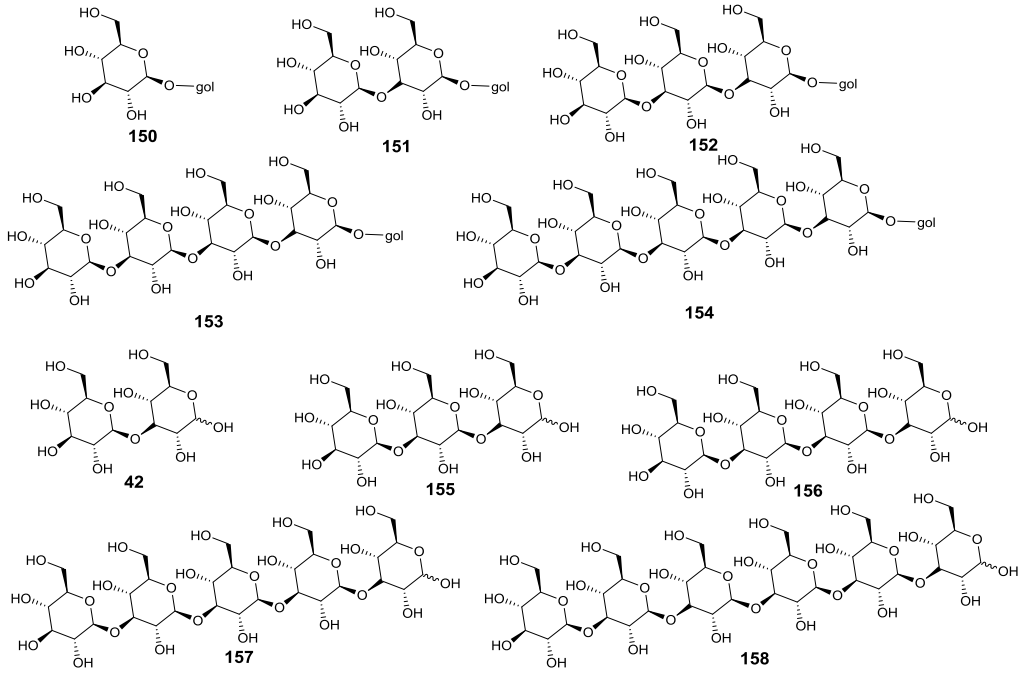


Figure 5.9: Proposed products of the glucosylation of glycerol with *Euglena* cell lysate

MALDI-ToF of the peracetylated products indicates that there are multiple products produced in the reaction. These are proposed to be β-1,3-glucans (**42, 155-158**), some of which terminate in glycerol (**150-154**).

5.2.6 Chemical synthesis of β-Glucosyl glycerols

In order to confirm that the product formed was indeed 2-*O*-β-glucosyl glycerol, the authentic compound was synthesised chemically. The chemical syntheses of all isomers of β-glucosyl glycerol have previously been described.[51] In order to make authentic standards of the peracetylated compound a statistical glycosylation was performed (Scheme 1). β-D-Glucose pentaacetate (**159**) was reacted with HBr in AcOH for 3 hours at 0 °C in DCM. The product was recrystallised from diethyl ether to give 2,3,4,6-tetra-*O*-acetyl-α-D-glucopyranosyl bromide (**160**). Glucosyl bromide **160** was then refluxed for one hour with glycerol in dry dichloroethane (DCE) in the presence of $HgBr_2$ and $Hg(CN)_2$, and the reaction was left to stir at room temperature over the weekend.[52] The glycerol was barely soluble in DCE, forming a tar in the reaction vessel with the mercury salts and Drierite. After working up, the reaction was directly acetylated with acetic anhydride in pyridine, yielding peracetylated mono- (**166-167**), di- (**168-169**) and tri- glucosylated glycerol (**165**), as judged by MALDI-ToF MS (Fig 5.10). These three species were separated by silica gel chromatography to single spots on TLC, which were confirmed by MALDI-ToF MS.

The chemically synthesised mono-glucosylated glycerol contained three products by NMR, which match the previously reported (2R/2S)-1-*O*-β-glucosyl glycerol (**167**, δ: 4.51(d, $J_{1,2}$ 8 Hz) and 4.52 (d, $J_{1,2}$ 8 Hz)) and the 2-*O*-β-glucosyl glycerol (**166**, δ: 4.63(d, $J_{1,2}$ 8 Hz)),[51] which could not be separated (Fig 5.11). Integration indicates a 10:6 ratio of **167**:**166**, and that the isomers of 1-*O*-glucosyl glycerol have a diastereomeric excess of 20%. The anomeric proton from the biologically synthesised product gave a doublet in ^{1}H NMR spectrum which matched the signal allocated to the 2-*O*-β-glucosyl glycerol synthesised chemically.

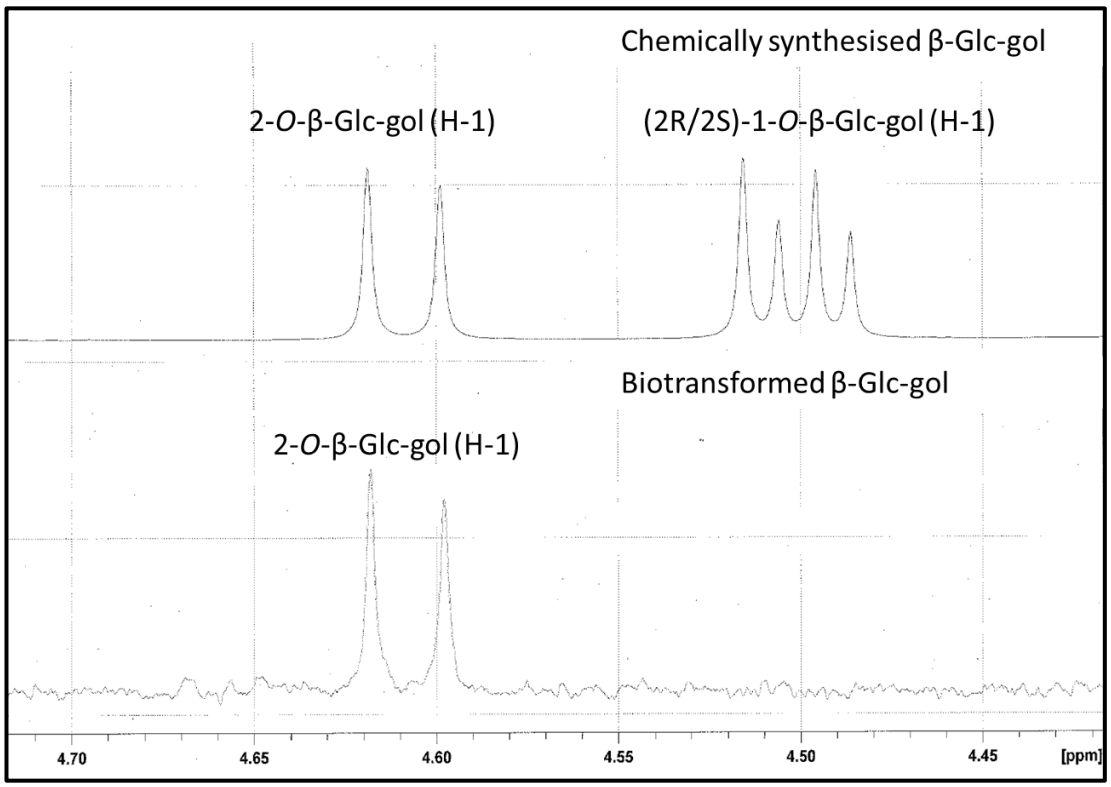


Figure 5.11: ^{1}H NMR analysis of the peracetylated mono-glucosyl glycerols

The purified peracetylated mono glucosyl glycerols formed from the chemical synthesis contains both 2-*O*-β-Glc-gol (**166**) and both (2R)- (**149**) and (2S)-1-*O*-β-Glc-gol (**148**), as judged by the chemical shift of the anomeric protons. The product of the biotransformation contains exclusively the 2-*O*-β-Glc-gol.

5.2.7 Conclusions on the synthesis of glucosyl glycerols

Glucosyl glycerols are attractive compounds, but the chemical synthesis is impractical and the natural biosynthesis too complicated. Using enzymatic biotransformation provides a commercially viable synthetic route for these compounds and the laminarin phosphorylase from *Euglena* has the potential for the production of large amounts of anomerically pure 2-*O*-β-glucosyl glycerol.

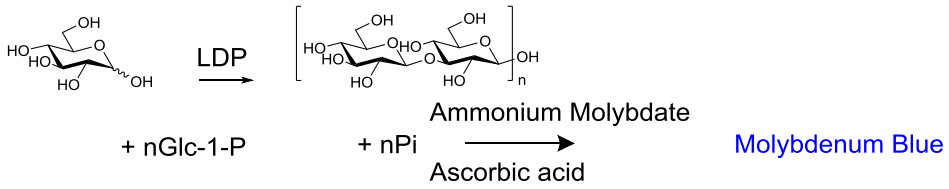
5.3 Partial purification of laminarin phosphorylase from *Euglena*

5.3.1 Purification strategies

In order to identify the laminarin phosphorylase form *Euglena* two purification strategies were developed (see Section 8.4.4) and proteomic analysis carried out on the most active fractions.

5.3.1.1 Strategy one

In strategy one (see Section 8.4.4.5), anion exchange chromatography (AIEX) was first performed on cleared *Euglena* cell lysate and two active fractions separated (Eg1.1/2, Table 5.1). There was very little recovery of the activity from the AIEX column, giving a very poor purification. The two active fractions were separately purified by hydrophobic interaction chromatography (HIC) which gave an increase in the purification, though again with loss of most of the activity (Eg2.1/2). After the final HIC purification Eg2.1 and Eg2.2 are both able to extend APTS-labelled laminaritriose, as judged by CE (Fig 5.12). These fractions were separately purified by gel filtration (GF). However no activity could be detected in the fractions. Instead the HIC fractions were precipitated with 5% TCA and run on SDS-PAGE. The region from 100-150 kDa was cut out, based on the previously identified size of LDP,[15] and proteomic analysis performed.



	Specific activity (µM Pi/min/mg)	% Recovery	Fold purification
Total (EgT)	0.056	100	1.0
AIEX1 (Eg1.1)	0.18	2.0	3.3
HIC on AIEX 1 (Eg2.1)	0.89	0.43	15.9
AIEX2 (Eg1.2)	0.074	2.7	1.3
HIC on AIEX2 (Eg2.2)	0.35	0. 37	6.2

Table 5.1: Purification of LDP using strategy one

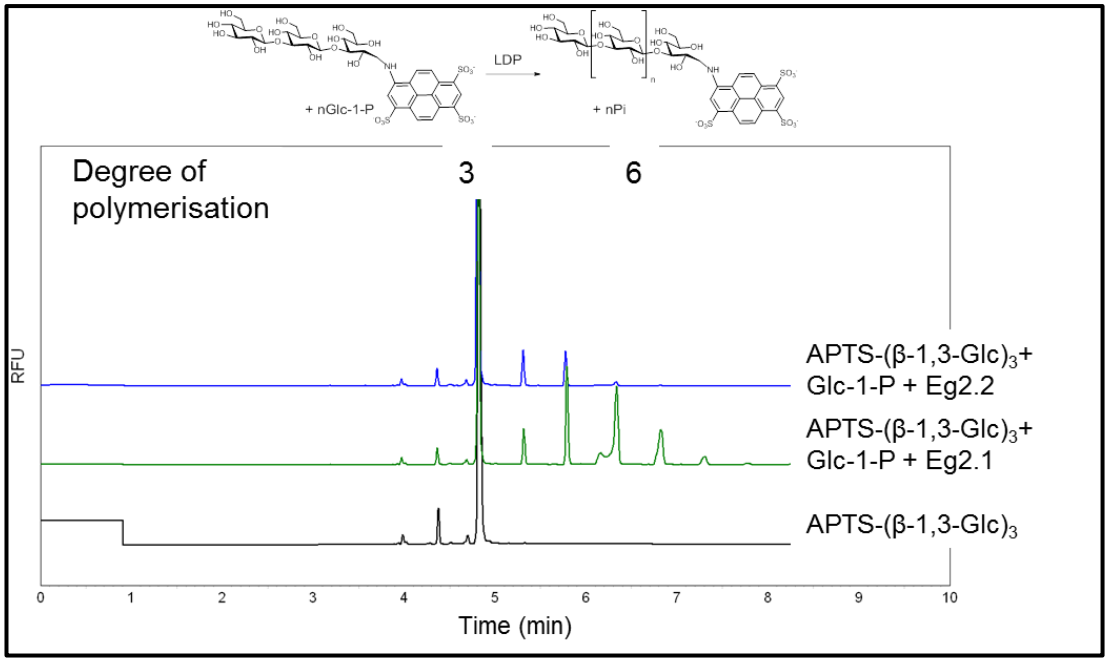
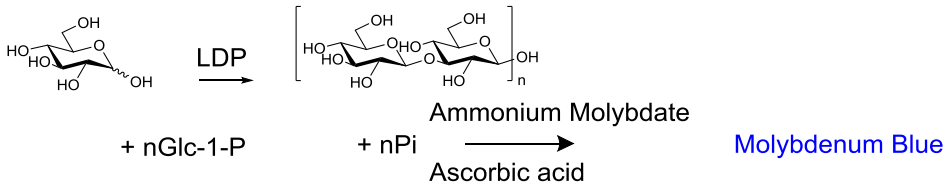


Figure 5.12: Activity of laminarin phosphorylase partially purified from *Euglena*

The LDP activities in the eluants from the HIC chromatography were confirmed by extension of APTS-(β-1,3-Glc)$_3$ (9.2 µM) by transferring Glc from Glc-1-P (20 mM) in buffer (20 mM MES pH 7.0) over 16 hours at 30 °C.

5.3.1.2 Strategy two

In strategy two (see Section 8.4.4.6) GF was performed (EgA), followed by HIC (EgB), both of which gave good purification and reasonable recovery (Table 5.2). The fraction with the highest relative activity was applied to the AIEX column. The activity at this point was only detectable after 16 hour incubation with the assay mixture, with the loss of activity possibly caused by inactivation of the enzyme under the high pH conditions (pH 8.5) of the AIEX. Proteomic analysis was performed on the most active fraction (EgD), as well as a fraction which showed no activity (EgC) but is expect to contain LDP, based on previous chromatographies.



	Specific activity (µM Pi/min/mg)	% Recovery	Fold purification
Total (EgT)	0.025	100	1.0
GF (EgA)	0.080	79.8	3.2
HIC (EgB)	0.297	3.1	12
AIEX1 (EgC)	NA	NA	NA
AIEX2 (EgD)	0.018	0.028	0.72

Table 5.2: Purification of LDP using strategy two
NA indicates activity was too low to be detected.

5.3.2 Proteomic analysis of partially purified laminarin phosphorylase

The protein fractions which were thought to contain laminarin phosphorylases (Eg2.1 and Eg2.2 gel slices and EgC and EgD) were subjected to proteomic analysis by tryptic digestion and peptide mass-fingerprinting. These were initially compared to the limited *Euglena* genome, protein and EST sequences in the NCBI. There were many peptides identified, but there were no proteins with homology to carbohydrate active enzymes and thus no candidates for the LDP in these sequences.

5.3.3 Conclusions on the purification of laminarin phosphorylases

Whilst it was not possible to completely purify the laminarin phosphorylase, fractions which showed greatly enhanced purity were achieved. This was in part due to the low alkaline stability of the activity. However there were many peptides identified which matched candidate enzymes, when better sequencing was interrogated (see Section 5.5.1, Appendix 3).

5.4 *Euglena* transcriptome sequencing

In order to identify candidate laminarin phosphorylases from the partially purified fractions good quality sequence data is needed. The genome of *Euglena* has not been sequenced due to its size and complexity. To assist in the identification of the laminarin phosphorylase a *de novo* transcriptome sequencing approach was utilised.

5.4.1 Sequencing of the *Euglena* transcriptome

5.4.1.1 Growth of Euglena

To obtain a broad set of transcripts the RNA of *Euglena gracilis* var. *saccharophila* was extracted from cells which had been grown under two radically different growth conditions.

The first growth condition was designed to force the cells to undergo photosynthesis and synthesis all the cellular components. This comprised a minimal media, solid agar containing no amino acids or carbon source. Cells were maintained at 21 °C with 12 hours light-dark cycle with ambient illumination. Cells from this condition are known as light grown cells.

The second growth condition included high nutrient media which provides the cells with all the basic cellular building blocks. This media contained amino acids, yeast extract and extra glucose. Cells were maintained at 30 °C in liquid culture with shaking at 200 rpm and light was excluded and cells were collected in mid-log phase growth. These cells were designated dark grown cells.

These growth conditions were chosen to give the widest diversity in growth types of *Euglena* in order to provide the broadest *de novo* transcriptome with which to understand the biosynthetic potential of this organism. Under both these conditions cells grew successfully and are expected to undergo paramylon synthesis, whereas under carbon starvation conditions heterotrophic cells rapidly degrade the paramylon granules.[53] The dark grown cells were used for the purification of the laminarin phosphorylase (see Section 5.3) and so both cell types are expected to contain the transcript for this enzyme.

5.4.1.2 Sequencing

mRNA was extracted from 10^6 cells grown under each of these conditions and equal amounts were pooled and sequenced using 100 base paired-end reads on one lane of Illumina HiSeq 2000, to give 26.5 Gbp and 11.9 Gbp of good quality reads, for light and dark grown cells respectively (Table 5.3). These sequences were assembled into contigs using the Trinity software package,[54] with approximately 290 and 140 fold read depth, for the light and dark samples.

	Light	Dark	Total
Total number of reads	264,808,150	118,608,486	383,414,636
Total number of nucleotides	26.5Gbp	11.9Gbp	38.3Gbp
Total number of contigs	233,748	231,176	
Mean length of contigs	391bp	364bp	
Total number of unique proteins	22,280	20,811	32,218
Mean length of unique proteins	456aa	401aa	

Table 5.3: Summary of *Euglena gracilis* transcriptome sequencing

The coding sequences were identified, to give 32,128 non-redundant protein fragments. Approximately one third of these were unique to the light grown cells, one third unique to the dark grown cells and one third shared between them (Fig 5.13).

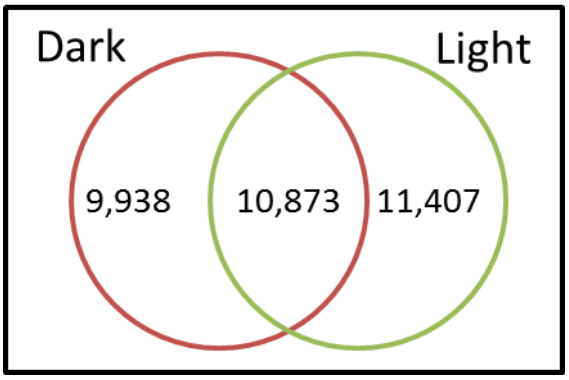


Figure 5.13: Sequence distribution in the light and dark grown cells

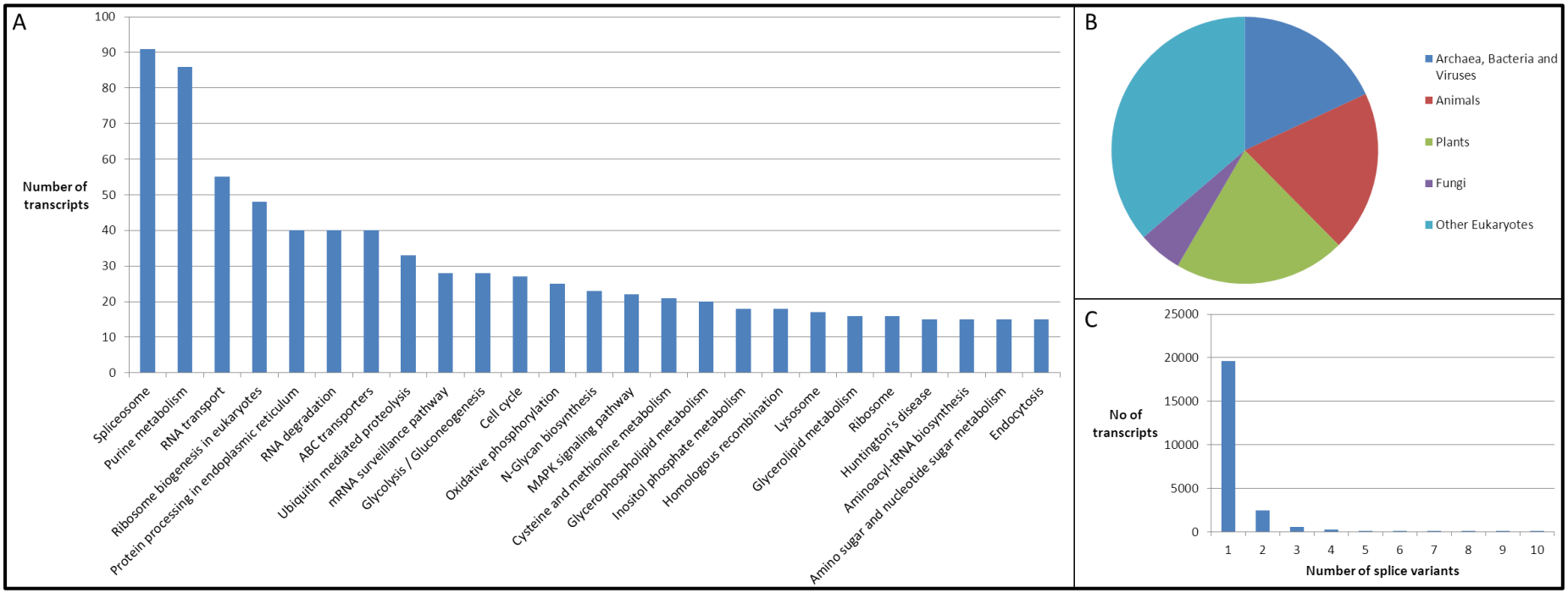


Figure 5.14: Annotation of the *Euglena* transcriptome

A. Top 25 GO ontology terms assigned to the transcriptome. **B**. Kingdom distribution of top hits of BLAST matches (>1 E^{-10}) of *E. gracilis* unique sequences. **C**. Distribution of numbers of splice variants estimated from Trinity component and sequence designation.

5.4.2 Annotation of the *Euglena* transcriptome

To annotate gene functions comparison of the proteins encoded on the contigs using BLASTP revealed 14,389 matches the UniRef100 database. Of these proteins 12,020 were classified into 157 Gene Ontology classifications (Fig 5.14.A). The species distribution of the top hits shows the diversity of sources of genetic material present in the *Euglena* genome and highlights its complex genetic history (Fig 5.14.B). It should be noted that whilst around 30% of the *Euglena* sequences deposited in the bioinformatics databases match the transcripts presented here, the rest do not. For example there are sequences for the HSP70 protein in this transcriptome, from Simpson *et al.*[55] and from the sequencing by the Mark C. Field laboratory, which only encodes 35 amino acids. These sequences only have 46% identity at the protein level over these 35 amino acids (Fig 5.15). This is likely to be because the strains used in other studies are not related to the one used in this work: caution should be taken when comparing sequence data from different algal isolates.

```
A          QLPANPENTIYAVKRLIGRKYEDKTVQSDKNLLAY
B          QSAMNPENTVFDAKRLIGRKIMDPDVQADMKNWPF
C          QVAMNPKNTVFDAKRLIGRKFTDATVQADMKHWPF
           *   **:**::  *******  *  **:* :   :
```

Figure 5.15: Comparison of the HSP70 protein sequence from different strains of *Euglena*

The HSP70 protein sequence from this transcriptome (**A**, lm.7137) was aligned with that obtained by genome sequencing (**B**, *E. gracilis* genome data obtained from the Mark C. Field laboratory sequence database) and that obtained from direct sequencing of the mRNA (**C**)[55] shows low identity (46%) between strains of this species. (**:** = homology, * = identity)

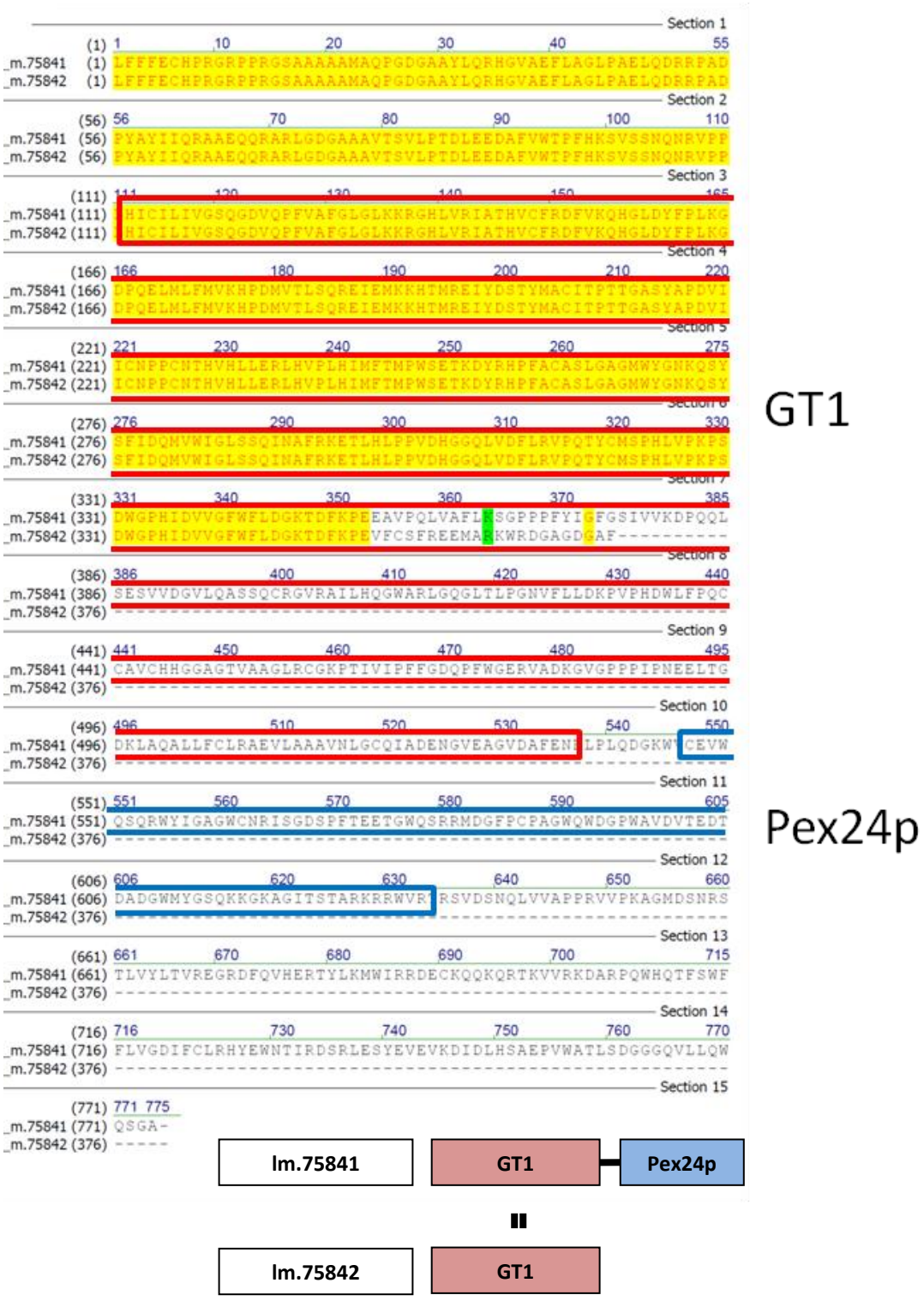


Figure 5.16: Examples of splice variants amongst the *Euglena* transcripts

lm.75841 and lm.75842 are identical for the first 354 amino acids, comprising a GT1 domain, but lm.75841 has a further 421 amino acids including a domain related to Pex24p, an integral peroxisomal membrane protein. Yellow indicates identity, green indicates similarity. Red box indicates GT1, blue box indicates Pex24p.

5.4.3 Splice variants amongst the *Euglena* transcripts

Many of the transcripts identified in the present study are alternative splice variants (Fig 5.14.C), information that would not be available from genome sequencing. For example, transcript lm.75841 and lm.75842 contain an identical N-terminus, coding for a glycosyltransferase, but the former has a C-terminal extension, encoding a peroxisomal protein which is not present in the latter (Fig 5.16). The splice variant abundance is at an approximately 11:1 ratio of long to short in the light sample, but in the dark sample no short sequence variant is detectable. This suggests that in the light the enzyme activity is required in both subcellular locations, but in the dark it is not required in the cytosol. *Euglena* appears to make use of alternative splicing to control sub-cellular targeting of a single gene product, as has been seen for many enzymes,[56] including glycolytic enzymes in fungi[57] and amino acid metabolic enzymes in plants.[58]

5.5 Carbohydrate-active enzymes in the *Euglena* transcriptome

In an attempt to identify the laminarin phosphorylases, all *Euglena* transcripts annotated as encoding carbohydrate-active enzymes (CAZys) were identified and classified into families, with the assistance of Bernard Henrissat at the AFMB in Marseille (Appendix 4).[59] The wealth of information provided by the transcriptome highlights many other aspects of carbohydrate metabolism carried out by this alga. *Euglena* is not reported to have a carbohydrate-based cell wall, although Trypanosomes, close relatives of *Euglena*, use sophisticated protein glycosylation machinery to achieve their highly flexible cell-surface glycoprotein coat, which is important for evasion of the immune system of infected hosts.[60] There are many more CAZys in the *Euglena* transcriptome than in the sequenced Euglenozoa, the human pathogens *Trypanosoma brucei* and *Leishmania braziliensis* (Fig 5.17). There are also more CAZys than most other sequenced algae, but fewer than the land plants, which require complex cell walls to support their growth.

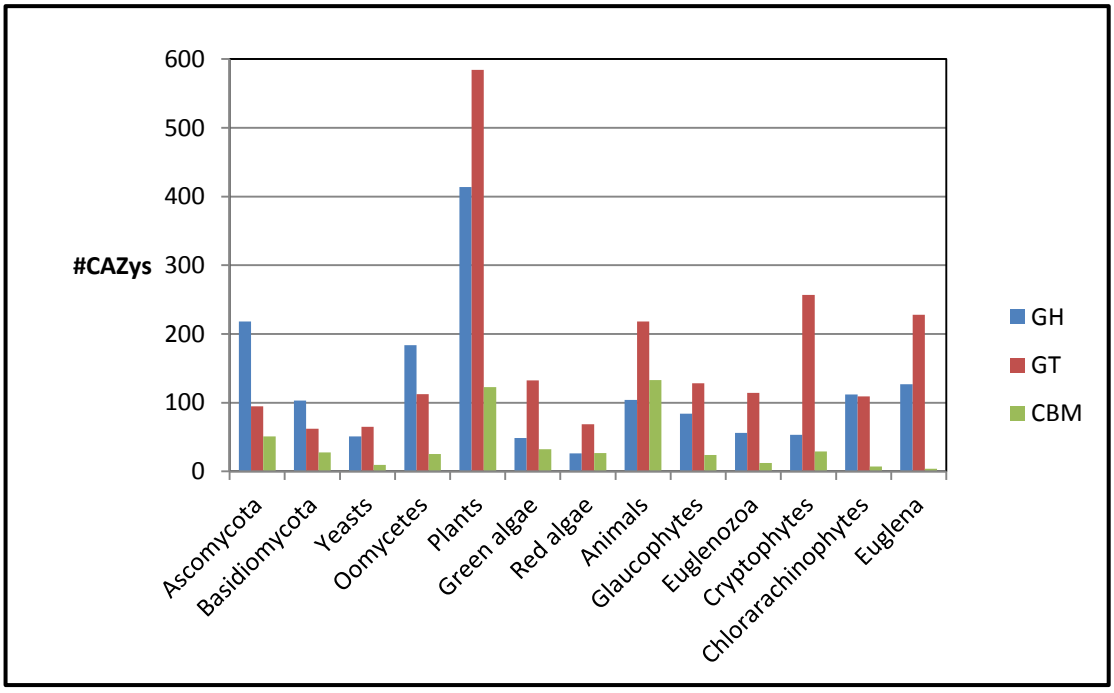


Figure 5.17: Average numbers of carbohydrate-active enzymes in different Phyla

The number of carbohydrate active enzymes annotated in the genomes of sequenced organisms was averaged amongst related organisms (Appendix 5). Most organisms have more glycosyltransferases then glycoside hydrolases, except amongst the saprophytic fungi and oomycetes.

5.5.1 Phosphorylases

Euglena is known to possess both trehalose[61] and laminarin[62] phosphorylases.

5.5.1.1 Trehalose phosphorylases

Bacterial trehalose phosphorylases are in family GH65 and have been well characterised.[63] There are sequences for four full length members of GH65 in the *Euglena* transcriptome (lm.12469, lm.41777, lm.50851 and lm.95741), which are related to the known bacterial trehalose phosphorylases. The first of these also encodes a β-phosphoglucomutase, which isomerises the β-Glc-1-P, released by the phosphorylase domain, into Glc-6-P, for general cellular metabolism (Fig 5.18).

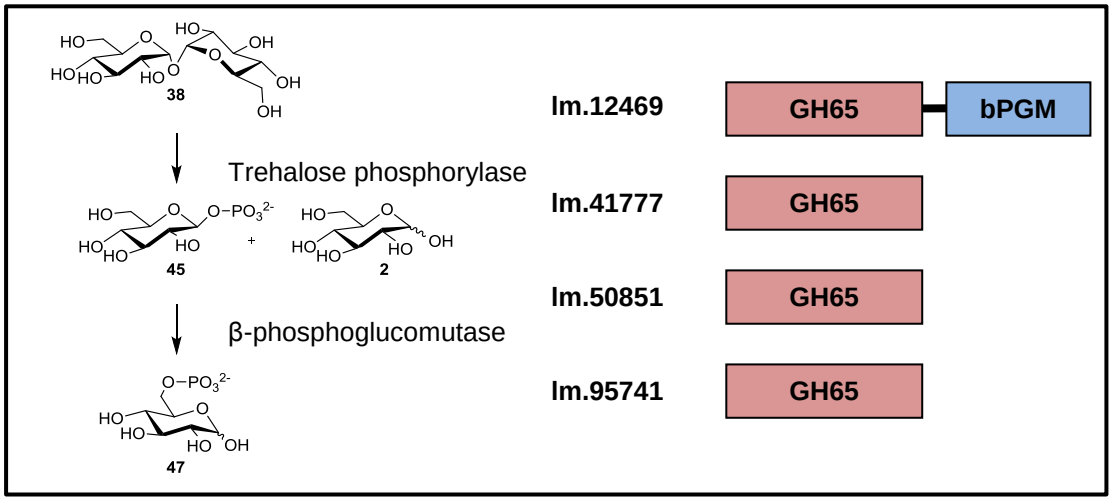


Figure 5.18: Trehalose phosphorylases encoded in the *Euglena* transcriptome

Trehalose (**38**) can be phosphorolysed in *Euglena* to produce β-Glc-1-P (**45**) and Glc (**2**). One of the enzymes (lm.12469) also contains a C-terminal β-phosphoglucomutase for the conversion of the β-Glc-1-P (**45**) into Glc-6-P (**47**).

5.5.1.2 Laminarin phosphorylases

β-1,3-Glucan phosphorylases from bacteria belong to family GH94 and act exclusively upon the disaccharide, laminaribiose.[12,13] No members of this family are present in the transcriptome, suggesting that *Euglena* utilises another enzyme family for this activity.

The proteomic analysis of the purified laminarin phosphorylase (see Section 5.3) yielded two potential candidate enzymes for the laminarin phosphorylase (dm.25642 and lm.26257), belonging to GH81 family of endo β-1,3-glucanases (see Appendix 3).[64]

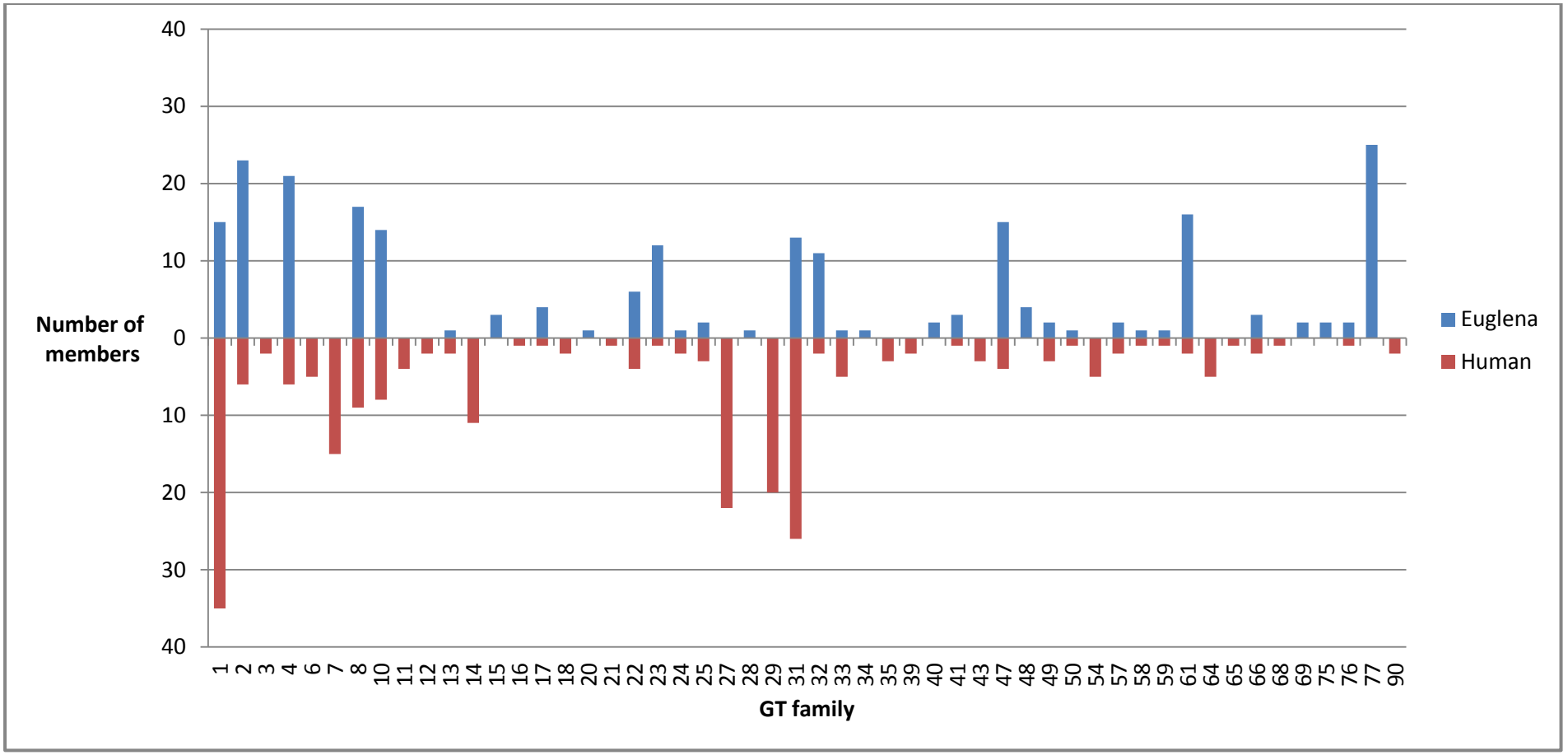


Figure 5.19: Comparison of glycosyltransferase family members in *Euglena* and humans

There are a similar number of glycosyltransferases in the human genome as in the *Euglena* transcriptome, but they are in radically different families.

5.5.2 Glycosyltransferase

By comparing the distribution of the *Euglena* CAZys with those in humans, hypotheses about the likely glycome of *Euglena* can be developed. There are approximately as many GTs in the *Euglena* transcriptome as are found in humans, but their distribution is markedly different (Fig 5.19).

5.5.2.1 Energy storage carbohydrates

As would be expected for an organism that does not produce or utilise α-1,4-glucans as an energy storage molecule, there are no members of GT3 or GT5, for their synthesis, or GT35 family of α-1,4-glucan phosphorylases for their degradation. Instead there are members of GT48 and GT2 for β-1,3-glucan synthesis which are absent from the human genome (see Section 5.5.5).

5.5.2.2 Glycolipids

The thylakoid membranes of plants contain galactolipids and sulfoquinovosyl diacyl glycerol, which are synthesised by members of CAZy families GT4[65] and GT28 (see Section 5.5.6).[66] These are absent from the human genome but present in the *Euglena* transcriptome, as would be expected for a photosynthetic organism. In addition there are two galactofuranosyl transferases (GT40) not present in humans. In *Leishmania* members of this family are involved in the biosynthesis of lipophosphoglycans.[67] The *Euglena* transcriptome also encodes the UDP-Gal mutase (lm.28797), required for the synthesis of the substrate for this enzyme. This may suggest that *Euglena* contains galactofuranose residues on their lipids or possibly proteins.[68]

5.5.2.3 Extra-cellular polysaccharides

Enzymes in families GT7, GT14, GT43 and GT64, which are involved in the synthesis of the human extra-cellular matrix glycosaminoglycans,[69] are entirely absent from the *Euglena* transcriptome. Instead There are members of diverse families for hemicellulose biosynthesis, such as GT77 α-galactosyl-, xylosyl- and arabinosyl-transferases[70] and GT34 α-galactosyl- and xylosyl-transferases,[71] which may be involved in the biosynthesis of extracellular polysaccharides (see Section 5.5.8).

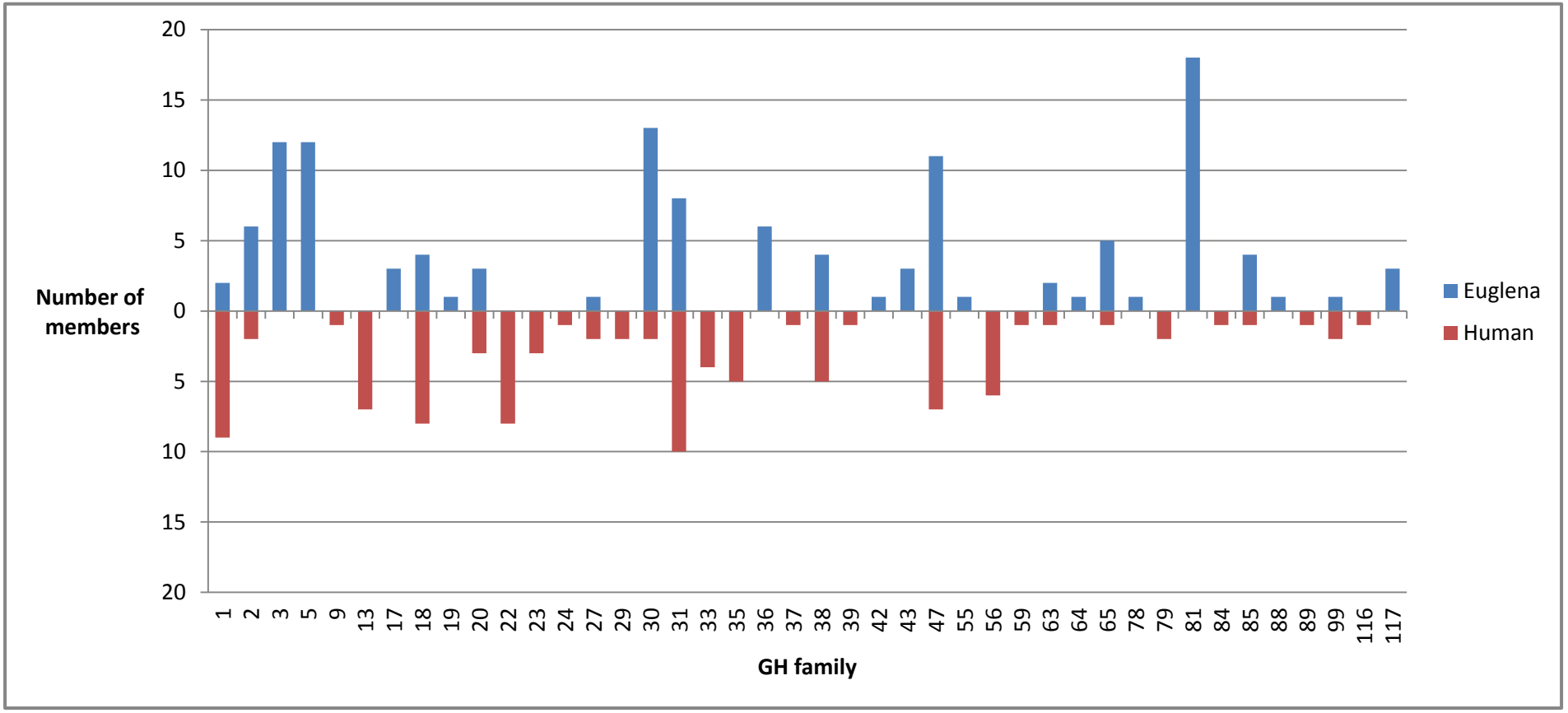


Figure 5.20: Comparison of glycoside hydrolase family members in *Euglena* and humans

There are fewer glycoside hydrolases in the human genome than in the *Euglena* transcriptome, and they are in radically different families.

5.5.3 Glycoside hydrolases

There are 30% more glycoside hydrolases in the *Euglena* transcriptome than are found in humans, and their distribution is markedly different (Fig 5.20).

5.5.3.1 Energy-storage carbohydrates

Critically, in the *Euglena* transcriptome there are no members of GH13 or GH77, which degrade α-1,4-glucans, but there are members of GH3, GH5, GH17, GH55, GH64, GH81, and an expansion of GH30 for β-1,3-glucan degradation (see Section 5.5.5). This is consistent with the use of β-1,3-glucans as the storage carbohydrate in *Euglena*, as opposed to the α-1,4-glucans used in bacteria, animals and plants.

5.5.3.2 Mammalian GH families, absent from the Euglena transcriptome

There do not appear to be any lysozyme family enzymes in *Euglena*, whereas these are prominent families in humans (GH22, GH23 and GH24). *Euglena* also lacks members of GH56, GH79, GH84 or GH89 for degradation of glycosaminoglycans, which forms a critical part of the extracellular matrix in animals. Whilst *Euglena* has been noted as containing sialic acid,[72] there are no members of the eukaryotic (GH33) or the viral (GH34, GH58 and GH83) sialidases, for the removal of this sugar (see Section 5.5.7.3).[73]

5.5.3.3 Complex polysaccharide hydrolases

In the *Euglena* transcriptome there are members of many families likely to be involved in the degradation of complex environmental polysaccharides, which are not encompassed by the human glycome. These include chitinases (GH19), galactosidases (GH56 and GH42), xylosidases (GH43) and rhamnosidases (GH78). There is a member of the GH88 family (lm.102467), which has only been characterised as hydrolysing unsaturated glucuronic acid, left after the action of a polysaccharide lyase. There is no evidence for any polysaccharide lyases in the transcriptome, suggesting this enzyme acts upon another sugar. There is also one full length member of GH117 (lm.76046), which has so far only been characterised as cleaving 3,6-anhydro-α-L-galactose from the non-reducing end of agar oligosaccharides.[74]

5.5.4 Carbohydrate binding modules

Carbohydrate binding modules (CBMs) are domains, normally on the same polypeptide as other CAZy domains, which bind to specific carbohydrates.[75] There are just four CBMs encoded in the *Euglena* transcriptome (Fig 5.21), compared to 126 in *Arabidopsis* and 40 in humans. *Trypanosoma* and *Leishmania*, the other sequenced Euglenozoa, only contain one CBM each, although these organisms live in the sugar rich environment of their animal hosts and thus they do not require complicated carbohydrate degradation mechanisms. All three sequenced Eugelnozoa contain a protein with a CBM48, which binds to α-1,4-glucans, and a protein kinase domain, which together make the AMP-activated protein kinase β-subunit,[76] which is activated upon autophosphorylation.[77] The precise target of this CBM in these organisms remains unclear, as they do not themselves use glycogen. There is a transcript in the *Euglena* transcriptome which encodes a GH18 chitinase with a chitin targeting CBM18. Expansins are used by plants for the loosening of their cell walls during growth.[78] The *Euglena* transcriptome contains a sequence with a possible expansin attached to a cellulose-binding CBM63, which is probably involved in the degradation of plant material,[79] as *Euglena* does not have a cell wall that would require expansin-mediated growth. There is also a single member of CBM57 malectin family in the *Euglena* transcriptome, which is thought to bind the Glc residues of the $Glc_3Man_9GlcNAc_2$ N-glycans in the endoplasmic reticulum.[80]

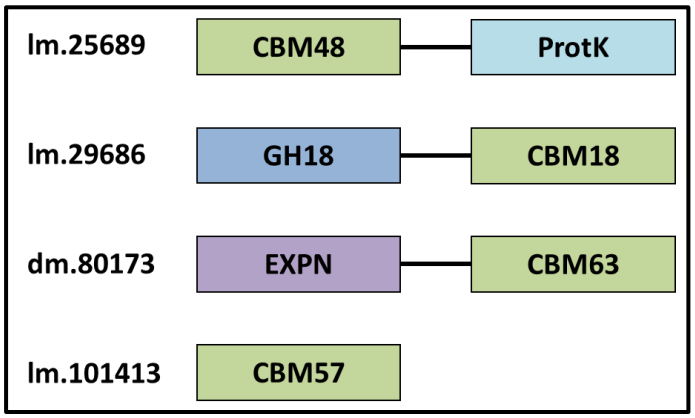


Figure 5.21: Domain architectures of the four CBM containing sequences in *Euglena*

CBM48 is characterised as binding α-1,4-glucans, and when associated with this protein kinase domain (ProtK), as in lm.25689, is the β-subunit of AMP-activated protein kinase. lm.29686 contains a chitin-binding CBM18 and a chitin degrading GH18. dm.80173 encodes a probable expansin (EXPN) and a cellulose directing CBM63. lm.101413 encodes a CBM57 with no other protein domains.

5.5.5 Paramylon metabolism

Paramylon granules are synthesised by the transfer of glucose from UDP-glucose by paramylon synthase, a membrane bound 670 kDa complex composed of at least seven different proteins as judged by SDS-PAGE.[81] Paramylon biosynthesis is reported to commence with the glucosylation of a protein,[82] in a similar way to glycogen biosynthesis. In the *Euglena* transcriptome there are many members of various CAZy families which act on β-glucans, and it is possible to propose the enzymes involved in paramylon metabolism (Fig 5.22).

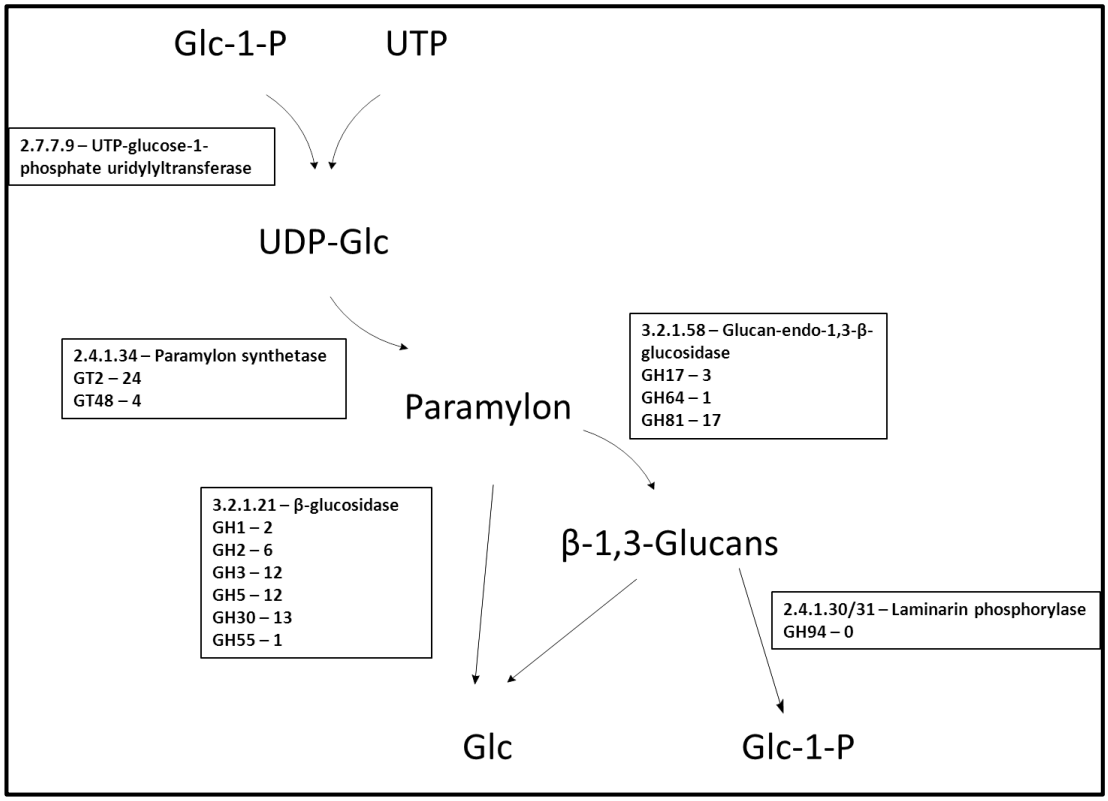


Figure 5.22: Enzymes proposed to be involved in paramylon metabolism

The number of members of the CAZy families which contain β-glucan metabolising activities encoded in the *Euglena* transcriptome.

5.5.5.1 β-Glucan synthesis

5.5.5.1.1 GT48

Plants[83] and fungi[84] synthesise the β-1,3-glucans in their cell walls using membrane-bound members of the GT48 family, which may represent candidate paramylon synthases in *Euglena*. Four sequences could be identified in the *Euglena* transcriptome which encode GT48 domains, three of which are closely related (94% identity), and one of which (lm.13336) is a short fragment, missing the first 1900 amino acids, and is probably not a full length transcript. These longer proteins are predicted to be around 300 kDa, much larger than the 16-80 kDa proteins which are reported to make up the paramylon synthase complex.[81] It is possible that these larger proteins undergo proteolytic cleavage to leave the lower molecular weight species identified in the purified complexes. Alternatively, these members of GT48 may be involved in paramylon biosynthesis but are not associated with these complexes or they may be involved in the synthesis of another β-glucan.

5.5.5.1.2 GT2

Enzymes of the GT2 family are used in the biosynthesis of β-1,3- and β-1,4-glycans such as callose, chitin and cellulose[59] and also represent good candidates for paramylon synthesis. This family is heavily represented in the *Euglena* transcriptome, with 16 unique members (Fig 5.23). The activities of the characterised members of GT2 do not cluster based upon their sequences and the *Euglena* enzymes are spread throughout the family making it impossible to predict their functions reliably. Whilst some of the sequences are clearly not full length, probably a factor of the sequencing rather than a true reflection of their natural characteristics, these proteins are of the expected size for members of the paramylon synthase complex.[81]

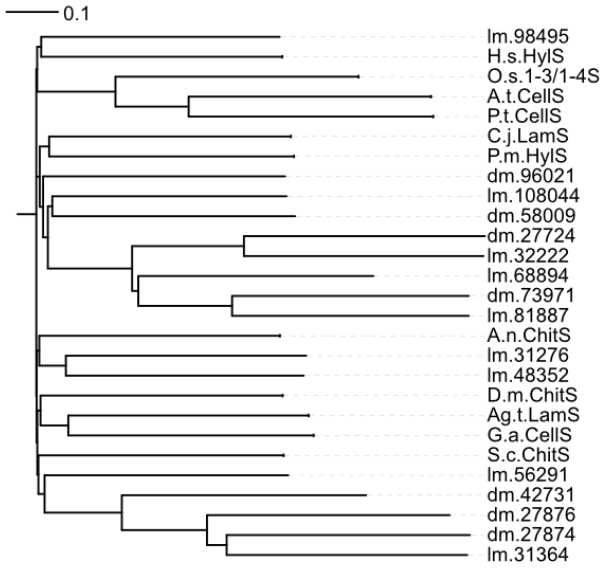


Figure 5.23: Phylogeny of family GT2 proteins

Euglena and well characterised GT2 family proteins were aligned. *Euglena* enzymes are spread throughout this family and their specificity cannot be predicted. H.s.: *Homo sapiens*. A.t.: *Arabidopsis thaliana*. P.t.: *Populus trichocarpa*. C.j.: *Campylobacter jejunei*. P.m.: *Pasteurella multocida*. A.n.: *Aspergillus nidulans*. D.m.: *Drosophila melanogaster*. Ag.t.: *Agrobacterium tumefaciens*. G.a.: *Glucanoacetobacter diazotrophicus*. S.c.: *Saccharomyces cerevisiae*. HylS: Hyalurone synthase. 1-3/1-4S: β-1,3/1,4-glucan synthase. CellS: Cellulose synthase. LamS: Laminarin synthase. ChitS: Chitin synthase.

5.5.5.1.3 GT75

Two components of the paramylon synthase complex were found to become glucosylated during granule biosynthesis.[81] A member of family GT75 from *Zea mays* is noted as being capable of self glucosylation, with the formation of a β-glucosyl arginine linkage.[85] It is conceivable that the two members of this family encoded in the *Euglena* transcriptome (lm.47975 and lm.47982) represent this initiator of paramylon granule biogenesis.

5.5.5.2 β-Glucan degradation

It is known that in *Euglena* paramylon is degraded by a series of endo- and exo-β-glucosidases to form glucose,[86] or by laminarin phosphorylase to form α-Glc-1-P (Section 5.5.1).[62] There are many glycosyl hydrolases present in the *Euglena* transcriptome that belong to a variety of GH families capable of degrading β-1,3-glucans.

5.5.5.2.1 GH81

There are transcripts encoding 17 members of GH81, which are well characterised endo-β-1,3-glucan hydrolases predominantly found in fungi, where they are required for cell division.[87] This family includes the best candidates for the laminarin phosphorylase identified by protein purification.

5.5.5.2.2 GH17

Three members of GH17, which is used by plants to degrade the cell wall β-glucans of pathogenic fungi,[88] are present in the *Euglena* transcriptome.

5.5.5.2.3 GH64

Members of GH64 has been found to specifically cleave five residues from the end of β-1,3-glucans.[89] There is one member of this family in the *Euglena* transcriptome (lm.63754) which is only distantly related to any sequenced organisms.

5.5.5.2.4 GH1

Members of families GH1 catalyse the hydrolysis of β-linked glucose and galactose from the non-reducing end of carbohydrates and are relatively non-specific.[90] There are two members of this family in the *Euglena* transcriptome.

5.5.5.2.5 GH2

GH2 includes diverse β-glycosidases, notably the *E. coli* lacZ gene.[91] Whilst β-glucosidase activity is not known in this family, activity towards GlcNAc and glucuronic acid is known and there are 6 members in the transcriptome of *Euglena*, suggesting it could have some involvement in paramylon degradation.

5.5.5.2.6 GH3

GH3 β-glycosidases are widely distributed in nature and act in a wide variety of roles, including cellulose degradation, cell wall remodelling and pathogen response.[92] There are 12 members of this family in the *Euglena* transcriptome and they may be involved in the release of storage carbohydrates, as they are known to be in barley.[93]

5.5.5.2.7 GH5 and GH30

GH5 is one of the largest and most complex CAZy families, which cleave a wide range of β-linked oligosaccharides.[94] There are 12 members of this family in the transcriptome of *Euglena*, although it is difficult to predict their exact substrate specificity. Several members of GH5 were transferred into the already extant GH30 after further sequence analysis,[95] to provide this family with β-xylosidase activity, as well as β-fucosidase and β-galactosidase activities. β-Glucosidase activity is also noted in family GH5[96] and as such the 13 members in the *Euglena* transcriptome may be good candidate paramylon hydrolases.

5.5.5.2.8 GH55

Fungal enzymes with exclusive β-glucosidase activities make up GH55.[97] This makes the single member of GH55 in the *Euglena* transcriptome the most likely candidate hydrolase for the degradation of paramylon, though all the other families noted may also be involved.

EC #	Reaction	Protein	Transcript #	Closest homologue	E-value
5.5.1.4	Inositol-3-phosphate synthase 1	INO1	lm.32631	inositol-3-phosphate synthase 1 [Saccoglossus kowalevskii]	0.00E+00
3.1.3.25	Myoinositol-1-phosphatase	IMPA1	lm.49313	hypothetical protein [Trichoplax adhaerens]	1.00E-60
2.7.8.11	Phosphatidylinositol synthase	PIS1	lm.37320	phosphatidylinositol synthase [Spathaspora passalidarum]	3.00E-50
			lm.79366	predicted protein [Phaeodactylum tricornutum]	1.00E-47
2.4.1.198	GPI-GlcNAc transferase	PIG-A	dm.79942	UDP-GlcNAc:PI α-1,6 GlcNAc-transferase [Trypanosoma cruzi]	0.00E+00
		PIG-C	none		
		PIG-H	dm.82217	glycosyltransferase [Ectocarpus siliculosus]	1.00E-16
		PIG-P	dm.82867	predicted protein [Populus trichocarpa]	4.00E-24
		PIG-Q	dm.48080	hypothetical protein Batrachochytrium dendrobatidis]	5.00E-36
		PIG-Y	none		
		DPM2	lm.108666	unknown [Picea sitchensis]	6.00E-09
3.5.1.89	GlcNAc-PI de-*N*-acetylase	PIG-L	dm.85635	GlcNAc-PI de-*N*-acetylase [Leishmania major]	2.00E-52
			dm.27659*	hypothetical protein [Paramecium tetraurelia strain d4-2]	3.00E-51
2.3.-.-	Inositol acyltransferase	PIG-W	lm.93529	uncharacterized protein [Brachypodium distachyon]	2.00E-64
2.4.1.-	*α*-(1-4)-Mannosyltransferase	PIG-M	lm.94033	GPI mannosyltransferase 1 [Dicentrarchus labrax]	2.00E-100
2.7.-.-	EtNP transferase	PIG-N	lm.94615	GPI ethanolamine phosphate transferase 1-like [Brachypodium distachyon]	3.00E-177
2.4.1.-	*α*-(1-6)-Mannosyltransferase II	PIG-V	lm.98056	dolichol-p-mannose mannosyltransferase [Selaginella moellendorffii]	4.00E-65
2.4.1.-	*α*-(1-2)-Mannosyltransferase III	PIG-B	dm.85690	predicted protein [Physcomitrella patens]	3.00E-86
2.7.-.-	EtNP transferase	PIG-O	lm.48401	hypothetical protein [Trichoplax adhaerens]	1.00E-94
		PIG-F	dm.104252	protein kinase [Dictyostelium fasciculatum]	9.00E-09
	GPI transamidase	PIG-K	dm.51731	predicted protein [Physcomitrella patens]	3.00E-122
		GAA1	lm.88011	predicted protein [Naegleria gruberi]	4.00E-25
		PIG-S	dm.90214	unnamed protein product [Vitis vinifera]	8.00E-21
		PIG-T	dm.67348	hypothetical protein [Batrachochytrium dendrobatidis]	8.00E-52
		PIG-U	lm.71943	gpi transamidase component [Colletotrichum gloeosporioides]	3.00E-31
	GPIdeacylase	PGAP1	lm.73955	GPI inositol-deacylase [Metarhizium anisopliae]	2.00E-36

Table 5.4: Annotation of enzymes involved in the biosynthesis of GPI anchors

Euglena transcripts were identified which encode enzymes involved in GPI anchor biosynthesis by homology with known enzymes.[98] Highlighted in yellow are transcripts only present in the dark cells and highlighted in green are those present only in the light grown cells. *Also contains an N-terminal mannosyltransferase domain.

5.5.6 Glycolipids

5.5.6.1 Glycophosphatidylinositol anchor

Glycophosphatidylinositol (GPI) anchors are glycolipids attached to proteins to anchor them to the membranes of cells.[99] Their synthesis is well characterised (Fig 5.24) and all of the enzymes involved, with the exception of the small accessory proteins PIG-Y, PIG-X and PIG-C, are readily identifiable in the *Euglena* transcriptome (Table 5.4).[98] The first enzyme in the biosynthesis of the inositol ring has been purified from *Euglena* cells previously,[100] but the range of anchored proteins and the structure of the carbohydrate remain unclear. The presence of a PIG-N homologue suggests the core may have an additional phosphatidylethanolamine added to the first mannose of the GPI anchors, as is the case with mammalian and yeast cells, though not in Trypanosomes.[101]

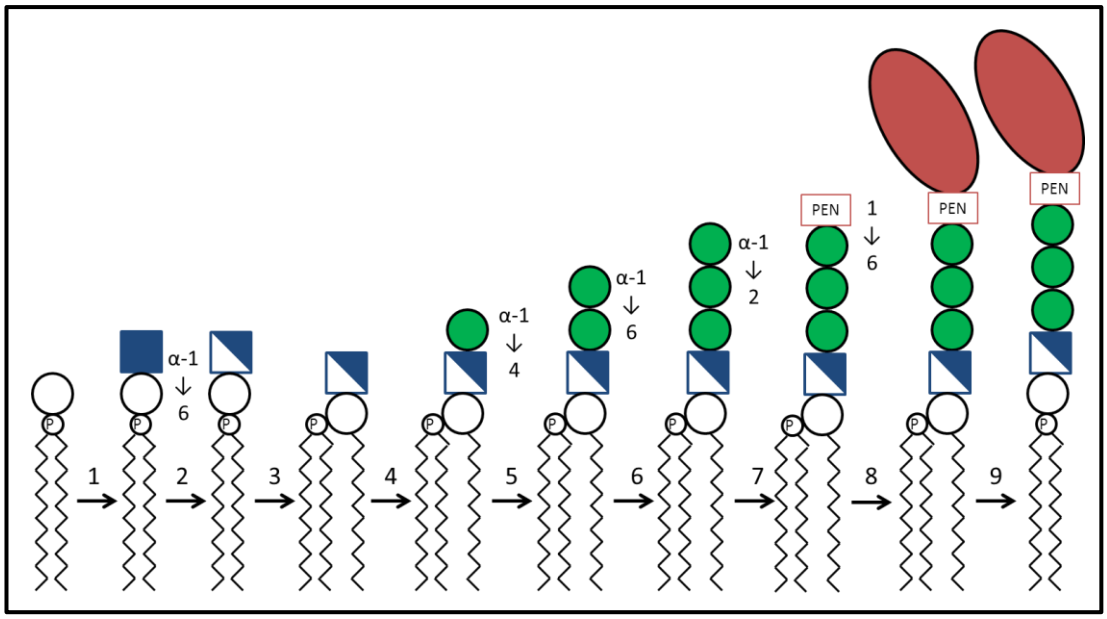


Figure 5.24: Biosynthesis of the glycophosphatidylinositol anchor

The core GPI anchor is synthesised in an ordered manner. 1. GlcNAc is added to the phosphatidylinositol, catalysed by a complex of PIG-A, C, H, P, Q, Y and DPM2. 2. The GlcNAc acetate is removed by PIG-L, leaving GlcN. 3. An acyl chain is attached to the inositol by PIG-W. 4. Man is then transferred on to the GlcN from dolichol-phosphate mannose by PIG-M. 5. Another Man is added to the first Man by PIG-V. 6. Another Man is added to the second Man by PIG-B. 7. A phosphatidylethanolamine (PEN) is added to the final Man. 8. A protein is attached to the distal PEN by a transamidase composed of PIG-K, S, T, U and GAA1. 9. The acyl chain is cleaved from the inositol by PGAP1. Zig-zag: acyl chain. P: phosphate. White circle: inositol. Blue square: GlcNAc. Diagonal blue square: GlcN. Green circle: Man. Red oval: protein.

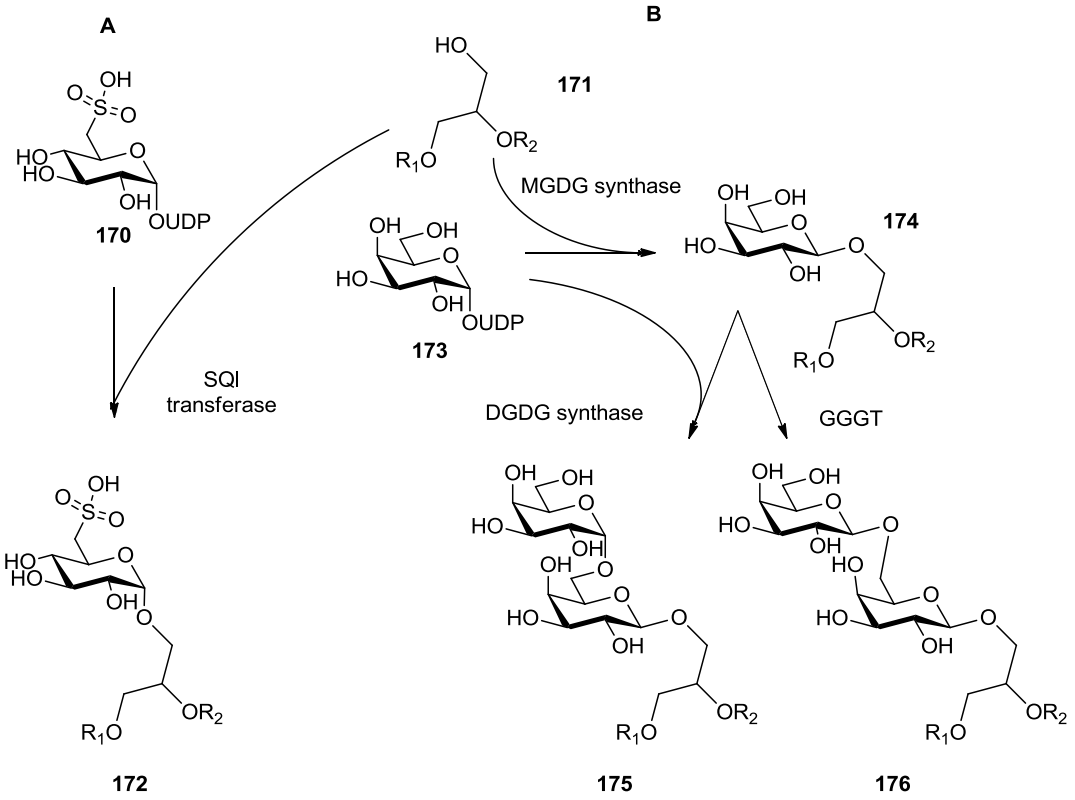


Figure 5.25: Biosynthesis of plastid glycolipids

A. Sulfoquinovosyl diacylglycerols (**172**) are formed by the transfer of SQ from UDP-SQ (**170**) on to diacylglycerol (**171**) by SQ transferase. Galactolipids β-MGDG (**174**) and αβ-DGDG (**175)** are formed by the transfer of galactose from UDP-Gal (**173**) on to diacylglycerol (**175**), by MGDG synthase, and then extended by the transfer of another Gal from UDP-Gal or, in the case of ββ-DGDG (**176**), by transfer from MGDG. R_1/R_2 = acyl chains

5.5.6.2 Plastid glycolipids

The thylakoid membranes of the chloroplasts, including those in *Euglena*,[102] contain galactosyl glycerides and sulfolipids,[103] which function to keep the membranes highly fluid (Fig 5.25). The sulfolipids appear to be present in *Euglena* cells grown in the dark,[104] suggesting a role in protoplasts. Sulfolipids are synthesised by the transfer of sulfoquinovose (SQ) from UDP-6-deoxy-6-sulfo-glucose on to diacylglycerides by SQ transferases, in family GT4.[65] There is a sequence for one homologue of this enzyme (lm.58009), related to bacterial isoforms, present in both the light and the dark transcriptomes. One sulfoquinovosyl transferase transcript (lm.20804) is related to the spikemoss enzyme and in the dark the transcript encodes the active domain with a 200 amino acids truncation of the N-terminus. The other transcript (lm. 33890), related to diatom homologues, is not present in the dark cells.

In plants β-monogalactosyldiacylglycerol (MGDG) is formed by the transfer of galactose from UDP-galactose on to diacyl glycerol by MGDG synthase, in family GT28.[66] αβ-Digalactosyldiacylglycerol (DGDG) is formed by transfer of another galactosyl moiety from UDP-galactose, by DGDG synthase (GT4) (Fig 5.24). Alternatively, galactose can be transfered from another MGDG, by galactolipid:galactolipid galactosyl transferase (GGGT), on to the 6'-OH of a molecule of MGDG to form β,β-DGDG.[105] Oligogalactosyl glycerides can also be formed by a GGGT homologue in *Arabidopsis*, although this enzyme has not yet been specifically identified.[66] The galactosyl glycerides are only present in *Euglena* cells that have photosynthetic plastids.[106] There is one MGDG synthase (dm.42760) in the *Euglena* transcriptome, but rather surprisingly this is only seen in the dark grown cells. There are two isoforms of DGDG synthase (lm84277 and lm.68525), the latter of which is only present in the light sequences.

Activity	EC no.	Transcripts		
dolichyldiphosphatase	3.6.1.43	lm.53772	dm.71782	
dolichol kinase	2.7.1.108	lm.96623		
UDP-GlcNAc-adolichol phosphate GlcNAc-1-p-transferase	2.7.8.15	lm.96341		
beta-1,4-N-acetylglucosaminyl transferase	2.4.1.141	lm.87840		
beta-1,4-mannosyl transferase	2.4.1.142	lm.92144		
alpha-1,3/alpha-1,6-mannosyltransferase	2.4.1.132	lm.79157		
alpha-1,2-mannosyltransferase	2.4.1.131	lm.67740		
Flippase		lm.83408		
alpha-1,3-mannosyltransferase	2.4.1.258	lm.68532		
alpha-1,2-mannosyltransferase	2.4.1.259/ 261	dm.60522	lm.71029	dm.60521
alpha-1,6-mannosyltransferase	2.4.1.260	lm.71029		
dolichol-phosphate mannosyltransferase	2.4.1.83	lm.31276		
dolichyl-phosphate beta-glucosyltransferase	2.4.1.117	lm.48352		
alpha-1,3-glucosyltransferase	2.4.1.267	lm.99883	dm.96080	
alpha-1,3-glucosyltransferase	2.4.1.265	lm.99883	dm.96080	
alpha-1,2-glucosyltransferase	2.4.1.256	lm.100691		
dolichyl-diphosphooligosaccharide-protein glycosyltransferase	2.4.99.18	dm.41728	lm.68366	dm.17756

Table 5.5: Annotation of enzymes involved in N-glycan biosynthesis

Euglena transcripts were identified which encode enzymes involved in N-glycan biosynthesis by similarity with known enzymes.[107] Highlighted in yellow are transcripts only present in the dark grown cells and highlighted in green are those present only in the light grown cells.

5.5.7 Protein glycosylation

5.5.7.1 N-Glycans

Protein glycosylation is one of the hallmarks of eukaryotic cells, but has also been found in bacteria, particularly pathogens, such as *Campylobacter*.[108] The biosynthesis of asparagine-linked glycans (N-glycans) proceeds by the synthesis of a lipid linked oligosaccharide precursor, which is transferred *en bloc* to proteins in the lumen of the endoplasmic reticulum (Fig 5.26.A).[109] *Euglena* is known to make the same precursor oligosaccharide as animals and fungi ($Glc_3Man_9GlcNAc_2Asn$),[110] whilst the related *Trypanosomes* have lost the ability to add the distal glucose residues to the precursor and *Leishmania* add two fewer mannose residues (Fig 5.26.B).[111] In *Euglena* the flagella-associated glycoproteins are also affected by tunicamycin, a known inhibitor of protein glycosylation,[112] but the exact structure of the glycans and the identity of glycosylated proteins still need to be established. Sequences for all of the enzymes required for the synthesis of the full N-glycan precursor are present in the *Euglena* transcriptome (Table 5.5). There are also three sequences for the GT66 oligosaccharyl transferases, which in *Trypanosomes* each show defined substrate specificities.[113]

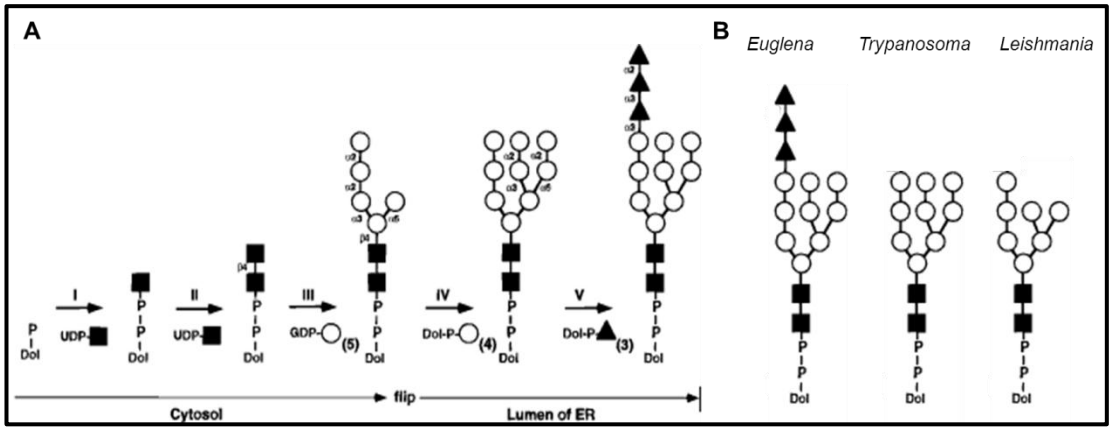


Figure 5.26: Biosynthesis of dolichol-linked oligosaccharide N-glycan precursor

A. The oligosaccharide precursor is synthesised on a dolichol pyrophosphate precursor. The first steps take place on the cytosolic face of the endoplasmic reticulum, using sugar nucleotides as the activated sugar donors. The $Man_5GlcNAc_2$ precursor is flipped into the lumen of the ER and the synthesis proceeds using dolichol phosphates as the sugar donors. The entire oligosaccharide is transferred *en bloc* to the protein. **B**. *Euglena* makes the full precursor, but the well characterised Euglenozoa do not. *Trypanosoma* does not add the final Glc residues and *Leishmania* does not add the final Man residues either. Black square: GlcNAc. White circle: Man. Black triangle: Glc. Dol: dolichol. P: phosphate. Reprinted from *Essentials of Glycobiology* with permission from CSHLP. Copyright (1999).[114]

5.5.7.2 O-GlcNAc

O-GlcNAc is involved in cell signalling,[115] including cross-talk with protein phosphorylation.[116] There are three putative GT41 O-GlcNAc transferases in the transcriptome, (lm.52466, dm.35031 and lm.92993) the latter of which is only present in the light grown cells. In humans there is only one transferase, whilst plants have two distinct enzymes.[117] There are however no homologues for the O-GlcNAcase encoded in the *Euglena* transcriptome, or in plants, which would reverse this signal, suggesting a non-canonical hydrolase carries out this reaction.

5.5.7.3 Sialic acids

Sialic acids are commonly found in animals at the ends of the glycans, and have roles including altering the cell surface charge and in specific cellular recognition.[107] They are also found in a range of pathogenic microorganisms, including parasitic protozoa[118] and bacteria,[119] which can synthesise or appropriate the sialic acid from the host.[120] *Euglena* has previously been reported to contain sialic acids, attached to glycolipids, based on immunostaining.[72] There is one transcript encoding N-acetylneuraminate pyruvate-lyase (dm.37271) which can reversibly add pyruvate to N-acetyl mannosamine to form N-acetyl neuraminic acid.[121] There are no obvious transcripts for enzymes which would covert this sialic acid to the usual CMP sugar-nucleotide, nor are there any candidate trans-sialidases for scavenging sialic acid from glycans in the environment. There are no obvious sialyl transferases in the *Euglena* transcriptome, such as GT12 or GT29. Whilst it is possible that this lyase is used in the degradation of exogenous sialic acid, it is also possible that *Euglena* uses a unique route for the addition of sialic acids to its glycans.

5.5.8 Xylose containing polysaccharides

It has been reported that *Euglena* has a xylose-containing protein associated with its flagella,[122] though neither the carbohydrate structure nor the protein to which it is attached have been elucidated. There are many enzymes in CAZy families with xylosyltransferase activities, which may represent good candidates for the biosynthesis of this flagella glycan.

5.5.8.1 GT61 family

In the *Euglena* transcriptome there are 19 members of the GT61 family of enzymes, for the transfer of α-xylose on to plant glycoproteins,[123] extra cellular O-GlcNAcylation in animals[124] and as probable candidates for arabinosylation of cereal xylans,[125] though this has yet to be shown biochemically. Many of the *Euglena* transcripts are related to the characterised members of GT61, but their precise activity is not clear (Fig 5.27). There are also many sequences which are only distantly related to this family and detailed biochemical analysis would be needed to either include or reject these enzymes from GT61.

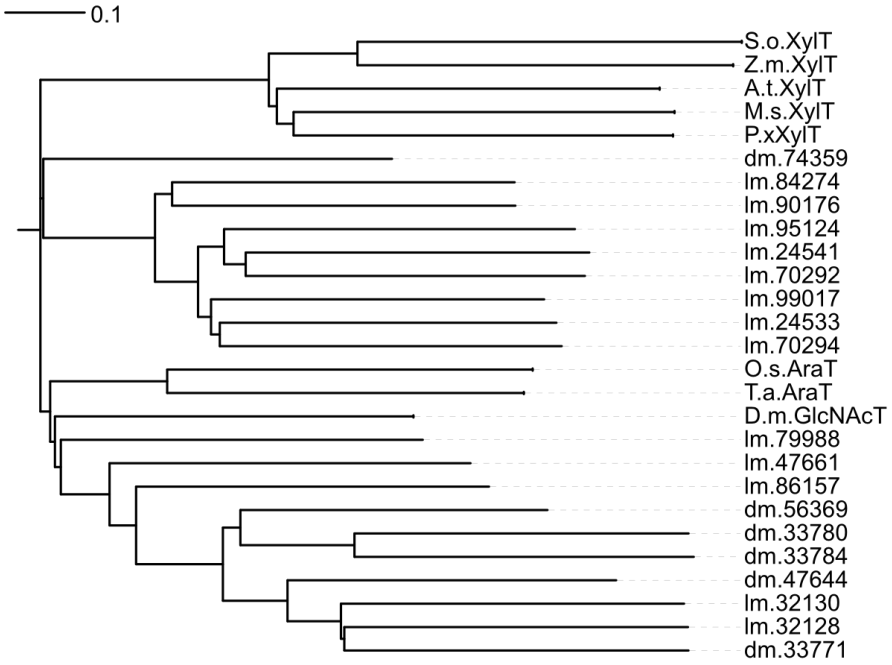


Figure 5.27 Phylogeny of family GT61 proteins

Members of family GT61 from *Euglena* and well characterised GT61 family proteins were aligned. S.o.: *Sachharum officinarum*. Z.m.: *Zea mays*. A.t.: *Arabidopsis thaliana*. M.s.: *Medicago sativa*. P.x: *Populus* hybrid. O.s.: *Oryza sativa*. T.a.: *Triticum aestivum*. D.m.: *Drosophila melanogaster*. GlcNAcT: N-acetyl glucosaminyltransferase. XylT: xylosyltransferase. AraT: arabinosyltransferase

5.5.8.2 GT77 family

Family GT77 is characterised as containing α-1,3-xylosyltransferase,[70] β-arabinosyltransferase and α-1,3-galactosyltransferase[126] activities. There are 23 sequences belonging to GT77 encoded in the *Euglena* transcriptome, which are likely to carry out these reactions, though it is not possible to predict their substrate specificity (Fig 5.28).

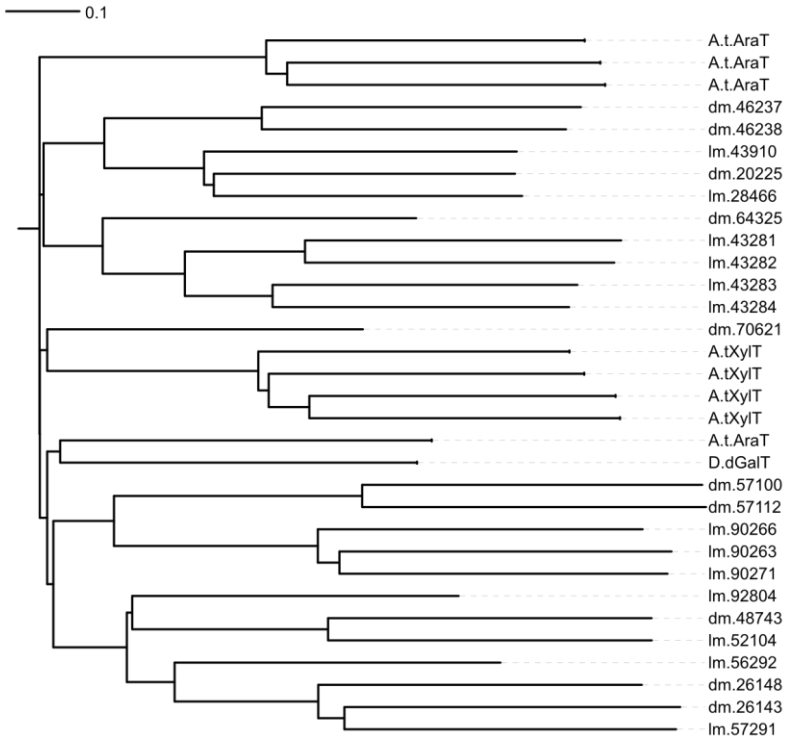


Figure 5.28: Phylogeny of family GT77 proteins

Members of family GT77 from *Euglena* and well characterised GT77 family proteins were aligned. A.t.: *Arabidopsis thaliana*. D.d.: *Dictyostelium discoideum*. AraT: arabinosyltransferase. XylT: xylosyltransferase. GalT: galactosyltransferase.

5.5.8.3 Other potential xylosyltransferases

Only one member of GT90 has been characterised, the xylosyltransferase of the *Cryptococcus* capsule.[127] There are two enzymes weakly related to GT90 encoded in the *Euglena* transcriptome, which may be involved in the biosynthesis of the surface xylose containing material. There is also an incomplete transcript of a member of GT34, predicted to be involved in xylosyl-[71] or galactosyl- transfer.[128]

5.5.8.4 Potential xylosyl hydrolases

There are three members of GH43 in the *Euglena* transcriptome which may be xylosidases[129] or arabinases.[130] There is also a single member of carbohydrate esterase family 3 (CE3), which are characterised as acetyl xylan esterases[131] and are involved in the degradation of plant cell wall acetylated polysaccharides.[132] Together this suggests *Euglena* has the capability of synthesising xylan-like material and may be able to degrade plant hemicellulose-related xylan.

5.5.9 Didomain CAZys

Whilst fusion of CBMs to other CAZYs is relatively common in nature, and contiguity of multiple glycoside hydrolase domains in a single protein is well known, it is much rarer to find glycosyl transferases as part of a protein containing other domains with carbohydrate processing activity.[133] Examples include the sea-squirt *Oikopleura dioica*,[133] which encodes a cellulose synthase (GT2) and a β-glucan hydrolase (GH6), and the O-GlcNAc transferases which are found in most eukaryotes and some bacteria, and encode tetratricopeptide repeats in addition to the GT41 GlcNAc-transferase domain.[134]

Two transcripts were identified in the *Euglena* transcriptome which encode proteins with two carbohydrate-active enzyme domains. The first protein (lm.71174) had a putative GT11 fucosyltransferase domain, and a putative GT15 mannosyltransferase domain (Fig 5.29). The active site of the former does not contain the second arginine in the HxRRxD motif,[135] whilst the latter contains the nucleophile and a zwitterionic ion-binding motif (see Appendix 6).[136] It is possible that this enzyme this enzyme maybe act to transfer both fucose and mannose to the same N-glycan core. Alternatively, this enzyme may transfer mannose on to a fucosylated glycan, to which it is directed by the GT11 domain, acting as a carbohydrate-binding module.

A second didomain protein (dm.47703) appears to contain a GT1 sugar transferase, most closely related to bacterial sterol β-glucuronic acid transferases, linked to a C-terminal GH78 α-rhamnosidase domain (Fig 5.29). Both domains appear to have an intact active site, suggesting both activities are viable (see Appendix 6).[137,138] In order to confirm the activity of this enzyme, synthetic gene constructs of the two domains were expressed in *E. coli*. Only the GH78 domain could be expressed successfully and was shown to be a rhamnosidase with kinetic parameters

consistent with other enzymes of this class (see Appendix 7). This didomain protein might conceivably be involved in cleaving rhamnose from a small molecule and adding a glucuronic acid moiety, a sugar addition that is known to facilitate sub-cellular relocalisation and drug detoxification.[139]

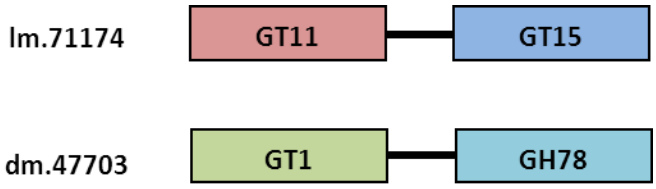


Figure 5.29 Domain architecture of didomain CAZys encoded in the *Euglena* transcriptome

lm.71174 encodes an N-terminal GT11 and a C-terminal GT15. dm.47703 encodes an N-terminal GT1 and a C-terminal GH78. See Appendix 6 for sequence alignments with well characterised members of these families.

5.5.1 Conclusions on carbohydrate metabolism in *Euglena*

The transcriptome of *Euglena* encodes a diverse array of carbohydrate active enzymes. We can predict the enzymes likely to be involved in synthesis and degradation of the β-glucan storage carbohydrate, based on similarity to enzymes used in plants and fungi for the synthesis of their cell walls. Validation is required, possibly by comparison of the proteins which make up the previously identified paramylon synthase complex[81] with the transcriptome presented here.

The *Euglena* transcriptome also encodes many enzymes involved in the synthesis and degradation of complex plant cell wall type carbohydrates, including galactosidases, xylosidases and rhamnosidases, suggesting they are capable of obtaining sugars from complex plant material. *Euglena* are not noted as having a rigid polysaccharide cell wall, and instead have a flexible proteinaceous coat. However, they do have xylan associated with their flagella[122] and the transcriptome shows that all the machinery for N-glycans and GPI anchored proteins is encoded in *Euglena*. This suggests that the surface of *Euglena* is coated in carbohydrates.

5.6 Conclusions on laminarin phosphorylase

The sequencing of *Euglena gracilis* has provided a wealth of information but has not yielded any stand-out candidates for the laminarin phosphorylases in the obvious CAZy families. However partial purification of the laminarin phosphorylase revealed two members of GH81 family endo-β-1,3-glucan hydrolases. Whilst this family is not known to contain any phosphorylase activity, work is on-going to characterise the activity of these proteins. Laminarin phosphorylases have great potential for biotechnology, both for the synthesis of immunomodulating β-glucans, and for the synthesis of less accessible glycans, such as β-Glc-gol. It is hoped that identification and further characterisation of these enzymes will allow their full potential to be realised.

Euglena are known to have unusual carbohydrate metabolism and the transcriptome data presented here provides a starting point for an in-depth study of the glycobiology of this class of alga. The complex N-glycan structures need to be elucidated but it is possible that this organism could be used to synthesise therapeutic human glycoproteins. Understanding the biosynthesis and degradation of the unusual storage polysaccharide, paramylon, is also of interest to fundamental biology and the carbohydrate active enzymes provides a basis for this research.

5.7 References

1 Bacic, A., Fincher, G. B. & Stone, B. A. *Chemistry, biochemistry, and biology of (1-3)-β-glucans and related polysaccharides*. (Elsevier Science, 2009).

2 Clarke, A. E. & Stone, B. A. Structure of the paramylon from *Euglena gracilis*. *Biochim. Biophys. Acta* **44**, 161-163 (1960).

3 Harada, T., Misaki, A. & Saito, H. Curdlan: a bacterial gel-forming β-1,3-glucan. *Arch. Biochem. Biophys.* **124**, 292-298 (1968).

4 Aspinall, G. O. & Carpenter, R. C. Structural investigations on the non-starchy polysaccharides of oat bran. *Carbohydr. Polym.* **4**, 271-282 (1984).

5 Anderson, F. B., Hirst, E. L., Manners, D. J. & Ross, A. G. The constitution of laminarin. Part III. The fine structure of insoluble laminarin. *J. Chem. Soc.*, 3233-3243 (1958).

6 Adeyeye, A., Jansson, P.-E., Lindberg, B. & Henrichsen, J. Structural studies of the capsular polysaccharide from *Streptococcus pneumoniae* type 37. *Carbohydr. Res.* **180**, 295-299 (1988).

7 Vandamme, E. J., De Baets, S. & Steinbüchel, A. *Polysaccharides II: polysaccharides from eukaryotes*. Vol. 6 (Wiley-VCH, 2002).

8 Huwyler, H. R., Franz, G. & Meier, H. β-1,3-Glucans in the cell walls of cotton fibres (*Gossypium arboreum* L.). *Plant Sci. Lett.* **12**, 55-62 (1978).

9 Painter, T. J. Structural evolution of glycans in algae. *Pure Appl. Chem.* **55**, 677-694 (1983).

10 Monfils, A. K., Triemer, R. E. & Bellairs, E. F. Characterization of paramylon morphological diversity in photosynthetic euglenoids (Euglenales, Euglenophyta). *Phycologia* **50**, 156-169 (2011).

11 Kiss, J. Z., Roberts, E. M., Brown, R. M., Jr. & Triemer, R. E. X-ray and dissolution studies of paramylon storage granules from *Euglena*. *Protoplasma* **146**, 150-156 (1988).

12 Kitaoka, M., Matsuoka, Y., Mori, K., Nishimoto, M. & Hayashi, K. Characterization of a bacterial laminaribiose phosphorylase. *Biosci., Biotechnol., Biochem.* **76**, 343-348 (2012).

13 Nihira, T. *et al.* Characterization of a laminaribiose phosphorylase from *Acholeplasma laidlawii* PG-8A and production of 1,3-β-D-glucosyl disaccharides. *Carbohydr. Res.* **361**, 49-54 (2012).

14 Marechal, L. R. β-1,3-Oligoglucan - orthophosphate glucosyltransferases from *Euglena gracilis*. 2. Comparative studies between laminaribiose- and β-1,3-oligoglucan phosphorylase. *Biochim. Biophys. Acta* **146**, 417-443 (1967).

15 Kitaoka, M., Sasaki, T. & Taniguchi, H. Purification and properties of laminaribiose phosphorylase (EC-2.4.1.31) from *Euglena-gracilis* Z. *Arch. Biochem. Biophys.* **304**, 508-514 (1993).

16 Manners, D. J. & Taylor, D. C. Acceptor specificity of laminaribiose phosphorylase from *Astasia ocellata*. *Biochem. J.* **94**, 17-19 (1965).

17 Klarzynski, O. *et al.* Linear β-1,3 glucans are elicitors of defense responses in tobacco. *Plant Physiol.* **124**, 1027-1038 (2000).

18 Gururaj, H. B., Giridhar, P. & Ravishankar, G. A. Laminarin as a potential non-conventional elicitor for enhancement of capsaicinoid metabolites. *Asian J. Plant Sci. Res.* **2**, 490-495 (2012).

19 Meszka, B. & Bielenin, A. Activity of laminarin in control of strawberry diseases. *Phytopathologia*, 15-23 (2011).

20 Nakao, Y. *et al.* Curdlan - properties and application to foods. *J. Food Sci.* **56**, 769-772 (1991).

21 Miwa, M., Nakao, Y. & Nara, K. in *Food Hydrocolloid.* (eds Katsuyoshi Nishinari & Etsushiro Doi) Ch. 14, 119-124 (Springer US, 1993).

22 Zhan, X.-B., Lin, C.-C. & Zhang, H.-T. Recent advances in curdlan biosynthesis, biotechnological production, and applications. *Appl. Microbiol. Biotechnol.* **93**, 525-531 (2012).

23 Kulicke, W.-M., Lettau, A. I. & Thielking, H. Correlation between immunological activity, molar mass, and molecular structure of different (1->3)-β-D-glucans. *Carbohydr. Res.* **297**, 135-143 (1997).

24 Moradali, M.-F., Mostafavi, H., Ghods, S. & Hedjaroude, G.-A. Immunomodulating and anticancer agents in the realm of macromycetes fungi (macrofungi). *Int. Immunopharmacol.* **7**, 701-724 (2007).

25 Kondo, Y. *et al.* Cytokine-related immunopotentiating activities of paramylon, a β-(1-3)-D-glucan from *Euglena gracilis*. *J. Pharmacobio-Dynam.* **15**, 617-621 (1992).

26 Chan, G., Chan, W. & Sze, D. The effects of β-glucan on human immune and cancer cells. *J. Hematol. Oncol.* **2**, 25 (2009).

27 Robbins, E. A. & Seeley, R. D. Cholesterol lowering effect of dietary yeast and yeast fractions. *J. Food Sci.* **42**, 694-698 (1977).

28 Koizumi, N. *et al.* Anti-HIV (human immunodeficiency virus) activity of sulfated paramylon. *Antiviral Res.* **21**, 1-14 (1993).

29 Franz, G. & Alban, S. Structure-activity relationship of antithrombotic polysaccharide derivatives. *Int. J. Biol. Macromol.* **17**, 311-314 (1995).

30 Deslandes, Y., Marchessault, R. H. & Sarko, A. Triple-helical structure of (1->3)-β-D-glucan. *Macromolecules* **13**, 1466-1471 (1980).

31 Fukuhara, G. & Inoue, Y. Oligosaccharide sensing with chromophore-modified curdlan in aqueous media. *Chem. Commun.* **46**, 9128-9130 (2010).

32 Koumoto, K., Kimura, T., Kobayashi, H., Sakurai, K. & Shinkai, S. Chemical modification of curdlan to induce an interaction with poly(C)(1). *Chem. Lett.*, 908-909 (2001).

33 Numata, M., Koumoto, K., Mizu, M., Sakurai, K. & Shinkai, S. Parallel vs. anti-parallel orientation in a curdlan/oligo(dA) complex as estimated by a FRET technique. *Org. Biomol. Chem.* **3**, 2255-2261 (2005).

34 Fukamizo, T. *et al.* Enzymatic hydrolysis of 1,3-1,4-β-glucosyl oligosaccharides by 1,3-1,4-β-glucanase from *Synechocystis* PCC6803: a comparison with assays using polymer and chromophoric oligosaccharide substrates. *Arch. Biochem. Biophys.* **478**, 187-194 (2008).

35 Rismani-Yazdi, H., Haznedaroglu, B. Z., Bibby, K. & Peccia, J. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics* **12** (2011).

36 De Groeve, M. R. M. *et al.* Development and application of a screening assay for glycoside phosphorylases. *Anal. Biochem.* **401**, 162-167 (2010).

37 Miyuki, K. Lilioside a from *Lilium longiflorum*: synthesis and absolute configuration. *Phytochemistry* **29**, 3559-3564 (1990).

38 Pocard, J. A., Smith, L. T., Smith, G. M. & Le Rudulier, D. A prominent role for glucosylglycerol in the adaptation of *Pseudomonas mendocina* SKB70 to osmotic stress. *J. Bacteriol.* **176**, 6877-6884 (1994).

39 Hincha, D. K. & Hagemann, M. Stabilization of model membranes during drying by compatible solutes involved in the stress tolerance of plants and microorganisms. *Biochem. J.* **383**, 277-283 (2004).

40 Takenaka, F. & Uchiyama, H. Synthesis of α-D-glucosylglycerol by α-glucosidase and some of its characteristics. *Biosci., Biotechnol., Biochem.* **64**, 1821-1826 (2000).

41 Nakano, H. *et al.* Synthesis of glycosyl glycerol by cyclodextrin glucanotransferases. *J. Biosci. Bioeng.* **95**, 583-588 (2003).

42 Takenaka, F. & Uchiyama, H. Effects of α-D-glucosylglycerol on the in vitro digestion of disaccharides by rat intestinal enzymes. *Biosci., Biotechnol., Biochem.* **65**, 1458-1463 (2001).

43 Kaneda, M., Mizutani, K., Takahashi, Y., Kurono, G. & Nishikawa, Y. Lilioside A and B, two new glycerol glucosides isolated from *Lilium longiflorum* thunb. *Tetrahedron Lett.* **15**, 3937-3940 (1974).

44 Hagemann, M. & Erdmann, N. Activation and pathway of glucosylglycerol synthesis in the cyanobacterium *Synechocystis* sp. PCC 6803. *Microbiology* **140**, 1427-1431 (1994).

45 Ohta, N. & Achiwa, K. Synthesis of biologically active galactosyl and glucosyl-glycerol derivatives. *Chemcial & Pharmaceutical Bulletin* **39**, 1337-1339 (1991).

46 Goedl, C., Sawangwan, T., Mueller, M., Schwarz, A. & Nidetzky, B. A high-yielding biocatalytic process for the production of 2-O-(α-D-glucopyranosyl)-sn-glycerol, a natural osmolyte and useful moisturizing ingredient. *Angew. Chem., Int. Ed.* **47**, 10086-10089 (2008).

47 Goedl, C., Sawangwan, T., Nidetzky, B. & Muller, M. Method for producing 2-O-glyceryl-α-D-glucopyranoside. US patent (2009).

48 Colombo, D. *et al.* 1-O-, 2-O- and 3-O-β-glycosyl-sn-glycerols: structure-anti-tumor-promoting activity relationship. *Bioorg. Med. Chem. Lett.* **6**, 1187-1190 (1996).

49 Luley-Goedl, C. & Nidetzky, B. Glycosides as compatible solutes: biosynthesis and applications. *Nat. Prod. Rep.* **28**, 875-896 (2011).

50 De Roode, M. *et al.* Optimization of production and downstream processing of the almond β-glucosidase-mediated glucosylation of glycerol. *Biotechnol. Bioeng.* **72**, 568-572 (2001).

51 Albini, F. M., Murelli, C., Patritti, G. & Rovati, M. A simple synthesis of glucosyl glycerols. *Synth. Commun.* **24**, 1651 - 1661 (1994).

52 Helferich, B. & Zirner, J. Zur Synthese von Tetraacetyl-hexosen mit freiem 2-Hydroxyl. Synthese einiger Disaccharide. *Chem. Ber. Recl.* **95**, 2604-2611 (1962).

53 Malkoff, D. B. & Buetow, D. E. Ultrastructural changes during carbon starvation in *Euglena gracilis*. *Experimental Cell Research* **35**, 58-& (1964).

54 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-U130 (2011).

55 Simpson, A. G. B. & Roger, A. J. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Mol. Phylogenet. Evol.* **30**, 201-212 (2004).

56 Danpure, C. J. How can the products of a single gene be localized to more than one intracellular compartment? *Trends Cell Biol.* **5**, 230-238 (1995).

57 Freitag, J., Ast, J. & Bolker, M. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature* **485**, 522-525 (2012).

58 Gebhardt, J. S., Wadsworth, G. J. & Matthews, B. F. Characterization of a single soybean cDNA encoding cytosolic and glyoxysomal isozymes of aspartate aminotransferase. *Plant Mol. Biol.* **37**, 99-108 (1998).

59 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-D238 (2009).

60 Mehlert, A., Zitzmann, N., Richardson, J. M., Treumann, A. & Ferguson, M. A. J. The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of *Trypanosoma brucei*. *Mol. Biochem. Parasit.* **91**, 145-152 (1998).

61 Belocopitow, E. & Maréchal, L. R. Trehalose phosphorylase from *Euglena gracilis*. *Biochim. Biophys. Acta - Enzymology* **198**, 151-154 (1970).

62 Marechal, L. R. & Goldemberg, S. H. Laminaribiose phosphorylase from *Euglena gracilis*. *Biochem. Biophys. Res. Commun.* **13**, 106-107 (1963).

63 Chaen, H. *et al.* Purification and characterization of thermostable trehalose phosphorylase from *Theymoanaeyobium brockii*. *J. Applied Glycosci.* **46**, 399-405 (1999).

64 Martín-Cuadrado, A.-B. *et al.* Characterization of the endo-β-1,3-glucanase activity of *S. cerevisiae* Eng2 and other members of the GH81 family. *Fungal Genet. Biol* **45**, 542-553 (2008).

65 Shimojima, M. Biosynthesis and functions of the plant sulfolipid. *Prog. Lipid Res.* **50**, 234-239 (2011).

66 Benning, C. & Ohta, H. Three enzyme systems for galactoglycerolipid biosynthesis are coordinately regulated in plants. *J. Biol. Chem.* **280**, 2397-2400 (2005).

67 Ryan, K. A., Garraway, L. A., Descoteaux, A., Turco, S. J. & Beverley, S. M. Isolation of virulence genes directing surface glycosyl-phosphatidylinositol synthesis by functional complementation of *Leishmania*. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 8609-8613 (1993).

68 De Lederkremer, R. M. & Colli, W. Galactofuranose-containing glycoconjugates in Trypanosomatids. *Glycobiology* **5**, 547-552 (1995).

69 Carlsson, P. & Kjellén, L. in *Heparin - A Century of Progress* Vol. 207 *Handbook of Experimental Pharmacology* (eds Rebecca Lever, Barbara Mulloy, & Clive P. Page) Ch. 2, 23-41 (Springer Berlin Heidelberg, 2012).

70 Egelund, J. *et al. Arabidopsis thaliana* RGXT1 and RGXT2 encode Golgi-localized (1,3)-α-D-xylosyltransferases involved in the synthesis of pectic rhamnogalacturonan-II. *Plant Cell* **18**, 2593-2607 (2006).

71 Faik, A., Price, N. J., Raikhel, N. V. & Keegstra, K. An Arabidopsis gene encoding an α-xylosyltransferase involved in xyloglucan biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7797-7802 (2002).

72 Preisfeld, A. & Ruppel, H. G. Detection of sialic acid and glycosphingolipids in *Euglena gracilis* (Euglenozoa). *Arch. Protistenkd.* **145**, 251-260 (1995).

73 Buschiazzo, A. & Alzari, P. M. Structural insights into sialic acid enzymology. *Curr. Opin. Chem. Biol.* **12**, 565-572 (2008).

74 Rebuffet, E. *et al.* Discovery and structural characterization of a novel glycosidase family of marine origin. *Environ. Microbiol.* **13**, 1253-1270 (2011).

75 Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769-781 (2004).

76 Gao, G. *et al.* Non-catalytic β- and γ-subunit isoforms of the 5′-AMP-activated protein kinase. *J. Biol. Chem.* **271**, 8675-8681 (1996).

77 Stapleton, D. *et al.* Mammalian 5'-AMP-activated protein kinase non-catalytic subunits are homologs of proteins that interact with yeast Snf1 protein kinase. *J. Biol. Chem.* **269**, 29343-29346 (1994).

78 Cosgrove, D. J. Loosening of plant cell walls by expansins. *Nature* **407**, 321-326 (2000).

79 Kerff, F. *et al.* Crystal structure and activity of *Bacillus subtilis* YoaJ (EXLX1), a bacterial expansin that promotes root colonization. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16876-16881 (2008).

80 Schallus, T. *et al.* Malectin: a novel carbohydrate-binding protein of the endoplasmic reticulum and a candidate player in the early steps of protein N-glycosylation. *Mol. Biol. Cell* **19**, 3404-3414 (2008).

81 Baumer, D., Preisfeld, A. & Ruppel, H. G. Isolation and characterization of paramylon synthase from *Euglena gracilis* (Euglenophyceae). *J. Phycol.* **37**, 38-46 (2001).

82 Tomos, A. D. & Northcote, D. H. Protein-glucan intermediate during paramylon synthesis. *Biochem. J.* **174**, 283-290 (1978).

83 Verma, D. & Hong, Z. Plant callose synthase complexes. *Plant Mol. Biol.* **47**, 693-701 (2001).

84 Bowman, S. M. & Free, S. J. The structure and synthesis of the fungal cell wall. *BioEssays* **28**, 799-808 (2006).

85 Singh, D. G. *et al.* β-Glucosylarginine: a new glucose-protein bond in a self-glucosylating protein from sweet corn. *FEBS Lett.* **376**, 61-64 (1995).

86 Vogel, K. & Barber, A. A. Degradation of paramylon by *Euglena gracilis*. *J. Protozool* **15**, 657-662 (1968).

87 Baladron, V. *et al.* Eng1p, an endo-1,3-β-glucanase localized at the daughter side of the septum, is involved in cell separation in Saccharomyces cerevisiae. *Eukaryot. Cell* **1**, 774-786 (2002).

88 Jondle, D. J., Coors, J. G. & Duke, S. H. Maize leaf β-1,3-glucanase activity in relation to resistance to *Exserohilum turcicum*. *Canadian Journal of Botany-Revue Canadienne De Botanique* **67**, 263-266 (1989).

89 Nishimura, T. *et al. Streptomyces matensis* laminaripentaose hydrolase is an 'inverting' β-1,3-glucanase. *FEBS Lett.* **499**, 187-190 (2001).

90 Withers, S. G. (CAZypedia, acessed 2013).

91 Withers, S. G. (CAZypedia, acessed 2013).

92 Harvey, A. J., Hrmova, M., De Gori, R., Varghese, J. N. & Fincher, G. B. Comparative modeling of the three-dimensional structures of family 3 glycoside hydrolases. *Proteins: Struct., Funct., Bioinf.* **41**, 257-269 (2000).

93 Hrmova, M. & Fincher, G. B. Barley β-D-glucan exohydrolases. Substrate specificity and kinetic properties. *Carbohydr. Res.* **305**, 209-221 (1997).

94 Aspeborg, H., Coutinho, P., Wang, Y., Brumer, H. & Henrissat, B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* **12**, 186 (2012).

95 St John, F. J., González, J. M. & Pozharski, E. Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups. *FEBS Lett.* **584**, 4435-4441 (2010).

96 Brunner, F. *et al.* A β-glucosidase/xylosidase from the phytopathogenic oomycete, *Phytophthora infestans*. *Phytochemistry* **59**, 689-696 (2002).

97 Pitson, S. M., Seviour, R. J., McDougall, B. M., Woodward, J. R. & Stone, B. A. Purification and characterization of 3 extracellular (1-3)-β-D-glucan glucohydrolases from the filamentous fungus *Acremonium persicinum*. *Biochem. J.* **308**, 733-741 (1995).

98 Orlean, P. & Menon, A. K. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. *J. Lipid Res.* **48**, 993-1011 (2007).

99 Ruhela, D., Banerjee, P. & Vishwakarma, R. A. Chemistry and biology of glycosylphosphatidylinositol molecules. *Curr. Sci.* **102**, 194-211 (2012).

100 RayChaudhuri, A. *et al.* L-myo-Inositol 1-phosphate synthase from plant sources - characteristics of the chloroplastic and cytosolic enzymes. *Plant Physiol.* **115**, 727-736 (1997).

101 Hong, Y. *et al.* Pig-n, a mammalian homologue of yeast Mcd4p, is ynvolved in transferring phosphoethanolamine to the first mannose of the glycosylphosphatidylinositol. *J. Biol. Chem.* **274**, 35099-35106 (1999).

102 Rosenber, A. Galactosyl diglycerides - their possible function in *Euglena* chloroplasts. *Science* **157**, 1191-& (1967).

103 Webb, M. S. & Green, B. R. Biochemical and biophysical properties of thylakoid acyl lipids. *Biochim. Biophys. Acta - Bioenergetics* **1060**, 133-158 (1991).

104 Davies, W. H., Mercer, E. I. & Goodwin, T. W. The occurrence and intracellular distribution of the plant sulpholipid in maize, runner beans, plant tissue cultures and *Euglena gracilis*. *Phytochemistry* **4**, 741-749 (1965).

105 van Besouw, A. & Wintermans, J. F. G. M. Galactolipid formation in chloroplast envelopes: I. Evidence for two mechanisms in galactosylation. *Biochim. Biophys. Acta - Lipid. Lipid. Met.* **529**, 44-53 (1978).

106 Hulanicka, D., Bloch, K. & Erwin, J. Lipid metabolism of *Euglena gracilis*. *J. Biol. Chem.* **239**, 2778-& (1964).

107 Varki, A. & Schauer, R. in *Essentials of Glycobiology* (eds A. Varki *et al.*) Ch. 14, (Cold Spring Harbor Laboratory Press, 2009).

108 Longwell, S. A. & Dube, D. H. Deciphering the bacterial glycocode: recent advances in bacterial glycoproteomics. *Curr. Opin. Chem. Biol.* **17**, 41-48 (2013).

109 Burda, P. & Aebi, M. The dolichol pathway of N-linked glycosylation. *Biochim. Biophys. Acta - Gen. Subjects* **1426**, 239-257 (1999).

110 de la Canal, L. & Parodi, A. J. Glycosylation of proteins in the protozoan *Euglena gracilis*. *Comp. Biochem. Physiol., B: Comp. Biochem.* **81**, 803-805 (1985).

111 Samuelson, J. *et al.* The diversity of dolichol-linked precursors to Asn-linked glycans likely results from secondary loss of sets of glycosyltransferases. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1548-1553 (2005).

112 Geetha-Habib, M. & Bouck, G. B. Synthesis and mobilization of flagellar glycoproteins during regeneration in *Euglena*. *The Journal of Cell Biology* **93**, 432-441 (1982).

113 Izquierdo, L., Mehlert, A. & Ferguson, M. A. The lipid-linked oligosaccharide donor specificities of *Trypanosoma brucei* oligosaccharyltransferases. *Glycobiology* **22**, 696-703 (2012).

114 Marth, J. D. in *Essentials of Glycobiology* (eds A. Varki *et al.*) Ch. 7, (Cold Spring Harbor Laboratory Press, 1999).

115 Love, D. C. & Hanover, J. A. The Hexosamine Signaling Pathway: deciphering the "O-GlcNAc Code". *Sci. STKE* **2005**, re13- (2005).

116 Zeidan, Q. & Hart, G. W. The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. *J. Cell Sci.* **123**, 13-22 (2010).

117 Olszewski, N. E., West, C. M., Sassi, S. O. & Hartweck, L. M. O-GlcNAc protein modification in plants: evolution and function. *Biochim. Biophys. Acta - Gen. Subjects* **1800**, 49-56 (2010).

118 Chava, A. K., Bandyopadhyay, S., Chatterjee, M. & Mandal, C. Sialoglycans in protozoal diseases: their detection, modes of acquisition and emerging biological roles. *Glycoconjugate J.* **20**, 199-206 (2003).

119 Severi, E., Hood, D. W. & Thomas, G. H. Sialic acid utilization by bacterial pathogens. *Microbiology* **153**, 2817-2822 (2007).

120 Schenkman, S., Jiang, M.-S., Hart, G. W. & Nussenzweig, V. A novel cell surface trans-sialidase of *Trypanosoma cruz*i generates a stage-specific epitope required for invasion of mammalian cells. *Cell* **65**, 1117-1125 (1991).

121 Comb, D. G. & Roseman, S. The sialic acids: I. the structure and enzymatic synthesis of N-acetylneuraminic acid. *J. Biol. Chem.* **235**, 2529-2537 (1960).

122 Bouck, G. B., Rogalski, A. & Valaitis, A. Surface organization and composition of *Euglena*. 2. Flagellar mastigonemes. *J. Cell Biol.* **77**, 805-826 (1978).

123 Strasser, R. *et al.* Molecular cloning and functional expression of β-1,2-xylosyltransferase cDNA from *Arabidopsis thaliana*. *FEBS Lett.* **472**, 105-108 (2000).

124 Sakaidani, Y. *et al.* O-Linked-N-acetylglucosamine on extracellular protein domains mediates epithelial cell-matrix interactions. *Nat. Commun.* **2** (2011).

125 Mitchell, R. A. C., Dupree, P. & Shewry, P. R. A novel bioinformatics approach identifies candidate genes for the synthesis and feruloylation of arabinoxylan. *Plant Physiol.* **144**, 43-53 (2007).

126 Ercan, A. *et al.* Molecular characterization of a novel UDP-galactose:fucoside α-3-Galactosyltransferase that modifies Skp1 in the cytoplasm of *Dictyostelium*. *J. Biol. Chem.* **281**, 12713-12721 (2006).

127 Klutts, J. S., Levery, S. B. & Doering, T. L. A β-1,2-xylosyltransferase from *Cryptococcus neoformans* defines a new family of glycosyltransferases. *J. Biol. Chem.* **282**, 17890-17899 (2007).

128 Chappell, T. G., Hajibagheri, M. A., Ayscough, K., Pierce, M. & Warren, G. Localization of an α-1,2-galactosyltransferase activity to the Golgi apparatus of *Schizosaccharomyces pombe*. *Mol. Biol. Cell* **5**, 519-528 (1994).

129 Ito, T. *et al.* Xylosidases associated with the cell surface of *Penicillium herquei* IFO 4674. *J. Biosci. Bioeng.* **96**, 354-359 (2003).

130 Flipphi, M. J., Panneman, H., van der Veen, P., Visser, J. & de Graaff, L. H. Molecular cloning, expression and structure of the endo-1,5-α-L-arabinase gene of Aspergillus niger. *Appl. Microbiol. Biotechnol.* **40**, 318-326 (1993).

131 Correia, M. A. *et al.* Crystal structure of a cellulosomal family 3 carbohydrate esterase from *Clostridium thermocellum* provides insights into the mechanism of substrate recognition. *J. Mol. Biol.* **379**, 64-72 (2008).

132 Zhang, J. X., Martin, J. & Flint, H. J. Identification of non-catalytic conserved regions in xylanases encoded by the xynB and xynD genes of the cellulolytic rumen anaerobe *Ruminococcus flavefaciens*. *Mol. Gen. Genet.* **245**, 260-264 (1994).

133 Medie, F. M., Davies, G. J., Drancourt, M. & Henrissat, B. Genome analyses highlight the different biological roles of cellulases. *Nat. Rev. Microbiol.* **10**, 227-U (2012).

134 Lubas, W. A., Frank, D. W., Krause, M. & Hanover, J. A. O-linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *J. Biol. Chem.* **272**, 9316-9324 (1997).

135 Takahashi, T. *et al.* A sequence motif involved in the donor substrate binding by α-1,6-fucosyltransferase: the role of the conserved arginine residues. *Glycobiology* **10**, 503-510 (2000).

136 Lobsanov, Y. D. *et al.* Structure of Kre2p/Mnt1p: a yeast α-1,2-mannosyltransferase involved in mannoprotein biosynthesis. *J. Biol. Chem.* **279**, 17921-17931 (2004).

137 Mulichak, A. M., Lu, W., Losey, H. C., Walsh, C. T. & Garavito, R. M. Crystal atructure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: interactions with acceptor and nucleotide ligands. *Biochemistry* **43**, 5170-5180 (2004).

138 Cui, Z., Maruyama, Y., Mikami, B., Hashimoto, W. & Murata, K. Crystal structure of glycoside hydrolase family 78 α-L-rhamnosidase from *Bacillus* sp. GL1. *J. Mol. Biol.* **374**, 384-398 (2007).

139 Tukey, R. H. & Strassburg, C. P. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu. Rev. Pharmacool. Toxicol.* **40**, 581-616 (2000).

Chapter 6 – Further details on the *Euglena* transcriptome

6.1 Introduction

6.1.1 *Euglena gracilis*

Euglena are common Eukaryotic algae, especially in organic rich freshwaters where they can be so populous as to give their characteristic colour to the water, such as the verdant green *Euglena viridis* or blood red *Euglena sanguinea*. *Euglena* were first recorded[1] by van Leeuwenhoek in his 1674 letter to the Royal Society,[2] although Harris is usually credited with the first description.[3] They have a characteristic shape and size, with a highly flexible cell surface (Fig 6.1).

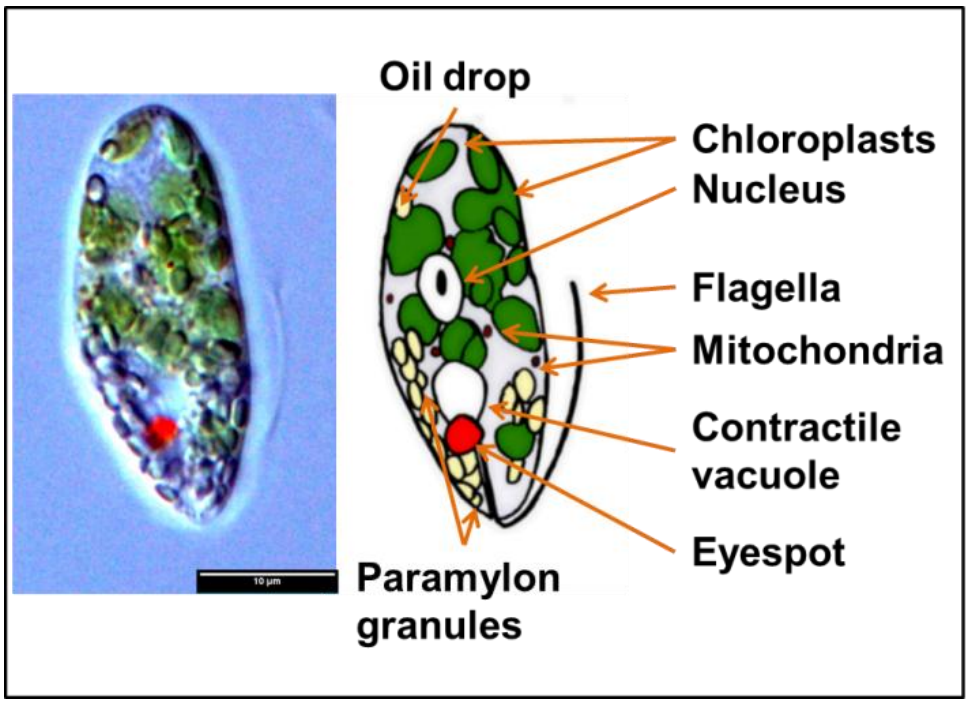


Figure 6.1: Physiology of *Euglena gracilis*

The Euglenoids are algae related to the pathogenic protozoa *Trypanosoma* and *Leishmania* (Fig 6.2) and are extremely difficult to classify, even using modern molecular techniques.[4] *Euglena* exhibit a highly flexible lifestyle, switching between photosynthesis, absorbing nutrients and eating other eukaryotes and bacteria.

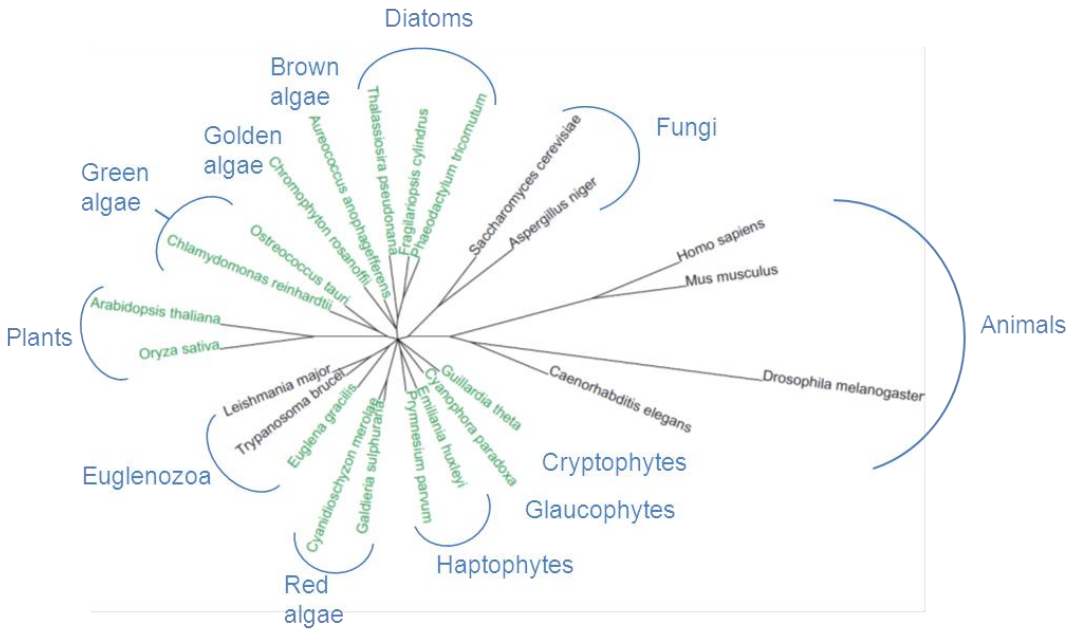


Figure 6.2: Phylogeny of *Euglena gracilis*

Unrooted tree showing *Euglena gracilis* in relation to sequenced algae and model organisms.[5] Organisms in green are photosynthetic.

Euglena are readily cultured and have been exploited in diverse fields including as a reporter for vitamin B12 production,[6] to study the ecotoxicity of zinc oxide nanoparticles,[7] in a neurocomputer based on their motion in a micro aquarium[8] and for the intracellular biosynthesis of ferri-hydrite nanoparticles.[9] Their behaviour has been studied using zero gravity flight[10] and the nature of the active gravitaxis mechanisms was studied under varying acceleration aboard the space shuttle Columbia.[11]

Euglena gracilis has been investigated for the biotechnological production of many high value compounds, including vitamins A, C, E and essential amino acids.[12] They are a good source of polyunsaturated fatty acids,[13] as well as making high levels of paramylon.[14] In many ways *Euglena* is a “super food” and is currently sold as a nutritional and cosmetic supplement by euglena Co., Ltd (http://www.euglena.jp). Under anaerobic conditions some strains of *Euglena gracilis* can form over 50% dry weight wax esters[15,16] and are able to produce these compounds under adverse conditions encountered in, for example, acid mine drains.[17] This makes them suitable for use in waste water treatment,[18] where they can produce these compounds for use in animal feed or as biofuels.

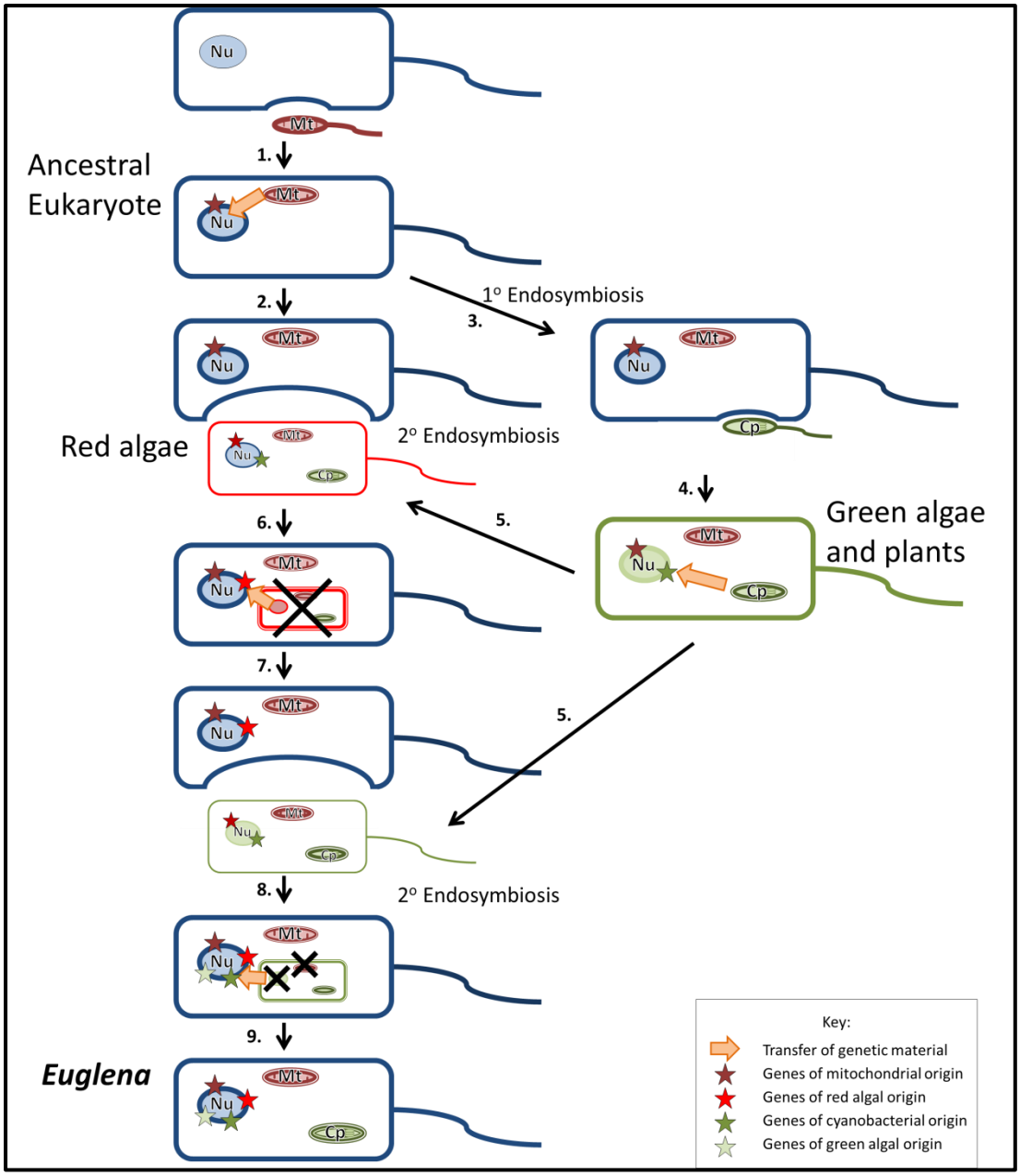


Figure 6.3: Sources of the *Euglena* genome

1. The ancestor of all eukaryotic cells formed an endosymbiotic relationship with a prokaryote to form the mitochondrion (Mt). 2. Most of the genetic material was transferred to the nucleus (Nu) whilst some remained in the mitochondria. 3. The ancestor of all plant cells subsequently took up a cyanobacterial cell to form the chloroplast (Cp) 4. Most of the genetic material was transferred to the nucleus. 5. The ancestral plant cells diversified to form green algae and plants, golden algae and red algae.[19] 6. A red algal cell was taken up by the ancestor of Euglenids and some of the DNA transferred to the nucleus.[20] 7. The red algae was then lost and the ancestor of photosynthetic *Euglena* formed an endosymbiotic relationship with a green algae,[21] which has been subsequently lost in several independent Euglenid lineages. 8. Nuclear and chloroplast DNA were subsequently transferred from the green algae to the nucleus of the Euglenid ancestor. 9. The nucleus and mitochondia were lost from the plant cell to leave the final chloroplast.

Genome sequencing of *Euglena* has not been possible to date, due to the large and complex genome (approximately 2 Gbp with 80% repetitive sequence – Mark Field, private communication), which has arisen from a series of endosymbiotic events during its evolution. The genetic material is derived from: the ancestral protozoa; the mitochondrion, related to alphaproteobacteria; a red algae endosymbiont which has since been lost;[20] the primary photosynthetic host, thought to be related to green alga;[22] and the primary photosynthetic endosymbiont, related to cyanobacteria, from the green alga (Fig 6.3). Some genetic material is retained in the chloroplast, which has been sequenced,[23] and the mitochondrion, which has unusually fragmented gene organisation.[24] Aside from normal eukaryotic epigenetic modification, including DNA methylation and histone acylation, the genome of *Euglena* also contains the modified nucleotide BaseJ (see Section 6.3.3), which is found amongst the kinetoplastids.[25] *Euglena* has the ability to extensively process mRNA during transcription, including removal of both classical[26] and atypical[27] introns as well as transplicing of small leader sequences,[28] altering the sequences before translation.

6.1.2 Rationale

The biochemistry and cellular biology of *Euglena* have been heavily studied for many years but recently the focus on model organisms has led to a reduction in research into this alga. The transcriptome sequencing of *Euglena* (see Section 5.4) has allowed us to avoid the difficulties involved in genome sequencing, often encountered with algae,[29] including large genome size,[30] highly repetitive sequences[31] and a high degree of DNA modification.[32] Aside from enzymes involved in carbohydrate metabolism (see Section 5.5) the transcriptome revealed the complex metabolism available to this highly flexible alga, which has helped us to understand the biosynthesis of the many medicinally and biotechnologically interesting compounds.

FM	lm.32046	dm.18553				
IFT88	lm.21777					
IFT52	lm.22088	lm.20473				
IFT57	dm.22343	lm.26251				
IFT22	lm.34442	lm.7956				
IFT140	lm.34442	dm.42150	lm.48576			
IFT172	dm.9347	lm.56041	lm.69878	dm.17483		
RIB72	dm.1605		dm.23873	lm.41593	lm.69421	
RSP3						
ε-tubulin						
BBS5	dm.8542					
AM						
Drebrin/ABP1	dm.38407					
Twinfilin	lm.40062	dm.58212				
Filamin	dm.96986		lm.55772	dm.66337	dm.22825	lm.54986
WASH	lm.58501					
Nadrin	lm.56666	lm.31604				
Sphingomyelin synthase related proteins						
Saposin	lm.76724	lm.87578				
Sphingomyelinase		dm.18553	lm.74949	dm.24189		

Table 6.1: *Euglena* transcripts encoding key genes associated with flagella motility (FM) or amoeboid motility (AM)[33]
Highlighted in yellow are transcripts only present in the dark grown cells.

6.2 Movement

Euglena are extremely active with a rapid swimming movement, propelled by their flagella, or by euglenoid movement, based on changes in their cytoskeleton (Fig 6.4). The euglenoid movement has unusual kinematics and hydrodynamics,[34] and is of interest to the biomimetic community, inspiring development of active envelopes. The key genes which define flagella and amoeboid movement[33] are all present in the transcriptome with the exceptions of the particular tubulin isoform used in flagella, a drebrin homologue involved in microtubule clustering, and the saposin activator of sphingomyelin remodelling (Table 6.1). This suggests that euglenoid movement, whilst morphologically distinct from amoeboid movement, uses similar molecular mechanisms.

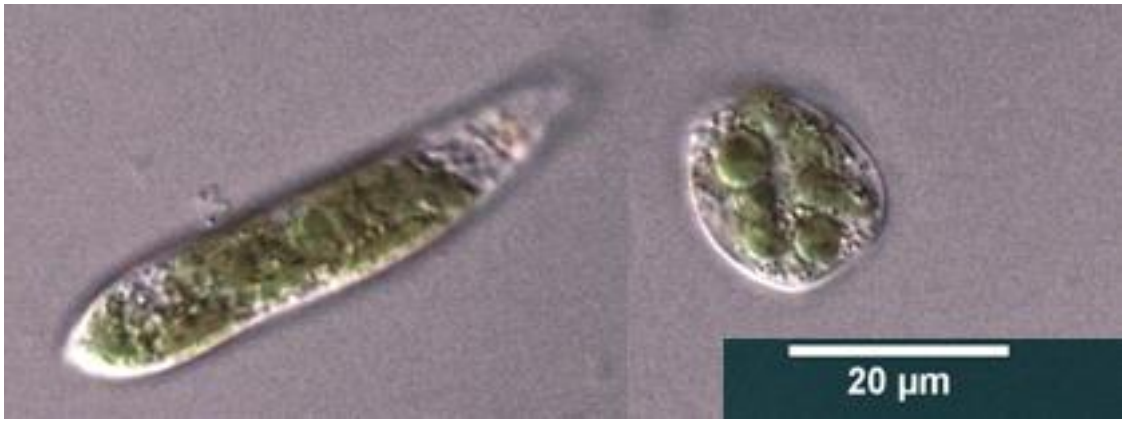


Figure 6.4: *Euglena* movement

Euglena have a highly flexible cell morphology and can change between long, thin, highly motile cells and static, round cells in a few seconds.

Gene	Transcripts				
Spo11					
Hop1	lm.96881				
Hop2	lm.36457				
Mnd1	lm.38224				
Dmc1	dm.98745				
Rad51	lm.97302	dm.8059	dm.33887		
Msh4,5	lm.100901	dm.51827	dm.50158	lm.57414	dm.73380
Msh2,6					
Mre11	dm.28705				
Rad50	lm.54833	dm.30793	lm.41294		
Rad52	dm.85580				
Mlh2	dm.41679	lm.85993	lm.94725		
Mlh1				dm.89596	
Pms1/2				lm.51341	
Mlh3					

Table 6.2: *Euglena* transcripts encoding key genes involved in meiosis[35]

Several of the proteins have close homology to more than one of the Eukaryotic homologues. Highlighted in yellow are transcripts only present in the dark grown cells.

6.3 Genetic control

Eukaryotes use diverse mechanism for genetic control and repair but there remain many unresolved issues around the lack of aging in *Euglena*.[36] During symmetric cell division all the genetic damage accumulated through the cell cycle must be repaired, and epigenetic modifications copied on to the next generation. By studying how this is accomplished in the immortal *Euglena*, greater understanding about human aging may lead to advances in medicine.

6.3.1 Sexual reproduction

There is no evidence in *Euglena* for meiosis or sexual reproduction, which complicates mutagenesis and genetic transformation, hampering the use of this alga in biotechnology. However genome sequencing and direct observation of other algae, including *Chlorella*[37] and *Ostreococcus,*[38] has revealed cryptic sexual reproduction. The transcriptome of *Euglena* contains all the genes required for meiosis, with the sole exception of the particular isoform of topoisomerase (TOPO VI) associated with meiosis (Table 6.2), although this enzyme is not actually required for meiosis.[39] There are transcripts encoding all four other classes of topoisomerase (IA, IB, II and III) which may be able to substitute for this enzyme, or homologous recombination may not occur during meiosis. Together this suggests that under certain conditions this alga may undergo meiosis and sexual reproduction.

6.3.2 DNA gyrase

The *Euglena* transcriptome also encodes one of each subunit of the bacterial topoisomerase, DNA gyrase. In plant cells these are targeted to the chloroplast and mitochondrion, where they have a role in controlling DNA topology,[40] and would be expected to have the same role in *Euglena*. The fluoroquinolone ofloxacin is a specific inhibitor of DNA gyrase and in *Euglena* is known to disrupt both chloroplast[41] and mitochondria.[42] We have found the fluoroquinolone derivative ciprofloxacin to inhibit growth of *Euglena* cells (IC_{50} = 8±3 µM) and to lead to permanent loss of chlorophyll (Figure 6.5), in a similar way to streptomycin treatment of this organism.[43] These results support the necessity for DNA gyrase in chloroplast growth.

Protein name	Activity	Pathway	No of homologues in *Euglena*
AGO	RNA slicer	miRNA, S-PTGS, tasi-RNA, chromatin	4
DCL	miRNA synthesis	All	4
RDR	RNA-dependant RNA-polymerase	S-PTGS, Transitivity, tasi-RNA, nat-siRNA, Chromatin	0
CMT	Cytosine DNA methyltransferase	Chromatin	2
K9 MeT	Histone methyl transferase	Chromatin	5
NRPD	DNA-dependant RNA-polymerase	Chromatin, nat-siRNA	8
HDA6	Histone deacetylase	Chromatin	8
SDE3	helicase	S-PTGS, Transitivity	8
HST	Exportin	miRNA	1
HYL1	dsRNA binding	miRNA	0
WEX	exonuclease	S-PTGS	0
SGS3	unknown	S-PTGS, Transitivity, tasiRNA, nat-siRNA	0
HEN1	sRNA-methyl transferase	All	0

Table 6.3: *Euglena* transcripts encoding key genes for RNA silencing

Number of transcripts for genes involved in gene silencing pathways in the *Euglena* transcriptome.[44] tasi: transacting siRNA. S-PTGS: sense posttranscriptional gene silencing. IR-PTGS: inverted repeat posttranscriptional gene silencing.

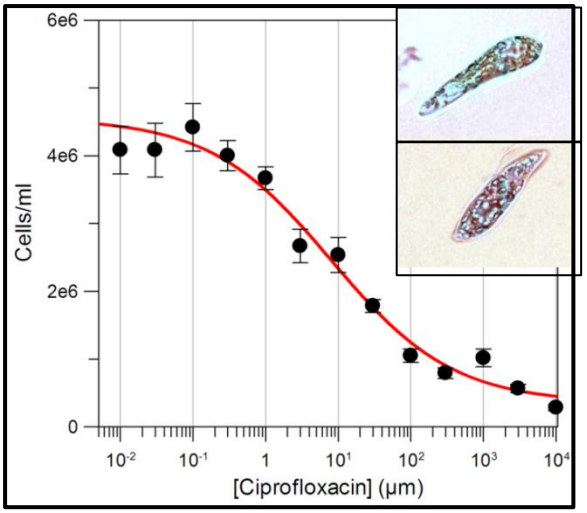


Figure 6.5: Ciprofloxacin inhibition of *Euglena* culture growth

Cells (10^4) were inoculated into high nutrient media containing ciprofloxacin and the number of cells counted after seven days (30 °C). Ciprofloxacin inhibited growth at concentrations above 1 µM, although there have still been around five doublings at inhibitor concentrations of 10 mM. The inset images show single cells after treatment with 0 mM (top) and 100mM (bottom) antibiotic. Ciprofloxacin causes loss of the chloroplasts and therefore chlorophyll, leaving the cells brown in colour.

6.3.3 Gene silencing

RNA-mediated gene silencing is a ubiquitous control mechanism found in bacteria,[45] plants[46] and animals.[47] Three main components make up the machinery in Eukaryotes: Dicer like (DCL) which cleave double stranded RNA;[48] Argonaute (AGO) which targets inactivation of sequences complementary to small RNAs (21-24nt) as part of the RNA-induced silencing complex (RISC);[49] and RNA-dependent RNA polymerase (RDRP) which amplifies the silencing.[50] Whilst some sequenced algae do not have any of the components, most retain some capability.[51] *Euglena* is known to possess some capacity for RNA-silencing: successful knockdown experiments of a gene involved in vitamin C biosynthesis[52] and a photoreceptor protein[53] have been performed in this organism. The transcriptome encodes four DCL proteins and four AGO proteins, along with many of the other components of the gene silencing machinery including helicases, histone methylases, histone deacetylases, and cytosine methylases (Table 6.3). There is no RDRP present which suggests that *Euglena* is incapable of producing trans-acting RNAs or utilising the aberrant transcript pathway.[44] It is possible another enzyme is capable of acting to amplify the signal, as is suggested for mammals and insects, which also lack the RDRP.[50] In *Arabidopsis* this amplification is involved in development, where the original double-stranded RNA might be diluted during growth,[54] which is conceivably not required in a single celled organism such as *Euglena.*

6.3.4 BaseJ

BaseJ (**179**) is a modified nucleotide, formed by hydroxylation and glucosylation of thymidine (Fig 6.6), uniquely found amongst the Euglenozoa, including the important human pathogens *Trypanosoma* and *Leishmania*.[55] It prevents RNA polymerases passing along the DNA strand, silencing the modified region of the genome, and is of key importance to the control of coat expression in *Trypanosomes*.[25] In *Leishmania* some of the BaseJ is not located around the telomeres, as it is in *Trypanosomes*,[56] and has been shown to prevent transcriptional readthrough, ensuring proper transcriptional termination. BaseJ is found throughout the genome of *Euglena*, where it makes up approximately 0.2% of the bases.[57]

The two proteins involved in the initial hydroxylation of thymidine are well studied in *Trypanosomes* and *Leishmania*.[58] There are matching homologues encoded in the *Euglena* transcriptome, though the JBP1 homologue (dm.72228), essential in *Leishmania*,[59] is not found in the transcriptome of the light grown cells. Unlike JBP1, JBP2 (dm.12798 in *Euglena*) does not actually bind BaseJ, but instead appears to initiate *de novo* BaseJ biosynthesis, whilst JBP1 amplifies this signal.[60] The enzymes for addition or removal of the glucose have not, to date, been identified in any organism. After removal of the glucose the HOMedU (**178**) could be either dehydroxylated or the base could be excised.

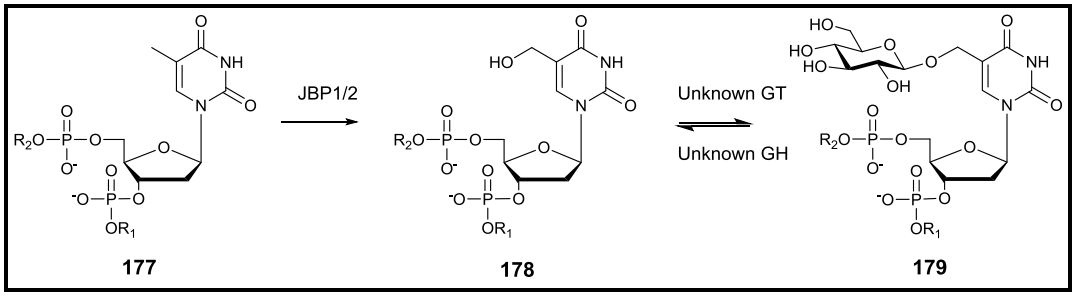


Figure 6.6: BaseJ metabolism

To form Base J (**179**) a thymidine (**177)** is hydroxylated by JBPs, to form hydroxymethyl-deoxy-uracil (**178**, HOMedU), and then glucosylated by an unidentified glucosyl transfrerase. The glucose is removed by an as yet unidentified glucosidase, leaving HOMedU. R_1 = 3' DNA strand, R_2= 5' DNA strand.

6.4 Primary metabolism

6.4.1 Core metabolic enzymes

Candidate genes were identified in the *Euglena* transcriptome for all of the enzymes in the primary metabolic pathways, including glycolysis and gluconeogenesis, citrate cycle, pentose phosphate pathway and fatty acid biosynthesis, based on homology to orthologs in the KEGG modules and pathways (see Appendix 8). There are also candidates for carotenoid biosynthesis and the Calvin cycle, which are required for photosynthesis. The completeness of these pathways confirms that meaningful sequencing depth and functional assignments have been achieved.

The *Euglenoid* chloroplast is derived from a green algal cell and has had substantial genetic rearrangement during evolution.[61] This has included the loss of several genes from the chloroplast genome to the nucleus, from where they can be identified in the *Euglena* transcriptome, including the protease clpP (lm.95241 and lm.15675), the membrane protein cemA (lm.59206) and photosystem 1 component ycf3 (lm.46611).

6.4.2 Lipids

Under anaerobic conditions *Euglena* converts pyruvate, produced during glycolysis, to acetyl-CoA. This is used to produce diverse final products, including ethanol, lactate and fatty acids. *Euglena* has all four of the enzymes necessary for the fermentation of pyruvate[62] to acetate in the absence of oxygen (Table 6.4). This is then used to make a wide variety of long-chain wax esters, which can reach over 50% of the cell dry weight under anaerobic conditions.[63]

pyruvate formate lyase	dm.62807		
pyruvate-ferredoxin oxidoreductase	lm.15157	lm.3147	
lactate dehydrogenase	lm.11577	dm.18437	lm.9457
pyruvatedecarboxylase	lm.8614		

Table 6.4: *Euglena* transcripts encoding enzymes for the anaerobic fermentation of pyruvate
Pyruvate formate lyase, highlighted in yellow, is only present in the dark grown cells.

Fatty acids (FA) identified by GCMS from *Euglena gracilis* as methyl esters.

	13:0	14:0	15:0	16:4	16:3	16:0	Cyp16	17:0	18:4	18:2	18:3	20:4	20:5	20:3	20:2	22:5	22:6
% FA	4.5 ±0.1	8.6 ±0.3	8.3 ±0.5	0.9 ±1.6	7.2 ±2.0	9.1 ±0.8	5.4 ±0.3	1.4 ±2.3	7.5 ±2.3	1.7 ±3.0	12.7 ±2.2	6.1 ±1.6	7.6 ±2.3	3.2 ±2.8	5.8 ±2.5	2.5 ±2.2	7.6 ±2.3

Wax esters (WE) identified by GCMS from *Euglena gracilis*. Isomers cannot be separated and are of the form $H(CH_2)_{x-1}\ COO(CH_2)_yH$.

x+y=	22	23	24	25	26	27	28	29	30	31	32	33	34	36
% WE	0.2 ±0.01	0.6 ±0.03	2.2 ±0.1	7.7 ±0.1	17.2 ±0.2	23.0 ±0.3	21.4 ±0.2	12.7 ±0.2	6.8 ±0.2	3.5 ±0.1	1.7 ±0.4	0.8 ±0.1	0.4 ±0.4	0.2 ±0.2

Triglycerides (TG) identified by GCMS from *Euglena gracilis*. Isomers cannot be separated and there are multiple desaturation variants.

$C_nH_xO_3$ n=	37	38	39	40	41	42
% TG	12.8 ±2.9	22.0 ±5.0	26.6 ±6.2	21.3 ±5.2	12.6 ±3.4	4.6 ±4.0

Table 6.5: Analysis of the hexane soluble fraction from *Euglena*

Most of the isolated fatty acids are multiply desaturated in *Euglena.* Wax esters make up over 95% of the components of the heptane fractions. The triglyceride components also contain phospholipids, after hydrolysis and methyl-esterification. Cyp16 = Cyclopropyl hexadecanoic acid.

6.4.2.1 Polyunsaturated fatty acids

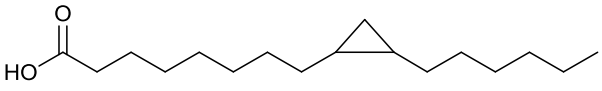
Polyunsaturated fatty acids, which have been credited with a wide variety of health benefits,[64] make up over 60% of the fatty acids in *Euglena* (Table 6.5). In order to introduce double bonds into the fatty acids, *Euglena* requires several desaturase enzymes, of which three have previously been sequenced and characterised.[65-67] In the *Euglena* transcriptome six putative desaturases can be identified (lm.37937, lm.17252, dm.49413, lm.14243, dm.2712 and dm.2753). It is difficult to predict specific roles for these enzymes, but it is tempting to suggest they each perform a specific desaturation on route to docosahexaenoic acid, the most unsaturated fatty acid identified in *Euglena*.[13]

6.4.2.2 Cyclopropane fatty acids

Cyclopropane fatty acids are synthesised by a variety of bacteria,[68] plants and protozoa, including cotton[69] and trypanosomatids.[70] They are made by methylation of an unsaturated phospholipid[68] and inhibit eukaryotic lipid desaturases.[71] Cyclopropane fatty acids are attractive for industry because they have the physical properties of unsaturated fatty acids, combined with the oxidative stability of saturated hydrocarbons and chemical reactivity from the strained ring.[72]

In the transcriptome there is one credible cyclopropane fatty acid synthase (lm.28488) related to bacterial enzymes. There is another candidate (dm.16838), only present in the dark, which is related to green algal enzymes and encodes an extra C-terminal domain of unknown function.

In order to see if *Euglena* produces cyclopropane fatty acids the heptane extract was compared to an authentic standard. Several fatty acids with odd numbered alkyl chains were identified, especially in the triglyceride components. Amongst the fatty acids is a compound with the same retention time and fragmentation pattern, in GC-MS, as authentic cyclopropyl hexadecanoic acid (**180**), although this cannot be resolved from cis-10-heptadecenoic acid (Table 6.5). This suggests *Euglena* could be explored as a source of this potentially high value compound.[73]



180

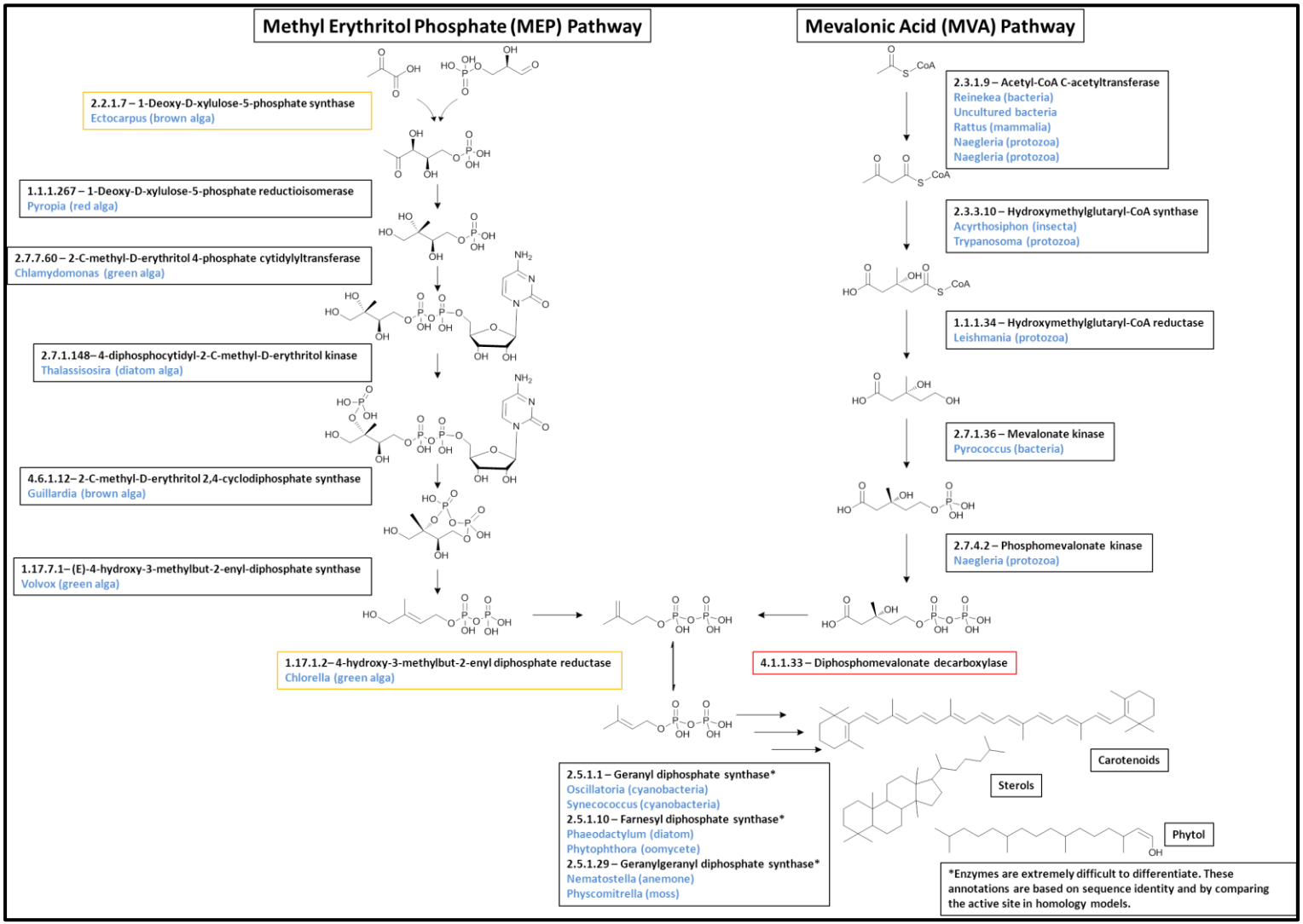


Figure 6.7: Isoprenoid biosynthetic pathway

Both the MEP and MVA pathways are present in *Euglena*. All enzymes could be identified in the transcriptomes, except the final decarboxylase of the MEP pathway, indicating an alternative enzyme must be used in *Euglena*. Listed in blue is the closest homologue for each isoform found in the NCBI non-redundant protein sequence collection, using BLASTP. The yellow boxes indicate enzymes for which sequences could only be found in the dark grown cells. See Appendix 9.1 for details of the biosynthetic enzymes discussed.

6.4.3 Isoprenoids

Isoprenoids are one of the most abundant natural products in algae,[74] with functions in primary metabolism, including as carotenoids and attached to chlorophyll, and as diverse secondary metabolites including polyethers[75] and terpenoids.[76]

6.4.3.1 Isoprene biosynthesis

There are two pathways to the formation of the isoprenoid precursor in plants: the cytosolic mevalonate (MVA) and the plastidial methylerythritol phosphate (MEP) pathways. Both pathways are known in *Euglena*, though unusually the MEP pathway only contributes to carotenoid biosynthesis and is not involved in phytol synthesis.[77]

All enzymes for both pathways can readily be identified in the transcriptome, with the exception of the final decarboxylase of the MVA pathway (Fig 6.7). The presence of the rest of the pathway in *Euglena* implies this last reaction must occur, and is catalysed by a non-canonical enzyme. The initial and final enzymes for the chloroplast-localised MEP pathway are only apparent in the transcripts from dark grown cells. This suggests that, under the conditions utilised in this study, the MEP pathway is only active in chloroplasts that are not supplying energy to the cells.

6.4.3.2 Polyprene biosynthesis

Linear polyprenyl pyrophosphates are formed by condensation of isoprene units, which is catalysed by a family of closely related prenyltransferases, and several members of this enzyme class are present in the transcriptome. In order to differentiate between the geranyl, farnesyl and geranylgeranyl diphosphate synthases structural models of the proteins were built using Swissmodel (Fig 6.8).[78] By comparing the active sites with the geranylgeranyl diphosphate (GGPP) bound to the human synthase[79] it is possible to propose likely enzyme specificities.

The likely geranyl diphosphate synthases (lm.32576 and lm.90469) show some interactions with isoprene 2 and 3 of the overlaid GGPP, which suggests they are unlikely to accommodate anything longer than geranyl diphosphate in the active site (Fig 6.8.A). There are steric clashes between the likely farnesyl diphosphate synthases (lm.19578 and dm.26131) with isoprene 4 of the superimposed GGPP, which would prevent the formation of GGPP by these enzymes. For the likely GGPP synthases (dm.10389 and dm.48870) the active site has the same fold as the genuine GGPP synthase, with no steric clashes to the overlaid substrate. These models are based on genuine synthases, and so biased towards the backbone shape of the template, and biochemical evidence would be required to definitively identify the specificity of these enzymes.

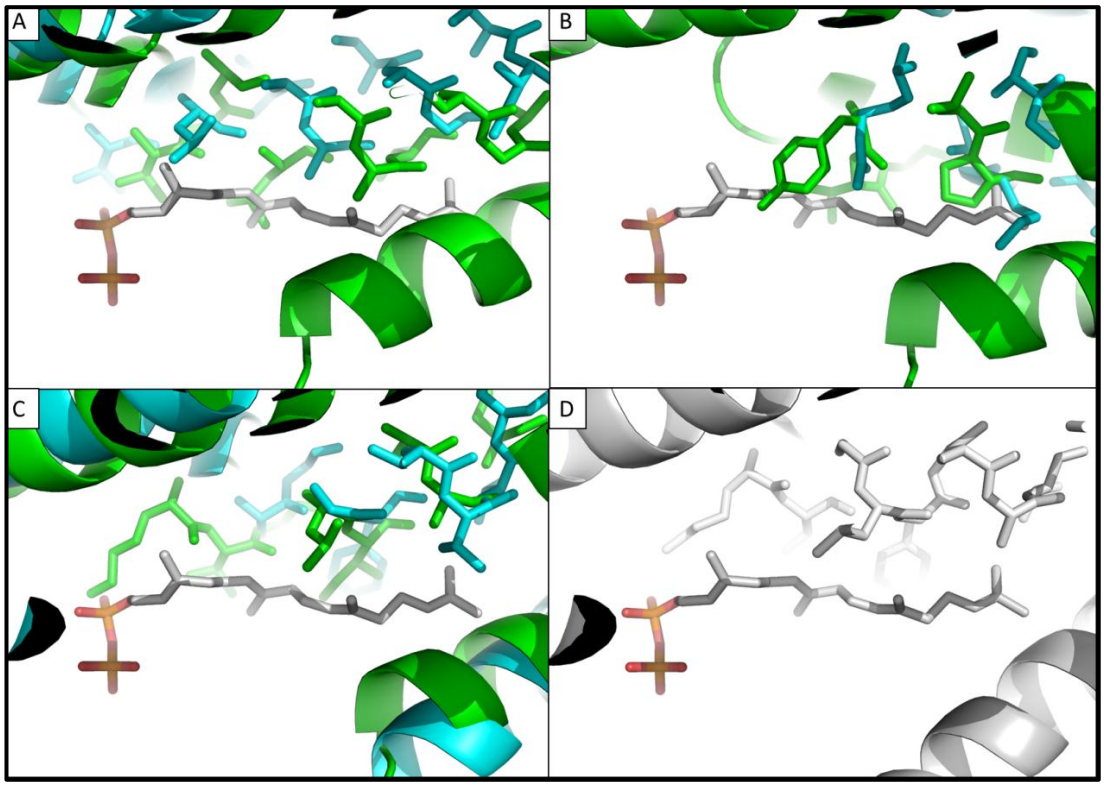


Figure 6.8: Polyprene synthase homology models

Geranylgeranyl diphosphate was overlaid from 2Q80[79] and the six prenyl transferases found in the *Euglena* transcriptome were modelled using Swissmodel in automated mode.[78] These were then aligned and the key amino acids in the interacting helix displayed as sticks. **A**. Models of likely geranyl diphosphate synthases lm.32576 (green) and lm.90469 (cyan). **B**. Models of likely farnesyl diphosphate synthases lm.19578 (green) and dm.26131 (cyan). **C**. Models of likely geranylgeranyl diphosphate synthases dm.10389 (green) and dm.48870 (cyan). **D**. Human GGPP synthase with GGPP bound in the active site (2Q80).

6.4.3.3 Carotenoids

Carotenoids, which play a major role in light absorption during photosynthesis, can be obtained at over 70 mg per litre culture from *Euglena*.[12] They are synthesised from geranylgeranyl diphosphate by phytoene synthase, followed by a series of cyclisations and oxidations. There is one sequence for a phytoene synthase in the *Euglena* transcriptome (lm.32898), present in both the light and the dark grown cells, which is related to the chloroplast-localised enzyme from the diatom *Thalassiosira pseudonana*.

6.4.3.4 Triterpenes

Triterpenes, including sterols and saponins, are synthesised from two molecules of farnesyl diphosphate which are joined by squalene synthase and cyclised by triterpene cyclase. Various cycloartenols, including lanosterol[80] and ergosterol,[81] have been identified in *Euglena*. There are sequences for one triterpene cyclase (dm.61818) in the *Euglena* transcriptome, which has a histidine in position to act as a base in the reaction (Fig 6.9),[82] leading to the prediction that the triterpene formed will have a cycloartenol skeleton. This suggests that *Euglena* is probably not able to make any other triterpene skeletons, such as β-amyrin or oxidosqualene. Nystatin, which complexes ergosterol, and miconazole, an inhibitor of ergosterol biosynthesis, both inhibit growth of *Euglena* (unpublished data) and ergosterol itself is the only sterol detectable, at around 0.1 ± 0.01 % of the total heptane soluble compounds, as identified in the lipid analysis (section 6.4.2). This suggests that ergosterol is the major, if not the only, sterol formed by *Euglena gracilis*.

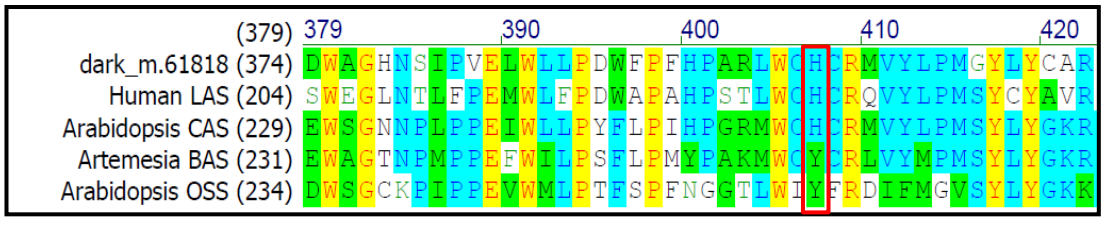


Figure 6.9: Alignment of cycloartenol synthases

The *Euglena* triterpene cyclase (dm.61818) was aligned against other triterpene synthases. The red box indicates the key residue for determining the scaffold produced – a histidine acting as a base in the case of cycloartenol and lanosterol synthase.[82] LAS: lanosterol synthase. CAS: Cyloartenol synthase. BAS: β-amyrin synthase. OSS: Oxidosqualene synthase.

6.4.3.5 Vitamin E

Tocopherols are fat soluble antioxidants produced in photosynthetic cells[83] and are an essential part of the human diet, as Vitamin E. They are synthesised from tyrosine and phytol, and are known to accumulate to over 7 mg per gram in *E. gracilis*.[84] In the transcriptome the key genes, dioxygenase and homogentisate phytyl transferase, are present, and strong candidates for the rest of the pathway are apparent (Fig 6.10). There are many amino transferases in the transcriptome which may act on tyrosine and there are many methyltransferases whose substrates are difficult to conclusively assign.

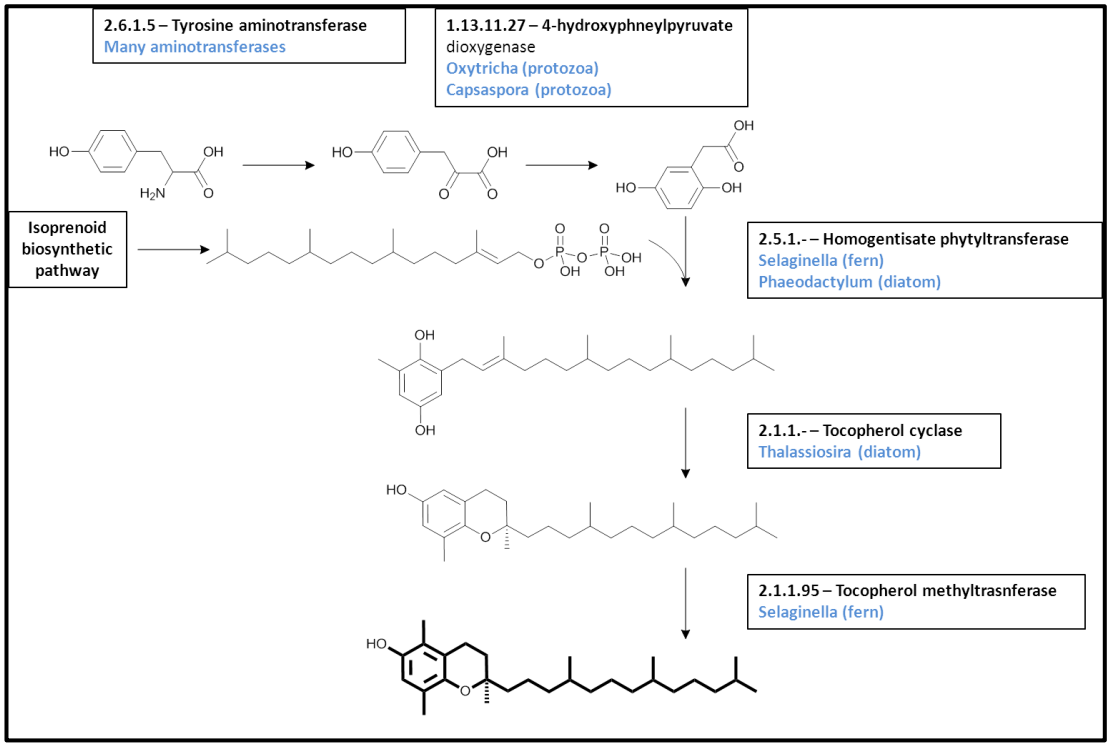


Figure 6.10: Tocopherol biosynthesis

Good candidates could be found in the *Euglena* transcriptome for each gene in the biosynthesis of tocopherols. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.2 for details of the biosynthetic enzymes discussed.

6.4.4 Amino acid biosynthesis

Like plants and other algae, *Euglena* is capable of synthesising all of the amino acids it requires for growth, and can produce high levels of all the essential amino acids for the human diet.

6.4.4.1 Lysine biosynthesis

Lysine cannot be synthesised by animals and must be obtained in the diet. Plants and bacteria utilise the diaminopimelate pathway, whilst the α-aminoadipate pathway[85] is only found in higher fungi and certain protozoa.[86] Biochemical analysis has shown that in *Euglena* lysine biosynthesis is through this latter pathway.[87] All of the enzymes for the α-aminoadipate pathway[88] can readily be identified in the transcriptome (Fig 6.11), whilst no homologues of the enzymes for the diaminopimelate pathway are present.

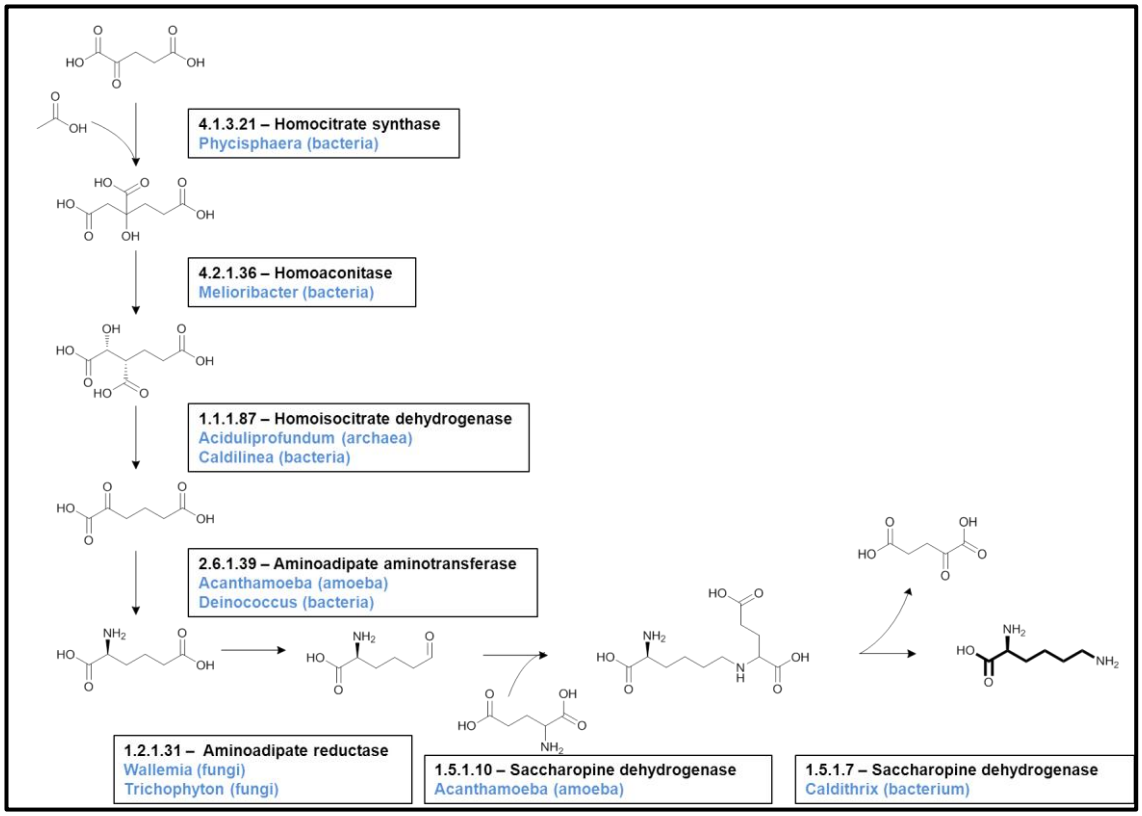


Figure 6.11: The α-aminoadipate pathway for biosynthesis of lysine

All enzymes for the α-aminoadipate pathway can be identified in the *Euglena* transcriptome. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.3 for details of the biosynthetic enzymes discussed.

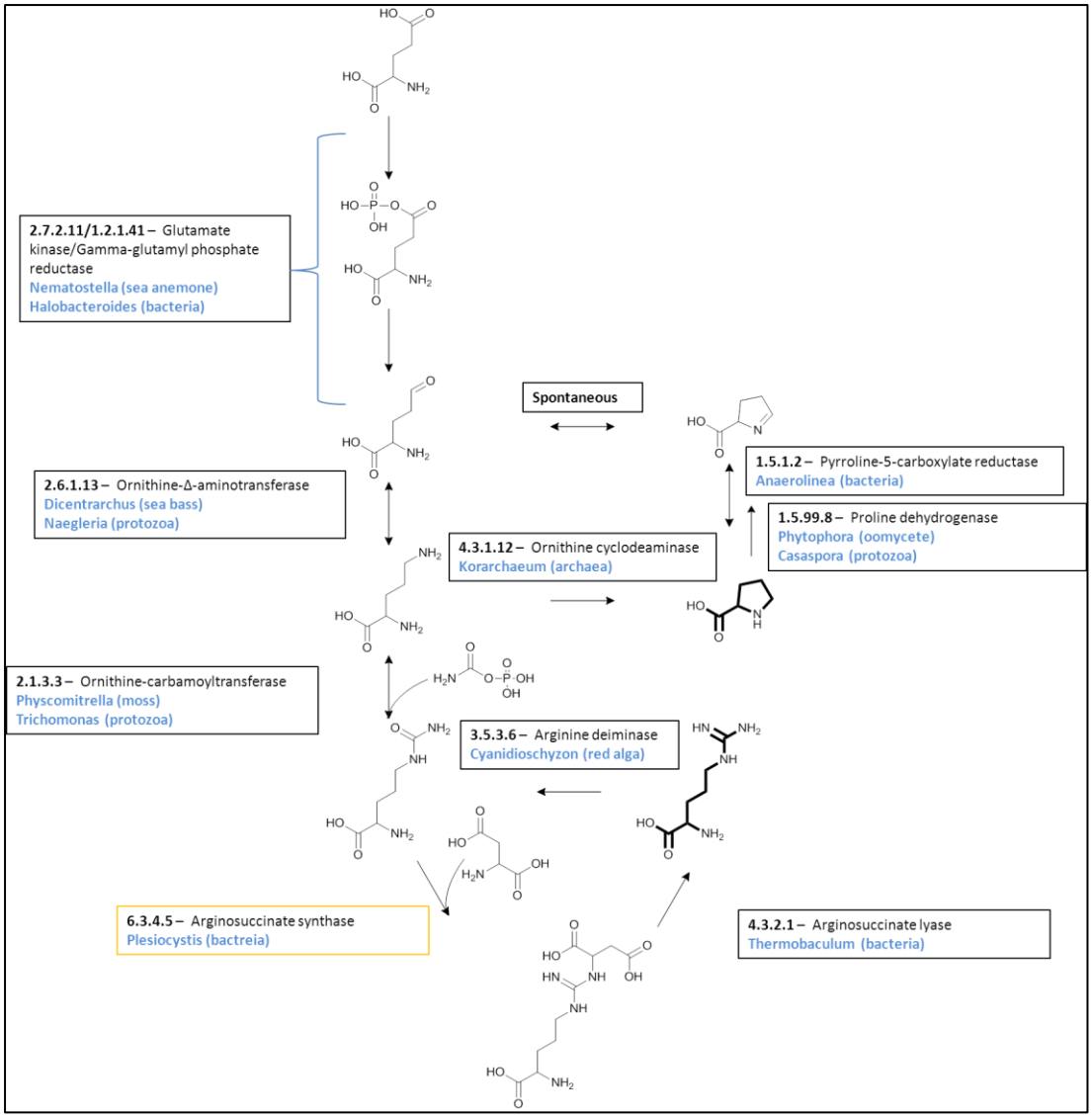


Figure 6.12: Arginine and proline biosynthesis

All of the enzyme for arginine synthesis from glutamate via ornithine are present in the transcriptome. The enzymes for proline biosynthesis, both by the reduction of pyrroline-5-carboxylate[89] and for the cylodeamination of ornithine,[90] are present in the transcriptome. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.4 for details of the biosynthetic enzymes discussed. The transcript for the arginosuccinate synthase is only found in the dark grown cells.

6.4.4.2 Arginine biosynthesis

Arginine is synthesised by the sequential transfer of nitrogen on to glutamate semialdehyde. All the enzymes required are apparent in the transcriptome from both light and dark cells, with the exception of the arginosuccinate synthase which is only apparent in the dark grown cells.

Arginine is used as an efficient nitrogen reservoir in *Euglena* accumulating to high levels, both free and in peptides, when cells are grown in excess nitrogen.[91] *Euglena* does not possess a urease and degradation is instead via the non-urease pathway, with sequential removal of ammonia by the reversible enzymes of arginine synthesis (Fig 6.12).[92] Nitrogen is rapidly released from arginine by transfer of nitrogen, sequentially forming citrulline, ornithine and glutamate semialdehyde, efficiently releasing the stored nitrogen.

6.4.4.3 Proline biosynthesis

Proline is synthesised in most organisms by the spontaneous cyclisation of glutamate semialdehyde followed by reduction.[89] In a few cases, such as in *Clostridium*, proline is synthesised from ornithine by cyclisation and deamination.[90] In the transcriptome a bacterial proline reductase and an archaeal ornithine cyclodeaminase are present, indicating both pathways probably operate in *Euglena*.

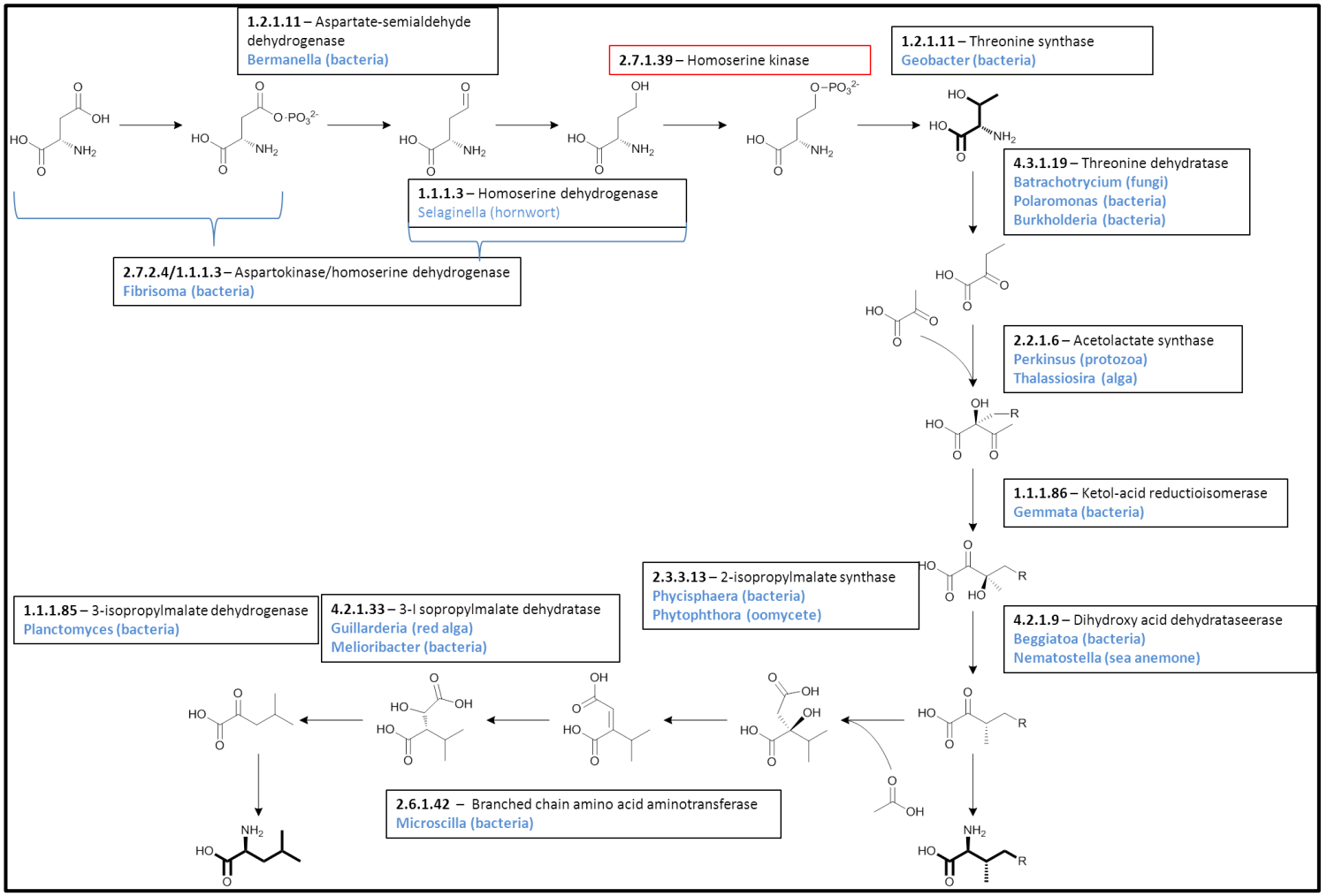


Figure 6.13: Threonine and branched-chain amino acid biosynthesis

The enzymes for threonine biosynthesis from aspartate are all present in the transcriptome except the homoserine kinase. For valine (R=H) and leucine biosynthesis two molecules of pyruvate are combined by acetolactate synthase. For isoleucine synthesis (R=Me) oxobutanoate, derived from threonine, and pyruvate are combined by acetolactate synthase. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.5 and 9.6 for details of the biosynthetic enzymes discussed.

6.4.4.4 Threonine biosynthesis

Threonine is one of the essential amino acids in the human diet. It is synthesised from aspartate via a series of phosphorylations and decarboxylations (Fig 6.13).[93] All the enzymes are present in the *Euglena* transcriptome, with the exception of the homoserine kinase, suggesting another enzyme carries out this reaction.

6.4.4.5 Branched-chain amino acid biosynthesis

The initial step in branched chain amino acid biosynthesis is the combination of two keto acids by acetolactate synthase (Fig 6.13).[94] Either, for the synthesis of leucine and valine, two molecules of pyruvate or, for isoleucine biosynthesis, one molecule of pyruvate and one molecule of 2-oxobutarate are joined. Three homologues of threonine dehydratase, which generates the 2-oxobutarate from threonine, can be found in the *Euglena* transcriptome and two have previously been identified biochemically in *Euglena*.[95] As only two were found previously it is possible that one of the transcripts could encode an enzyme involved in the interconversion of serine and pyruvate, and thus not show any activity as a dehydratase. Sequences for the other enzymes in the biosynthesis of these amino acids can be found in the *Euglena* transcriptome. The isopropylmalate dehydratase sequences encode contiguous large and small subunits, rather than encoding them as separate polypeptides, as is the case in most organisms.[96]

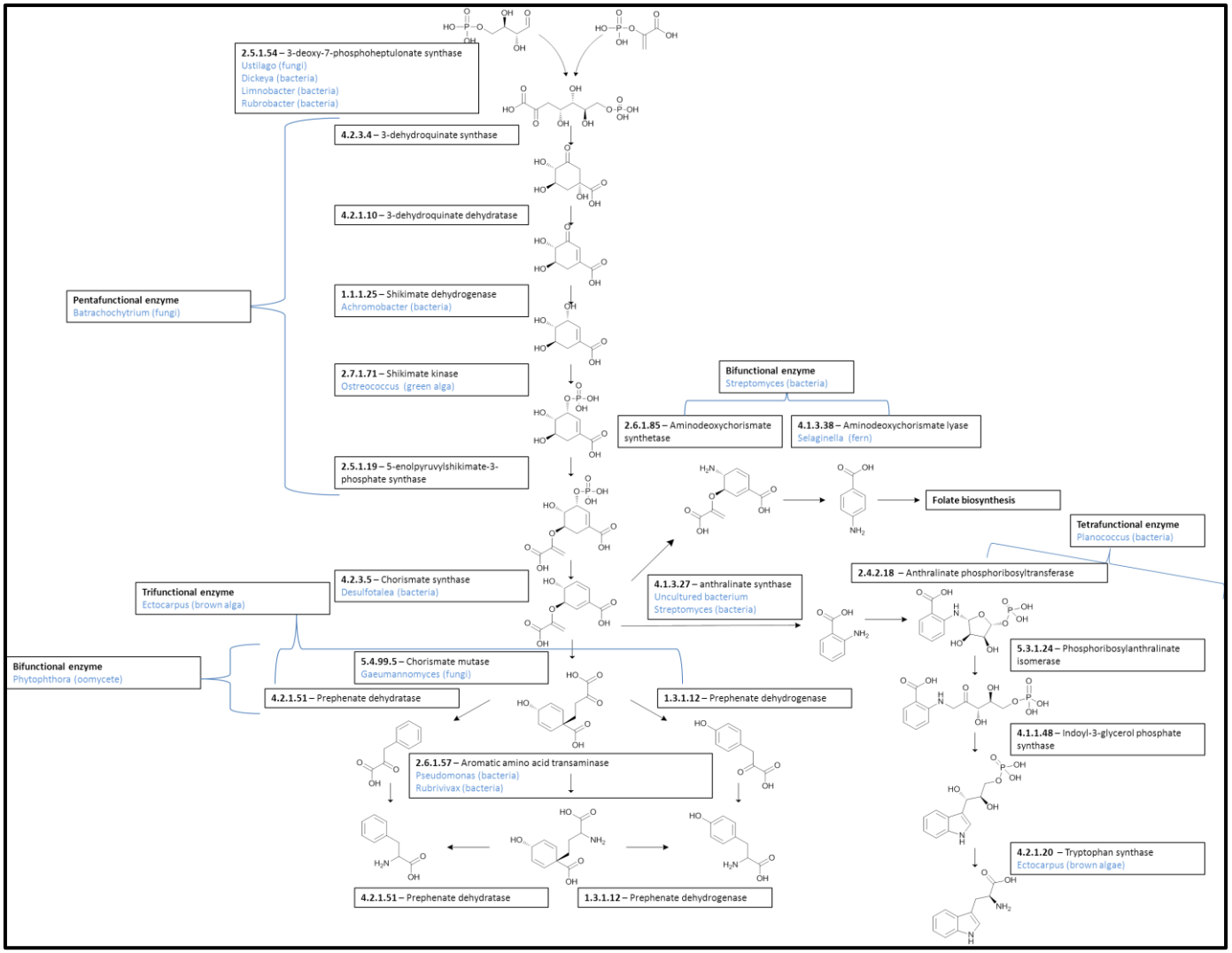


Figure 6.14: Aromatic amino acid biosynthesis

In *Euglena* the five central enzymes of the shikimate pathway are encoded as a single polypeptide, as is seen in fungal enzymes. The pathway branches at chorismate: towards folate biosynthesis, via 4-aminobenzoate, catalysed by a bifunctional enzyme; towards tryptophan biosynthesis, via 2-aminobenzoate; or to prephenate via chorismate mutase, which is present either as a single domain protein or as part of a trifunctional protein with dehydratase and dehydrogenase domains. Aromatic amino acid transaminases, for the final step in tyrosine and phenylalanine synthesis, are present in the transcriptome. The biosynthesis of tryptophan is carried out by a tetrafunctional enzyme. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.7 for details of the biosynthetic enzymes discussed.

6.4.4.6 Aromatic amino acid biosynthesis

Chorismate, the precursor to aromatic amino acids, is synthesised by the shikimate pathway in seven steps. There are four isoforms of the first committed enzyme in the pathway, DHAP synthase, one of which is only apparent in the dark grown cells. The next five reactions can be catalysed either by separate enzymes, as in plants,[97] or by a pentafunctional enzyme, as in fungi (Fig 6.14).[98] Phylogenetic analysis suggests the fusion of these five genes occurred in an ancient eukaryotic ancestor and was subsequently rearranged or lost.[99] Previously cytosolic and plastid isoforms of 5-enolpyruvylshikimate-3-phosphate synthase have been identified in *Euglena*,[100] which suggested the presence of both types of pathway. The pentafunctional enzyme can be identified in the *Euglena* transcriptome, as can individual shikimate synthase and shikimate kinase domains, although no sequences for the other single domains are present.

Chorismate is then converted into tyrosine and phenylalanine, via prephenate. Bifunctional chorismate mutase-prephenate dehydratase[101] and chorismate mutase-prephenate dehydrogenase[102] are common. However, the trifunctional chorismate mutase-prephenate dehydratase-dehydrogenase encoded in the *Euglena* transcriptome, has only been characterised from the hyperthermophile *Archaeoglobus fulgidu*I. In this archaeon it is hypothesised that this domain structure protects the labile intermediate under extreme conditions,[103] which is clearly not the situation with *Euglena*.

The biosynthesis of tryptophan from chorismate, via anthranilate, is carried out by four enzymatic reactions.[104] These are normally located on separate chains in bacteria, though there is sometimes one gene fusion.[105] In *Euglena* a single transcript encodes all four enzymes, which has previously been partially sequenced,[106] and for which proteomic evidence supports the domain architecture (see Appendix 10). The final synthase is composed of a hetero-tetramer in bacteria[107] and a homo-dimer in fungi.[108] The tetra-functional enzyme encoded in the *Euglena* transcriptome only contains the β-subunit, whilst the α-chain is encoded on a separate transcript. This suggests tryptophan biosynthesis is catalysed by a hetero-tetramer composed of two copies of the multi-function protein and two copies of the α-chain.

6.4.4.7 Tetrapyrrole biosynthesis

Heme, chlorophyll and vitamin B12, all contain a tetrapyrrole which is synthesised from δ-aminolevulinic acid (ALA). Heterotrophs tend to synthesize ALA from glycine and succinyl-CoA in the Shemin pathway,[109] whilst photoautotrophs make ALA from glutamate in the C5 pathway.[110] *Euglena* is known to utilise both routes,[111] and the transcriptome shows a bacterial derived Shemin pathway and a green algae-related C5 pathway, presumably obtained with the chloroplast (Fig 6.15). ALA is then converted to porphobilinogen by ALA dehydratase, and hydroxymethylbilane synthase joins four porphobilinogens to make the core tetrapyrrole.

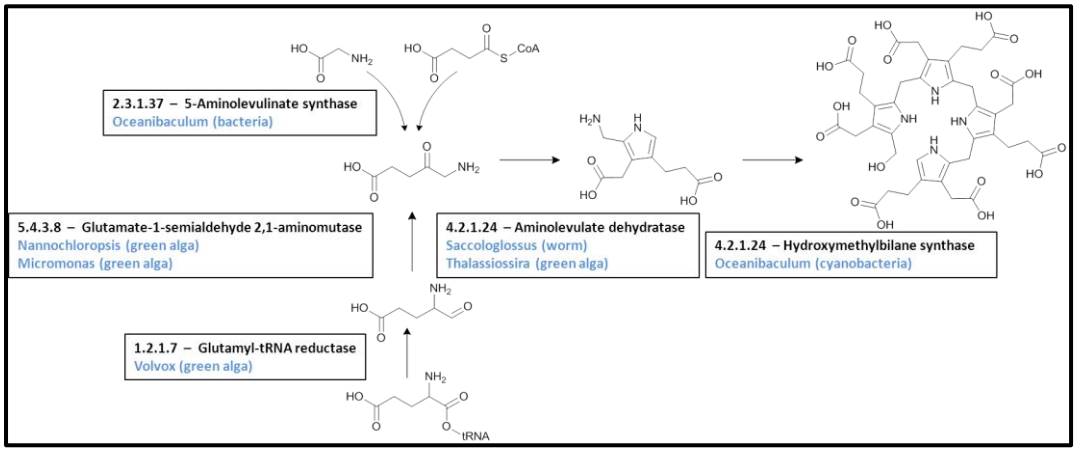


Figure 6.15: Tetrapyrrole biosynthesis

Tetrapyrroles are synthesised from δ-aminolevulinic acid. In the Shemin pathway this is synthesised from glycine and succinyl-CoA by 5-aminolevulinate synthase and from glutamate by reduction and isomerisation, in the C5 pathway. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.8 for details of the biosynthetic enzymes discussed.

6.4.5 Hydrogenase

Since the 1940s it has been known that certain algae, such as *Scendesmus*[112] and *Chlamydomonas*,[113] produce hydrogen under illumination, using an hydrogenase containing iron cofactors. One sequence from the *Euglena* transcriptome (dm.82403) encodes a putative hydrogenase candidate, though this alga has not previously been noted to produce hydrogen. There are no candidates for the accessory proteins,[114,115] which are required for organelle-localised hydrogenase, and the enzyme does not have a strong signal peptide,[116] suggesting this is a cytosolic enzyme. Conceivably this enzyme is therefore an accessory protein for the correct assembly of cytosolic iron sulphur proteins, which are close homologues to the iron only hydrogenases.[117]

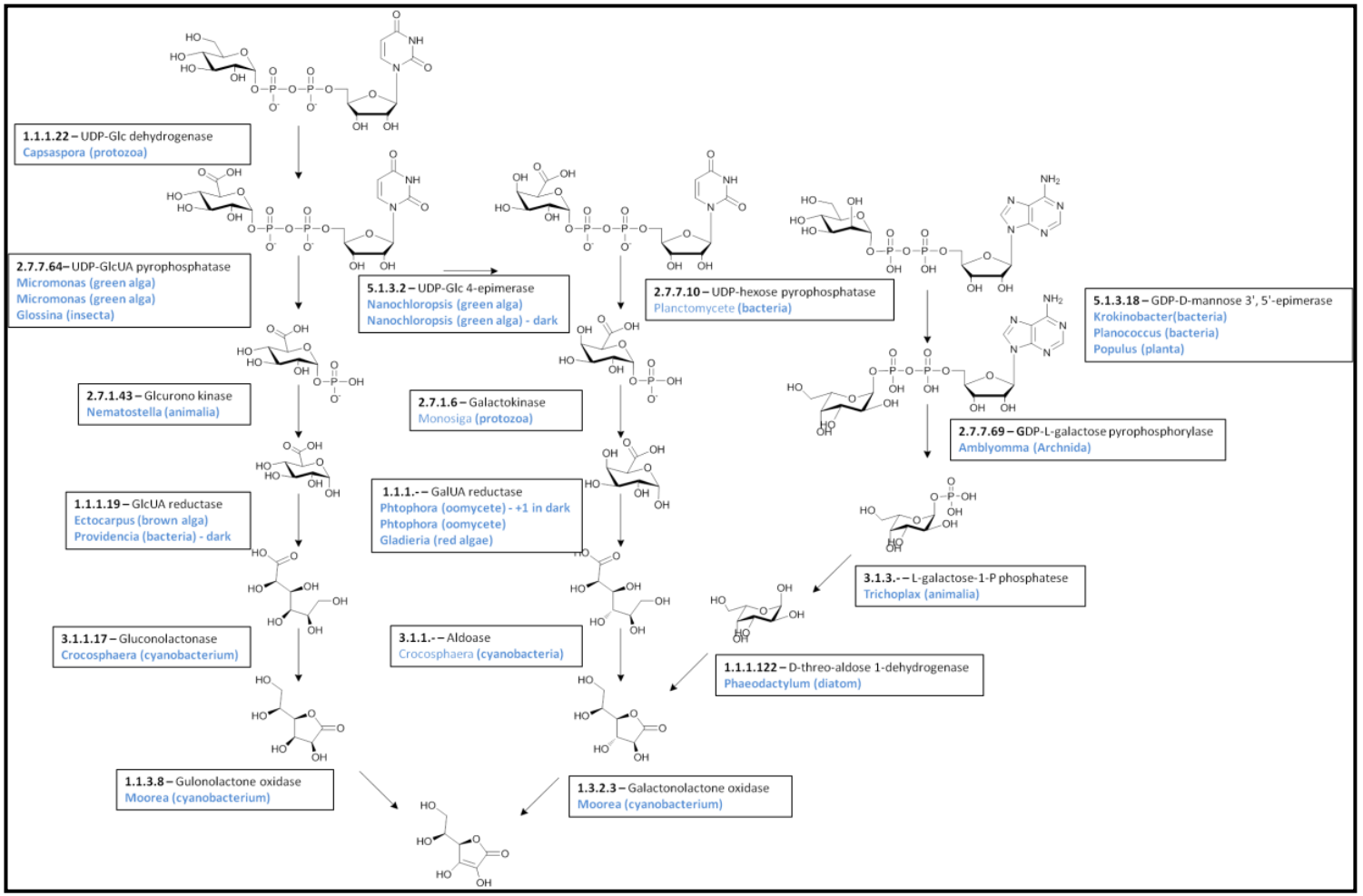


Figure 6.16: Ascorbic acid biosynthesis

Ascorbic acid is synthesised from sugar nucleotides. Candidate genes are present for every reaction in the biosynthesis of ascorbate but several represent weak assignments, particularly as prediction of sugar based substrates is notoriously difficult. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.9 for details of the biosynthetic enzymes discussed.

6.5 Redox control

Small molecule antioxidants have a major role in many cellular processes, including maintenance of the internal redox balance and xenobiotic detoxification. Whilst the major redox modulating thiol is glutathione, many parasites use unusual compounds which represent good drug targets.[118] These include bacillithiol in certain bacilli,[119] mycothiol in actinobacteria[120] and ovothiol and trypanothione in Kinetoplastids.[121]

6.4.1 Vitamin C

Ascorbate is the most important antioxidant in photosynthesising organisms, reducing reactive oxygen species formed during photosynthesis.[122] This reaction is catalysed by ascorbate peroxidase[123] which, in *Euglena*, is localised in the cytosol rather than the chloroplast, and has a unique dimeric form.[124] Ascorbic acid is synthesised either via L-gulonolactone, as in animals,[125] or via L-galactonolactone, as in plants.[126] Radioactive-tracer experiments from the 1970s show *Euglena* possesses both pathways,[127] though the L-galactonolactone pathway is apparently dominant. More recently the aldonolactonase gene has been identified in *Euglena*[52] and the galacturonic acid reductase has been characterised,[128] but the remaining enzymes in the pathway remain to be identified. There are candidate genes present in the transcriptome for all enzymes in these pathways (Fig 6.16), although several of these are only tentative assignments, based on nearest homologues of several similar sequences. There are also transcripts encoding the enzymes for the conversion of mannose into ascorbate, as is proposed for higher plants,[129] which is unprecedented in *Euglena*.

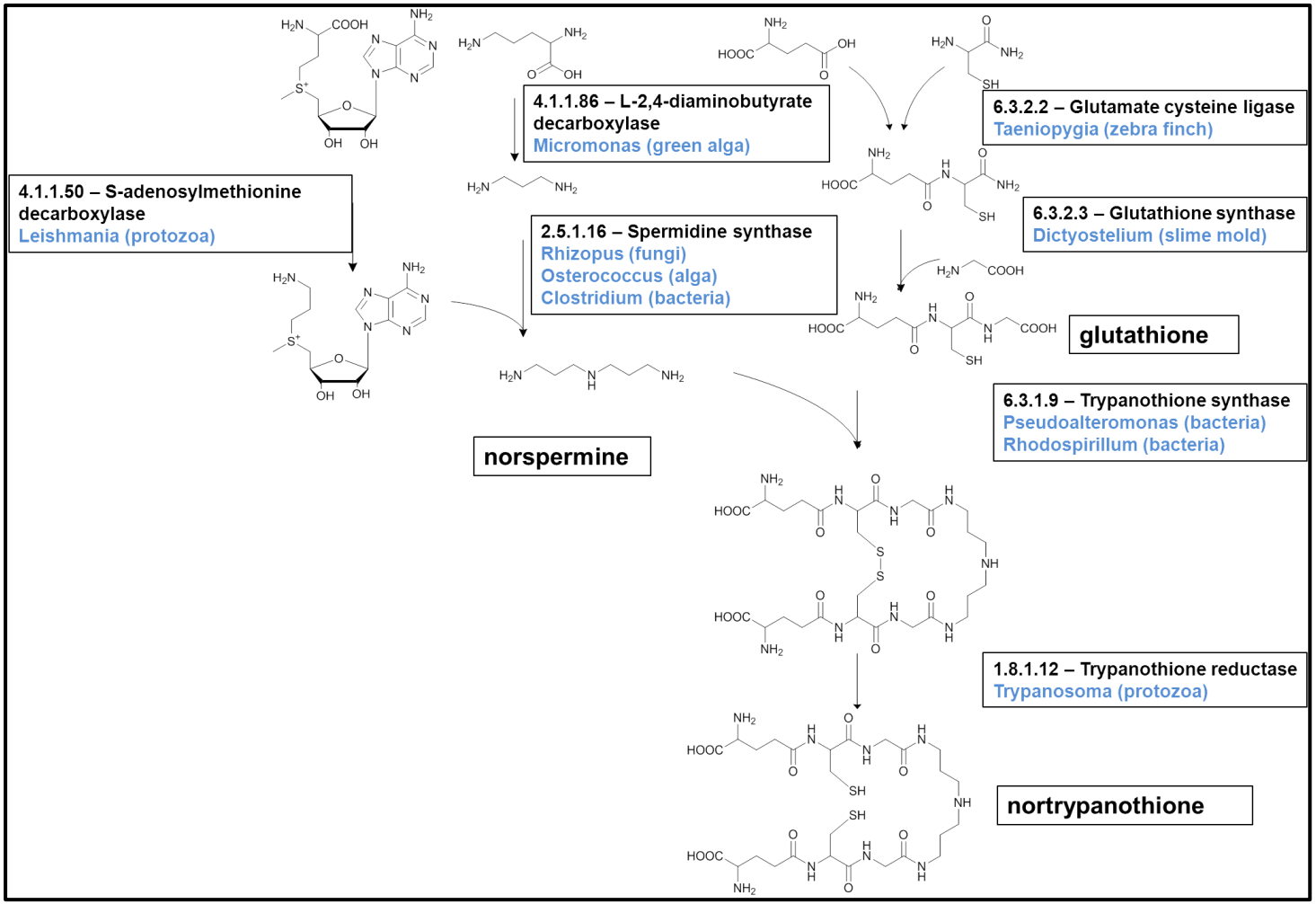


Figure 6.17: Proposed nortrypanothione biosynthetic pathway

Trypanothione is synthesised by joining one molecule of glutathione to each end of spermine. It is proposed that nortrypanothione is synthesised in the same way, attaching glutathione to norspermine, synthesised from diaminobutyrate. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.10 for details of the biosynthetic enzymes discussed.

6.4.2 Trypanothione

Independent reductases with exclusive activity towards either glutathione or trypanothione, of the tested thiols, have previously been purified from *Euglena*,[130] but the presence of the trypanothione substrate has not been demonstrated in this organism.

Trypanothione is formed from spermidine with one glutathione moiety at each end and can form intra and intermolecular disulphides. The conjugation of glutathione to spermidine to form trypanothione can be performed by one or two enzymes. In the two enzyme process, such as in *Crithidia*, glutathione is initially joined to spermidine by glutathionylspermidine synthase, followed by addition of another glutathione by trypanothione synthase.[131] In *Trypanosoma* a single enzyme catalyses the addition of glutathione to both ends of the spermidine.[132]

In the *Euglena* transcriptome the biosynthetic pathways for glutathione and polyamines are present, and there are two sequences which more closely match the trypanothione, rather than the glutathionylspermidine synthase of *Crithidia*, although they are most closely related to bacterial glutathionyl spermidine synthases (Fig 6.17). It is unclear whether these enzymes each carry out a single glutathione addition, or perform both conjugations.

6.4.3 Ovothiol

Ovothiols are a group of highly reducing antioxidant mercaptohistidines, which accumulate to very high levels in the eggs of certain marine invertebrates, including sea urchins, scallops and starfish,[133] where they act to scavenge H_2O_2 released during fertilisation.[134] They are also found in *Trypanosomes* and *Leishmania*[121] where they are involved in the oxidation of trypanothione. The only known enzyme for the biosynthesis of ovothiol, OvoA, contains a sulfoxide synthase at the N-terminus and an *N*-methyl transferase at the C-terminus.[135] The *Euglena* transcriptome encodes one sequence containing both domains, only found in the dark cells, and there are also separate sequences related to the N-terminus and to the C-terminus in both light and dark grown cells (Fig 6.18).

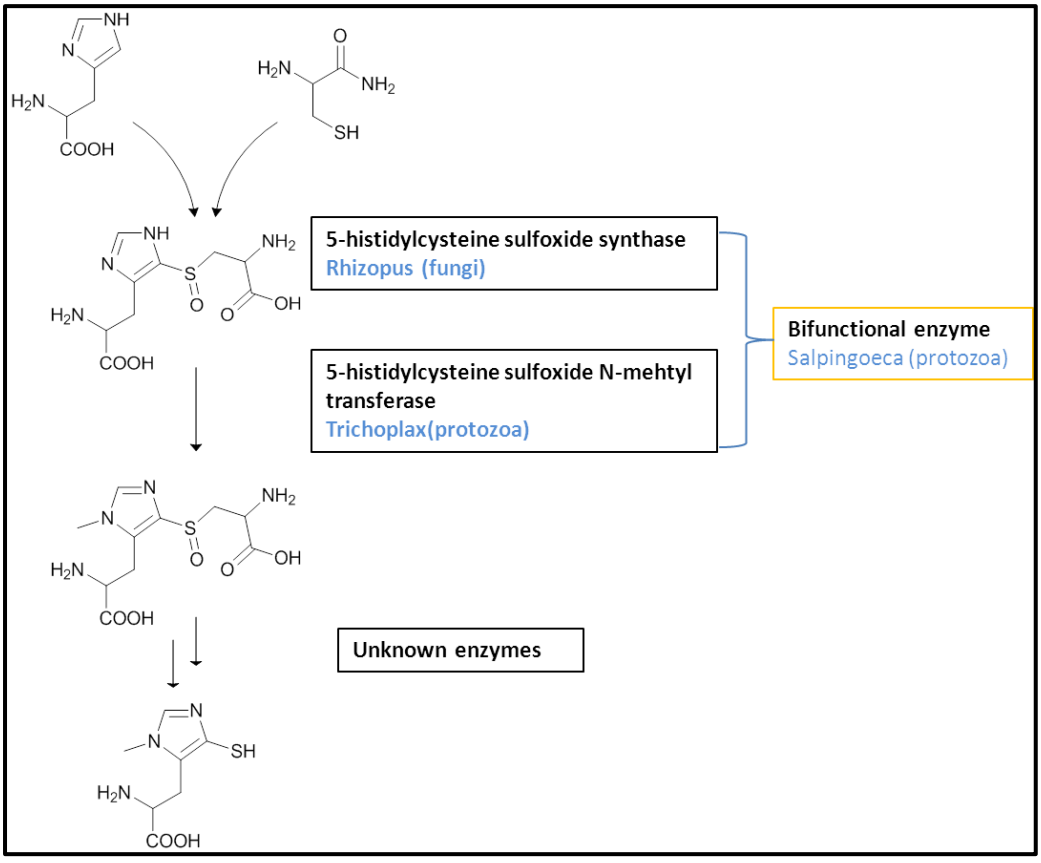


Figure 6.18: Ovothiol biosynthetic pathway

The only known enzyme for the biosynthesis of ovothiol contains a sulfoxide synthase and an *N*-methyl transferase. Sequences for one enzyme of each of these domains are present in the *Euglena* transcriptome, and one sequence containing both domains, only found in the dark grown cells. The genus of the closest homologue in the NCBI for each isoform is shown in blue. See Appendix 9.11 for details of the biosynthetic enzymes discussed.

6.4.4 Thiol analysis

In order to definitively identify trypanothione as having a physiological role in *Euglena* HPLC analysis was performed on extracts from cells that had been grown in high nutrient media with ambient light. The *Euglena* extracts clearly show the ubiquitous glutathione and cysteine peaks. There is a small peak co-eluting with trypanothione, suggesting this compound has some physiological relevance (Fig 6.19). LC-MS revealed this fraction contained both trypanothione and a compound with a molecular weight 14 Da lower (see section 8.5.3). MS2 shows that this corresponds to a one carbon shorter spermidine backbone than is usual in trypanothione, which we suggest as nortrypanothione. *Euglena* is known to make the norspermine and norspermidine from 1,3-diaminopropane, which is proposed to be derived from diaminobutyric acid.[136] This would require some flexibility in the polyamine chain length in the trypanothione synthase, as is seen in *T. cruzi*, which makes homo-trypanothione, with a one carbon longer spermidine, when exogenously supplied with cadavarine.[137]

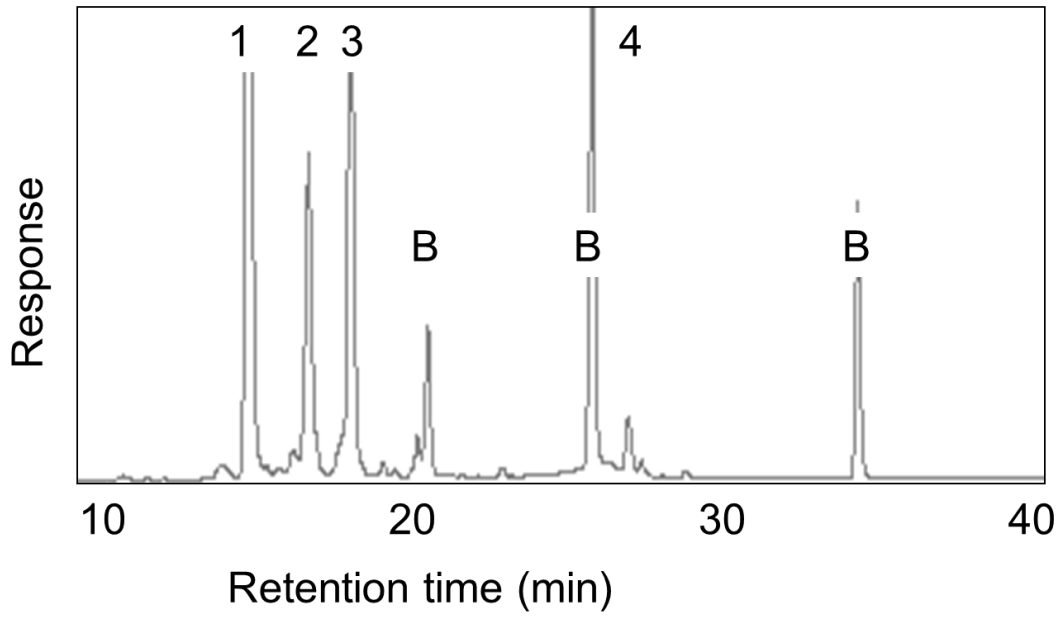
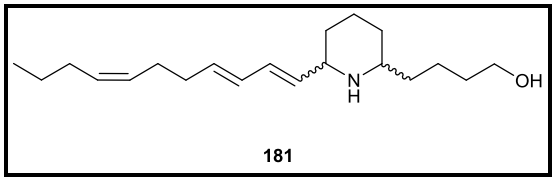


Figure 6.19 Analysis of the thiol content of *Euglena gracilis*

Thiols were labelled with monobromobimane and analysed by HPLC. Cysteine (1) and glutathione (3) can be clearly identified. There is a small peak (4) which coelutes with authentic trypanothione and a peak which elutes in the same region of the chromatogram as ovothiol (2).[121] These were confirmed by MS. B: background peaks.

6.5 Secondary metabolites

Only one secondary metabolite, the alkaloid euglenophycin (**181**), has been isolated and characterised from a *Euglena* (*E. sanguinea*), based on its toxicity to fish.[138] All four stereoisomers were identified, with one of the *cis* conformations being the major isomer isolated. There is no evidence for the production of this compound by *E. gracilis* (Paul Zimba, private communication) and the biosynthetic pathway leading to this compound has not been explored.



181

6.5.1 Indole alkaloids

Many plant secondary metabolites are based on the strictosidine skeleton. This contains an indole ring, derived from tryptophan by decarboxylation, joined to secologanin by strictosidine synthase.[139] In the *Euglena* transcriptome there is one sequence for a tryptophan decarboxylase, which may be involved in tryptophan degradation. Two potential strictosidine synthases are encoded in the transcriptome, though they are only around 20% homologous to the characterised *Rauvolfia* enzyme and appear to have an active site lacking the necessary glutamate to be a true synthase (Fig 6.20).[140] Instead they are likely to be paraoxonases, a highly conserved class of lactonases which act on diverse substrates including organophosphates, such as sarin,[141] drugs, such as lovastatin, and fatty acid lactones, which are their likely natural substrates.[142]

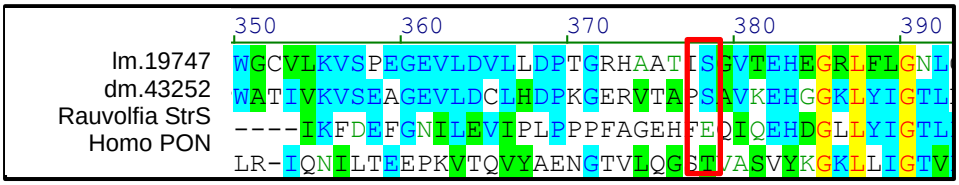


Figure 6.20: Alignment of candidate *Euglena* strictosidine synthases to known enzymes

The key glutamate for defining a strictosidine synthase (StrS)[140] over a paraoxonase (PON) is highlighted and indicates these enzymes may not be strictosidine synthases.

6.5.2 Polyketides

Polyketides are a wide range of compounds, including tetracycline antibiotics[143] and toxins such as the ciguatoxin,[144] formed by repeated condensation and dehydration of carboxylic acids, activated in the form of Coenzyme A thioesters. Broadly speaking polyketide synthases (PKSs) can be large multidomain proteins (type I) or composed of discreet proteins with individual functions (type II), although other architectures are possible.[145]

In order to identify polyketide synthases the translated *Euglena* transcriptome was searched for ketosynthase domains, the key catalytic domain, using BLASTP (see Appendix 11.1).[146] Fourteen potential PKSs were identified as having this domain: six sequences encode multidomain proteins indicative of type I PKSs; three sequences encode one catalytic domain and an acyl carrier domain; one sequence carries two catalytic domains; four sequences only encode individual ketosynthase domains. The latter may be true type II PKS modules or fragments of type I modules, caused by failure to obtain full length transcripts. They may also be enzymes with other functions that this search technique reveals, such as fatty acid synthases.

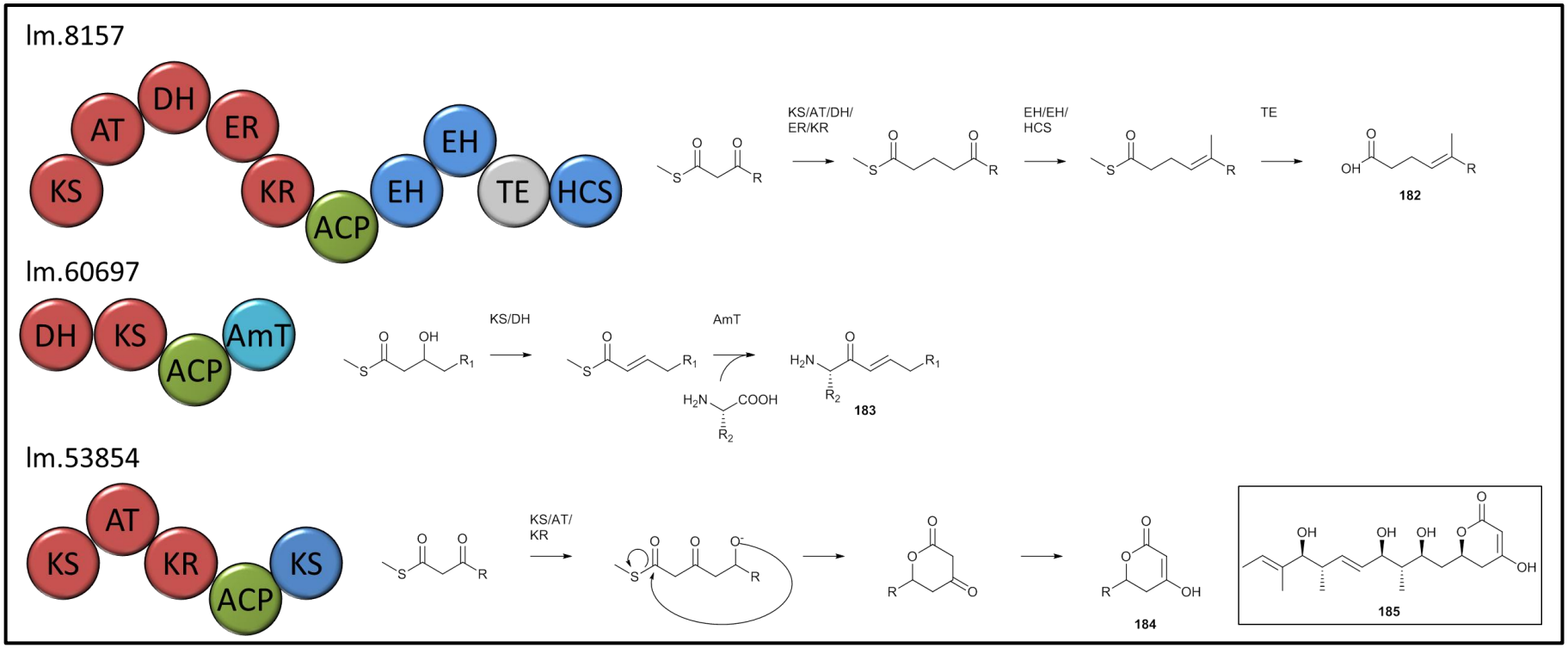


Figure 6.21: Polyketide synthases encoded in the *Euglena* transcriptome

Domain architectures of three candidate polyketide synthases which have multiple domains. Domains were identified using the Conserved Domain Database.[147] Reactions and partial products of the polyketide synthases were predicted, based on the arrangement of the domain architectures. lm.8157 has similar domain architecture to an erythromycin biosynthetic gene module, followed by three modules, which in bacillaene biosynthesis are three discrete proteins, which together add one branching methyl group, predicted to give a product similar to **182**.[148] Decarboxylative condensation of an amino acid to the polyketide in lm.60697, eliminating the thioester, would leave a free amine to give **183**. Reduction of a ketone, in lm.53854, can lead to cyclisation and elimination of the thioester, giving rise to a cyclic lactone (**184**), similar to the structure of the band 1 toxins in a fungal pathogen, *Alternaria citri* (**185**).[149]

Attempts to predict the structures of the polyketides synthesised by these enzymes, based on domain architectures, using the web based prediction programs SBSPKS[150] and PKS/NRPS Analysis Web-site[151] were not successful. This is probably due to the evolutionary distance from the bacterial and fungal species that these pieces of software were designed to deal with. Analysis of these enzymes, using DELTA-BLAST, which searches based on the domain architecture, rather than primary sequence, has allowed some product prediction (Fig 6.21).

The largest polyketide synthase encoded in the *Euglena* transcriptome (lm.8157) contains, aside from a fully reducing PKS module, two enoyl hydratases and an HMGCoA synthase. These proteins have been characterised, as single domain proteins, in some bacterial gene clusters, such as PksG, H and I in bacillaene biosynthesis, to add a β-methyl branch to polyketides (Fig 6.22).[148] Whilst the association of other components of the complete synthase cannot be predicted, this domain architecture suggests the formation of methyl branched alkene, which could be part of a polyketide, or alternatively may be included in a fatty acid.

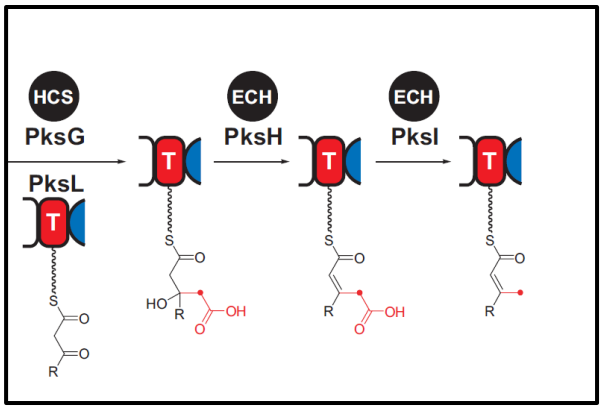


Figure 6.22: Insertion of a β-methyl branch on PksL in the biosynthesis of Bacillaene

The HMGCoA synthase and two enoyl hydratase domains in the *Euglena* transcript lm.8157 match the three domains on PksG, H and I which insert a β-methyl branch during Baciallaene biosynthesis. Reprinted from Butcher *et. al*.[148] Copyright (2007) National Academy of Sciences, USA.

One of the *Euglena* transcripts (lm.60697) encodes a small multidomain PKS containing an aminotransferase domain. These domains are found in diverse enzymes, including sphingolipid[152] and biotin biosynthesis,[153] where they catalyse decarboxylative condensations, releasing the product. In the *Euglena* enzyme this will add a nitrogen atom to the polyketide, together with the amino acid side chain (Fig 6.21).

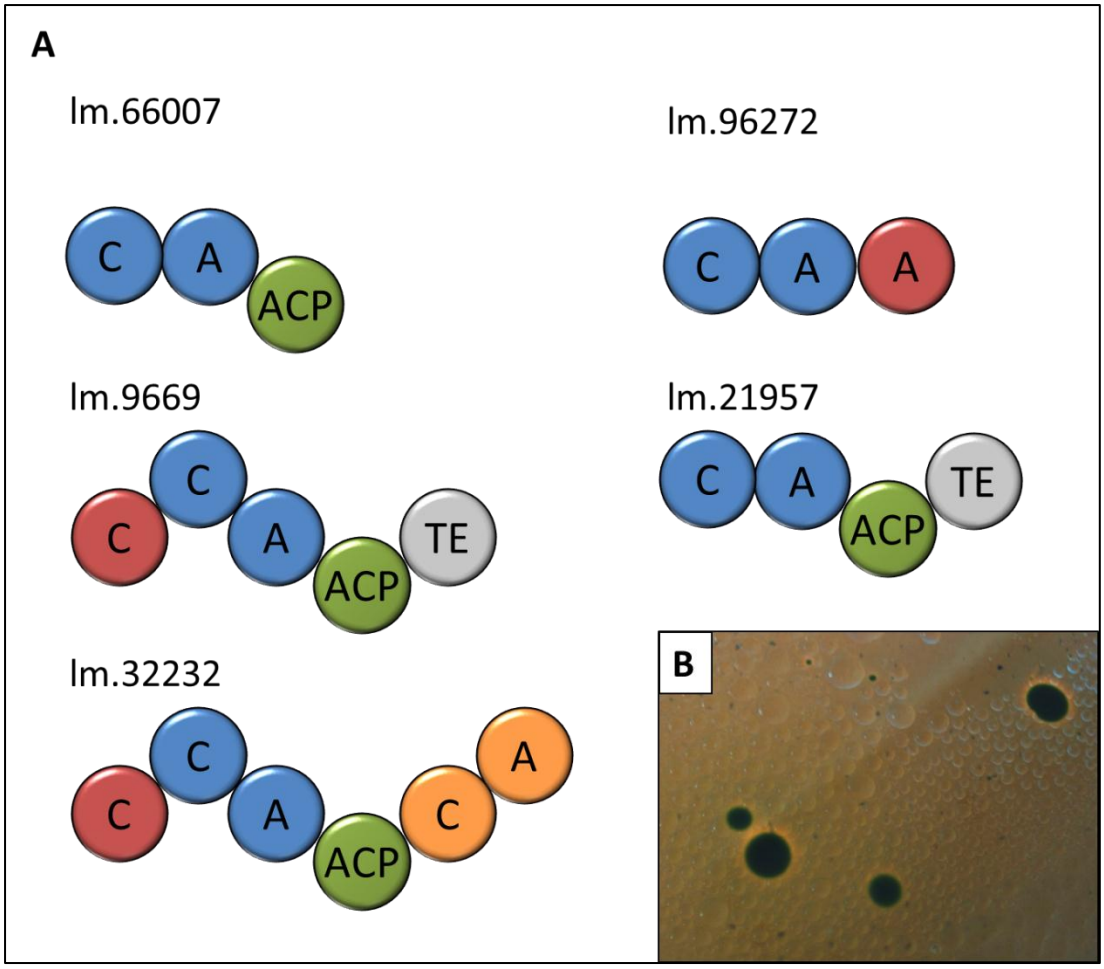


Figure 6.23: Nonribosomal peptide synthases encoded in the *Euglena* transcriptome

A. Domain architectures of three representative candidate non-ribosomal peptide synthases from *Euglena*. Domains were identified using the Conserved Domain Database.[147] **B**. Siderophore production by *E. gracilis*. Siderophore indicator plates were made and algae were plated out to get single colonies. The halo around these colonies, particularly vivid when viewed with an orange filter, indicates the production of a siderophore.

The third highlighted polyketide synthase from *Euglena* (lm.53854) contains a dehydratase and no thioesterase domain. If instead cyclisation cleaves the thioester, this could lead to a structure related to the toxin produced by the phytopathogen *Alternaria citri*.[149]

6.5.3 Nonribosomal peptides

Nonribosomal peptides synthases (NRPS) are modular multidomain proteins[154,155] which join amino acids to form small molecules with a diversity of function from siderophores, such as enterobactin,[156] to antibiotics, such as actinomycin,[157] and toxins, such as microcystin.[158] No nonribosomal peptides have been isolated from Euglenoids to date. The *E. gracilis* transcriptome was searched for sequences encoding adenylation domains (A), which found 19 transcripts, and condensation domains (C) which identified 16 transcripts (see Appendix 11.2). Five transcripts encode proteins with both A and C domains (Fig 6.23.A), but this search strategy also revealed many sequences which are most closely related to enzymes involved in fatty acid biosynthesis and two transcripts for L-aminoadipate-semialdehyde dehydrogenase, involved in lysine biosynthesis. Notably three adenylation-domain encoding transcripts also have ankyrin domains, involved in protein-protein interactions,[159] which may implicate formation of multi-protein NRPSs. The prediction programs, SBSPKS and NRPS/PKS Analysis Web-site, are again incapable of dealing with these sequences. In order to obtain iron from the environment *Euglena* produce siderophores (Fig 6.23.B), which may be synthesised using NRPSs, analogously with the *E. coli* siderophore enterobactin.[156]

6.6 Conclusions of *Euglena* transcriptome sequencing

Euglena are sometimes considered to be related to both animals and plants. In the *Euglena* transcriptome reported here there are some metabolic pathways that are more similar to plants and some that are more similar to animals, and often more than one pathway exists for the biosynthesis of the same compound. There are many unusual and unique enzyme architectures used by *Euglena*, including many novel gene fusions, which suggest there is some selective pressure in this organism, enhancing the acquisition of exotic enzymology.

No polyketides or non-ribosomal peptides have been found to date in this species, but there are transcripts apparent for the complex secondary metabolite synthases needed to make such compounds, as has been found from an increasing range of algae as genome sequencing data becomes available.[160] We have also identified sequences for the enzymes necessary to make unusual fatty acids and thiols and confirmed the existence of these compounds in this organism.

This first transcriptome of *Euglena* has shown that even in relatively well studied organisms there are many unexpected and unusual metabolic pathways and compounds that may prove useful in biotechnological applications.

6.7 References

1 Dobell, C. & van Leeuwenhoek, A. *Antony van Leeuwenhoek and his "Little animals": being some account of the father of protozoology and bacteriology and his multifarious discoveries in these disciplines; collected, translated, and edited, from his printed works, unpublished manuscripts, and contemporary records*. (Bale., 1932).

2 Van Leeuwenhoek, A. Letter 6. *Phil. Trans. R. Soc.* **9**, 178-182 (1674).

3 Harris, J. Some microscopical observations of vast numbers of aimalcula seen in water *Phil. Trans.* **19**, 254-259 (1695).

4 Linton, E. W. *et al.* Reconstructing Euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of euglenaria gen. nov (Euglenophyta). *Protist* **161**, 603-619 (2010).

5 Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475-W478 (2011).

6 Guttman, H. N. & Funk, H. B. Vitamin B12-controlled biotin avidity in autotrophic *Euglena gracilis*. *Nature* **213**, 103-104 (1967).

7 Brayner, R. *et al.* ZnO nanoparticles: synthesis, characterization, and ecotoxicological studies. *Langmuir* **26**, 6522-6528 (2010).

8 Ozasa, K., Lee, J., Song, S., Hara, M. & Maeda, M. *Euglena*-based neurocomputing with two-dimensional optical feedback on swimming cells in micro-aquariums. *Appl. Soft Comput.* **13**, 527-538 (2013).

9 Brayner, R. *et al.* Intracellular biosynthesis of superparamagnetic 2-lines ferri-hydrite nanoparticles using *Euglena gracilis* microalgae. *Colloids Surf., B* **93**, 20-23 (2012).

10 Strauch, S. M., Richter, P., Schuster, M. & Häder, D. P. The beating pattern of the flagellum of *Euglena gracilis* under altered gravity during parabolic flights. *J. Plant Physiol.* **167**, 41-46 (2010).

11 Häderi, D.-P., Rosumi, A., Schäfer, J. & Hemmersbach, R. Gravitaxis in the flagellate *Euglena gracilis* is controlled by an active gravireceptor. *J. Plant Physiol.* **146**, 474-480 (1995).

12 Takeyama, H. *et al.* Production of antioxidant vitamins, β-carotene, vitamin C, and vitamin E, by two-step culture of *Euglena gracilis* Z. *Biotechnol. Bioeng.* **53**, 185-190 (1997).

13 Korn, E. D. The fatty acids of *Euglena gracilis*. *J. Lipid Res.* **5**, 352-362 (1964).

14 Rodríguez-Zavala, J. S., Ortiz-Cruz, M. A., Mendoza-Hernández, G. & Moreno-Sánchez, R. Increased synthesis of α-tocopherol, paramylon and tyrosine by *Euglena gracilis* under conditions of high biomass production. *J. Applied Microbiol..* **109**, 2160-2172 (2010).

15 Inui, H., Miyatake, K., Nakano, Y. & Kitaoka, S. Wax ester fermentation in *Euglena gracilis*. *FEBS Lett.* **150**, 89-93 (1982).

16 Tucci, S., Vacula, R., Krajcovic, J., Proksch, P. & Martin, W. Variability of wax ester fermentation in natural and bleached *Euglena gracilis* strains in response to oxygen and the elongase inhibitor flufenacet. *J. Eukaryot. Microbiol.* **57**, 63-69 (2010).

17 Dasgupta, S., Fang, J., Brake, S. S., Hasiotis, S. T. & Zhang, L. Biosynthesis of sterols and wax esters by *Euglena* of acid mine drainage biofilms: implications for eukaryotic evolution and the early Earth. *Chem. Geol.* **306–307**, 139-145 (2012).

18 Park, K. Y., Lim, B.-R. & Lee, K. Growth of microalgae in diluted process water of the animal wastewater treatment plant. *Water Sci. Technol.* **59**, 2111-2116 (2009).

19 Moreira, D., Le Guyader, H. & Philippe, H. The origin of red algae and the evolution of chloroplasts. *Nature* **405**, 69-72 (2000).

20 Maruyama, S., Suzaki, T., Weber, A. P. M., Archibald, J. M. & Nozaki, H. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in Euglenids. *BMC Evol. Biol.* **11** (2011).

21 Gockel, G. & Hachtel, W. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* **151**, 347-351 (2000).

22 Martin, W., Somerville, C. C. & Loiseauxdegoer, S. Molecular phylogenies of plastid origins and algal evolution. *J. Mol. Evol.* **35**, 385-404 (1992).

23 Hallick, R. B. *et al.* Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* **21**, 3537-3544 (1993).

24 Spencer, D. & Gray, M. Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Mol. Genet. Genomics.* **285**, 19-31 (2011).

25 Borst, P. & Sabatini, R. Base J: discovery, biosynthesis, and possible functions. *Annu. Rev. Microbiol.* **62**, 235-251 (2008).

26 Breckenridge, D. G., Watanabe, Y.-i., Greenwood, S. J., Gray, M. W. & Schnare, M. N. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 852-856 (1999).

27 Tessier, L.-H., Chan, R. L., Keller, M., Weil, J.-H. & Imbault, P. The *Euglena gracilis* rbcS gene contains introns with unusual borders. *FEBS Lett.* **304**, 252-255 (1992).

28 Tessier, L. H. *et al.* Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *EMBO J.* **10**, 2621-2625 (1991).

29 Rismani-Yazdi, H., Haznedaroglu, B. Z., Bibby, K. & Peccia, J. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics* **12** (2011).

30 Shoguchi, E. *et al.* Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* **23**, 1399-1408 (2013).

31 Jaeckisch, N. *et al.* Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. *PLoS ONE* **6**, e28012 (2011).

32 Steele, R. E. & Rae, P. M. M. Ordered distribution of modified bases in the DNA of a dinoflagellate. *Nucleic Acids Res.* **8**, 4709-4726 (1980).

33 Fritz-Laylin, L. K. *et al.* The genome of *Naegleria gruberi* illuminates early Eukaryotic versatility. *Cell* **140**, 631-642 (2010).

34 Arroyo, M., Heltai, L., Millan, D. & DeSimone, A. Reverse engineering the Euglenoid movement. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17874-17879 (2012).

35 Ramesh, M. A., Malik, S.-B. & Logsdon Jr, J. M. A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early Eukaryotic origin of meiosis. *Curr. Biol.* **15**, 185-191 (2005).

36 Goto, K. & Beneragama, C. K. Circadian clocks and antiaging: do non-aging microalgae like *Euglena* reveal anything? *Ageing Res. Rev.* **9**, 91-100 (2010).

37 Blanc, G. *et al.* The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* **22**, 2943-2955 (2010).

38 Grimsley, N., Pequin, B., Bachy, C., Moreau, H. & Piganeau, G. Cryptic sex in the smallest Eukaryotic marine green alga. *Mol. Biol. Evol.* **27**, 47-54 (2010).

39 McKim, K. S. *et al.* Meiotic synapsis in the absence of recombination. *Science* **279**, 876-878 (1998).

40 Wall, M. K., Mitchenall, L. A. & Maxwell, A. *Arabidopsis thaliana* DNA gyrase is targeted to chloroplasts and mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7821-7826 (2004).

41 Ebringer, L., Polonyi, J. & Krajcovic, J. Influence of ofloxacin on chloroplasts and mitochondria in *Euglena gracilis*. *Arzneim.-Forsch.* **43-2**, 777-781 (1993).

42 Polónyi, J., Ebringer, L., Dobias, J. & Krajčovič, J. Giant mitochondria in chloroplast-deprived *Euglena gracilis* late after *N*-succinimidylofloxacin treatment. *Folia Microbiol.* **43**, 661-666 (1998).

43 Ebringer, L., Mego, J. L., Jurasek, A. & Kada, R. Action of streptomycins on chloroplast system of *Euglena gracilis*. *J. Gen. Microbiol.* **59**, 203-& (1969).

44 Brodersen, P. & Voinnet, O. The diversity of RNA silencing pathways in plants. *Trends Genet.* **22**, 268-280 (2006).

45 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181-190 (2010).

46 Baulcombe, D. RNA silencing in plants. *Nature* **431**, 356-363 (2004).

47 Berezikov, E. & Plasterk, R. H. A. Camels and zebrafish, viruses and cancer: a microRNA update. *Human Molecular Genetics* **14**, R183-R190 (2005).

48 Liu, Q., Feng, Y. & Zhu, Z. Dicer-like (DCL) proteins in plants. *Funct. Integr. Genomics* **9**, 277-286 (2009).

49 Hutvagner, G. & Simard, M. J. Argonaute proteins: key players in RNA silencing. *Nat. Rev. Mol. Cell Biol.* **9**, 22-32 (2008).

50 Baulcombe, D. C. Amplified silencing. *Science* **315**, 199-200 (2007).

51 Cerutti, H., Ma, X., Msanne, J. & Repas, T. RNA-mediated silencing in algae: biological roles and tools for analysis of gene function. *Eukaryot. Cell* **10**, 1164-1172 (2011).

52 Ishikawa, T. *et al.* The pathway via D-galacturonate/L-galactonate is significant for ascorbate biosynthesis in *Euglena gracilis*. *J. Biol. Chem.* **283**, 31133-31141 (2008).

53 Ntefidou, M., Iseki, M., Watanabe, M., Lebert, M. & Häder, D.-P. Photoactivated adenylyl cyclase controls phototaxis in the flagellate *Euglena gracilis*. *Plant Physiol.* **133**, 1517-1521 (2003).

54 Fahlgren, N. *et al.* Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in *Arabidopsis*. *Curr. Biol.* **16**, 939-944 (2006).

55 van Leeuwen, F. *et al.* β-D-Glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 2366-2371 (1998).

56 van Luenen, H. G. A. M. *et al.* Glucosylated hydroxymethyluracil, DNA Base J, prevents transcriptional readthrough in *Leishmania*. *Cell* **150**, 909-921 (2012).

57 Dooijes, D. *et al.* Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. *Nucleic Acids Res.* **28**, 3017-3021 (2000).

58 DiPaolo, C., Kieft, R., Cross, M. & Sabatini, R. Regulation of *Trypanosome* DNA glycosylation by a SWI2/SNF2-like protein. *Molecular Cell* **17**, 441-451 (2005).

59 Genest, P.-A. *et al.* Formation of linear inverted repeat amplicons following targeting of an essential gene in *Leishmania*. *Nucleic Acids Res.* **33**, 1699-1709 (2005).

60 Cliffe, L. J. *et al.* JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. *Nucleic Acids Res.* **37**, 1452-1462 (2009).

61 Hrdá, Š., Fousek, J., Szabová, J., Hampl, V. V. & Vlček, Č. The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in Euglenids. *PLoS ONE* **7**, e33746 (2012).

62 Mus, F., Dubini, A., Seibert, M., Posewitz, M. C. & Grossman, A. R. Anaerobic acclimation in *Chlamydomonas reinhardtii*: Anoxic gene expression, hydrogenase induction, and metabolic pathways. *J. Biol. Chem.* **282**, 25475-25486 (2007).

63 Matsuda, F., Hayashi, M. & Kondo, A. Comparative profiling analysis of central metabolites in *Euglena gracilis* under various cultivation conditions. *Biosci., Biotechnol., Biochem.* **75**, 2253-2256 (2011).

64 Ruxton, C. H. S., Reed, S. C., Simpson, M. J. A. & Millington, K. J. The health benefits of omega-3 polyunsaturated fatty acids: a review of the evidence. *J. Hum. Nutr. Diet.* **17**, 449-459 (2004).

65 Wallis, J. G. & Browse, J. The Δ8-desaturase of *Euglena gracilis*: an alternate pathway for synthesis of 20-carbon polyunsaturated fatty acids. *Arch. Biochem. Biophys.* **365**, 307-316 (1999).

66 Meyer, A. *et al.* Biosynthesis of docosahexaenoic acid in *Euglena gracilis*: biochemical and molecular evidence for the involvement of a Δ4-fatty acyl group desaturase. *Biochemistry* **42**, 9779-9788 (2003).

67 Pollak, D. W. *et al.* Isolation of a delta 5 desaturase gene from *Euglena gracilis* and functional dissection of its HPGG and HDASH motifs. *Lipids* **47**, 913-926 (2012).

68 Grogan, D. W. & Cronan, J. E. Cyclopropane ring formation in membrane lipids of bacteria. *Microbiol. Mol. Biol. Rev.* **61**, 429-441 (1997).

69 Yu, X.-H., Rawat, R. & Shanklin, J. Characterization and analysis of the cotton cyclopropane fatty acid synthase family and their contribution to cyclopropane fatty acid synthesis. *BMC Plant Biol.* **11**, 97 (2011).

70 Fish, W. R., Holz Jr, G. G., Beach, D. H., Owen, E. & Anekwe, G. E. The cyclopropane fatty acid of trypanosomatids. *Mol. Biochem. Parasit.* **3**, 103-115 (1981).

71 Fabrias, G., Gosalbo, L., Quintana, J. & Camps, F. Direct inhibition of (Z)-9 desaturation of (E)-11-tetradecenoic acid by methylenehexadecenoic acids in the biosynthesis of *Spodoptera littoralis* sex pheromone. *J. Lipid Res.* **37**, 1503-1509 (1996).

72 Christine, W. W. in *Topics in Lipid Chemistry* Vol. 1 (ed F.D. Gunstone) 1-49 (Logos press, 1970).

73 Kai, Y. & Pryde, E. H. Production of branched-chain fatty acids from *Sterculia* oil. *J. Am. Oil Chem. Soc.* **59**, 300-305 (1982).

74 Maschek, J. A. & Baker, B. J. in *Algal Chemical Ecology* (ed C.D. Amsler) 1-24 (Springer, 2007).

75 Fernandez, J. J., Souto, M. L. & Norte, M. Marine polyether triterpenes. *Nat. Prod. Rep.* **17**, 235-246 (2000).

76 Fenical, W. & Paul, V. Antimicrobial and cytotoxic terpenoids from tropical green algae of the family Udoteaceae. *Hydrobiologia* **116-117**, 135-140 (1984).

77 Kim, D., Filtz, M. R. & Proteau, P. J. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in *Euglena gracilis*. *J. Nat. Prod.* **67**, 1067-1069 (2004).

78 Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201 (2006).

79 Kavanagh, K. L., Dunford, J. E., Bunkoczi, G., Russell, R. G. G. & Oppermann, U. The crystal structure of human geranylgeranyl pyrophosphate dynthase reveals a novel hexameric arrangement and inhibitory product binding. *J. Biol. Chem.* **281**, 22004-22012 (2006).

80 Anding, C., Brandt, R. D. & Ourisson, G. Sterol biosynthesis in *Euglena gracilis* Z. *Eur. J. Biochem.* **24**, 259-263 (1971).

81 Stern, A. I., Schiff, J. A. & Klein, H. P. Isolation of ergosterol from *Euglena gracilis*: distribution among mutant strains. *J. Eukaryot. Microbiol.* **7**, 52-55 (1960).

82 Thoma, R. *et al.* Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432**, 118-122 (2004).

83 Munne-Bosch, S. & Alegre, L. The function of tocopherols and tocotrienols in plants. *Crit. Rev. Plant Sci.* **21**, 31-57 (2002).

84 Ogbonna, J. C. Microbiological production of tocopherols: current state and prospects. *Appl. Microbiol. Biotechnol.* **84**, 217-225 (2009).

85 Velasco, A. M., Leguina, J. I. & Lazcano, A. Molecular evolution of the lysine biosynthetic pathways. *J. Mol. Evol.* **55**, 445-459 (2002).

86 Torruella, G., Suga, H., Riutort, M., Pereto, J. & Ruiz-Trillo, I. The evolutionary history of lysine biosynthesis pathways within Eukaryotes. *J. Mol. Evol.* **69**, 240-248 (2009).

87 Vogel, H. J. Lysine biosynthesis in *Chlorella* and *Euglena*: phylogenetic significance. *Biochim. Biophys. Acta* **34**, 282-283 (1959).

88 Mark Zabriskie, T. & D. Jackson, M. Lysine biosynthesis and metabolism in fungi. *Nat. Prod. Rep.* **17**, 85-97 (2000).

89 Vogel, H. J. & Davis, B. D. Glutamic γ-semialdehyde and Δ1-pyrroline-5-carboxylic acid, intermediates in the biosynthesis of proline. *J. Am. Chem. Soc.* **74**, 109-112 (1952).

90 Costilow, R. N. & Laycock, L. Ornithine cyclase (deaminating): purification of a protein that converts ornithine to proline and definition of the optimal assay conditions. *J. Biol. Chem.* **246**, 6655-6660 (1971).

91 Park, B. S., Hirotani, A., Nakano, Y. & Kitaoka, S. The physiological role and catabolism of arginine in *Euglena gracilis*. *Agric. Biol. Chem.* **47**, 2561-2567 (1983).

92 Hill, D. L. & Vaneys, J. Relationship between arginine citrulline and ornithine in *Tetrahymena pyriformis*. *J. Protozool* **12**, 259-265 (1965).

93 Kaplan, M. M. & Flavin, M. Threonine biosynthesis: on the pathway in fungi and bacteria and the mechanism of the isomerization reaction *J. Biol. Chem.* **240**, 3928-3933 (1965).

94 Grimminger, H. & Umbarger, H. E. Acetohydroxy acid synthase I of *Escherichia coli*: purification and properties. *J. Bacteriol.* **137**, 846-853 (1979).

95 Oda, Y., Nakano, Y. & Kitaoka, S. Occurrence and some properties of two threonine dehydratases in *Euglena gracilis*. *J. Gen. Microbiol.* **129**, 57-61 (1983).

96 Drevland, R. M., Waheed, A. & Graham, D. E. Enzymology and evolution of the pyruvate pathway to 2-oxobutyrate in *Methanocaldococcus jannaschii*. *J. Bacteriol.* **189**, 4391-4400 (2007).

97 Maeda, H. & Dudareva, N. in *Annual Review of Plant Biology, Vol 63* Vol. 63 *Annual Review of Plant Biology* (ed S. S. Merchant) 73-105 (2012).

98 Duncan, K., Edwards, R. M. & Coggins, J. R. The pentafunctional AROM enzyme of *Saccharomyces cerevisiae* is a mosaic of monofunctional domains. *Biochem. J.* **246**, 375-386 (1987).

99 Richards, T. A. *et al.* Evolutionary origins of the Eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryot. Cell* **5**, 1517-1531 (2006).

100 Reinbothe, C., Ortel, B., Parthier, B. & Reinbothe, S. Cytosolic and plastid forms of 5-enolpyruvylshikimate-3-phosphate synthase in *Euglena gracilis* are differentially expressed during light-induced chloroplast development. *Mol. Gen. Genet.* **245**, 616-622 (1994).

101 Zhang, S. *et al.* Chorismate mutase-prephenate dehydratase from *Escherichia coli*: study of catalytic and regulatory domains using genetically engineered proteins. *J. Biol. Chem.* **273**, 6248-6253 (1998).

102 Christopherson, R. I. & Morrison, J. F. Chorismate mutase-prephenate dehydrogenase from Escherichia coli: positive cooperativity with substrates and inhibitors. *Biochemistry* **24**, 1116-1121 (1985).

103 Lim, S. *et al.* Characterization of a key trifunctional enzyme for aromatic amino acid biosynthesis in *Archaeoglobus fulgidus*. *Extremophiles* **13**, 191-198 (2009).

104 Crawford, I. P. Gene rearrangements in evolution of tryptophan synthetic pathway. *Bacteriol. Rev.* **39**, 87-120 (1975).

105 Cohn, W., Kirschner, K. & Paul, C. N-(5-Phosphoribosyl)anthranilate isomerase-indoleglycerol-phosphate synthase. 2. Fast-reaction studies show that a fluorescent substrate analog binds independently to two different sites. *Biochemistry* **18**, 5953-5959 (1979).

106 Schwarz, T. *et al.* Multifunctional tryptophan-synthesizing enzyme: the molecular weight of the *Euglena gracilis* protein is unexpectedly low. *J. Biol. Chem.* **272**, 10616-10623 (1997).

107 Miles, E. W. in *Adv. Enzymol. Relat. Areas Mol. Biol.* 93-172 (John Wiley & Sons, Inc., 2006).

108 Matchett, W. H. & DeMoss, J. A. The subunit structure of tryptophan synthase from *Neurospora crassa*. *J. Biol. Chem.* **250**, 2941-2946 (1975).

109 Shemin, D. & Rittenberg, D. The utilization of glycine for the synthesis of a porphyrin. *J. Biol. Chem.* **159**, 567-568 (1945).

110 Beale, S. I. & Castelfranco, P. A. Biosynthesis of delta-aminolevulinic-acid in higher-plants.1. Accumulation of delta-aminolevulinic-acid in greening plant-tissues. *Plant Physiol.* **53**, 291-296 (1974).

111 Iida, K., Mimura, I. & Kajiwara, M. Evaluation of two biosynthetic pathways to δ-aminolevulinic acid in *Euglena gracilis*. *Eur. J. Biochem.* **269**, 291-297 (2002).

112 Gaffron, H. & Rubin, J. Fermentative and photochemical production of hydrogen in algae. *J. Gen. Physiol.* **26**, 219-240 (1942).

113 Hemschemeier, A., Fouchard, S., Cournac, L., Peltier, G. & Happe, T. Hydrogen production by *Chlamydomonas reinhardtii*: an elaborate interplay of electron sources and sinks. *Planta* **227**, 397-407 (2008).

114 Posewitz, M. C. *et al.* Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. *J. Biol. Chem.* **279**, 25711-25720 (2004).

115 Putz, S. *et al.* Fe-hydrogenase maturases in the hydrogenosomes of *Trichomonas vaginalis*. *Eukaryot. Cell* **5**, 579-586 (2006).

116 Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005-1016 (2000).

117 Balk, J., Pierik, A. J., Netz, D. J. A., Mihlenhoff, U. & Lill, R. The hydrogenase-like Nar1p is essential for maturation of cytosolic and nuclear iron-sulphur proteins. *EMBO J.* **23**, 2105-2115 (2004).

118 Krauth-Siegel, R. L. & Leroux, A. E. Low-molecular-mass antioxidants in parasites. *Antioxid. Redox Signal.* **17**, 583-607 (2012).

119 Newton, G. L. *et al.* Bacillithiol is an antioxidant thiol produced in Bacilli. *Nat. Chem. Biol.* **5**, 625-627 (2009).

120 Koledin, T., Newton, G. & Fahey, R. Identification of the mycothiol synthase gene (mshD) encoding the acetyltransferase producing mycothiol in actinomycetes. *Arch. Microbiol.* **178**, 331-337 (2002).

121 Ariyanayagam, M. R. & Fairlamb, A. H. Ovothiol and trypanothione as antioxidants in trypanosomatids. *Mol. Biochem. Parasit.* **115**, 189-198 (2001).

122 Noctor, G. & Foyer, C. H. Ascorbate and glutathione: keeping active oxygen under control. *Annu. Rev. Plant Physiol. PLant Mol. Biol.* **49**, 249-279 (1998).

123 Ishikawa, T. & Shigeoka, S. Recent advances in ascorbate biosynthesis and the physiological significance of ascorbate peroxidase in photosynthesizing organisms. *Biosci., Biotechnol., Biochem.* **72**, 1143-1154 (2008).

124 Ishikawa, T. *et al. Euglena gracilis* ascorbate peroxidase forms an intramolecular dimeric structure: its unique molecular characterization. *Biochem. J.* **426**, 125-134 (2010).

125 Isherwood, F. A., Chen, Y. T. & Mapson, L. W. Synthesis of L-ascorbic acid in plants and animals. *Biochem. J.* **56**, 1-15 (1954).

126 Smirnoff, N. & Wheeler, G. L. Ascorbic acid in plants: biosynthesis and function. *Crit. Rev. Biochem. Mol. Biol.* **35**, 291-314 (2000).

127 Shigeoka, S., Nakano, Y. & Kitaoka, S. The biosynthetic pathway of L-ascorbic acid in *Euglena gracilis* Z. *J. Nutr. Sci. Vitaminol.* **25**, 299-307 (1979).

128 Ishikawa, T. *et al.* Functional characterization of D-galacturonic acid reductase, a key enzyme of the ascorbate biosynthesis pathway, from *Euglena gracilis*. *Biosci., Biotechnol., Biochem.* **70**, 2720-2726 (2006).

129 Wolucka, B. A. & Van Montagu, M. GDP-mannose 3′,5′-epimerase forms GDP-L-gulose, a putative intermediate for the *de novo* biosynthesis of vitamin C in plants. *J. Biol. Chem.* **278**, 47483-47490 (2003).

130 Montrichard, F., Le Guen, F., Laval-Martin, D. L. & Davioud-Charvet, E. Evidence for the co-existence of glutathione reductase and trypanothione reductase in the non-trypanosomatid Euglenozoa: *Euglena gracilis* Z. *FEBS Lett.* **442**, 29-33 (1999).

131 Tetaud, E. *et al.* Cloning and characterization of the two enzymes responsible for trypanothione biosynthesis in *Crithidia fasciculata*. *J. Biol. Chem.* **273**, 19383-19390 (1998).

132 Comini, M., Menge, U. & Flohe, L. Biosynthesis of trypanothione in *Trypanosoma brucei brucei*. *Biol. Chem.* **384**, 653-656 (2003).

133 Turner, E., Klevit, R., Hager, L. J. & Shapiro, B. M. Ovothiols, a family of redox-active mercaptohistidine compounds from marine invertebrate eggs. *Biochemistry* **26**, 4028-4036 (1987).

134 Turner, E., Hager, L. J. & Shapiro, B. M. Ovothiol replaces glutathione-peroxidase as a hydrogen-peroxide scavenger in sea-urchin eggs. *Science* **242**, 939-941 (1988).

135 Braunshausen, A. & Seebeck, F. P. Identification and characterization of the first ovothiol biosynthetic enzyme. *J. Am. Chem. Soc.* **133**, 1757-1759 (2011).

136 Villanueva, V. R., Adlakha, R. C. & Calvayrac, R. Biosynthesis of polyamines in *Euglena gracilis*. *Phytochemistry* **19**, 787-790 (1980).

137 Hunter, K. J., Quesne, S. A. L. & Fairlamb, A. H. Identification and biosynthesis of *N*1,*N*9-bis(glutathionyl)aminopropylcadaverine (homotrypanothione) in *Trypanosoma cruzi*. *Eur. J. Biochem.* **226**, 1019-1027 (1994).

138 Zimba, P. V., Moeller, P. D., Beauchesne, K., Lane, H. E. & Triemer, R. E. Identification of euglenophycin – a toxin found in certain Euglenoids. *Toxicon* **55**, 100-104 (2010).

139 O'Connor, S. E. & Maresh, J. J. Chemistry and biology of monoterpene indole alkaloid biosynthesis. *Nat. Prod. Rep.* **23**, 532-547 (2006).

140 Hicks, M. A. *et al.* The evolution of function in strictosidine synthase-like proteins. *Proteins: Struct., Funct., Bioinf.* **79**, 3082-3098 (2011).

141 Davies, H. G. *et al.* The effect of the human serum paraoxonase polymorphism is reversed with diazoxon, soman and sarin. *Nat. Genet.* **14**, 334-336 (1996).

142 Draganov, D. I. *et al.* Human paraoxonases (PON1, PON2, and PON3) are lactonases with overlapping and distinct substrate specificities. *J. Lipid Res.* **46**, 1239-1247 (2005).

143 Pickens, L. B. & Tang, Y. Decoding and engineering tetracycline biosynthesis. *Metab. Eng.* **11**, 69-75 (2009).

144 Murata, M., Legrand, A. M., Ishibashi, Y., Fukui, M. & Yasumoto, T. Structures and configurations of ciguatoxin from the moray eel Gymnothorax javanicus and its likely precursor from the dinoflagellate Gambierdiscus toxicus. *J. Am. Chem. Soc.* **112**, 4380-4386 (1990).

145 Shen, B. *et al.* in *Polyketides* Vol. 955 *ACS Symposium Series* Ch. 11, 154-166 (American Chemical Society, 2007).

146 Johnson, M. *et al.* NCBIBLAST: a better web interface. *Nucleic Acids Res.* **36**, W5-W9 (2008).

147 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225-D229 (2011).

148 Butcher, R. A. *et al.* The identification of bacillaene, the product of the PksX megacomplex in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1506-1509 (2007).

149 Kono, Y. *et al. Alternaria citri* toxins. 3. Plant pathotoxins from *Alternaria citri* - stereochemistry of the major and minor toxins. *Phytochemistry* **25**, 69-72 (1986).

150 Anand, S. *et al.* SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* **38**, W487-W496 (2010).

151 Bachmann, B. O. & Ravel, J. in *Complex enzymes in microbial natural product biosynthesis, part A: overview articles and peptides* Vol. 458 *Methods in Enzymology* (ed D. A. Hopwood) 181-217 (2009).

152 Ikushiro, H., Hayashi, H. & Kagamiyama, H. Bacterial serine palmitoyltransferase: a water-soluble homodimeric prototype of the eukaryotic enzyme. *Biochim. et Biophys. Acta - Proteins Proteom.* **1647**, 116-120 (2003).

153 Webster, S. P. *et al.* Mechanism of 8-amino-7-oxononanoate synthase: spectroscopic, kinetic, and crystallographic studies. *Biochemistry* **39**, 516-528 (1999).

154 Finking, R. & Marahiel, M. A. Biosynthesis of nonribosomal peptides. *Annu. Rev. Microbiol.* **58**, 453-488 (2004).

155 Hur, G. H., Vickery, C. R. & Burkart, M. D. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Nat. Prod. Rep.* **29**, 1074-1098 (2012).

156 Walsh, C. T., Liu, J., Rusnak, F. & Sakaitani, M. Molecular studies on enzymes in chorismate metabolism and the enterobactin biosynthetic pathway. *Chem. Rev.* **90**, 1105-1129 (1990).

157 Schauwecker, F., Pfennig, F., Schröder, W. & Keller, U. Molecular cloning of the actinomycin synthetase gene cluster from *Streptomyces chrysomallus* and functional heterologous expression of the gene encoding actinomycin synthetase II. *J. Bacteriol.* **180**, 2468-2474 (1998).

158 Noguchi, T. *et al.* Genetic analysis of the microcystin biosynthesis gene cluster in *Microcystis* strains from four bodies of eutrophic water in Japan. *The Journal of General and Applied Microbiology* **55**, 111-123 (2009).

159 Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. Y. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci.* **13**, 1435-1448 (2004).

160 Sasso, S., Pohnert, G., Lohr, M., Mittag, M. & Hertweck, C. Microalgae in the postgenomic era: a blooming reservoir for new natural products. *FEMS Microbiol. Rev.* **36**, 761-785 (2012).

Chapter 7 – Future directions

The chemical synthesis of defined oligosaccharides is technically challenging, but the use of enzymes can expedite this. Phosphorylases have proved a useful class of enzymes for the synthesis of long oligosaccharides from relatively cheap starting materials. New product specificities need to be discovered or engineered to fully realise the potential of phosphorylases for carbohydrate synthesis.

7.1 PHS2

7.1.1 Further crystallography of PHS2

The crystal structures of PHS2 presented herein have not definitively elucidated the active site of this enzyme. The binding of acarbose indicates there are certain differences in the active site compared to the other α-1,4-glucan phosphorylase structures available. This suggests further studies are required; to produce more refined structures for the Glc-1-P binding site and to bind oligosaccharides in the active site channel. Of particular interest are the oligosaccharides related to the substrate SHG, the structure of which has not been fully elucidated.[1]

PHS1 is the plastid localised α-1,4-glucan phosphorylase, and has a role in synthesis and degradation of starch granules.[2] The sequence is similar to PHS2 (63%), except for the plastid targeting signal sequence and a large insert in the active site which alters the substrate specificity.[3] Crystal structures of PHS1 should help us to understand the different substrate specificities for these two related enzymes and help to understand the role PHS1 plays in plants.

7.1.2 Engineering of PHS2

The PHS2 surface glucan-binding site elucidated in this work is unique to this protein. In order to understand the relative importance of this to the natural activity of this enzyme, mutations of this binding site should be assayed *in planta*. The binding, activity and substrate recognition of this enzyme on glycan arrays should also be compared, for wild type and mutants in this site.[4]

This class of enzymes is competent to transfer only one molecule of mannose, xylose, glucosamine and glucuronic acid onto each acceptor terminus. If the efficiency with which these sugars can be transferred can be improved,

using mutagenesis and protein engineering, it might be possible to synthesise novel polysaccharides. However it is probable that the acceptor site will also need to be substantially altered before efficient synthesis of long α-1,4-linked glycans can be achieved. These glycans are likely to have interesting properties and may be compatible with cyclising enzymes, making modified cyclodextrins and cycloamyloses.

Overall this enzyme has limited use in the synthesis of novel glucans as it requires maltotetraose as the minimum acceptor. This is likely because of the requirement for binding of a long polymer to overcome the unfavourable binding during the reaction. If this unfavourable binding can be reduced, or the relative binding strength of the acceptor binding to the -2 position be increased, then shorter glucans may be possible. It is proposed that some α-1,4-glucan phosphorylases can use acceptors as short as maltose[5] and comparison between these enzymes and PHS2 might suggest ways to improve the flexibility.

7.1.3 Starch-like surfaces

The starch-like surface synthesised with PHS2 is a good analogue of natural starch granules which can be probed with a range of enzymes. Proteins which interact with the starch granule can be assayed with this surface, in classical SPR experiments, to ascertain binding kinetics. For example, Glucan-Water Dikinase, which transfers phosphate from ATP onto the surface of starch granules, undergoes autophosphorylation, surface binding and phosphate transfer, but it is unclear at what stage in the catalytic cycle binding to the surface occurs.[6] The surfaces developed in this study represent a viable system for the assay of this enzyme and working towards this analysis is on-going.

There are several isoforms of starch synthases[7] and the precise role of each is not well understood. The binding kinetics, in the presence and absence of substrate, and their activity could be analysed on the starch like surface using SPR. This will provide substantial insights into the roles each of these isoforms plays and enhance our biological understanding of starch metabolism. Enzymes which naturally degrade starch granules are normally assayed in solution which is not a good analogue of the insoluble granule. This surface should be used as a more informative assay to provide data on the inherent degradation of the surface, not the release of products.

The use of SPR is severely limited by throughput and expense, particularly when assaying surface-modifying enzymes. The surface acoustic wave technology (SAW instruments GmbH) may allow the direct detection of the rearrangement and crystallisation of glucans on the surface, and their disruption by non-hydrolytic enzymes. Bio-Layer Interferometry systems, such as the Octet (ForteBio), can be used with a higher throughput and removes the risks associated with polysaccharide synthesis in a microfluidics device. The synthesis of starch-like surfaces will be explored using these systems for the future analysis of starch active enzymes.

These glucan surfaces can also be used as biocompatible coatings and in nanotechnology. By combining controlled display of the acceptor and phosphorylase catalysed vine-twinning polymerisation around hydrophobic polymers, the controlled orientation of insulated polymers, such as carbon nano-tubes,[8] may be achieved.

7.2 Cellodextrin phosphorylase

7.2.1 Further crystallography of CDP

The structure of CDP presented in this work is at 2.3 Å resolution and contains extra density at the phosphate binding site, which is modelled as a chloride ion. Improvements in resolution and alterations in the crystallisation conditions might help to resolve phosphate in this site. It has not been possible to achieve structures of CDP with substrates bound, but work towards this goal is on-going. Without substrate bound structures it is not possible to understand the amino acids which give rise to substrate specificity, but comparison with related enzymes can help to understand substrate interactions.

7.2.2 Engineering of CDP

Family GH94, to which CDP belongs, contains phosphorylases with different substrate specificity. Translating the differences between disaccharide phosphorylases onto CDP would allow the generation of a novel polymerising phosphorylase. For example, the residues which interact with the GlcNAc moiety in chitobiose phosphorylase are exactly the same in CDP, with the exception of Tyr804, which would clash with the *N*-acetyl group (Fig 7.1). In chitobiose phosphorylase it is replaced with valine and this modification might allow the CDP to act upon chitin. CDP has been shown to slowly catalyse the

transfer of Xyl, Man and GlcN from their respective phosphates. In order to improve this enzyme for the transfer of these sugars, random or site-saturation mutagenesis may allow the generation of efficient mannan, xylan and chitosan phosphorylases.

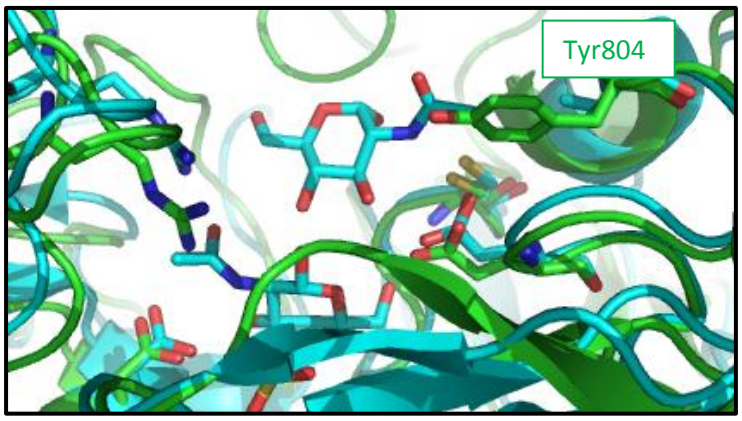


Figure 7.1: Substrate binding site of CDP and chitobiose phosphorylase

CDP (green) has a tyrosine which overlays with the acetate on the chitobiose in the chitobiose phosphorylase structure (cyan, 1V7X).

7.2.3 Cellulose-like surfaces

Efficient enzymatic degradation of cellulose is highly valuable for biomass conversion but there are limitations with the current analytical tools. CDP is able to generate long cellulose polymers in solution, which rapidly precipitate. Initial attempts to synthesise these polymers on SPR surfaces, analogously with the PHS2 experiments, were unsuccessful. However, it is hoped that by modifying the display techniques or using alternative systems, this technology can be explored. These surfaces will then be used to analyse cellulose binding and degrading enzymes.

When cellulose is formed in solution it rapidly precipitates to form insoluble material, which under the right conditions can form defined structures, such as cellulose whiskers[9] or enzymatically synthesised spherulites.[10] β-lipid glucosides are reasonable acceptors for CDP,[11] allowing production of novel glycolipids, which rapidly precipitate.[12] Preliminary data suggest that by carrying out the CDP catalysed extension above the critical micelle concentration of the lipid glucans, the micelles are coated in a layer of insoluble cellulose forming nanospheres. These materials could be used to make biocompatible coatings or to synthesise carbohydrate nanoparticles. The CDP catalysed synthesis of insoluble cellulose could be used more widely to coat surfaces with a nanoscale, biocompatible and insoluble carbohydrate layer.

7.3 Laminarin phosphorylase

7.3.1 Identification of LDP

The laminarin phosphorylase from *Euglena* has been investigated for a number of years for the biosynthesis of β-1,3-glucans. The polymerising phosphorylase is a highly desirable tool for the synthesis of β-glucans for use in biomedicine and nanomaterials. Identification and heterologous expression of the polymerising phosphorylase remains a high priority. LDP will be the first phosphorylase identified in whichever GH family it belongs to, as there are no known β-phosphorylase families encoded in the *Euglena* transcriptome. From the evidence in this work the phosphorylase is most likely to be a member of GH81, though we have been unable to confirm this. Once heterologous expression is achieved the enzyme can be thoroughly explored for the production of defined β-glycans, including non-glucose based polymers. In particular the synthesis of β-glucosyl glycerols should be explored in more detail using the purified enzyme.

7.3.2 Crystallography of LDP

Since the LDP will be in a novel family, structural studies will be a high priority, in order to determine what factors control the phosphorolytic activity. There are currently no structures of GH81, the best candidate found in this work, and thus elucidation of any structure in this family is valuable. Once structural information is available the key residues involved in carbohydrate recognition and specificity can be elucidated. This will inform mutagenesis and chimerisation studies to enhance substrate promiscuity and help to develop tools for the synthesis of defined β-1,3-glycan polymers.

7.4 Other phosphorylases

Phosphorylases are highly versatile enzymes which can readily be used to synthesis long polysaccharide chains. However, there are only a limited number currently known and re-engineering these to access other polysaccharides has limitations. Instead the diversity of activities already present in nature should be fully mined before novel activities are artificially developed.

7.4.1 Identification of the missing glucan phosphorylases

There are no known phosphorylases for β-trehalose, but this linkage has not so far been found in Nature and thus it might not be possible to find a β-1,1-phosphorylase. On the other hand α-1,6 and β-1,6-glucans are rather common in Nature, produced, for example, by bacteria in dental plaque[13] and in the yeast cell wall,[14] respectively, although no phosphorylases have yet been discovered for them. Thus, it is expected that phosphorylases may exist for these linkages and could be found by, for example, assaying organisms which can use linear dextran as a sole carbon source. Identification of these enzymes would complete the suite of phosphorylase linkages.

7.4.2 Identification of non-glucose phosphorylases

Most of the phosphorylases so far identified act on homoglucans. These make up the major energy storage and structural polysaccharides used by organisms but there are numerous other polysaccharides found in nature: fructans are storage polymers of fructose found in some plants and algae;[15] plant cell walls contain many complex polymers including xylans and uronic acid polymers;[16] animals cells are covered in layers of charged polysaccharides such as chondroitin and heparin;[17] many types of seaweed also have complex and charged polymers in their cell wall.[18] With the large amount of the compounds produced and degraded it is likely there are many more phosphorylases yet to be identified. In order to find them microbes which can use these alternative polysaccharides as carbon sources could be screened. Since many of these compounds are useful in biotechnology, food and medicine the possibility of synthesising defined components, as offered by phosphorylases, is highly attractive.

7.4.3 Polymerising phosphorylases

Many of the phosphorylases so far studied act on disaccharides. This limits their use in biotechnology and the possibility of converting them into polymerising phosphorylases is highly attractive. The comparison between the structures of CDP and CBP shows that substantial changes might be necessary to allow polymerisation. Initial studies have probed this, opening up the active site of CBP to allow anomerically substituted glucose to act as an acceptor[19] and increasing the length of kojioligosaccharides produced by a GH94 phosphorylase.[20]

Comparison of phosphorylases in the same family, but with different acceptor lengths, will inform mutations and chimerisations to produce polymerising enzymes. Chitobiose phosphorylase could be altered by chimerising with CDP and nigerose phosphorylase should be compatible with kojibiose phosphorylase. It is likely that disaccharide phosphorylases will continue to yield to biochemical analysis more readily than their polymerising counterparts. The ability to convert them into polymerising enzymes is highly desirable for polysaccharide synthesis.

7.4.4 Relationship between hydrolases and phosphorylases

Most of the phosphorylases are in glycosyl hydrolases families, suggesting they have evolved as a more biochemically efficient degradation tool. There are many more hydrolases characterised and logic dictates there must be a hydrolase (or phosphorylase) for every sugar linkage in nature. If hydrolases can be engineered into phosphorylases then every linkage can be readily synthesised from simple sugar phosphates suing soluble enzymes.

To understand the difference between phosphorylases and hydrolases detailed structural analysis of closely related enzymes is required. Amylosucrase, sucrose hydrolase and sucrose phosphorylase are all members of GH13. The phosphorylase structure shows a slight deviation in one active site loop and there are notably His234 and Leu341 in the active site, not Ile and Phe, as in the hydrolase and transglycosylase structures (Fig 7.2).[21] As there is no phosphate bound it is not possible to say with confidence the impact of these alterations, but it is likely that the histidine is important for the coordination of the phosphate in the active site. Further crystallography is required to understand the phosphate localisation and exclusion of H_2O. Mutagenesis and chimerisaton studies can be used to study the interconversion of hydrolases and phosphorylases.

There are α-1,6-hydrolases in family GH13 and it might be possible to convert these into phosphorylases, based on the principles elucidated. There are however no enzymes in this family which act on any other sugars.

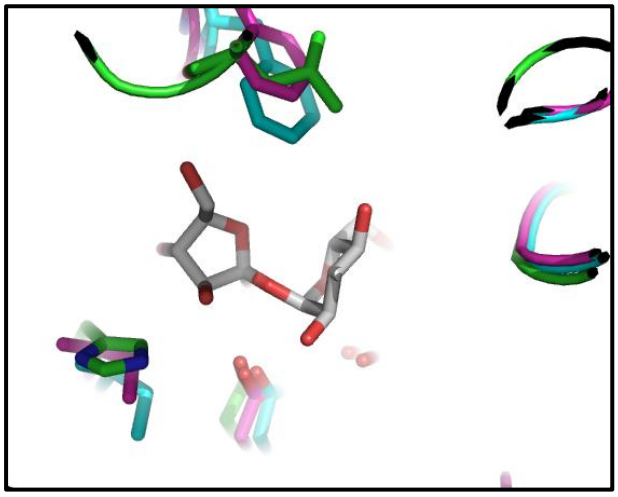


Figure 7.2: Family GH13 active site

Sucrose phosphorylase (1R7A, green)[21] has a similar active site to sucrose hydrolase (3CZK, magenta)[22] and amylosucrase (1G5A, blue).[23] However one loop has moved and there are two important amino acid changes – His234 and Leu341.

Trehalose phosphorylase and trehalose hydrolase are both present in GH65. The structure of the hydrolase is not yet known and comparison of these structures may highlight the important structural differences in this family. There are no other hydrolases in this family, however, so there is only limited application of the principles elucidated for the development of phosphorylases in this family

There are no other hydrolase families which also contain phosphorylases, so the rules developed from families GH13 and 65 would need to then be tested in other families. It is likely that the exclusion of water is the most important aspect when altering these enzymes and the development of glycosynthases has provided some direction for this work.[24]

7.5 *Euglena gracilis*

7.5.1 Further sequencing of *Euglena*

The *de novo* transcriptome sequencing of *Euglena* presented in this work only provides a snapshot of metabolism under two conditions. Genome sequence of *Euglena gracilis* is needed to see all the genes in this alga, not just expressed transcripts. Additional transcriptome data sets from a variety of conditions, such as oxidative stress and anaerobosis, will reveal the strategies used to adapt to these conditions. Additional sequencing efforts will be needed for the various strains of *Euglena gracilis* currently under investigation, as it is apparent that individually sequenced genes do not match the transcriptome presented in this study. There will also be opportunities to build on this sequencing for related Euglenoids, such as the toxin producing *Euglena sanguinea*,[25] and non-photosynthetic *Astasia*.[26]

7.5.2 Genetic control mechanisms in *Euglena*

The elements for classical gene silencing appear to be present in *Euglena*, with the exception of the RDRP. There is much on-going work in plants and animals on the various mechanisms by which RNA silencing occurs and the various genes that are targeted. *Euglena* is not closely related to any of the model organisms and thus might reveal novel mechanisms and targets.

Incorporation of BaseJ into DNA is used in unusual gene silencing routes in *Trypanosoma* and *Leishmania*. *Euglena* also utilises this base, but it has not been thoroughly investigated. With the transcriptome sequences it is clear that BaseJ metabolism is substantially the same in these organisms, but until genome sequencing of *Euglena* has been completed there is limited possibility to study this epigenetic modification. Key experiments include immunoprecipitation and sequencing of BaseJ containing chromatin, identification of the glycosyltransferases and glycoside hydrolases involved in BaseJ metabolism, and down regulation of the enzymes involved in BaseJ turnover, possibly by RNAi or mutagenesis.

7.5.3 Natural products from *Euglena*

It is apparent that *Euglena* expresses complex natural product synthases. It has not been possible to identify any natural products to date, though it is probable that one of the nonribosomal peptides is involved in iron sequestration. The role of the other complex natural products remains unknown and identification of these products is on-going (Fig 7.3).

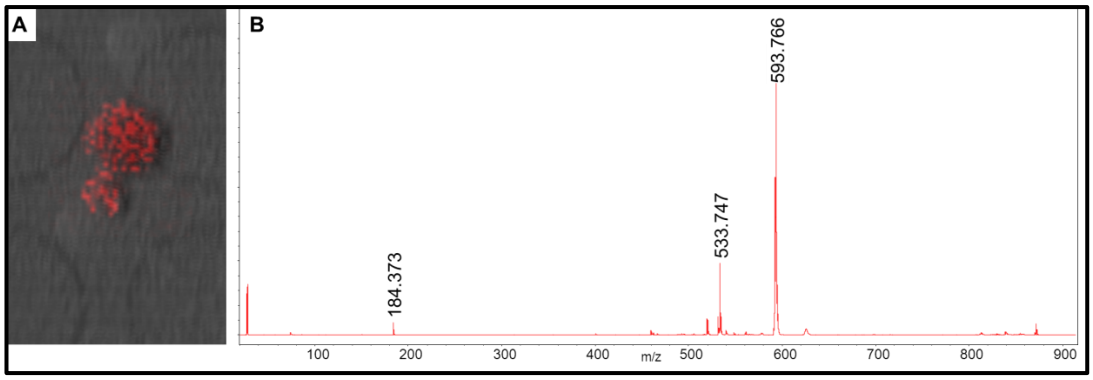


Figure 7.3: *Euglena* natural products

A. MALDI-imaging MS of *Euglena* cultures from high nutrient media agar revealed a peak at m/z = 870.553 only on or close to the colonies. **B**. Ms/ms fragmentation of this ion, after methanol extraction of the colony, does not allow conclusive structure solution.

7.5.4 Biotechnological uses of *Euglena*

Euglena are highly attractive organisms for biotechnology. They have all the attributes of Eukaryotes and are exceptionally easy to cultivate. However, in order to fully utilise *Euglena* in biotechnology, there are a number of hurdles to overcome.

Euglena have successfully been transformed and there is some evidence for homologous recombination.[27] This need to be confirmed as it represents a useful technique for introducing targeted mutations into the *Euglena* genome. The chloroplast can also be transformed, although there is some complex gene processing which may complicate successful expression in this organelle.[28] The nature of their protein glycosylation and ease of growth makes *Euglena* ideal for expression of therapeutic glycoproteins, once reliable transformation protocols are developed. If homologous recombination can be achieved then they can be humanised by replacing all the tailoring enzymes with human isoforms.

It is possible to mutate genes in *Euglena* but the procedure is complicated and has not been widely used.[29] Gene silencing, using electroporation-mediated introduction of dsRNA, has been used to deactivate genes and is stable over several generations.[30] Stable mutants will allow the detailed understanding of all the genes involved in various metabolic pathways and the effects of specific perturbations. If, as seems likely, *Euglena* is not haploid then all copies of a gene need to be inactivated. In other organisms this is normally achieved by mutation of one copy and repeated rounds of sexual recombination in order to achieve homozygotes. To accomplish this in *Euglena* sexual reproduction must be induced. Since all the genes necessary appear to be present in the transcriptome (see Section 6.3.1), and it is known that under phosphate limited conditions the number of chromosomes is halved,[31] this seems an achievable aim.

It has long been recognised that *Euglena* contains many important metabolic pathways with unusual features. The transcriptome provides sequences for all the enzymes and will allow their combination into designed pathways in other organisms. The unusual protein architectures will facilitate metabolic engineering and provide extra tools in the synthetic biology toolbox.

7.6 Final thoughts

A great deal of the biology that is currently studied focuses on topics which are already well characterised. By concentrating on phosphorylases which have already been the subject of research, PHS2, CDP and to a certain extent LDP, I have fallen into this trap. The *de novo* sequencing of the well characterised alga *Euglena gracilis* has revealed an unexpected metabolic capability, including unexpected natural product synthases and novel enzyme architectures. All of the analysis has focused on <40% of the sequences which match Gene Ontology classifications, leaving >60% of the sequences with no functional predictions. This highlights the breadth of biology in just this one alga, which has only been partially unveiled, even with the power of modern sequencing technology. There are whole worlds to explore in every drop of pond water, and we have barely begun to scratch the surface.

7.7 References

1 Fettke, J., Malinova, I., Eckermann, N. & Steup, M. Cytosolic heteroglycans in photoautotrophic and in heterotrophic plant cells. *Phytochemistry* **70**, 696-702 (2009).

2 Satoh, H. *et al.* Mutation of the plastidial α-glucan phosphorylase gene in rice affects the synthesis and structure of starch in the endosperm. *Plant Cell* **20**, 1833-1849 (2008).

3 Mori, H., Tanizawa, K. & Fukui, T. A chimeric alpha-glucan phosphorylase of plant type-L and type-H isozymes - functional-role of 78-residue insertion in type-L isozyme. *J. Biol. Chem.* **268**, 5574-5581 (1993).

4 Pedersen, H. L. *et al.* Versatile high-resolution oligosaccharide microarrays for plant glycobiology and cell wall research. *J. Biol. Chem.* **287**, 39429-39438 (2012).

5 Singh, S. & Sanwal, G. G. Multiple forms of α-glucan phosphorylase in banana fruits - properties and kinetics. *Phytochemistry* **15**, 1447-1451 (1976).

6 Mikkelsen, R., Mutenda, K. E., Mant, A., Schürmann, P. & Blennow, A. α-Glucan, water dikinase (GWD): a plastidic enzyme with redox-regulated and coordinated catalytic activity and binding affinity. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1785-1790 (2005).

7 Edwards, A. *et al.* Specificity of starch synthase isoforms from potato. *Eur. J. Biochem.* **266**, 724-736 (1999).

8 Yang, L. Q. *et al. In situ* synthesis of amylose/single-walled carbon nanotubes supramolecular assembly. *Carbohydr. Res.* **343**, 2463-2467 (2008).

9 Eichhorn, S. J. Cellulose nanowhiskers: promising materials for advanced applications. *Soft Matter* **7**, 303-315 (2011).

10 Kobayashi, S. *et al.* Formation and structure of artificial cellulose spherulites via enzymatic polymerization. *Biomacromol.* **1**, 168-173 (2000).

11 Hai Tran, G., Desmet, T., De Groeve, M. R. M. & Soetaert, W. Probing the active site of cellodextrin phosphorylase from *Clostridium stercorarium*: kinetic characterization, ligand docking, and site-directed mutagenesis. *Biotechnol. Progr.* **27**, 326-332 (2011).

12 Tran, H. G. *et al.* Biocatalytic production of novel glycolipids with cellodextrin phosphorylase. *Bioresour. Technol.* **115**, 84-87 (2012).

13 Monsan, P. *et al.* Homopolysaccharides from lactic acid bacteria. *Int. Dairy J.* **11**, 675-685 (2001).

14 Klis, F. M. Review: cell wall assembly in yeast. *Yeast* **10**, 851-869 (1994).

15 Hendry, G. A. F. Evolutionary origins and natural functions of fructans - a climatological, biogeographic and mechanistic appraisal. *New Phytologist* **123**, 3-14 (1993).

16 Mohnen, D. Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* **11**, 266-277 (2008).

17 Bernfield, M. *et al.* Functions of cell surface heparan sulfate proteoglycans. *Annu. Rev. Biochem.* **68**, 729-777 (1999).

18 Domozych, D. S. *et al.* The cell walls of green algae: a journey through evolution and diversity. *Front. Plant Sci.* **3**, 82-82 (2012).

19 De Groeve, M. R. M. *et al.* Construction of cellobiose phosphorylase variants with broadened acceptor specificity towards anomerically substituted glucosides. *Biotechnol. Bioeng.* **107**, 413-420 (2010).

20 Yamamoto, T., Nishimoto, T., Chen, H. & Fukuda, S. Improvments of the enzymetic properties of kojibiose phosphorylase by random mutagenesis and chimerization. *J. Appl. Glycosci.*, 123-129 (2006).

21 Sprogøe, D. *et al.* Crystal structure of sucrose phosphorylase from *Bifidobacterium adolescentis*. *Biochemistry* **43**, 1156-1162 (2004).

22 Kim, M.-I., Kim, H.-S., Jung, J. & Rhee, S. Crystal structures and mutagenesis of sucrose hydrolase from *Xanthomonas axonopodis* pv. *glycines*: insight into the exclusively hydrolytic amylosucrase fold. *J. Mol. Biol.* **380**, 636-647 (2008).

23 Skov, L. K. *et al.* Amylosucrase, a glucan-synthesizing enzyme from the α-amylase family. *J. Biol. Chem.* **276**, 25273-25278 (2001).

24 Hidaka, M. *et al.* Structural explanation for the acquisition of glycosynthase activity. *J. Biochem.* **147**, 237-244 (2010).

25 Zimba, P. V., Rowan, M. & Triemer, R. Identification of Euglenoid algae that produce ichthyotoxin(s). *Journal of Fish Diseases* **27**, 115-117 (2004).

26 Gockel, G. & Hachtel, W. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* **151**, 347-351 (2000).

27 Krajcovic, J. *et al.* Development of an effective transformation system for the nuclear genome of the flagellate *Euglena gracilis*. *Curr. Opin. Biotechnol.* **22**, S45-S45 (2011).

28 Doetsch, N. A., Favreau, M. R., Kuscuoglu, N., Thompson, M. D. & Hallick, R. B. Chloroplast transformation in *Euglena gracilis*: splicing of a group III twintron transcribed from a transgenic psbK operon. *Curr. Genet.* **39**, 49-60 (2001).

29 Shneyour, A. & Avron, M. A method for producing, selecting, and isolating photosynthetic mutants of *Euglena gracilis*. *Plant Physiol.* **55**, 142-144 (1975).

30 Iseki, M. *et al.* A blue-light-activated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. *Nature* **415**, 1047-1051 (2002).

31 Epstein, H. T. & Allaway, E. Properties of selectively starved *Euglena*. *Biochim. et Biophys. Acta - Nucleic Acids Prot. Synth.* **142**, 195-207 (1967).

Chapter 8 – Materials and methods

8.1 General methods

Unless otherwise noted, all chemicals were purchased form Sigma-Aldrich Company Ltd.

8.1.1 – Molecular biology methods

8.1.1.1 – Transformation of E. coli

Chemically competent *E. coli* were prepared by collecting 5 ml of mid-log phase culture by centrifugation (16,000 g), washing and incubating these cells in $CaCl_2$ (100 mM, 1 ml) for 2 hours at 4 °C. Cells were diluted into 40% (v/v) glycerol and stored at -80 °C until required. After resuspending cells in of $CaCl_2$ (100 mM, 250 µl) 50 µl aliquots were incubated at 4 °C with approximately 200 ng of plasmid DNA for 30 min. Cells were then subjected to a 30 s heat shock at 42 °C and immediately returned to 4 °C. After 2 min, 500 µl of LB medium was added to the cells. The cultures were plated on LB agar immediately, for AmpR selection, or after 1 hour incubation at 37 °C for other antibiotic resistances, and incubated for 16 hours at 37 °C. Single colonies were subsequently grown in 10 ml LB with appropriate antibiotic selection and 1 ml of a single successful culture was placed at -80 °C in 40% (v/v) glycerol for long term storage. Carbenicillin was used at 100 µg/ml and chloramphenicol at 35 µg/ml.

8.1.1.2 – Agarose gel electrophoresis

Agarose gels (1% w/v), containing 1x SYBR Safe DNA gel stain (Invitrogen), were made and run in TBE (58 mM TrisHCl, 89 mM boric acid, 3 mM EDTA). Commercial loading buffer (NEB) was added to the samples before loading and the gels were run (100 V) until the dyes reached the lower portion of the gels. Gels were visualised with a DR88 Dark Reader (Clare Chemical Research) or Gene Genius (Syngene) transilluminators. Gel extraction was performed using QiaQuick Gel Extraction Kits (Qiagen)

8.1.1.3 – PCR, sequencing and alignments

PCR was performed using Pfusion high fidelity RNA polymerase (Finnzymes) in a thermocycler (GStorm). DNA sequencing reactions were performed with BigDye v3.1 (Applied Biosystems) according to the manufacturer instructions. The reaction products were analysed by Genome Enterprises Limited. Sequence alignments were performed using AlignX (Invitrogen) and ClustalW2 (www.ebi.ac.uk) and phylogenies displayed using iTOL web.[1]

8.1.2 – Purification and analysis of recombinant proteins from *E. coli*

8.1.2.1 – Purification of proteins

Cell pellets were thawed and resuspended in 50 ml of lysis buffer (50 mM HEPES, pH 7.5, 100 mM NaCl, 1x cOmplete, EDTA-free Protease Inhibitor Cocktail Tablet (Roche), 0.02 mg/ml DNaseI). Cells were lysed using a cell disruptor (one shot mode, 25 kpsi, Constant Systems) and the cell debris removed by centrifugation at 30,000 g (30 min, 4 °C). Protein was purified at 4 °C using an ÄKTA Xpress FPLC system (GE Healthcare). The supernatant was passed through a HiTrap Ni-NTA column (5 ml, GE Healthcare), washed with wash buffer (50 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 0.03 M imidazole) and eluted with elution buffer (50 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 0.5 M imidazole (BioUltra)). Further purification was performed by gel filtration on a Superdex S75 26/60 column (GE Healthcare) using GF buffer (50 mM HEPES, pH 7.5, 100 mM NaCl, 3.2 ml/min) with collection of protein containing peaks with A_{280} over 100 mAU and slope steeper than 100 mAU/min.

8.1.2.2 Denaturing polyacrylamide gel electrophoresis (SDS-PAGE)

Polyacrylamide gels were made as a single concentration (8-12%) of acrylamide (37.5:1 mono: bis, Severn Biotech) in TrisHCl (0.75 M, pH 8.8, 0.1% SDS), with a 5% acrylamide stacking gel in TrisHCl (140 mM, pH 6.8, 0.1% SDS), and run in SDS Running Buffer (25 mM Tris, 192 mM Glycine, 0.1% SDS). Alternatively 4-20% RunBlue precast gels (Expedeon), run in RunBlue Running Buffer (40 mM Tricine, 60 mM Tris, 0.1% SDS, 2.5 mM Sodium Bisulfite, pH 8.2), were used. Samples were loaded in loading buffer (SDS Running Buffer, 10% (v/v) glycerol, 0.1 mg/ml bromophenol blue, 0.1 mg/ml β-mercaptoethanol) after incubating at >95 °C for 5 min and run until the dye reached the end of the gel. Gels were removed from the cassettes and stained for protein using InstantBlue (Expedeon).

8.1.2.3 – Dynamic light scattering (DyLS)

Samples of protein (20 µl) were filtered (0.1 µm, Millipore) by centrifugation and loaded into a 12 µl quartz cuvette (Wyatt). Measurements were taken over 5 s at 25 °C in a DynaPro (Wyatt) and averaged over 10 results.

8.1.3 – Phosphorylase activity assays

All kinetic analyses were performed in the synthetic direction, measuring release of Pi from Glc-1-P. The concentration of released Pi was measured colourimetrically using a method modified from De Groeve et al.[2] Colour solution (75 µl, 0.1 M HCl, 13.6 mM sodium ascorbate) was added to the stopped enzyme reaction (25 µl), containing ammonium molybdate (200 mM), in a microtitre plate. After incubating for 5 min at 21 °C stop solution (75 µl, 68 mM sodium citrate, 2 % acetic acid) was added and A_{620} was measured. pH variation was performed in a mixed buffer (0.1 M each Tris, MES, citrate) with pH determined to 0.02 units accuracy at 21 °C. Temperature variation was performed in a thermocycler (GStorm).

8.1.4 – Protein crystallography

Structure determination using protein crystallography was carried out with the assistance of C. Stevenson and D. Lawson (John Innes Centre), and with the support of the Diamond Light Source and the assistance of the on-site beam line scientists.

8.1.4.1 – Protein crystallisation screens

Initial protein crystallisation was performed using the sitting drop vapour diffusion method in a 96 well plate format (MRC 96 well crystallization plate, Molecular Dimensions). 50 µl of well solutions ($AmSO_4$ and PEG (Qiagen) and JCSG, Classics and PACT (Molecular Dimensions) screens) were robotically dispensed (freedom evo, Tecan). Drops were made using a robot (OryxNano, Douglas instruments) to mix an equal volume (0.3 µl) of well solution and filtered (0.1 µm, Millipore) protein solution (10 mg/ml) in GF buffer. Plates were placed in a crystal hotel (CrystalPro, TriTek Corp.) at 20 °C and monitored over the course of 4 weeks, when the best conditions were selected for further screening in 24-well format using with 1 ml well solution and drops, composed of 1 µl protein and 1 µl well solution.

8.1.4.2 – Preparation and transport of protein crystals to Diamond Light Source (DLS)

Crystals were soaked with ligands and cryoprotectant as necessary before being mounted in CryoLoops (Hampton Research). They were then flash cooled in liquid nitrogen to -173 °C and transported to the Diamond Light Source synchrotron (Oxford, UK) in Unipuck cassettes. There they were robotically mounted on to the goniostats at and maintained at -173 °C with a Cryojet cryocooler (Oxford Instruments).

8.1.4.3 – Data processing

The resultant data were integrated using either MOSFLM[3] or XDS[4] and subsequently scaled by SCALA.[5] All downstream processing was carried out using the CCP4 software suit.[6] After initial structures were obtained subsequent rounds of rebuilding with COOT[7] and refinement with REFMAC5[8] were used to produce the final model.

8.1.5 APTS labelling of sugars

Sugars were labelled for electrophoresis with 8-aminopyrene-1,3,6-trisulfonic acid (APTS) according to the PACE method.[9] APTS (0.5 mg, 0.2 M) in aqueous acetic acid (5 µl, 30%) was mixed with $NaBH_3CN$ (0.5 mg, 0.8 M) in THF (5 µl). Reducing carbohydrate (1 mg) was dissolved in the mixture and incubated at 37 °C for 18 hr or 70 °C for 2 hr. The sample was then loaded on to a 30% 38:2 mono:bis acrylamide (Merck) Tris-borate (100 mM, pH 8.2) gel and separated by electrophoresis at 400 V in Tris-borate buffer (100 mM, pH 8.2), water cooled to room temperature. The carbohydrate band (upper) was excised and the labelled sugar was extracted into purified water by grinding using a ceramic bead in a bead mill (Fast Prep FP 120, Thermo) and washing with MQ-H_2O (3x 30 ml). The borate was removed, either by desalting using a PD-10 column (GE Healthcare) or by evaporating three times from methanol, and the product was quantified using APTS absorption at 455 nm (17160 1/M/cm).[10]

8.1.6 Capillary electrophoresis with laser induced fluorescence (CE-LIF)

After enzymatic reactions had been carried out using APTS-labelled carbohydrates, samples were placed in boiling water for 5 min and centrifuged for 1 min at 16,000 g, to inactivate and remove proteins. The samples were made up to at least 50 µl and loaded on to an N-CHO coated capillary (50.2 cm, 50 µm) in a PA800 ProteomeLab (Beckman Coulter) by injection at 0.5 psi for 20 s. They were then separated in running buffer (25 mM LiOAc, 0.4% polyethylene oxide, pH 4.75, 20 °C) at 30 kV for 7 to 45 min and detected using LIF (excitation at 488 nm and detection at 520 nm).[11]

8.1.7 General chemistry protocols

8.1.7.1 Nuclear magnetic resonance (NMR)

^{1}H NMR was performed at 25 °C (400 MHz, Bruker), referenced to the solvent peaks and analysis was performed using TopSpin (Bruker).

8.1.7.2 Mass spectrometry (MS)

Unless otherwise noted ESI-MS was performed on a Synapt G2 HDMS (Waters), run in MS/sensitivity/positive mode with standard settings. Spectra under the LC peak were combined and deconvoluted using the MaxEnt 1 tool in MassLynx 4.1 (Waters). MALDI-Tof was performed on an Ultraflex Tof/ToF (Bruker) in positive mode, using 2,4,6-trihydroxyacetophenone as the matrix and 200 spectra were collected and combined.

8.1.7.3 Thin layer chromatography (TLC)

TLC was performed on Merck silica gel 60 F_{254} and visualised by dipping in EtOH:H_2SO_4 3:1 and charring.

8.1.8 Atomic force microscopy (AFM)

AFM was performed by A. P. Gunning (Institute of Food Research) on an MFP-3D BIO (Asylum Research) using AC160TS cantilevers (Olympus) with a nominal spring constant of ~42 N/m oscillated at a frequency 10% below resonance (typically around 320 kHz). Images were acquired at a scan rate of 1 Hz in air using AC mode.

8.1.9 Transmission electron microscopy (TEM)

8.1.9.1 Microscopy

TEM was performed with the assistance of K. Findlay (BioImaging, John Innes Centre) on a Tecnai20 microscope (FEI), at 200 kV, and imaged using an AMT digital camera (Deben UK Ltd.).

8.1.9.2 Uranyl acetate (UA) staining

Particles were deposited on TEM Grids (Pd/Cu with plastic/carbon coating) and negatively stained with uranyl acetate (UA, 2%, 5 µl) for 5 min before imaging

8.1.9.3 PATAg staining

The periodic acid/thiocarbhydrazide/silver (PATAg) staining protocol[12] was used in order to specifically stain AuNPs for the carbohydrates. Briefly particles were deposited on the TEM grids and inverted on to periodic acid (10 mg/ml) for 20 min. After washing with MQ-H_2O the grids were placed on thiocarbohydrazide (2 mg/ml in 20% (v/v) acetic acid) for 2 hours and washed with decreasing aqueous acetic acid (20%, 10%, 5%) and finally MQ-H_2O. Grids were then floated on silver proteinate (10 mg/ml, 8% Ag) for 30 min in the dark and washed with MQ-H_2O. Grids were kept in the dark before imaging.

8.1.10 Proteomics

Proteomic analysis was performed by G. Saalbach (Proteomics Facility, John Innes Centre). Proteins for analysis were separated by SDS-PAGE and gel slices excised. In-gel digestion with trypsin was performed, after carbamidomethylation of the cysteines, and the peptides released were loaded at high flow rate onto a reverse-phase trapping column (0.3 mm x 1 mm, containing 5 µm C18 100 Å PepMap packing, LC Packings) and eluted through a reverse-phase capillary column (75 µm x 150 mm column, containing Symmetry C18 300 Å packing, Waters Ltd.) directly into the nano-electrospray ion source of a quadrupole time-of-flight mass spectrometer (Q-ToF2, Micromass UK Ltd.). Alternatively MALDI-ToF MS was performed using α-cyano-4-hydroxycinnamic acid as the matrix. Fragment ion spectra were searched using the MASCOT search tool (Matrix Science Ltd.), allowing oxidation of methionine.

8.2 Specific protocols for PHS2

8.2.1 Cloning of PHS2

The *PHS2* gene was amplified from *Arabidopsis thaliana* Col-0 cDNA (kind gift of C. Ruzanski) by PCR using the following primers:

At*PHS2* TOPOF (eurofins MWG):	CACCATGGCAAACGCCAATGGAAAAGC
At*PHS2* R*(Sigma):	TTAGGGAACAGGACAAGCCTC

The forward primer contains the TOPO cloning overhang (underlined).

The band corresponding to 2621 bp, when compared to the marker (Fig 8.1), was extracted from the gel and cloned into pET151 using directional TOPO cloning (Invitrogen) forming plasmid pET151-PHS2 (Fig 8.2).

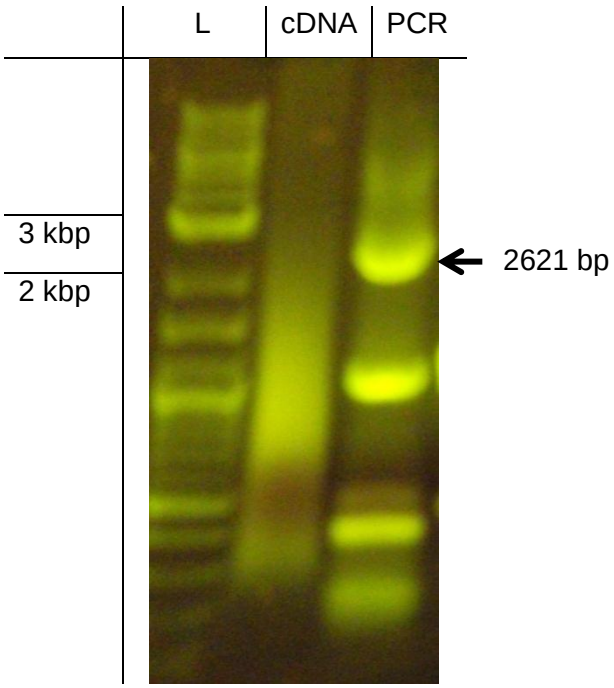


Figure 8.1: PCR amplification of *PHS2*

PHS2 was amplified by PCR from *Arabidopsis* cDNA. A 10 µl reaction was run on a 1% (w/v) agarose gel. The band corresponding to 2621 bp was extracted and the other bands, corresponding to non-specific priming, were discarded. L: 1 kb DNA Ladder (NEB). cDNA: total cDNA. PCR: total reaction.

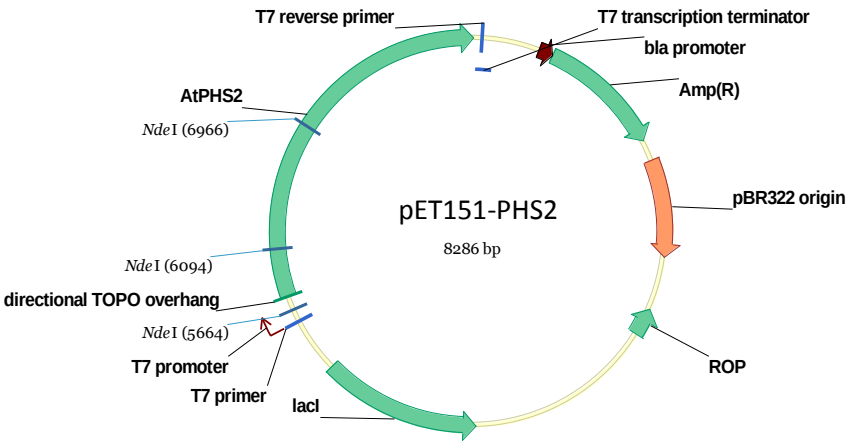


Figure 8.2: Plasmid map of pET151-PHS2 showing NdeI restriction sites

The plasmid pET151-PHS2 was transformed from the TOPO reaction mixture (1 µl) into OneShot TOP10 chemically competent *E. coli* (50 µl, Invitrogen), and successful transformants were selected for on LB agar using carbenicillin. Successful construction of the plasmid was confirmed by growing overnight cultures and extracting plasmids from mid log phase *E. coli* cells (3 ml) using Qiaprep Spin Miniprep Kit (Qiagen), eluting with MQ-H_2O (50 µl). These were analysed by restriction digest using NdeI (NEB) (Fig 8.3), and finally by DNA sequencing (Appendix 1).

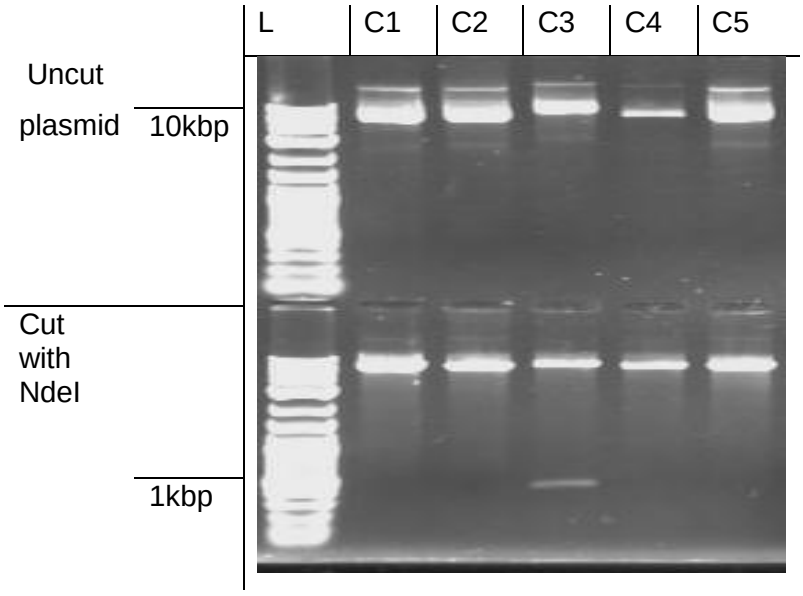


Figure 8.3: Restriction digest analysis of pET151-PHS2 plasmid from single clones

Plasmids were extracted from 5 colonies (C1-5) and either run without digestion or after 2 hour digestion with NdeI. Only C3 gave a band at around 1 kb, compared to 1 kb DNA Ladder (L, NEB), corresponding to 872 bases, though no band corresponding to 430 bp is visible.

8.2.2 Optimisation of PHS2 expression

For protein expression four *E. coli* strains (BL21, BL21CodonPlus (Agilent), Rosetta-2 and Rosetta-2 GroEL/ES (Novagen)) were transformed with pET151-PHS2 and cultured in Luria-Bertani medium containing carbenicillin and, except for BL21, chloramphenicol. Cultures (10 ml) were grown at 37 °C and samples taken from uninduced cell and 2 and 4 hours after induction with IPTG (0.5 mM) and analysed by SDS-PAGE (Fig 8.4).

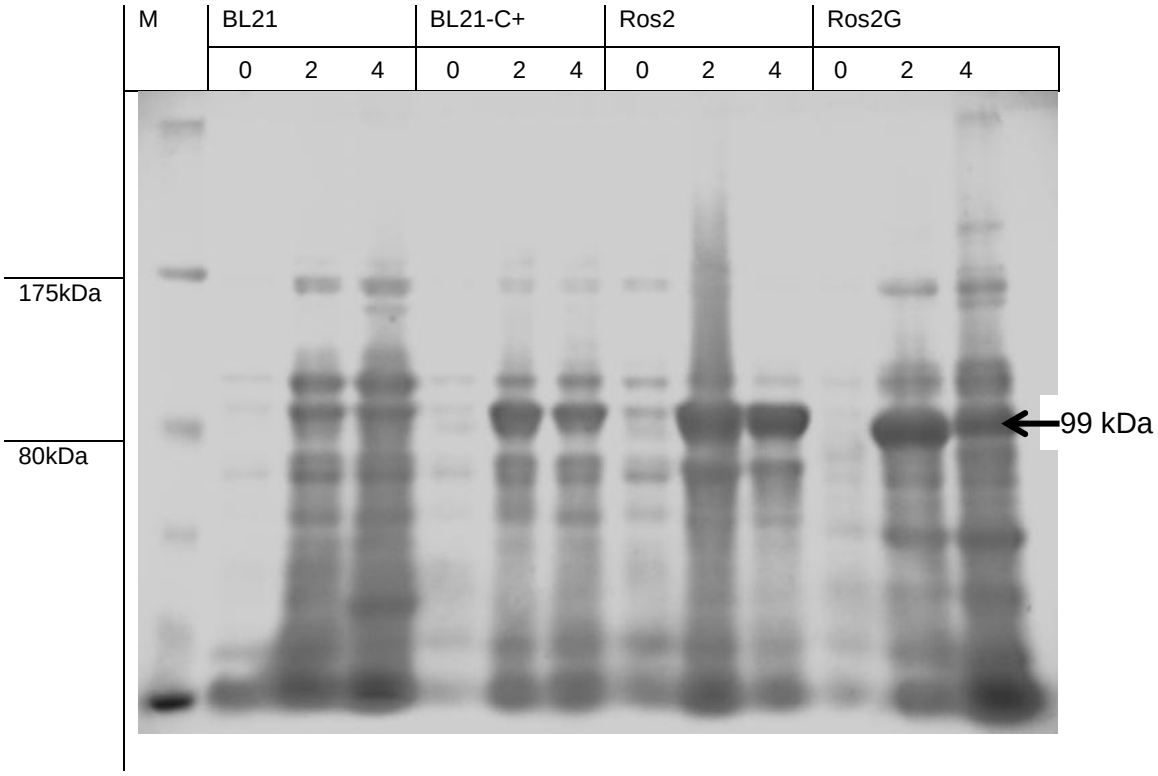


Figure 8.4: SDS-PAGE analysis of *E. coli* strains expressing PHS2

Samples were taken from the four strains of *E. coli* before, 2 and 4 hours after induction with IPTG. M: Broad Range Prestained Ladder (NEB), BL21: BL21, BL21C+: BL21 CodonPlus, Ros2: Rosetta-2, Ros2G: Rosetta-2 GroEL/ES.

The best expressing strain, as judged by the intensity of the band at 99 kDa, was Rosetta-2, which was then selected for expression of PHS2. Cultures of LB (8 x 1 litre) were inoculated with these cells and grown at 37 °C until OD_{600} reached 0.1 when they were transferred to 16 °C for 1 hour. IPTG was then added to a final concentration of 1 mM. After 16 hours the cells were harvested at 7,500 g and frozen at -80 °C until required. The cells were lysed and the protein purified using the standard protocol (section 8.1.2) to give a total yield of 2.5 mg of purified PHS2 per litre culture (see section 3.2), which was concentrated to 40 mg/ml with an Amicon Ultra-15 50 kDa spin concentrator (Millipore). Protein concentration was estimated by absorption at 280 nm (Nanovue, GE Healthcare) using an extinction coefficient of 145190 1/M/cm calculated from the amino acid composition.[13] Intact mass gave a molecular weight of 98943.05 Da (Fig 8.5), in close agreement with the calculated mass of the PHS2 monomer with the tag, and with no pyridoxal phosphate bound, of 98942.5 Da.

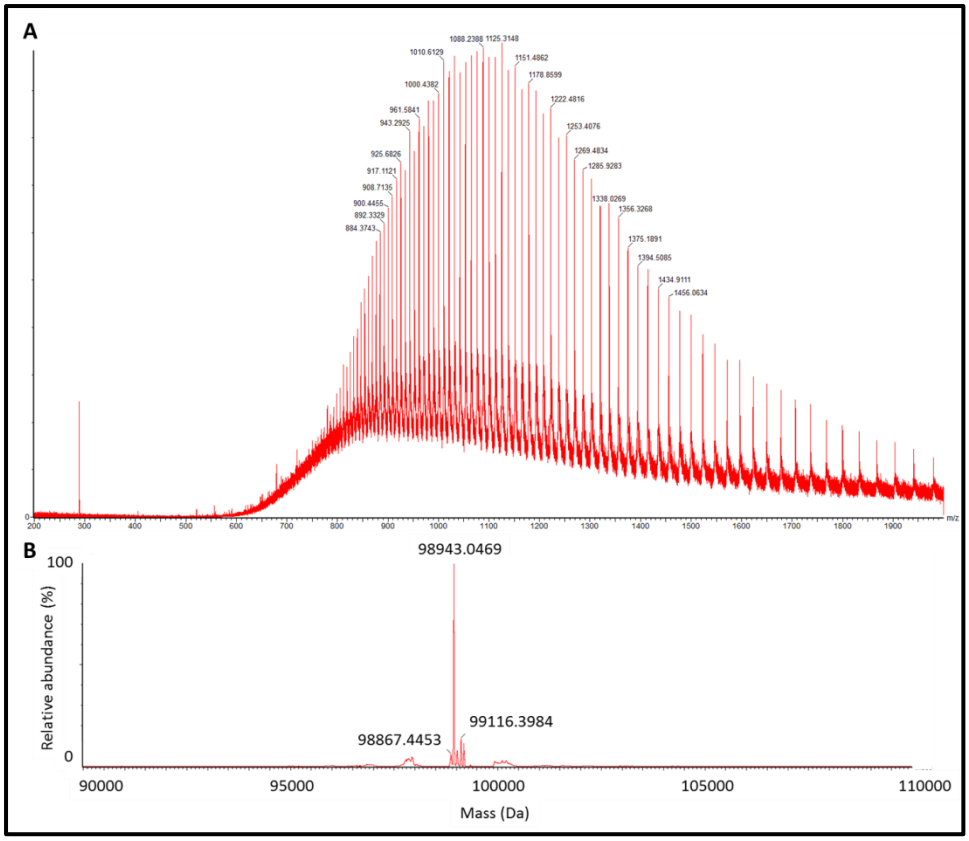


Figure 8.5: Intact mass of PHS2

PHS2 (200 pmol) was loaded on to a C4 reverse phase column (BEH300 C4 1.7 µm, 1 x 50 mm Waters) attached to an Acquity UPLC and eluted with a gradient from 10-80% acetonitrile in 0.1% formic acid in 13 min, with ESI-MS analysis. **A**. Spectrum under main peak **B**. Mass calculated from spectrum.

8.2.3 Acceptor length specificity of PHS2

Assays were performed in triplicate in the range 0.05-3 mM acceptor glucan, using 10 mM Glc-1-P in 12.5 µl MES buffer (20 mM, pH 6.0) at 21 °C using 8 µg/ml PHS2.

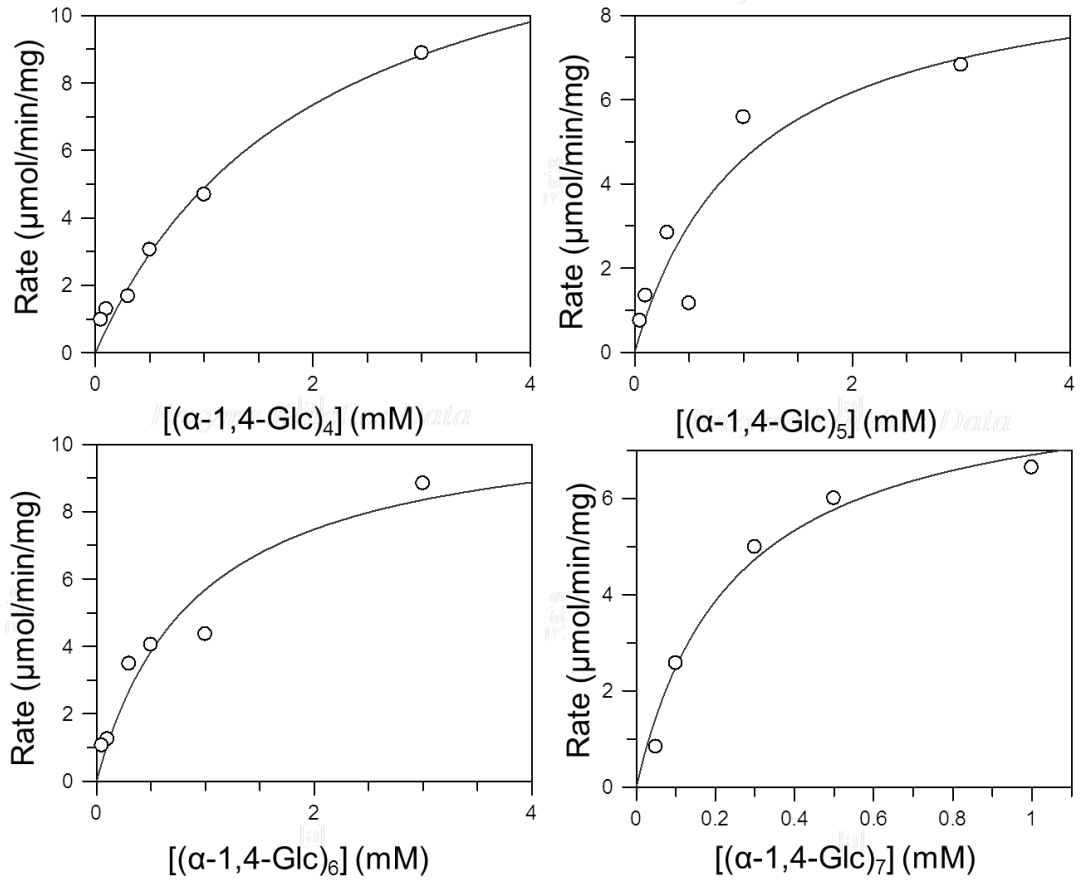


Figure 8.6: Acceptor length specificity of PHS2

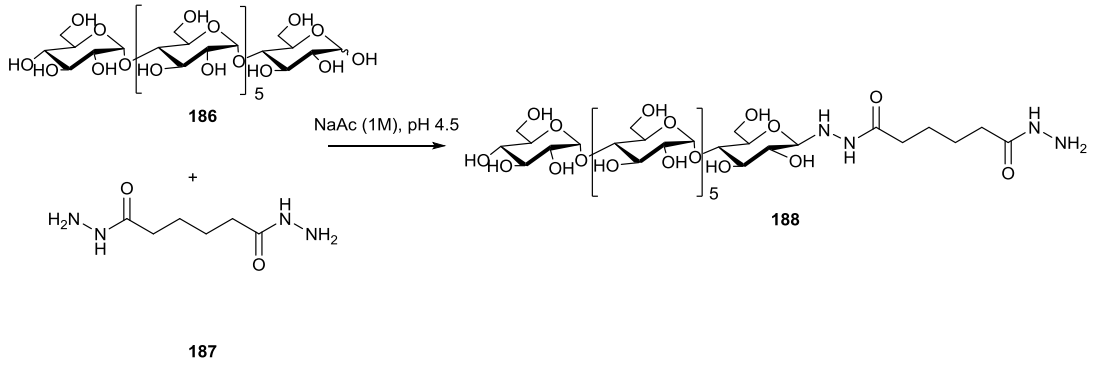
The K_M^{app} for maltooligosaccharides was calculated from the kinetics data using GraFit (Erithacus Software Ltd).

8.2.4 SPR experiments

SPR experiments were performed in a BiacoreX instrument (Biacore, GE Healthcare) at 25 °C using a continuous flow rate of MES buffer (20 µl/min, 40 mM, pH 6.0, degassed under vacuum) and injections were performed only after the baseline had stabilised.

8.2.4.1 Maltoheptaose surface

$(\alpha\text{-}1,4\text{-}Glc)_7$ (**186**, 10 mg, 8 µmol) was reacted with adipic acid dihydrazide (**187**, ADH, 17.4 mg, 100 µmol) in sodium acetate (500 µl, 1 M, pH 4.5), at 37 °C for 4 hours. The product (**188**, $(\alpha\text{-}1,4\text{-}Glc)_7$-ADH) was purified by gel permeation chromatography (20 mM NH_4HCO_3, pH 7.0, 0.5 ml/min, TSK HW40S) on a series 200 HPLC (Perkin Elmer), equipped with an refractive index detector. The peaks were analysed by TLC (iPrOH: H_2O, 3:1) and MALDI-ToF. The fractions containing product **188** were combined and lyophilised to yield 7.9 mg.



Scheme 2: The coupling of maltoheptaose to adipic acid dihydrazide

The surface was made by reacting 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (**68**, EDC, 1 µl, 0.2 M) and *N*-hydroxysuccinimide (**69**, NHS, 1 µl, 0.05 M) in DMF with (1-mercapto-11-undecyl)carboxymethyl hexa(ethylene glycol) $[HSC_{11}\text{-}(EG)_6OCH_2COOH]$ (**67**, SAM linker, 0.25 µmol, neat, Prochimia) at room temperature for 30 min and adding $(\alpha\text{-}1,4\text{-}Glc)_7$-ADH (**188**, 1 µmol) for 30 min, forming $(\alpha\text{-}1,4\text{-}Glc)_7$-ADH-SAM linker. The sensor surface was prepared by cleaning a commercial gold chip (Sensor Chip Au, Biacore BR-1005-42) using 5 µl Piranha solution (3:1 H_2SO_4 (conc): H_2O_2 (aq., 30%)) and washing exhaustively with purified water before drying with compressed air.

The (α-1,4-Glc)$_7$-ADH-SAM linker solution was then applied to the cleaned gold sensor surface and the thiol was left to immobilise for 2 hours and form a self-assembled monolayer (SAM) on the surface.[14] To this surface was added (1-mercapto-11-undecyl)tri(ethylene glycol) [HSC$_{11}$-(EG)$_3$OH] (**72**, SAM spacer, 0.8 µmoles, Assemblon) in DMF (4 µl), ensuring homogeneity of the surface and preventing discreet regions forming. The chip was left overnight for the SAM to assemble, before washing and loading the surface into the carrier.

8.2.4.2 Fractionated glucan (FG)

Glucidex IT12 (Roquette) has 3% branching (as judged by ^{1}H NMR integration of the broad peak at δ: 4.90 ppm (α-1,6) compared with the broad peak at δ: 5.33 ppm (α-1,4) (Fig 8.6)) and an average molecular weight of 19 kDa (DP ~110). It is known to contain low molecular weight maltooligosaccharides,[15] which would not act as acceptors for PHS2. Fractionated Glucan (FG) was obtained by washing these contaminants through a 30 kDa MWCO spin filter (Amicon) by washing with MQ-H$_2$O (3x 3 ml, with 0.5 ml retained each time), and lyophilising the retained fraction. FG (10 mg) was reacted with adipic acid dihydrazide as above and characterised by ^{1}H NMR spectroscopy (Fig 8.7). The reducing terminal anomeric signals [δ: 5.14 ppm (d, $J_{1\text{-}2}$=3.7 Hz, H1$_\alpha$) and δ: 4.57 ppm (d, $J_{1\text{-}2}$=8.0 Hz, H1$_\beta$)] disappeared completely and a new signal appeared, corresponding to the glycoside anomeric proton in the FG-ADH [δ: 4.03 ppm (d, J=9.3 Hz, H1$_\beta$)].

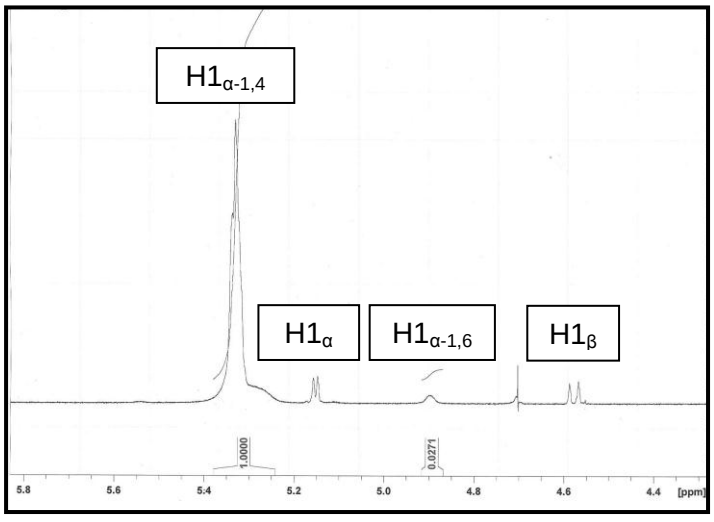


Figure 8.7: ^{1}H NMR of Glucidex IT12

Glucidex was analysed by ^{1}H NMR in D$_2$O and the integration of the anomeric protons shows that about 3% of the non-reducing residues are α-1,6 linked.

8.2.4.3 Low FG density SPR surface

SAM spacer (**72**, 98 mM) and SAM linker (**68**, 2 mM) in DMF (10 µl) was placed on the gold sensor chip surface and left overnight at 4 °C to assemble into a homogenous SAM. EDC (**68**, 10 µl, 0.2 M) and NHS (**69**, 10 µl, 0.05 M) in water were mixed and added to the surface and incubation was continued at room temperature for 30 min. FG-ADH (**188**, 10 µl, 10 mg/ml) in water was added to the surface and incubation was continued for a further 30 min at room temperature. The excess liquid was removed and the surface was washed with MQ-H_2O before inserting into the carrier.

8.2.4.4 High FG density surface

The higher sugar density surface, with a higher FG loading, was made in the same way as the maltoheptaose surface. Briefly, FG (10 mg) was coupled to the SAM linker (**67**, 0.25 µmol) in solution and then applied to the surface. The SAM spacer (**72**, 0.8 µmol) was added after 2 hours.

8.2.5 Fractionated glucan gold nanoparticles (FGNPs)

8.2.5.1 Synthesis of gold nanoparticles

Gold nanoparticles (AuNPs) were prepared by the method of Turkevich.[16] Gold (III) chloride ($AuCl_3 \cdot 3H_2O$, 12.5 mg in 100 ml MQ-H_2O) was reduced with trisodium citrate ($Na_3C_6H_5O_7 \cdot 2H_2O$, 50 mg in 50 ml MQ-$H_2O$), combined with stirring at 60 °C before heating to 85 °C for 150 min. Once the AuNPs had cooled to room temperature they were filtered through a disc filter (0.2 µm, Millipore). They were confirmed to be uniform in size and shape by TEM and the UV absorption maxima (Perkin Elmer Lambda 25 UV/Vis Spectrophotometer) at 524 nm, as expected.[16]

8.2.5.2 Immobilisation of FG on AuNPs

Approximately 1 µmole of FG-ADH-SAM linker, as was made for the high FG density SPR surface, was added to 10 ml of AuNPs and left to immobilise on the surface for 2 hours. 4 µmole of SAM spacer molecule was added and the solution was left to assemble overnight at 4 °C to form FGNPs. The FGNPs were purified by centrifugation and resuspension with sonication into fresh MQ-H_2O six times, and stored until ready for use at $5x10^{12}$ particles/ml in MQ-H_2O at 4 °C.

8.2.5.3 Enzymatic reactions on FGNPs

PHS2 mediated extension was performed by diluting the FGNPs to $1x10^{12}$ particles/ml in a solution containing Glc-1-P (10 mM), PHS2 (10 µg/ml) and buffer (MES, 10 mM, pH 6.5) and incubating at 37 °C for 60 min. The reaction was stopped by addition of ProteinaseK (20 µg/ml in 20 mM MES, pH 7.5) and incubating at 37 °C for 5 min followed by inhibition of the ProteinaseK by addition of PMSF (1 mM).

A fraction of the suspension was diluted two fold in enzyme buffers (0.3 mg/ml PAA in 100 mM MES pH 6.5 or 1 mg/ml AG in 100 mM MES, pH 4.5) and incubated for 60 min at 37 °C. The enzymes were inhibited by addition of acarbose (1 mM), a known inhibitor of both enzymes (PPA K_i = 2.1 µM,[17] AG (Aspergillus) K_i = 1 pM[18]), and the particles were deposited, stained and imaged immediately afterwards.

8.3 Specific protocols for cellodextrin phosphorylase

8.3.1 Optimisation of CDP expression

8.3.1.1 Initial cloning of CDP

The plasmid Gs-pET15b-CDP from Genscript was transformed into freshly prepared chemically competent *E. coli* BL21 and plated on to carbenicillin agar plates. Eight colonies were picked from these plates and grown in LB cultures (10 ml). These were subcultured, grown to OD_{600} of 0.6 and induced overnight at 30 °C using IPTG (1 mM) and the cleared cell lysate was assayed for activity using CE (Fig 8.8).

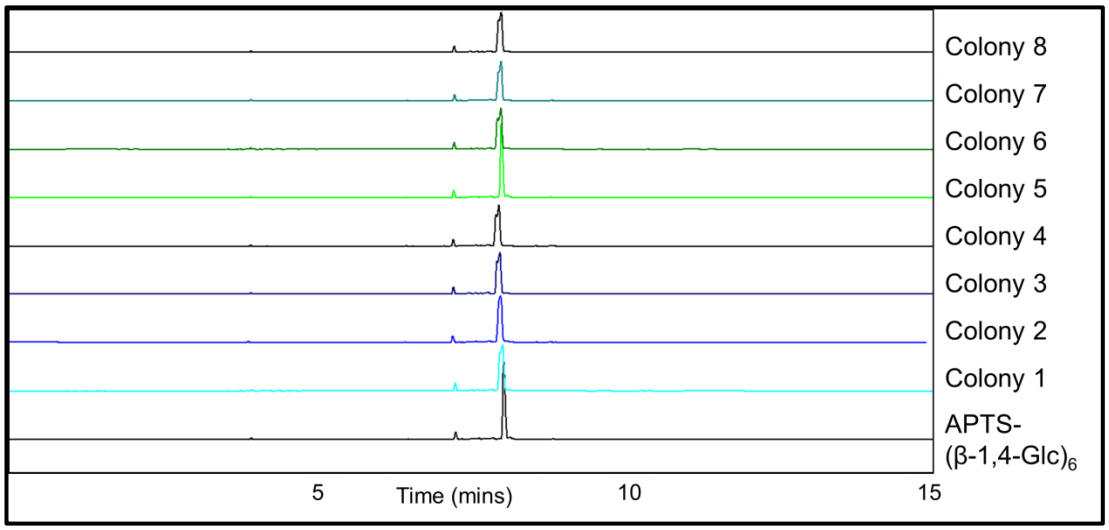


Figure 8.8: Activity assays of initial CDP expression clones

Cleared *E. coli* cell lysate (10 µl) from 8 colonies was assayed for CDP activity by incubating with Glc-1- P (10 mM) and 0.72 µM APTS-$(\beta\text{-}1,4\text{-}Glc)_6$ for 72 hr at 37 °C in HEPES (50 mM, pH 7.5), and the products were separated by carbohydrate electrophoresis. All samples only showed peaks consistent with the starting material APTS-$(\beta\text{-}1,4\text{-}Glc)_6$.

As none of the cultures contained any CDP activity it was concluded that the Gs-pET15b-CDP had the gene encoding CDP in the incorrect orientation for expression. To rectify this, the gene was excised and religated into the same backbone, and the correct orientation screened for by protein expression and CDP activity.

The plasmid Gs-pET15b-CDP was digested for 4 hr with BamHI (Roche) according to the manufacturer instructions. The resulting mixture was separated by electrophoresis in a 0.8 % agarose gel (Fig 8.9) and both bands were excised and extracted from the gel using QIAquick gel extraction kit (Qiagen).

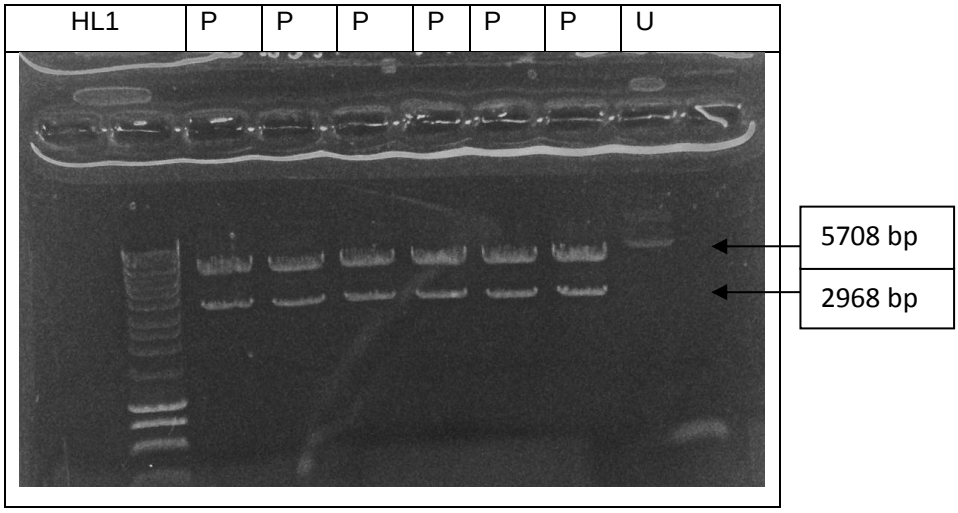


Figure 8.9: Restriction digest of pET15b-CDP

The plasmid Gs-pET15b-CDP (0.8 µg) (Lane U) was digested with BamHI (10 U) in Roche buffer B (20 µl) to give two bands (Lane P) corresponding to 5708 bp (backbone) and 2968 bp (insert), against Hyper Ladder 1 (Lane HL1) (NEB).

After extraction, the upper band had the terminal phosphate removed using Calf Intestinal Phosphatase (10 U, NEB) according to manufacturer instruction and the enzyme was removed using QIAquick PCR cleanup kit (Qiagen). The upper band (6 µl) and the lower band (2 µl) were religated using T4 DNA ligase (5 U, Fermentas) in T4 DNA ligase buffer. After 15 min the product (1 µl) was transformed into freshly prepared chemically competent TOP10 *E. coli* (50 µl) and transferred to LB medium (10 ml) containing carbenicillin. After 16 hours at 37 °C the plasmids were extracted using PureYield plasmid miniprep system (Promega). The purified plasmid mixture (1 µl) was transformed into freshly prepared chemically competent *E. coli* BL21 (50 µl) and plated on to carbenicillin containing LB agar. 10 colonies were picked into 10 ml LB and the subcultures were induced, once OD_{600} reached 0.6, with IPTG (1 mM) for 5 hours at 30 °C and analysed for expression of CDP by SDS-PAGE (Fig 8.10).

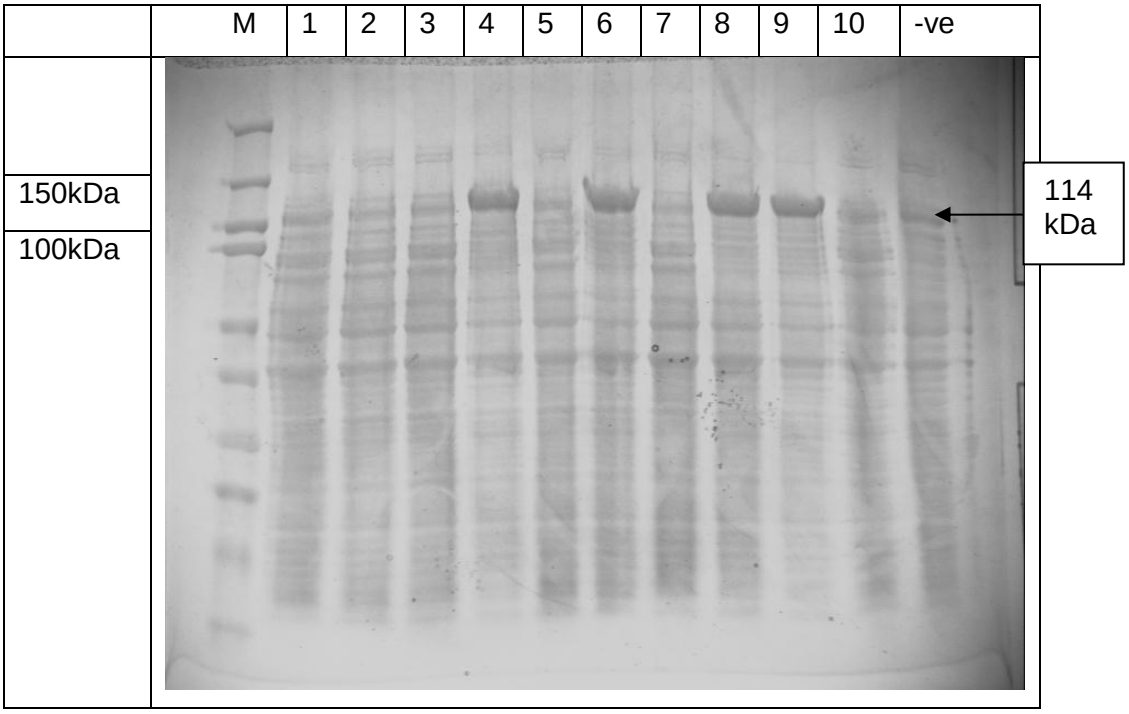


Figure 8.10: Trial expression of CDP

Whole cell lysate of colonies 1-10 (20 µl) were run on SDS-PAGE. M: Kaleidoscope protein standards (BioRad). –ve: induced cells not containing a plasmid.

Colony 4 was selected as containing plasmid pET15b-CDP (Fig 8.11), which was purified from the cells by mini prep as before.

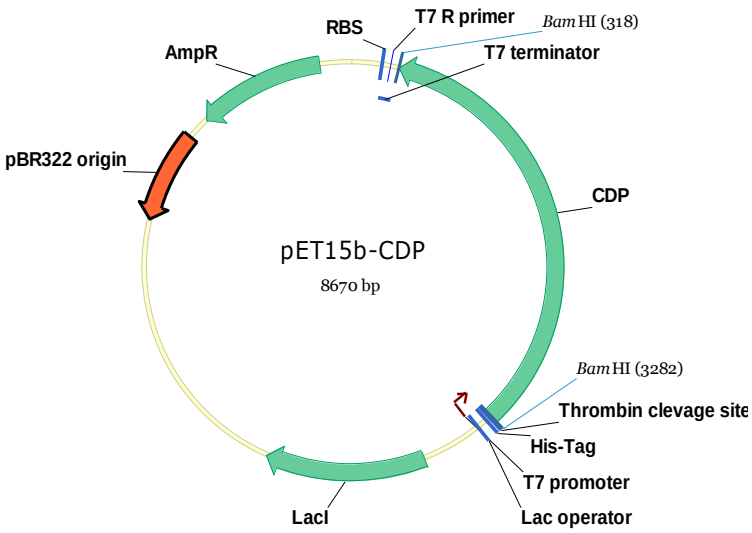


Figure 8.11: Plasmid map of pET15b-CDP showing BamHI restriction sites.

8.3.1.2 Expression of CDP

An overnight culture of Colony 4 cells (4x 1 ml) was used to inoculate 4x 1 litre cultures of LB which were grown at 37 °C, transferred to 30 °C and induced with 1 mM IPTG. After 4 hours the cells were harvested and frozen at -80 °C until required. A total yield of 10 mg/l culture of purified CDP was obtained and concentrated to 40 mg/ml with an Amicon Ultra-15 30 kDa spin concentrator (Millipore). Protein concentration was estimated at a wavelength of 280 nm using an extinction coefficient of 117010 1/M/cm, calculated from the amino acid composition,[13] and protein quality was analysed by SDS-PAGE.

8.3.2 Kinetic assays of CDP

8.3.2.1 Inhibition of CDP

Molybdate was found to be a good inhibitor of CDP (Fig 8.12) at the concentration used in the phosphorylase activity assay. This was carried out in the same way as for PHS2.

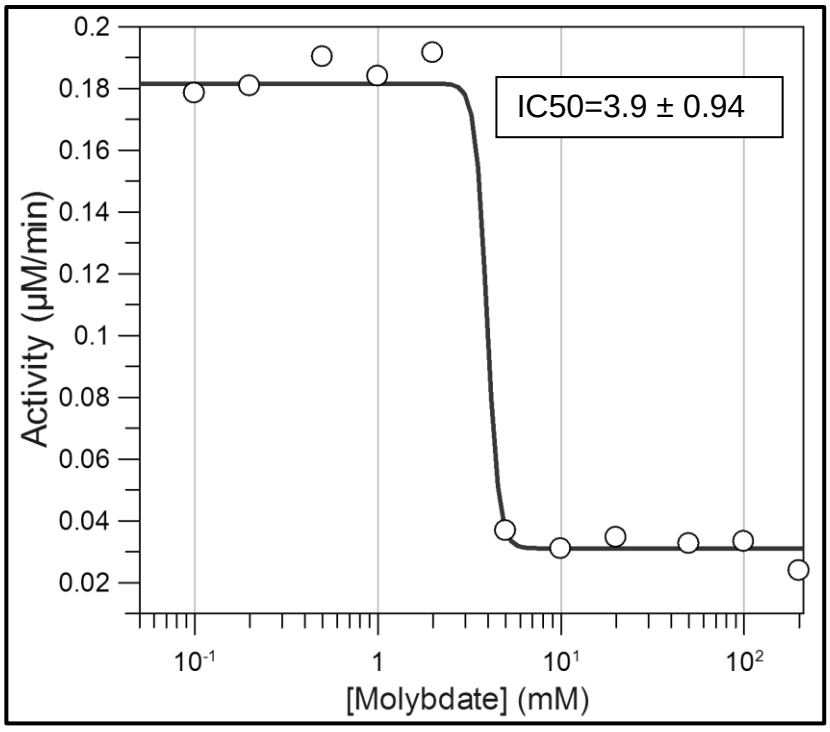


Figure 8.12: Inhibition of CDP by molybdate

Using Glc (10 mM) as the acceptor and Glc-1-P (10 mM) as the donor, the inhibition of CDP (50 µg/ml) by ammonium molybdate was assayed in the range 0.1-200 mM. Plotted using GraFit (Erithacus Software Ltd).

8.3.2.2 Donor specificity of CDP

The phosphorylase activity assay was used to measure addition of monosaccharides from their respective sugar-1-Ps on to Glc (10 mM) in one hour at 37 °C (Fig 8.13). There was a high background detected from Gal-1-P in the absence of enzyme, indicating release of phosphate due to hydrolysis, and Xyl-1-P is known to be a suitable donor for CDP.[19] Thus it was concluded that this assay was not suitable for measuring the donor specificity of CDP, and CE was used instead.

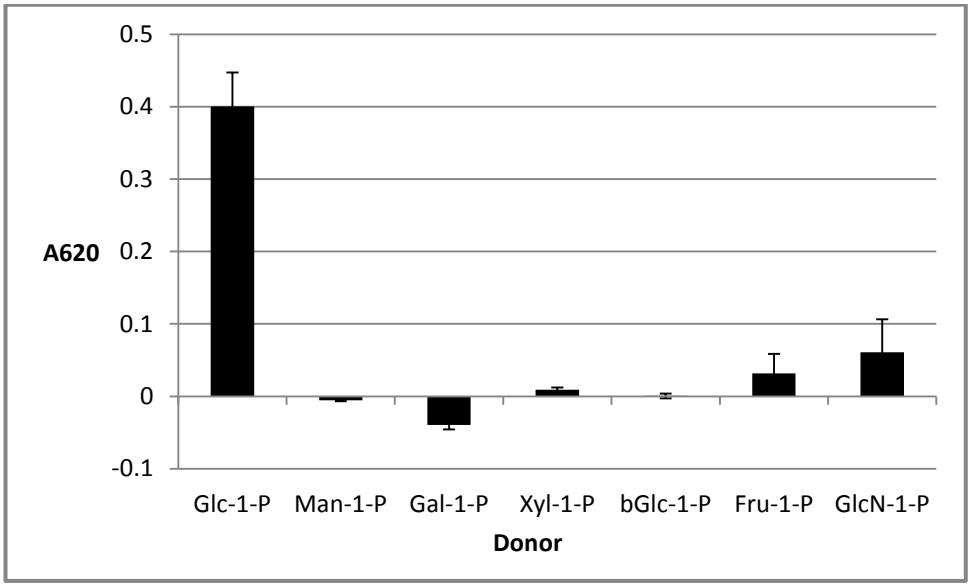


Figure 8.13: Donor specificity of CDP

Using standard conditions the amount of phosphate released from various sugar phosphates was assayed. It appears that only Glc-1-P is a suitable donor, though this does not agree with literature.

8.3.2.3 Acceptor specificity of CDP

The ability for CDP (50 µg/ml) to transfer Glc from Glc-1-P (10 mM) was assayed at 21 °C for 20 min in HEPES (50 mM, pH 7.5). Various acceptors were assayed over a range of concentrations to obtain kinetic parameters (Fig 8.14).

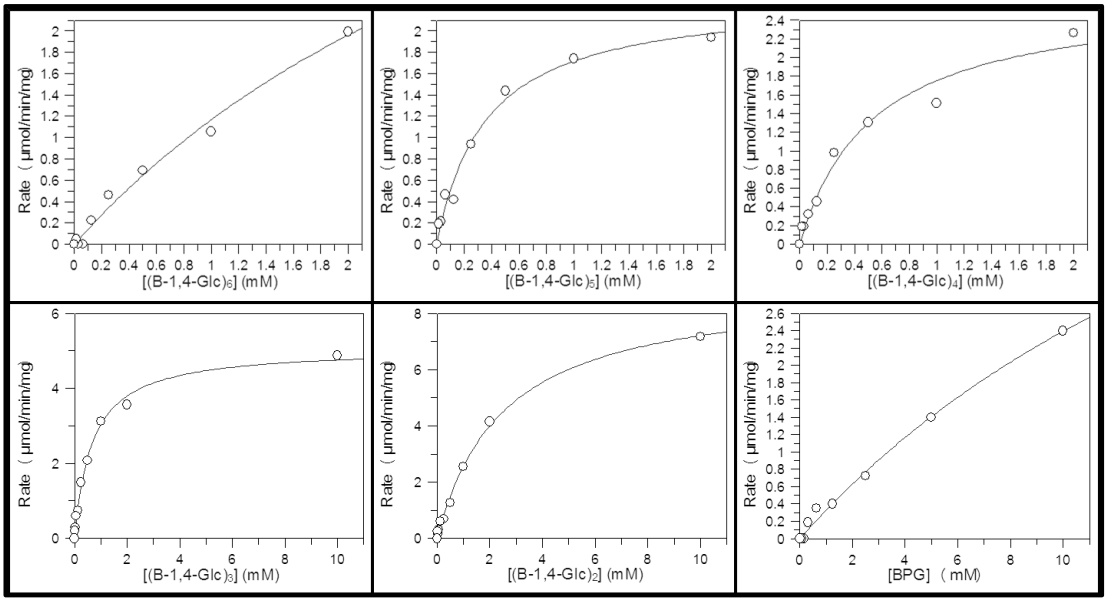
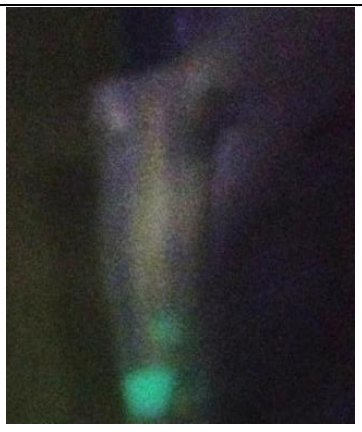


Figure 8.14: Acceptor length specificity of CDP

The enzyme activity was measured over a range of acceptor concentrations. The K_m^{app} was calculated for each using GraFit (Erithacus Software Ltd).

8.3.2.4 CE assays of CDP

Standard assays were carried out thereafter using CDP (5 µg/ml) at 40 °C with Glc-1-P (10 mM) and APTS-labelled acceptor (2 µM) in HEPES (50 mM, pH 7.5). After extended polymerisation the labelled acceptors became insoluble (Fig 8.15).



APTS-(β-1,4-Glc)$_6$ in solution	APTS-(β-1,4-Glc)$_6$ extended by CDP+Glc-1-P, after centrifugation

Figure 8.15: Insoluble product of CDP catalysed extension of APTS-(β-1,4-Glc)$_6$

After the CDP reaction on the APTS-(β-1,4-Glc)$_6$ had been terminated the APTS fluorescence had become insoluble and associated with the side of the tube, as visualised under UV light (254 nm), indicating that the glucans are insoluble.

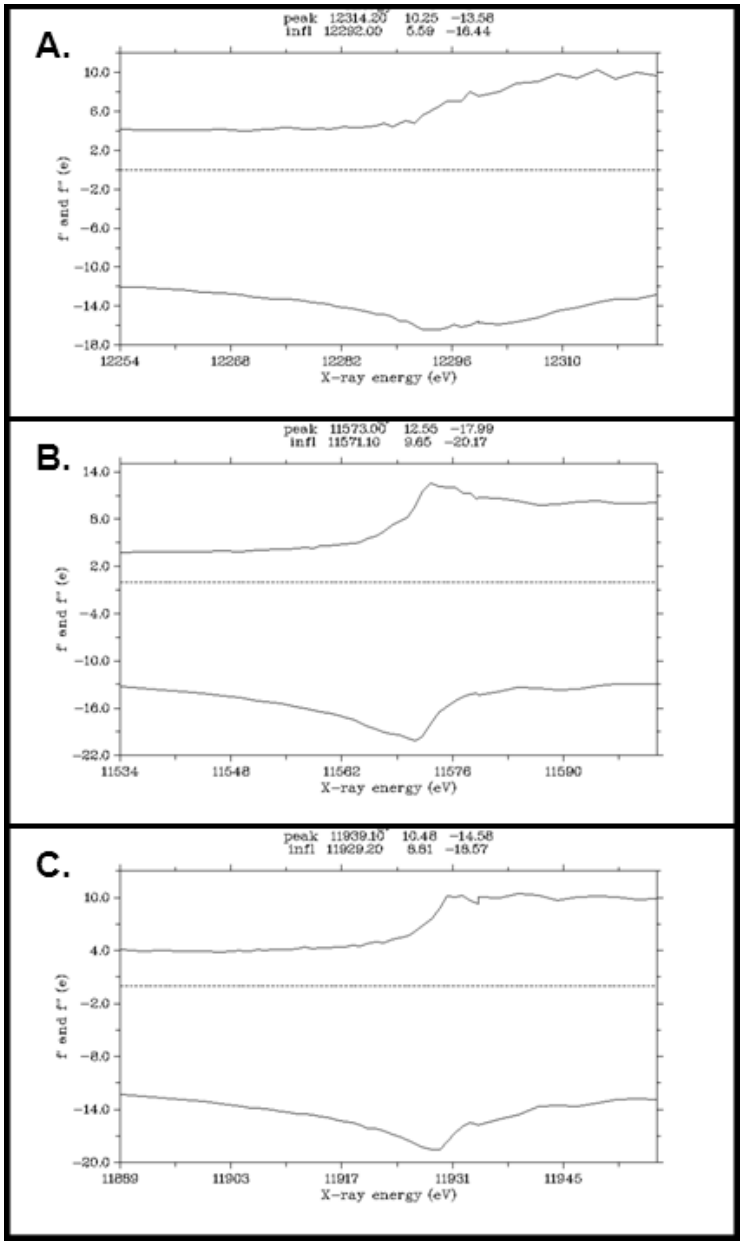


Figure 8.16: X-ray energy scans at the L_{III} absorption edge for heavy metal derivatives of CDP crystals

A. Thiomersal soaked crystals showed a weak edge at the expected 12290 eV.
B. Tetrachloroplatinate soaked crystals show a good absorption edge at 11568 eV.
C. Tetrachloroaurate soaked crystals show a good absorption edge at 11925 eV.

8.3.3 X-ray crystallography of CDP

8.3.3.1 Crystallisation of CDP

After setting up crystallisation screens hits, the CDP crystallisation solution was optimised to contain PEG 3350 (20% (w/v), KCl (300 mM), buffered to pH 7.5 (HEPES, 100 mM). However, CDP rapidly precipitated when mixed with this solution. The protein was pre-precipitated by mixing an equal volume of CDP crystallisation solution with CDP (10 mg/ml) in GF buffer and incubating on ice for 15 min. After centrifugation at 16,000 g for 1 min, crystallisation drops were set up by mixing 1 µl of the soluble protein with 1 µl of the crystallisation solution.

8.3.3.2 Heavy metal derivatisation

Saturated CDP cryoprotectant solutions (CDP crystallisation solution containing 20% glycerol) of thiomersal, potassium tetrachloroaurate and potassium tetrachloroplatinate were diluted 10x in cryoprotectant and used to soak crystals for 18 hours. After back soaking in cryoprotectant for 5 min, crystals were flash cooled in liquid nitrogen and transported to DLS for data collection. Prior to diffraction data collection, fluorescence scans were recorded across the appropriate L_{III} X-ray absorption edges from the heavy atom derivatised crystals; in all cases the expected absorption edges were detected (Fig 8.16). Subsequently diffraction data were collected on each of these at X-ray energies that were 50 eV above the corresponding absorption edges. The resultant data sets were processed to resolutions in the range 3.0-4.0 Å, but in every case the anomalous signal was too weak to facilitate structure solution.

8.3.3.3 Selenomethionine derivative of CDP

Cells containing plasmid pET15b-CDP were grown in SeMet minimal media[20] (SeMet MM: M9 salts, 0.2% glucose, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 10 mg/l thiamine, 20 ml of amino acid stock) containing 100 µg/ml carbenicillin. The amino acid stock contains Arg, Asp, Glu, Gln, His, Ile, Leu, Lys, Phe, Ser, Thr, Try, Tyr and Val (2 mg/ml). An overnight LB culture of cells (5 ml) was collected by centrifugation (4000 g) and washed with SeMet MM (3x 5 ml). These cells were then grown in SeMet MM (1 litre) at 37 °C until OD_{600} reached 0.6, when Lys, Phe and Thr (100 mg), Ile, Leu and Val (50 mg) and SeMet (60 mg) were added and the temperature adjusted to 30 °C. After a further 45 min, cells were induced with IPTG (1 mM) and after 16 hours the cells were collected by centrifugation and stored at -80 °C until required. The protein was purified using the standard protocol (section 8.1.2) to give a final yield of 12 mg of purified protein (Fig 8.17). This crystallised under the same conditions as the native protein and produce sufficient anomalous diffraction to facilitate structure solution.

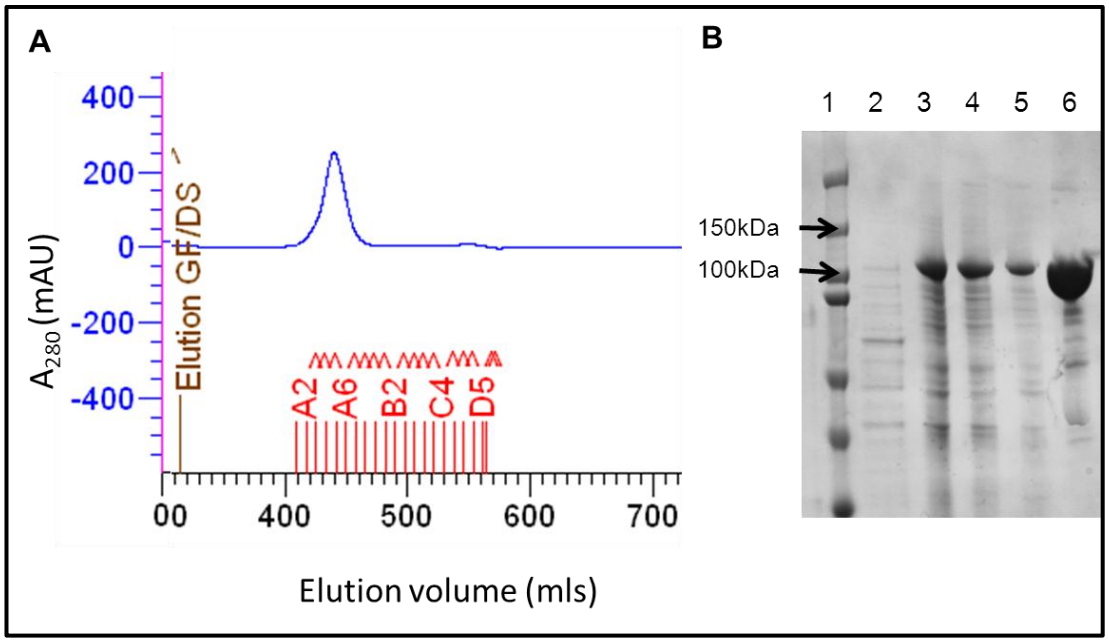


Figure 8.17: Purification of SeMet derivatised CDP

A. GF purification of SeMet-CDP after Ni-IMAC shows a single species, eluting at 120 ml. **B**. SDS-PAGE of protein purification. Lane 1: Kaleidoscope protein standards (Bio-Rad). Lane 2: total cell extract of uninduced cells. Lane 3: total cell extract of induced cells. Lane 4: insoluble cell material. Lane 5: soluble cell lysate. Lane 6: purified protein, showing a major band running just over 100 kDa.

8.4 Specific protocols for laminarin phosphorylase

8.4.1 Growth of *Euglena*

Euglena gracilis var. *saccharophila* Klebs was obtained from the culture collection of alga and protozoa (http://www.ccap.ac.uk/) (strain 1224/7a).

High nutrient media. For growth in the dark, cells were grown in the recommended 1x EG + 1x JM (Euglena gracilis medium plus Jaworski's Medium, replacing "Lab-Lemco" with Tryptone), supplemented with 15 g/l, glucose at 30 °C and shaken at 200 rpm in the dark for 7 days between sub-culturing. Dark grown cells were harvested by centrifugation at 800 g for 10 min with rapid decantation of the supernatant. This yielded over 10 g/l fresh weight cells which were frozen at -80 °C until required.

Low nutrient media. For growing the cells autotrophically the media was adjusted to contain no amino acids or sugars as energy sources. The medium consisted of $CaCl_2$ (0.1 g/l), $NaOAc{\cdot}(H_2O)_3$ (1 g/l) and 1x JM. Cultures were grown at 21 °C under ambient light and grew extremely slowly, taking 4 weeks between sub-culturing.

Agar plates were made by supplementing the media with 15 g/l agar (Duchefa) and cells were collected from the surface by resuspending in MQ-H_2O.

8.4.2 *Euglena* lysis

20 g of *Euglena* cells, which had been grown in the high nutrient media, were resuspended in 40 ml *Euglena* lysis buffer (10 mM HEPES NaOH, pH 7.0, 1 tablet cOmplete protease inhibitor cocktail (Roche), 1 mg DNase). Cells were broken using a cell disruptor (Constant Systems, 10 kpsi), centrifuged for 20 min at 30,000 g and the supernatant was filtered through 0.2 µm disc filter (Millipore) and frozen as 5 ml aliquots.

8.4.3 Synthesis of β-Glc-gol

8.4.3.1 Enzymatic synthesis of β-Glc-gol

Cleared *Euglena gracilis* cell lysate (100 µl) was added to 10 ml of reaction buffer (20 mM MES, pH 6.5) containing Glc-1-P (**29**, 100 mg, 0.38 mmol) and glycerol (**144**, 1 g, 11 mmol) and incubated at 30 °C for 4 days. Anion exchange chromatography was performed on the reaction mixture (DOWEX 1X2-400 OH^-). This mixture was concentrated *in vacuo* and dried by co-evaporation with toluene (3x 10 ml). The resulting solid, containing products **161**-**165**, was dissolved in pyridine (30 ml) and excess Ac_2O (10 ml) and the reaction was stirred overnight. The excess Ac_2O was destroyed by adding MeOH (10 ml) and the solvents were evaporated. Residual AcOH was removed by co-evaporation with toluene (2 x 10 ml). The residual solid was dissolved in EtOAc and washed with water (3 x 20 ml) and aqueous HCl (1 M, 10 ml), neutralised with aqueous $NaHCO_3$ (sat., 30 ml), washed with aqueous NaCl (sat., 20 ml) and dried over anhydrous $MgSO_4$ (1 g). The volatiles were removed by evaporation at reduced pressure. MALDI-ToF MS indicated $(Glc)_n$-gol where n=0-5 and $(Glc)_n$ where n=1-6, had been formed (Fig 5.8). The products were then separated using automated silica gel chromatography (Biotage) with a 25 g silica gel cartridge, using a Hex:EtOAc gradient (20-100% over 8 column volumes). The fractions with product were identified using TLC (EtOAc:Hex 2:1) and pooled accordingly.

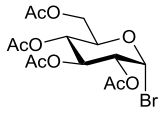
Fraction 1 Rf = 0.60 ESI-MS: m/z β-Glc-gol – (**150**)	Calculated $[C_{21}H_{30}O_{14}]Na^+$: 529.1533 Found: 529.1527
Fraction 2 Rf = 0.49 ESI-MS: m/z $(β\text{-Glc})_2$-gol – (**151**)	calculated $[C_{33}H_{46}O_{22}]Na^+$: 817.2378 Found: 817.2357
Fraction 3 Rf = 0.43 MALDI-ToF MS: m/z $(β\text{-Glc})_3$-gol – (**152**)	Calculated $[C_{45}H_{62}O_{30}]Na^+$: 1105.3223 Found: 1105.39, Calculated $[C_{45}H_{62}O_{30}]K^+$: 1121.2963 Found: 1121.39

All fractions which only contained products of Rf 0.6 were recombined, evaporated to dryness and the chromatography repeated. The purest fraction was then analysed by NMR. ^{1}H-NMR ($CDCl_3$) showed more than one compound was present, but a significant peak [δ: 4.60(d, $J_{1\text{-}2}$ 8 Hz)] matched the spectrum of the synthetic product (Fig 5.11).

8.4.3.2 Chemical synthesis of β-Glc-gol

A mixture of peracetylated (2R)-1-, (2S)-1- and 2-*O*-β-Glc-gol (**165**-**169**) was made according to Scheme 1 (section 5.2.6).

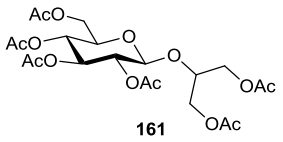
8.4.3.2.1 Chemical synthesis of *O*-2,3,4,6-tetra-*O*-acetyl-α-D-glucopyranosyl bromide (**189**)



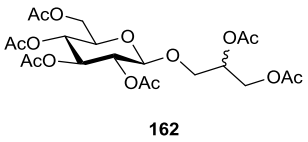
189

33% HBr in AcOH (25 ml, 102.5 mmol) was added to a solution of β-D-glucose pentaacetate (**159**, 10 g, 25.6 mmol, Fluka) in dry DCM (distillation, 10 ml) and the mixture was stirred at 0 °C for 3 hours. The mixture was washed with H_2O/ice, neutralised with aqueous $NaHCO_3$ (sat., 20 ml), washed with water (3 x 20 ml) and the organic layer was dried over anhydrous $MgSO_4$ (1 g). The product was recrystallised from diethyl ether to afford bromide **189**. ^{1}H NMR (400 MHz, $CDCl_3$) δ: 6.61 (1H, d, $J_{1,2}$ = 4.0 Hz, H-1), 5.56 (1H, dd, $J_{2,3}$ = $J_{3,4}$ = 9.5 Hz, H-3), 5.16 (1H,dd, $J_{3,4}$ = $J_{4,5}$ = 9.5 Hz, H-4), 4.84 (1H, dd, $J_{1,2}$ = 4.0 Hz, $J_{2,3}$ = 9.5 Hz, H-2), 4.35 – 4.28 (2H, m, H-5, H-6_a), 4.13 (1H, m, H-6_b), 2.11 (3H, s, OAc), 2.10 (3H, s, OAc), 2.06 (3H, s, OAc), 2.03 (3H, s, OAc).

8.4.3.2.2 Chemical synthesis of *O*-2,3,4,6-tetra-*O*-acetyl-β-D-glucopyranosyl-(1→2)-*O*-1,3-di-*O*-acetyl-glycerol (**161**) and *O*-2,3,4,6-tetra-*O*-acetyl-β-D-glucopyranosyl-(1→1)-*O*-2,3-di-*O*-acetyl-glycerol (**162**)



161



162

Glycerol (**144**, 2.24 g, 24.3 mmol) was dried by co-evaporation with toluene (3 x 20 ml) and mixed with dry DCE (distillation, Drierite, 30 ml) to which was added $HgBr_2$ (263 mg, 0.73 mmol), $Hg(CN)_2$ (1.474 g, 5.8 mmol)[21] and glucosyl bromide **189** (2 g, 4.9 mmol). The reaction was refluxed for 2 hours and stirred at room temperature for 3 days. The solution was washed with aqueous KI (sat., 20 ml) and water (3 x 20 ml), dried over anhydrous $MgSO_4$ (1 g) and the volatiles were evaporated under reduced pressure. The crude reaction mixture was

dissolved in pyridine (30 ml) and excess Ac_2O (10 ml) was added. The mixture was stirred overnight at RTP and the excess Ac_2O was destroyed by adding MeOH (10 ml). The solvents were evaporated and the residual AcOH was removed by co-evaporating with toluene (2 x 10 ml). The residue was dissolved in EtOAc and washed with water (3 x 20 ml) and aqueous HCl (1 M, 10 ml), neutralised with aqueous $NaHCO_3$ (sat., 30 ml), washed with aqueous NaCl (sat., 20 ml) and dried over anhydrous $MgSO_4$ (1 g). The volatiles were removed by evaporation at reduced pressure. The products were identified as per-acetylated mono-, di- and tri- glucosylated glycerol (**165**-**169**) by MALDI-ToF MS (Fig 5.10). The products were then separated using automated silica gel chromatography (Biotage) with a 25 g silica gel cartridge, using a Hex:EtOAc gradient (20-100% over 8 column volumes).

^{1}H-NMR ($CDCl_3$) was performed on one fraction with the single product with Rf 0.6, showing 6 signals for the anomeric proton which approximately matched literature signals for peracetylated (2R/2S)-β-D-glucopyranosyl-1-*O*-glycerol (**148**, **149**) and β-D-glucopyranosyl-2-*O*-glycerol (**147**).

^{1}H NMR (400 MHz, $CDCl_3$) δ: 5.20 – 5.15 (2.6 H, m, H-3',H2 (**148**+**149**)), 5.08 – 4.96 (3.2 H, m, H-4', H2'), 4.63 (0.6 H, d, $J_{1,2}$ = 8 Hz, H1' (**147**)), 4.52 (1 H, d, $J_{1,2}$ = 8 Hz, H1' (**148**+**149**)), 4.30 – 4.03 (8.8 H, m, H1, H2 (**147**), H3 (**147**), H5', H6'), 3.97 – 3.91 (1 H, m, H3a (**148**+**149**)), 3.72 – 3.65 (2.6 H, m, H5', H3b (**148**+**149**)), 2.07 – 1.98 (29 H, m, OAc)

MALDI-ToF-MS: m/z (**147,148,149**)
Calculated $[C_{21}H_{30}O_{14}]Na^+$: 529.1533
Found: 528.98

8.4.4 Laminarin phosphorylase purification

In order to purify the laminarin phosphorylase a range of purification strategies were explored and two multi-step protocols were developed.

8.4.4.1 Assay for laminarin phosphorylase

To assay for laminarin phosphorylase 6.25 µl of a 2x assay buffer (200 mM MES, pH 7.0, 20 mM Glc, 20 mM Glc-1-P) was incubated with a 6.25 µl assay sample in a microtitre plate at 30 °C for 5 min to 16 hours as appropriate and the phosphorylase activity assay was then performed.

The most active fractions were confirmed by reacting 50 µl of purified enzyme in 100 µl assay mixture (9.2 µM APTS-laminaritriose, 20 mM Glc-1-P, 20 mM MES, pH 7.0) at 30 °C for appropriate time. After heating to 95 °C in a boiling water bath the samples was centrifuged at 16,000 g for 2 min and capillary electrophoresis was performed.

8.4.4.2 Hydrophobic interaction chromatography (HIC)

Four different 1 ml hydrophobic interaction columns (GE Healthcare) were tested for purification of the LDP from the *Euglena* lysate (Fig 8.18). *Euglena* cleared cell lysate (5 ml) was mixed with $(NH_4)_2SO_4$ (5 ml, 2 M) and loaded on to the HIC columns, pre-equilibrated with $(NH_4)_2SO_4$ (1 M) in HIC Buffer (5 mM MES, pH 6.0) on an AKTA Xplor FPLC (GE healthcare). The protein was eluted by decreasing $(NH_4)_2SO_4$ to 0 M over 20 column volumes and 1 ml fractions were collected.

It was determined that the butyl column gave the best separation of LDP activity from total protein and so was used in further purification steps, using a gradient of 1-0 M $(NH_4)_2SO_4$ in MES (5 mM, pH 6.0) to elute.

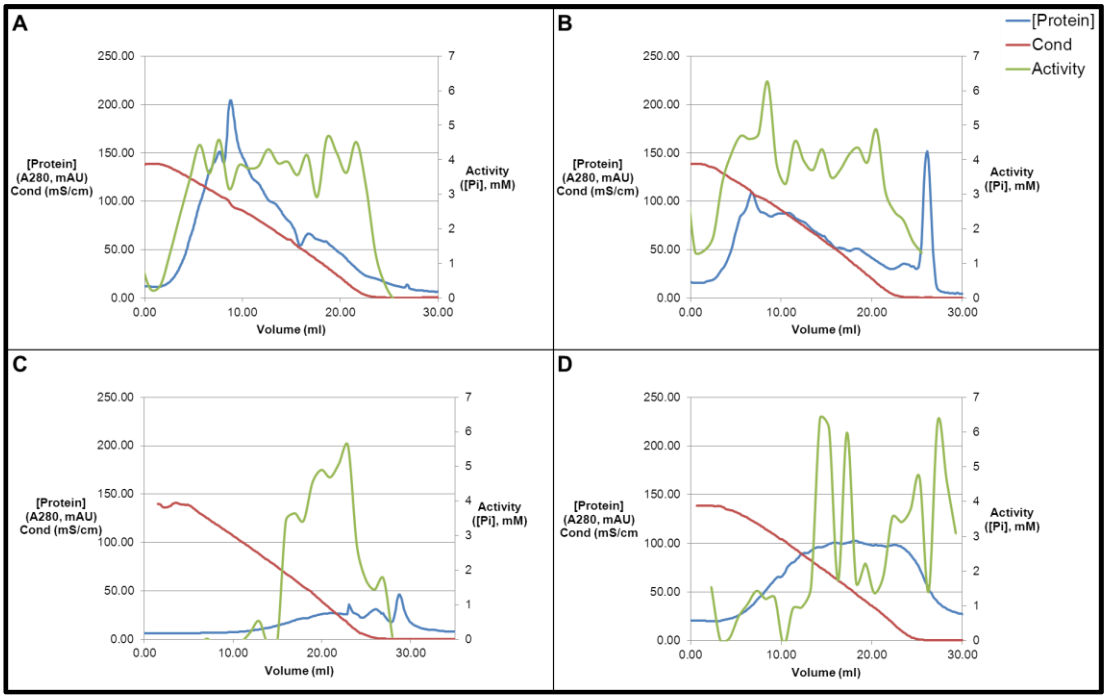


Figure 8.18: HIC purification of LDP

Separation of LDP from *Euglena* cell lysate on four HIC columns (butyl (**A**), octyl (**B**), phenyl low FF (**C**) and phenyl high FF (**D**)) eluting with a 1-0 M $(NH_4)_2SO_4$ gradient in MES (5 mM, pH 6.0). Conductivity (Cond, measure in mS/cm) indicates the reduction in $(NH_4)_2SO_4$ concentration. The amount of Pi released from the assay in 30 min was measured to give an indication of fractions containing LDP activity. Blue = protein, measured as A_{280}. Green = activity, measured as Pi released in the assay. Red = salt concentration, measured as conductivity.

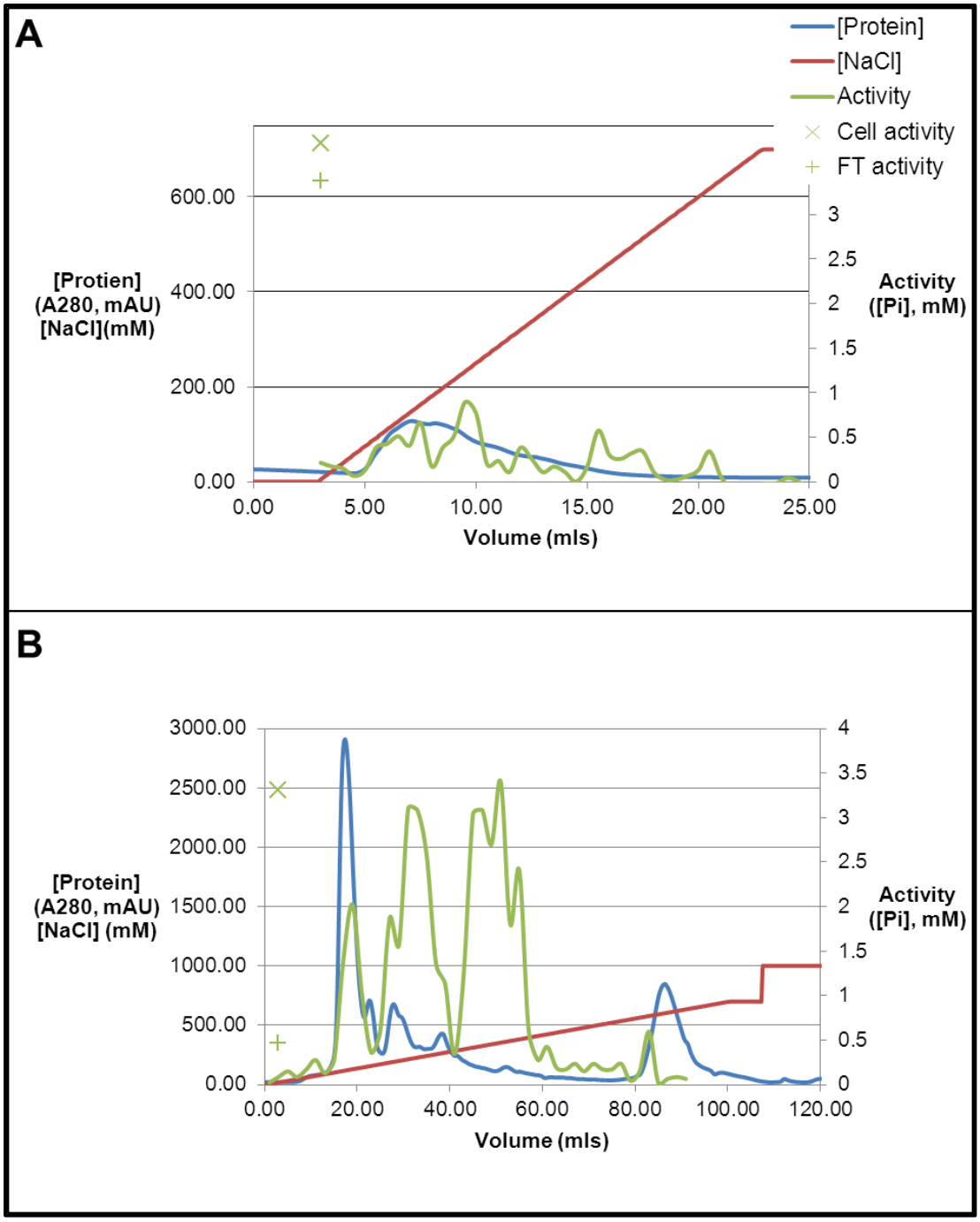


Figure 8.19: Ion exchange chromatography purification of LDP

A. CIEX separation of LDP from *Euglena* cell lysate, on a 1 ml Sepharose SP column. The activity in the flow-through (FT) is similar to the activity in the whole cell lysate. **B**. AIEX separation of LDP from *Euglena* cell lysate, on a 5 ml HiTrap Sepharose Q column. The activity separated from the total protein. Blue = protein, measured as A_{280}. Green = activity, measured as Pi released in the assay. Red = salt concentration, measured as conductivity.

8.4.4.3 Ion exchange chromatography

The suitability of ion exchange chromatography for the purification of the laminarin phosphorylase form *Euglena* lysate was tested.

8.4.4.3.1 Cation exchange chromatography (CIEX)

Cleared cell lysate (5 ml) was mixed with an equal volume of CIEX Buffer (25 mM MES NaOH, pH 5.8). This was loaded on to a 1 ml HiTrap Sepharose SP (GE Healthcare) CIEX column, pre-equilibrated with CIEX Buffer and eluted with a gradient of NaCl (0-700 mM) in CIEX Buffer over 20 column volumes, collecting 0.5 ml fractions (Fig 8.19.A).

The activity was not retained on the column at all and so no further attempts were made to use CIEX to purify LDP.

8.4.4.3.2 Anion exchange chromatography (AIEX)

Cleared cell lysate (5 ml) was mixed with an equal volume of AIEX Buffer (25 mM TrisHCl, pH 8.5). This was loaded on to a 5 ml HiTrap Sepharose Q (GE Healthcare) AIEX column, pre-equilibrated with AIEX Buffer and eluted with a gradient of NaCl (0-700 mM) in AIEX Buffer over 20 column volumes, collecting 2 ml fractions (Fig 8.19.B).

The activity was retained on the column and eluted as two distinct peaks of activity, away from the largest protein peaks.

8.4.4.3.3 Optimisation of pH for AIEX

In order to determine the best pH for separation of LDP activity using AIEX 20 g of Euglena cells were lysed into in 50 ml lysis buffer containing 5 mM TrisHCl, pH 8.0. Cells were lysed and the lysate cleared, as above (section 8.4.2), and four 10 ml aliquots were mixed with equal volumes of TrisHCl (50 mM, pH 7.5, 8.0, 8.5, or 9.0). Samples were loaded on to a pre-equilibrated 5 ml Hi Trap SepaharoseQ column (GE Healthcare) and eluted with a NaCl gradient (0-700 mM) in 50 mM TrisHCl at the appropriate pH (Fig 8.20).

It was determined that a gradient of 0-700 mM NaCl at pH 8.5 gave the best separation of activity (released Pi from the assay) from protein (as measured by A280) and this pH was used in further purification. 10% of the fraction volumes of HEPES NaOH (1 M, pH 7.0) were added to each collection tube to neutralise the fractions as they were collected.

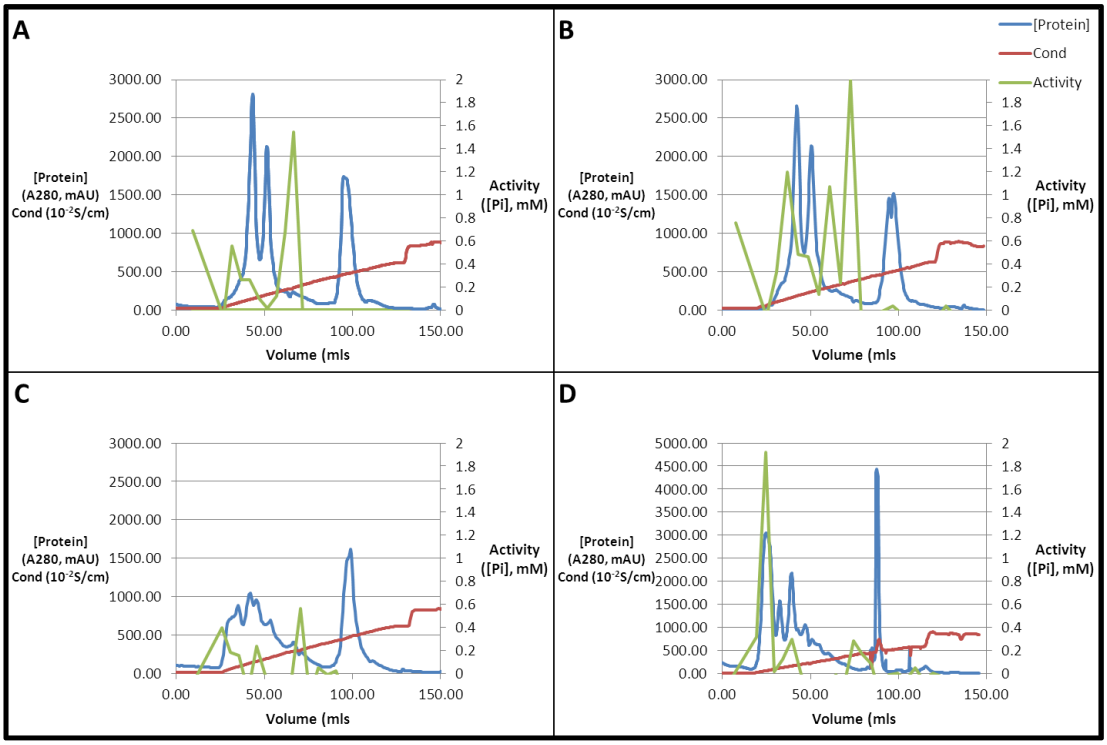


Figure 8.20: AIEX purification of LDP at various pHs

Separation of LDP from *Euglena* cell lysate, on a 5 ml HiTrap Sepharose Q column (50 mM TrisHCl) at pH 7.5 (**A**), 8.0 (**B**), 8.5 (**C**) and 9.0 (**D**), eluted with a 0-700 mM gradient of NaCl. Blue = protein, measured as A_{280}. Green = activity, measured as Pi released in the assay. Red = salt concentration, measured as conductivity.

8.4.4.4 Gel filtration (GF)

40 g of Euglena cleared cell lysate was further clarified by centrifuging at 100,000 g for 1 hour and the supernatant was concentrated by spin filtration, with 100 kDa cutoff (Millipore), to 8 ml. The whole rententate was applied to a Hi-Load 26/60 Superdex 200 gel filtration column (GE healthcare), pre-equilibrated in GF buffer and eluted at 3.2 ml/min (Fig 8.21). 5 ml fractions were collected between 100 and 300 ml. The activity eluted with the highest protein concentration, but efficiently removed around 75% of the total protein under mild conditions.

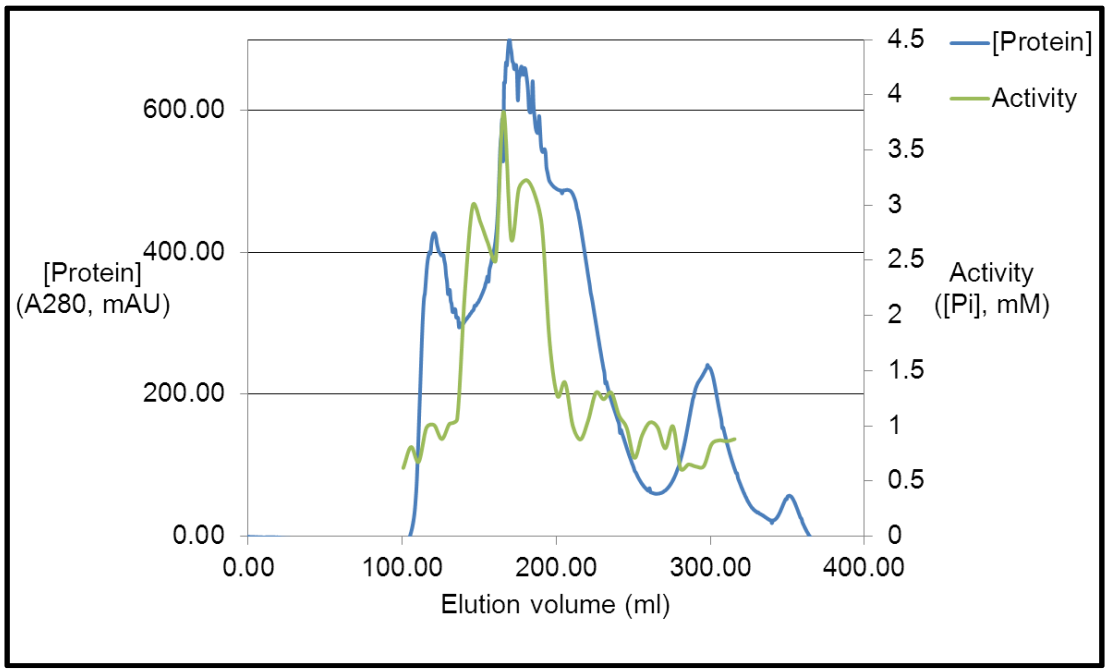


Figure 8.21: GF purification of LDP

Separation of LDP from *Euglena* cell lysate, on a Hi-Load 26/60 Superdex 200 column, eluted with GF buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 3.2 ml/min). Blue = protein, measured as A_{280}. Green = activity, measured as Pi released in the assay.

8.4.4.5 First purification protocol

Laminarin phosphorylase was purified by AIEX and fractions 38-40 ml (Eg1.1) and 60-62 ml (Eg1.2) were separately purified by HIC (Fig 8.22). The fraction at 18 ml from each HIC (Eg2.1 and Eg2.2) was precipitated with 50% TCA, run on SDS-PAGE and proteomic analysis performed on the region 100-150 kDa (Appendix 3).

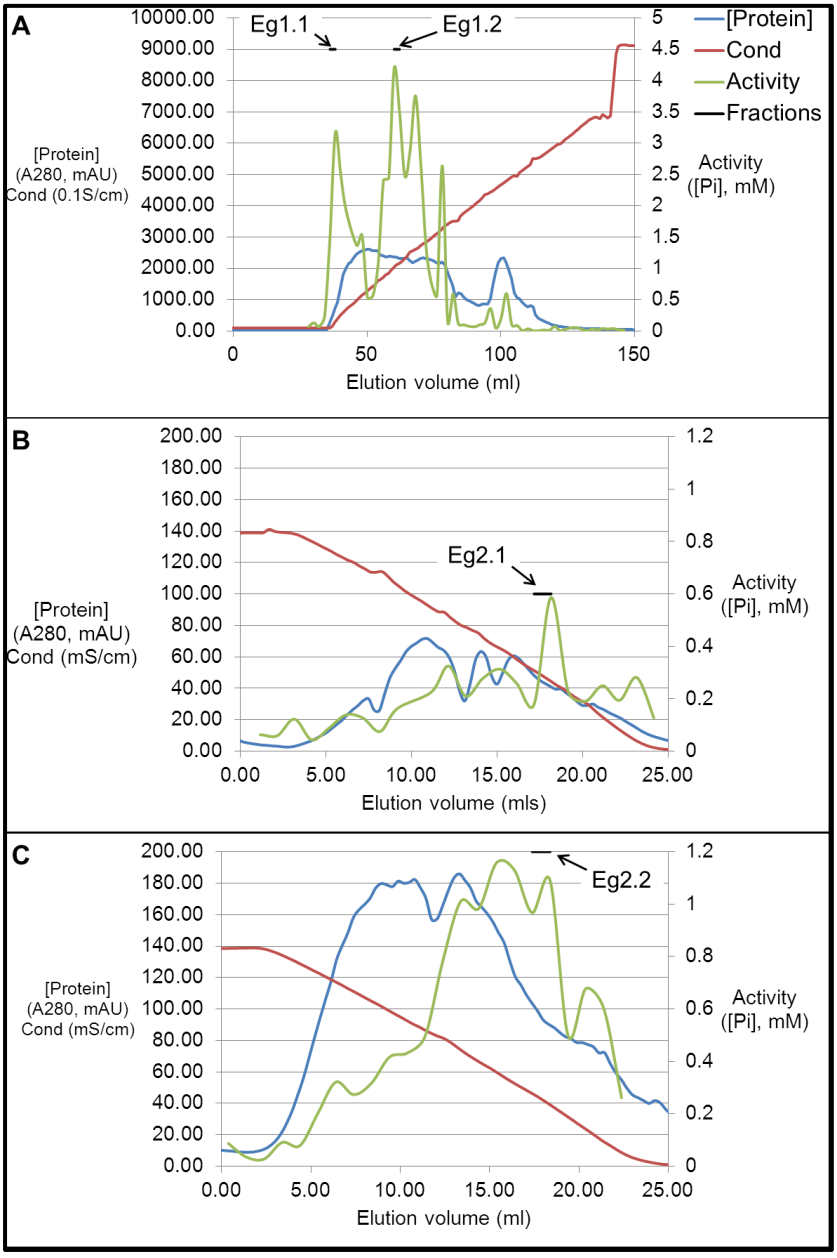


Figure 8.22: Purification of LDP by the first purification protocol

A. AIEX, using a 5 ml Sepahrose Q column, in HEPES (50 mM, pH 8.5) eluted with a 0-700 mM NaCl gradient. Fractions at 38-40 and 60-62 (Eg1.1 and 1.2) showed the highest activity. **B**. HIC, using a butyl column eluted with a 1-0 M $(NH_4)_2SO_4$ gradient in MES (5 mM, pH 6.0), on Eg1.1. The most active fraction was at 18 ml (Eg 2.1). **C**. HIC on Eg1.2. The most active fraction was at 18 ml (Eg2.2).

8.4.4.6 Second purification protocol

Laminarin phosphorylase was first purified by GF and the most active fraction was at 145-150 ml (EgA) (Fig 8.23). HIC was performed on this fraction and the activity eluted at 12-12.5 ml (EgB). Very little activity was detected after AIEX but fractions at 19-20 ml and 30-31 ml (EgC and D) were collected, based on previous elution profiles, for proteomics (Appendix 3).

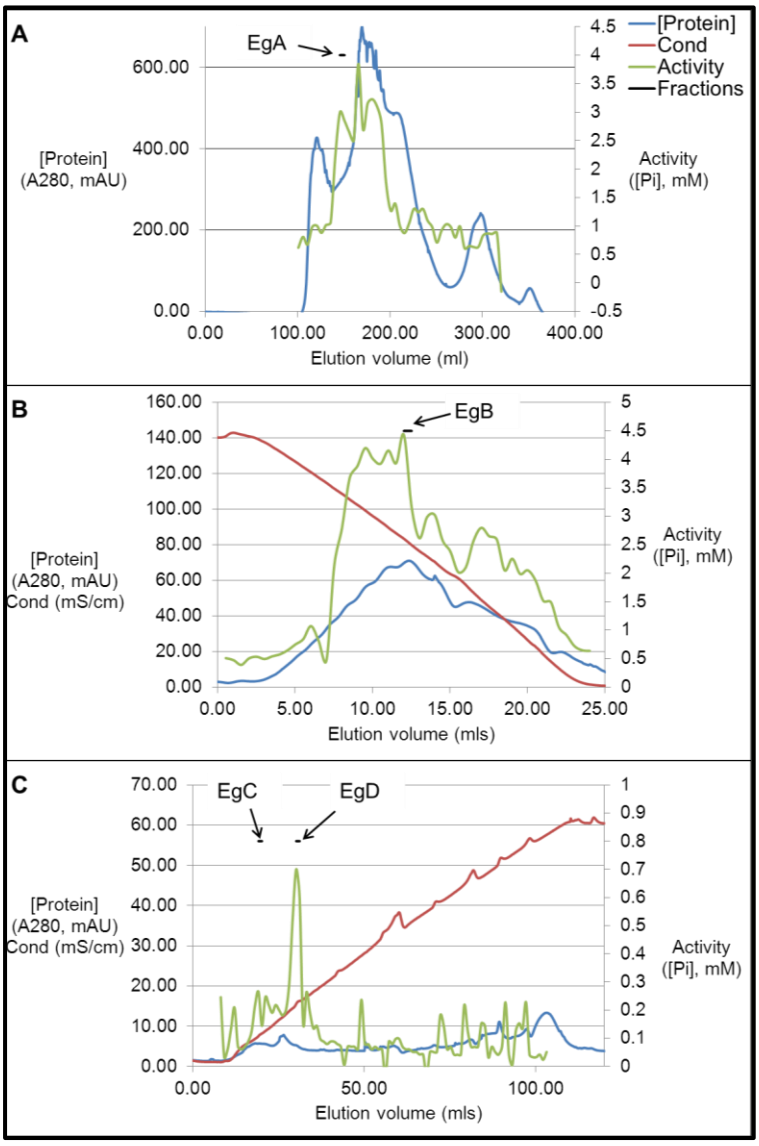


Figure 8.23: Purification of LDP by the second purification protocol

A. Separation of LDP from *Euglena* cell lysate, on a Hi-Load 26/60 Superdex 200 column (50 mM HEPES, pH 7.5, 150 mM NaCl). The most active fraction was at 145-150 ml (EgA). **B**. HIC, using a butyl column eluted with a 1-0 M $(NH_4)_2SO_4$ gradient in MES (5 mM, pH 6.0), on EgA. The most active fraction (EgB), was at 12-12.5 ml. **C**. AIEX on EgB, using a 5 ml Sepahrose Q column, in HEPES (50 mM, pH 8.5) eluted with a 0-700 mM NaCl gradient. Very little activity was detected, but two fractions were collected (EgC and EgD).

8.5 Specific protocols for *Euglena* transcriptome sequencing

8.5.1 RNA extraction and sequencing

Transcriptome assembly and automated annotation was performed by M. Trick (John Innes Centre).

Total RNA was isolated from 10^6 cells grown in: high nutrient media in the dark; and low nutrient media agar in ambient light, using RNeasy Minikit (Qiagen) with the assiatnce of Tom Turner (JIC), and stored at -80 °C. Purification of the mRNA, library preparation and sequencing on an Illumina HiSeq 2000 platform was performed by Source Bioscience (Nottingham, UK). A total of 313,205,944 and 137,689,062 100 base, paired-end reads were obtained from the light- and dark-grown samples respectively. The light-grown sequence data was down-sampled to 100 million pairs of reads, and the entire dataset of 68,844,531 read-pairs from the dark-grown sample were used for separate *de novo* assemblies using Trinity[30] version r2013-02-25, executed in parallel on two 256GB cluster nodes. 233,748 and 231,176 raw assemblies were generated from light and dark respectively. Likely coding sequences were extracted from these using the Perl utility script supplied with the Trinity distribution, transcripts_to_best_scoring_ORFs.pl, producing 45,126 and 47,607 candidate ORFs of lengths greater than 100 amino acids. These were combined and then a non-redundant set of 32,128 peptides produced using CD-HIT[22] (v4.5.4) with an identity threshold of 0.95 and a word length of 5. For functional annotation, the set was queried against the UniRef100 protein database using BLASTP with an E-value threshold of 10^{-10}. Sequence identifiers of the best hits were harvested and used to programmatically collect from databases, via SOAP-based web services, GO terms and KEGG pathway objects, and also used to enumerate kingdom-level taxonomic classifications for the species of origin. Transcript levels for isoforms were estimated from the reads using the RSEM wrapper script supplied in the Trinity distribution to implement the RNA-Seq by Expectation Maximisation[23] methods.

8.5.2 Annotation of CAZys

The annotation of CAZys was performed by B. Henrissat (Centre National de la Recherche Scientifique, Marseille, France). Putative carbohydrate-active enzymes were identified and classified using the methods used for updating the carbohydrate-active enzymes database (CAZy; www.cazy.org) as described previously.[24] Briefly, the translated protein sequences of *Euglena* were compared to the full-length sequences derived from CAZy using BLAST.[25] The sequences that gave an E-value <0.1 were then subjected to a second BLAST search against a library made with the constitutive modules of glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases and carbohydrate esterases and their associated carbohydrate-binding modules (CBM). In parallel, the sequences were subjected to a HMMer search[26] using hidden Markov models built for each CAZy module family. A protein was considered reliably assigned when it was placed in the same family by the two methods. Difficult cases were resolved by manual inspection of conserved features such as catalytic residues.

8.5.4 Fatty acid analysis of *Euglena* extracts

Cells grown in the high nutrient media at 21 °C with ambient light (100 mg) were vortexed with fatty acid extraction/methylation mixture (500 µL, methanol:toluene:2,2-dimethoxypropane:sulphuric acid 33:14:2:1) and heptane (250 µL). Samples were heated at 80 °C for 1 hour and 1 µl of the heptane fraction was injected on a 7HG-G015-02 GC capillary 30 m x 250 µm (Zebron) with 1 ml/min helium over a temperature gradient from 140-380 °C in 30 min attached to a Network GC (Agilent) and analysed by mass spectrometry (Inert, Agilent). Compounds were identified using Chemstation to query the NIST library and by manual identification.

8.5.3 Identification of thiols in *Euglena*

Purification of thiols from *Euglena* cells was performed by R. Dusi and C. Hamilton (University of East Anglia, Norwich, UK). Cells grown in the high nutrient media, at 21 °C with ambient light (20 mg), were collected by centrifugation at 13,000 g, flash cooled in liquid nitrogen and stored at -80 °C until analysis. Cells were extracted into 50% acetonitrile at 60 °C and the thiols were labelled with monobromobimane[27] (Invitrogen) and separated on a ACE AR C18 4.6 x 250 mm, 5 µm (HiChrom) by HPLC (JASCO) with fluorescence detection (JASCO FP2020 Plus) using excitation at 385 nm and emission at 460 nm. The gradient used started from 0.225% aqueous acetic acid, pH 4.0 to 90% aqueous methanol over 50 min, held for 2 min, reduced to 0% over 2 min and held for 10 min with a flow rate of 1 ml/min.[28]

Fluorescent peaks were collected and further analysed by L. Hill (John Ines Centre) using LC-MS on a Surveyor HPLC attached to a DecaXPplus ion trap MS (Thermo). Separation was on a Kinetex XB-C18 50 x 2.1 mm, 2.6 µm column (Phenomenex) running the following gradient of acetonitrile (B) versus 0.1% formic acid in water (A) at 300 µl/min and 30 °C: 0 min, 2% B; 10 min, 30% B; 20min, 95% B; 22 min, 95% B; 22.5 min, 2% B; 26 min, 2% B. Bimane-labelled thiols were detected by UV at 390 nm, and positive ESI-MS. MS spectra were collected from m/z 250-2000 and MS2 spectra of the most abundant ions were collected at an isolation width of m/z 4.0 and 35% collision energy. Spray chamber conditions were 50 units sheath gas, 5 units aux gas, 350 °C capillary temperature, and a spray voltage of 3.8 kV using a steel needle kit.

8.5.5 Identification of siderophore production by *Euglena*

The chrome azurol S (CAS) assay was used to show the production of siderophores.[29] High nutrient media agar was made containing CAS (60 mg/l), hexadecyltrimethylammonium bromide (73 mg/l) and additional $FeCl_3$ (2.7 mg/l) and cells were diluted such that approximately 10 cells were plated per petri dish. After two weeks a halo, indicating the presence of a siderophore, was clearly visible.

8.6 References

1 Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475-W478 (2011).

2 De Groeve, M. R. M. *et al.* Development and application of a screening assay for glycoside phosphorylases. *Anal. Biochem.* **401**, 162-167 (2010).

3 Leslie, A. G. The integration of macromolecular diffraction data. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **62**, 48-57 (2006).

4 Kabsch, W. XDS. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **66**, 125-132 (2010).

5 Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **62**, 72-82 (2006).

6 Bailey, S. The CCP4 suite - programs for protein crystallography. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **50**, 760-763 (1994).

7 Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **60**, 2126-2132 (2004).

8 Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **53**, 240-255 (1997).

9 Goubet, F., Jackson, P., Deery, M. J. & Dupree, P. Polysaccharide analysis using carbohydrate gel electrophoresis: a method to study plant cell wall polysaccharides and polysaccharide hydrolases. *Anal. Biochem.* **300**, 53-68 (2002).

10 Evangelista, R. A., Liu, M.-S. & Chen, F.-T. A. Characterization of 9-aminopyrene-1,4,6-trisulfonate derivatized sugars by capillary electrophoresis with laser-induced fluorescence detection. *Anal. Chem.* **67**, 2239-2245 (1995).

11 Prifti, E., Goetz, S., Nepogodiev, S. A. & Field, R. A. Synthesis of fluorescently labelled rhamnosides: probes for the evaluation of rhamnogalacturonan II biosynthetic enzymes. *Carbohydr. Res.* **346**, 1617-1621 (2011).

12 Thiery, J. P. Demonstration of polysaccharides in thin sections by electron microscopy. *J. Microsc.-Oxf.* **6**, 987-1018 (1967).

13 Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption-coefficient of a protein. *Protein Sci.* **4**, 2411-2423 (1995).

14 Zhi, Z. L. *et al.* A versatile gold surface approach for fabrication and interrogation of glycoarrays. *ChemBioChem* **9**, 1568-1575 (2008).

15 Kilburn, D., Claude, J., Schweizer, T., Alam, A. & Ubbink, J. Carbohydrate polymers in amorphous states: an integrated thermodynamic and nanostructural investigation. *Biomacromol.* **6**, 864-879 (2005).

16 Enustun, B. V. & Turkevich, J. Coagulation of colloidal gold. *J. Am. Chem. Soc.* **85**, 3317-3328 (1963).

17 Brzozowski, A. M. *et al.* Structural analysis of a chimeric bacterial α-amylase. High-resolution analysis of native and ligand complexes. *Biochemistry* **39**, 9099-9107 (2000).

18 Sauer, J. *et al.* Glucoamylase: structure/function relationships, and protein engineering. *Biochim. et Biophys. Acta - Prot. Struc. Mol. Enzymol.* **1543**, 275-293 (2000).

19 Shintate, K., Kitaoka, M., Kim, Y. K. & Hayashi, K. Enzymatic synthesis of a library of β-(1->4) hetero-D-glucose and D-xylose-based oligosaccharides employing cellodextrin phosphorylase. *Carbohydr. Res.* **338**, 1981-1990 (2003).

20 Doublié, S. in *Methods Enzymol.* Vol. Volume 276 (ed Charles W. Carter, Jr.) 523-530 (Academic Press, 1997).

21 Helferich, B. & Zirner, J. Zur Synthese von Tetraacetyl-hexosen mit freiem 2-Hydroxyl. Synthese einiger Disaccharide. *Chem. Ber. Recl.* **95**, 2604-2611 (1962).

22 Li, W. & Godzik, A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

23 Li, B. & Dewey, C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

24 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-D238 (2009).

25 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

26 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).

27 Newton, G. L. *et al.* Distribution of thiols in microorganisms: mycothiol is a major thiol in most actinomycetes. *J. Bacteriol.* **178**, 1990-1995 (1996).

28 Koledin, T., Newton, G. & Fahey, R. Identification of the mycothiol synthase gene (mshD) encoding the acetyltransferase producing mycothiol in actinomycetes. *Arch. Microbiol.* **178**, 331-337 (2002).

29 Schwyn, B. & Neilands, J. B. Universal chemical assay for the detection and determination of siderophores. *Anal. Biochem.* **160**, 47-56 (1987).

Appendices

Appendix 1 – PHS2 sequencing

Plasmid pET151-PHS2 was sequenced using the primers detailed below. This gave faithful sequence over the full length of the PHS2 insert (5764-8286 bp).

PHS2 Sequencing primers

AtPHS2 SEQ F1	ATGGTTTGAGGTACAGACATGG
AtPHS2 SEQ F2	GGAAATGGTCAGAGTTTCCA
AtPHS2 SEQ F3	AGGTGGTTACGTTTCTGCAG
AtPHS2 SEQ F4	AGCGGAGATGCTAATACCCG
AtPHS2 SEQ R1	AGGCAGATTTAGGGTGGCC
AtPHS2 SEQ R2	GCTGCCTTCAGTGGTGCTC
AtPHS2 SEQ R3	TCGAGGTGTGATGCCATTAG
AtPHS2 SEQ R4	GCTACAGTGACATTGTAGTTTGG

```
PHS2-pET151      CCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGA      5442
ATPHS2_SEQ_R1                                                 ACAAGCGCTCATGA

PHS2-pET151      GCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAA      5502
ATPHS2_SEQ_R1    GCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAA

PHS2-pET151      CCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCT      5562
ATPHS2_SEQ_R1    CCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCT

PHS2-pET151      CGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCC      5622
ATPHS2_SEQ_R1    CGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCC

PHS2-pET151      TCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCATCATCACCATCAC      5682
ATPHS2_SEQ_R1    TCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGCATCATCACCATCAC

PHS2-pET151      CATGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGGAAAACCTGTATTTT      5742
ATPHS2_SEQ_R1    CATGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGGAAAACCTGTATTTT


PHS2-pET151      CAGGGAATTGATCCCTTCACCATGGCAAACGCCAATGGAAAAGCTGCGACTAGTTTACCG      5802
ATPHS2_SEQ_R1    CAGGGAATTGATCCCTTCACCATGGCAAACGCCAATGGAAAAGCTGCGACTAGTTTACCG
ATPHS2_SEQ_R2    CAGGGAATTGATCCCTTCACCATGGCAAACGCCAATGGAAAAGCTGCGACTAGTTTACCG

PHS2-pET151      GAGAAAATCTCGGCTAAGGCGAATCCGGAGGCCGATGATGCTACGGAGATCGCTGGGAAT      5862
ATPHS2_SEQ_R1    GAGAAAATCTCGGCTAAGGCGAATCCGGAGGCCGATGATGCTACGGAGATCGCTGGGAAT
ATPHS2_SEQ_R2    GAGAAAATCTCGGCTAAGGCGAATCCGGAGGCCGATGATGCTACGGAGATCGCTGGGAAT

PHS2-pET151      ATCGTCTACCACGCCAAGTACAGTCCACATTTCTCTCCATTGAAGTTCGGGCCTGAGCAA      5982
ATPHS2_SEQ_R1    ATCGTCTACCACGCCAAGTACAGTCCACATTTCTCTCCATTGAAGTTCGGGCCTGAGCAA
ATPHS2_SEQ_R2    ATCGTCTACCACGCCAAGTACAGTCCACATTTCTCTCCATTGAAGTTCGGGCCTGAGCAA

PHS2-pET151      GCTCTCTACGCTACCGCAGAGAGTCTTCGCGATCGTCTCATTCAGCTGTGGAATGAGACT      6042
ATPHS2_SEQ_R1    GCTCTCTACGCTACCGCAGAGAGTCTTCGCGATCGTCTCATTCAGCTGTGGAATGAGACT
ATPHS2_SEQ_R2    GCTCTCTACGCTACCGCAGAGAGTCTTCGCGATCGTCTCATTCAGCTGTGGAATGAGACT

PHS2-pET151      TATGTTCATTTTAACAAAGTTGATCCAAAACAAACTTATTACTTGTCAATGGAGTATCTC      6102
ATPHS2_SEQ_R1    TATGTTCATTTTAACAAAGTTGATCCAAAACAAACTTATTACTTGTCAATGGAGTATCTC
ATPHS2_SEQ_R2    TATGTTCATTTTAACAAAGTTGATCCAAAACAAACTTATTACTTGTCAATGGAGTATCTC

PHS2-pET151      CAAGGTCGTGCTTTGACCAATGCCATTGGGAATTTGAACCTTCAAGGTCCATATGCTGAT      6162
ATPHS2_SEQ_R1    CAAGGTCGTGCTTTGACCAATGCCATTGGGAATTTGAACCTTCAAGGTCCATATGCTGAT
ATPHS2_SEQ_R2    CAAGGTCGTGCTTTGACCAATGCCATTGGGAATTTGAACCTTCAAGGTCCATATGCTGAT

PHS2-pET151      GCACTGCGTACGCTGGGTTATGAGCTTGAGGAGATAGCTGAGCAGGAGAAAGATGCAGCT      6222
ATPHS2_SEQ_R1    GCACTGCGTACGCTGGGTTATGAGCTTGAGGAGATAGCTGAGCAGGAGAAAGATGCAGCT
ATPHS2_SEQ_R2    GCACTGCGTACGCTGGGTTATGAGCTTGAGGAGATAGCTGAGCAGGAGAAAGATGCAGCT

PHS2-pET151      CTAGGAAATGGTGGGTTAGGGAGACTTGCCTCGTGTTTCTTGGATTCGATGGCCACCCTA      6282
ATPHS2_SEQ_R2    CTAGGAAATGGTGGGTTAGGGAGACTTGCCTCGTGTTTCTTGGATTCGATGGCCACCCTA

PHS2-pET151      AATCTGCCTGCTTGGGGTTATGGTTTGAGGTACAGACATGGGTTGTTTAAGCAAATAATC      6342
ATPHS2_SEQ_R2    AATCTGCCTGCTTGGGGTTATGGTTTGAGGTACAGACATGGGTTGTTTAAGCAAATAATC

PHS2-pET151      ACAAAGAAAGGTCAAGAAGAGATTCCAGAGGACTGGCTTGAGAAATTCAGCCCATGGGAA      6402
ATPHS2_SEQ_R2    ACAAAGAAAGGTCAAGAAGAGATTCCAGAGGACTGGCTTGAGAAATTCAGCCCATGGGAA

PHS2-pET151      ATTGTGAGGCACGACGTGGTATTCCCTGTCAGATTTTTCGGCAAGGTGCAAGTAAATCCG      6462
AtPHS2_SEQ_F1    ATTGTGAGGCACGACGTGGTATTCCCTGTCAGATTTTTCGGCAAGGTGCAAGTAAATCCG
ATPHS2_SEQ_R2    ATTGTGAGGCACGACGTGGTATTCCCTGTCAGATTTTTCGGCAAGGTGCAAGTAAATCCG
```

```
PHS2-pET151     GATGGATCAAGGAAATGGGTAGATGGTGATGTTGTACAAGCTCTTGCTTATGACGTGCCA  6522
AtPHS2_SEQ_F1   GATGGATCAAGGAAATGGGTAGATGGTGATGTTGTACAAGCTCTTGCTTATGACGTGCCA
ATPHS2_SEQ_R2   GATGGATCAAGGAAATGGGTAGATGGTGATGTTGTACAAGCTCTTGCTTATGACGTGCCA

PHS2-pET151     ATCCCGGGATATGGCACAAAGAACACAATCAGTCTCCGTCTCTGGGAAGCAAAAGCTAGA  6582
AtPHS2_SEQ_F1   ATCCCGGGATATGGCACAAAGAACACAATCAGTCTCCGTCTCTGGGAAGCAAAAGCTAGA
ATPHS2_SEQ_R2   ATCCCGGGATATGGCACAAAGAACACAATCAGTCTCCGTCTCTGGGAAGCAAAAGCTAGA

PHS2-pET151     GCTGAGGATCTTGATCTTTTTCAGTTCAACGAAGGAGAATATGAATTGGCTGCACAGCTT  6642
AtPHS2_SEQ_F1   GCTGAGGATCTTGATCTTTTTCAGTTCAACGAAGGAGAATATGAATTGGCTGCACAGCTT
ATPHS2_SEQ_R2   GCTGAGGATCTTGATCTTTTTCAGTTCAACGAAGGAGAATATGAATTGGCTGCACAGCTT

PHS2-pET151     CATTCTCGAGCTCAACAGATTTGCACTGTTTTATATCCAGGAGATGCTACCGAGAATGGG  6702
AtPHS2_SEQ_F1   CATTCTCGAGCTCAACAGATTTGCACTGTTTTATATCCAGGAGATGCTACCGAGAATGGG
ATPHS2_SEQ_R2   CATTCTCGAGCTCAACAGATTTGCACTGTTTTATATCCAGGAGATGCTACCGAGAATGGG

PHS2-pET151     AAGTTATTACGGTTAAAACAGCAGTTCTTTCTCTGCAGTGCTTCGCTTCAGGATATTATA  6762
AtPHS2_SEQ_F1   AAGTTATTACGGTTAAAACAGCAGTTCTTTCTCTGCAGTGCTTCGCTTCAGGATATTATA
ATPHS2_SEQ_R4   AAGTTATTACGGTTAAAACAGCAG-TCTTTCTCTGCAAGTGCTCGCTTCAGGATATTATA

PHS2-pET151     TCAAGATTTCACGAGAGGAGCACCACTGAAGGCAGCCGGAAATGGTCAGAGTTTCCAAGT  6822
AtPHS2_SEQ_F1   TCAAGATTTCACGAGAGGAGCACCACTGAAGGCAGCCGGAAATGGTCAGAGTTTCCAAGT
ATPHS2_SEQ_R4   TCAAGATTTCACGAGAGGAGCACCACTGAAGGCAGCCGGAAATGGTCAGAGTTTCCAAGT

PHS2-pET151     AAAGTTGCTGTTCAAATGAATGACACACACCCAACTCTTGCAATACCTGAGCTCATGCGA  6882
AtPHS2_SEQ_F1   AAAGTTGCTGTTCAAATGAATGACACACACCCAACTCTTGCAATACCTGAGCTCATGCGA
ATPHS2_SEQ_F2                           ACACCCGACTCATGCGATACCTGAGCTCATGCGA
ATPHS2_SEQ_R4   AAAGTTGCTGTTCAAATGAATGACACACACCCAACTCTTGCAATACCTGAGCTCATGCGA

PHS2-pET151     TTGCTAATGGATGACAATGGACTTGGATGGGATGAGGCTTGGGATGTGACATCAAAGACC  7942
AtPHS2_SEQ_F1   TTGCTAATGGATGACAATGGACTTGGATGGGATGAGGCTTGGGATGTGACATCAAAGACC
ATPHS2_SEQ_F2   TTGCTAATGGATGACAATGGACTAGGATGGGATGAGGCTTGGGATGTGACATCAAAGACC
ATPHS2_SEQ_R4   TTGCTAATGGATGACAATGGACTTGGATGGGATGAGGCTTGGGATGTGACATCAAAGACC

PHS2-pET151     GTTGCTTACACCAATCACACTGTCCTTCCTGAAGCGTTGGAGAAATGGTCACAATCTTTG  7002
AtPHS2_SEQ_F1   GTTGCTTACACCAATCACACTGTCCTTCCTGAAGCGTTGGAGAAATGGTCACAATCTTTG
ATPHS2_SEQ_F2   GTTGCTTACACCAATCACACTGTCCTTCCTGAAGCGTTGGAGAAATGGTCACAATCTTTG
ATPHS2_SEQ_R4   GTTGCTTACACCAATCACACTGTCCTTCCTGAAGCGTTGGAGAAATGGTCACAATCTTTG

PHS2-pET151     ATGTGGAAGCTTCTTCCTCGTCATATGGAAATAATAGAAGAGATTGACAAGAGGTTTGTT  7062
AtPHS2_SEQ_F1   ATGTGGAAGCTTCTTCCTCGTCATATGGAAATAATAGAAGAGATTGACAAGAGGTTTGTT
ATPHS2_SEQ_F2   ATGTGGAAGCTTCTTCCTCGTCATATGGAAATAATAGAAGAGATTGACAAGAGGTTTGTT
ATPHS2_SEQ_R3   ATGTGGAAGCTTCTTCCTCGTCATATGGAAATAATAGAAGAGATTGACAAGAGGTTTGTT
ATPHS2_SEQ_R4   ATGTGGAAGCTTCTTCCTCGTCATATGGAAATAATAGAAGAGATTGACAAGAGGTTTGTT

PHS2-pET151     CAAACCATTCGCGATACAAGAGTTGATCTGGAGGATAAGATTTCAAGTTTGAGCATCTTA  7122
AtPHS2_SEQ_F1   CAAACCATTCGCGATACAAGAGTTGATCTGGAGGATAAGATTTCAAGTTTGAGCATCTTA
ATPHS2_SEQ_F2   CAAACCATTCGCGATACAAGAGTTGATCTGGAGGATAAGATTTCAAGTTTGAGCATCTTA
ATPHS2_SEQ_R3   CAAACCATTCGCGATACAAGAGTTGATCTGGAGGATAAGATTTCAAGTTTGAGCATCTTA
ATPHS2_SEQ_R4   CAAACCATTCGCGATACAAGAGTTGATCTGGAGGATAAGATTTCAAGTTTGAGCATCTTA

PHS2-pET151     GATAACAATCCACAAAAGCCTGTGGTGAGAATGGCTAACTTATGTGTTGTATCCTCGCAT  7182
AtPHS2_SEQ_F1   GATAACAATCCACAAAAGCCTGTGGTGAGAATGGCTAACTTATGTGTTGTATCCTCGCAT
ATPHS2_SEQ_F2   GATAACAATCCACAAAAGCCTGTGGTGAGAATGGCTAACTTATGTGTTGTATCCTCGCAT
ATPHS2_SEQ_R3   GATAACAATCCACAAAAGCCTGTGGTGAGAATGGCTAACTTATGTGTTGTATCCTCGCAT
ATPHS2_SEQ_R4   GATAACAATCCACAAAAGCCTGTGGTGAGAATGGCTAACTTATGTGTTGTATCCTCGCAT

PHS2-pET151     ACGGTGAATGGCGTTGCTCAGTTACACAGTGATATCTTGAAGGCTGAGTTATTCGCAGAC  7242
AtPHS2_SEQ_F1   ACGGTGAATGGCGTTGCTCAGTTACACAGTGATATCTTGAAGGCTGAGTTATTCGCAGAC
ATPHS2_SEQ_F2   ACGGTGAATGGCGTTGCTCAGTTACACAGTGATATCTTGAAGGCTGAGTTATTCGCAGAC
ATPHS2_SEQ_R3   ACGGTGAATGGCGTTGCTCAGTTACACAGTGATATCTTGAAGGCTGAGTTATTCGCAGTC
ATPHS2_SEQ_R4   ACGGTGAATGGCGTTGCTCAGTTACACAGTGATATCTTGAAGGCTGAGTTATTCGCAGAC

PHS2-pET151     TATGTCTCTATATGGCCAAACAAGTTTCAAAACAAGACTAATGGCATCACACCTCGAAGG  7302
AtPHS2_SEQ_F1   TATGTCTCTATATGGCCAAACAAGTTTCAAAACAAGACTAATGGCATCACACCTCGAAGG
ATPHS2_SEQ_F2   TATGTCTCTATATGGCCAAACAAGTTTCAAAACAAGACTAATGGCATCACACCTCGAAGG
ATPHS2_SEQ_R3   TATGTCTCAATATGGGATTGCAGG
ATPHS2_SEQ_R4   TATGTCTCTATATGGCCAAACAAGTTTCAAAACAAGACTAATGGCATCACACCTCGAAGG

PHS2-pET151     TGGTTACGTTTCTGCAGCCCTGAGCTCAGTGATATAATCACAAAGTGGTTAAAGACTGAC  7362
AtPHS2_SEQ_F1   TGGTTACGTTTCTGCAGCCCTGAGCTCAGTGATATAATCACAAAGTG-CTAAAGACTGAC
ATPHS2_SEQ_F2   TGGTTACGTTTCTGCAGCCCTGAGCTCAGTGATATAATCACAAAGTGGTTAAAGACTGAC
ATPHS2_SEQ_F3                                                         AC
ATPHS2_SEQ_R4   TGGTTACGTTTCTGCAGCCCTGAGCTCAGTGATATAATCACAAAGTGGTTAAAGACTGAC

PHS2-pET151     AAATGGATTACCGATCTTGACCTACTTACCGGTCTTCGCCAGTTTGCGGACAATGAAGAA  7422
AtPHS2_SEQ_F1   AAATGGATTACCGATCTTGACCTACTACCGGTCCTTCGCCAGTTTGCG-ACAATGAAG-A
ATPHS2_SEQ_F2   AAATGGATTACCGATCTTGACCTACTTACCGGTCTTCGCCAGTTTGCGGACAATGAAGAA
ATPHS2_SEQ_F3   AAATGGATTACCGATCTTGACCTACTTACCGGTCTTCGCCAGTTTGCGGACAATGAAGAA
ATPHS2_SEQ_R4   AAATGGATTACCGATCTTGACCTACTTACCGGTCTTCGCCAGTTTGCGGACAATGAAGAA

PHS2-pET151     CTCCAATCTGAATGGGCTTCTGCAAAGACAGCCAATAAGAAACGTTTGGCTCAATATATA  7482
ATPHS2_SEQ_F2   CTCCAATCTGAATGGGCTTCTGCAAAGACAGCCAATAAGAAACGTTTGGCTCAATATATA
ATPHS2_SEQ_F3   CTCCAATCTGAATGGGCTTCTGCAAAGACAGCCAATAAGAAACGTTTGGCTCAATATATA
ATPHS2_SEQ_R4   CTCCAATCTGAATGGGCTTCTGCAAAGACAGCCAATAAGAAACGTTTGGCTCAATATATA

PHS2-pET151     GAGCGTGTGACTGGTGTGAGTATCGATCCAACAAGCTTATTTGACATACAAGTTAAGCGT  7542
ATPHS2_SEQ_F2   GAGCGTGTGACTGGTGTGAGTATCGATCCAACAAGCTTATTTGACATACAAGTTAAGCGT
ATPHS2_SEQ_F3   GAGCGTGTGACTGGTGTGAGTATCGATCCAACAAGCTTATTTGACATACAAGTTAAGCGT
ATPHS2_SEQ_R4   GAGCGTGTGACTGGTGTGAGTATCGATCCAACAAGCTTATTTGACATACAAGTTAAGCGT

PHS2-pET151     ATCCACGAATACAAGAGGCAGCTGATGAACATTCTTGGAGTAGTATACAGATTCAAGAAA  7602
ATPHS2_SEQ_F2   ATCCACGAATACAAGAGGCAGCTGATGAACATTCTTGGAGTAGTATACAGATTCAAGAAA
ATPHS2_SEQ_F3   ATCCACGAATACAAGAGGCAGCTGATGAACATTCTTGGAGTAGTATACAGATTCAAGAAA
ATPHS2_SEQ_F3   ATCCACGAATACAAGAGGCAGCTGATGAACATTCTTGGAGTAGTATACAGATTCAAGAAA
ATPHS2_SEQ_R4   ATCCACGAATACAAGAGGCAGCTGATGAACATTCTTGGAGTAGTATACAGATTCAAGAAA
```

```
PHS2-pET151     CTAAAGGAGATGAAGCCTGAGGAGAGGAAGAAAACAGTTCCTCGTACTGTCATGATTGGG    7662
ATPHS2_SEQ_F2   CTAAAGGAGATGAAGCCTGAGGAGAGGAAGAAAACAGTTCCTCGTACTGTCATGATTGGG
ATPHS2_SEQ_F3   CTAAAGGAGATGAAGCCTGAGGAGAGGAAGAAAACAGTTCCTCGTACTGTCATGATTGGG
ATPHS2_SEQ_R4   CTAAAGGAGATGAAGCCTGAGGAGAGGAAGAAAACAGTTCCTCGTACTGTCATGATTGGG

PHS2-pET151     GGTAAAGCATTTGCCACCTATACAAATGCAAAACGGATAGTGAAGCTGGTGAATGATGTT    7722
ATPHS2_SEQ_F2   GGTAAAGCATTTGCCACCTATACAAATGCAAAACGGATAGTGAAGCTGGTGAATGATGTT
ATPHS2_SEQ_F3   GGTAAAGCATTTGCCACCTATACAAATGCAAAACGGATAGTGAAGCTGGTGAATGATGTT
ATPHS2_SEQ_R4   GGTAAAGCATTTGCCACCTATACAAATGCAAAACGGATAGTGAAGCTGGTGAATGATGTT

PHS2-pET151     GGTGATGTTGTTAACAGCGATCCAGAGGTCAACGAATACCTAAAGGTGGTATTTGTTCCA    7782
ATPHS2_SEQ_F2   GGTGATGTTGTTAACAGCGATCCAGAGTCAACGAATACCTTAAAGGTGGTATTTGTTCCA
ATPHS2_SEQ_F3   GGTGATGTTGTTAACAGCGATCCAGAGGTCAACGAATACCTAAAGGTGGTATTTGTTCCA

PHS2-pET151     AACTACAATGTCACTGTAGCGGAGATGCTAATACCCGGAAGTGAGCTATCTCAACACATC    7842
ATPHS2_SEQ_F2   AACTACAATGTCACTGTAGCGGAGATGCTAATACCCGGAAGTGAGCTATCTCAACACATC
ATPHS2_SEQ_F3   AACTACAATGTCACTGAAGCGGAGATGCTAATACCCGGAAGTGAGCTATCTCAACACATC

PHS2-pET151     AGCACAGCAGGCATGGAGGCAAGTGGTACCAGCAATATGAAATTCGCTCTCAACGGTTGT    7902
ATPHS2_SEQ_F2   AGCACAGCAG-CATGAAGGCAAGTGGTACCAGCATTATGAAATTCGCTCTCAACGGTTGT
ATPHS2_SEQ_F3   AACACAACATGCATGGAGGCAAGTGGTACCAGCAATATTAAATTCGCTCTCAACGGTTGT

PHS2-pET151     CTTATTATAGGAACCCTTGATGGGGCTAATGTTGAGATAAGAGAGGAGGTTGGCGAAGAA    7962
ATPHS2_SEQ_F2   CTTATTATAG-AACCCTTGATGGGGCTATGG-TGAGATAAGAGAGGAGTTTG-CTGAGA
ATPHS2_SEQ_F3   CTTATTATAGGAACCCTTGATGGGGCTAATGTTGAGATAAGAGAAGAGGTTGGCGACAAA
ATPHS2_SEQ_F4   CTTATTATAGGAACCCTTGATGGGGCTAATGTTGAGATAAGAGAGGAGGTTGGCGAAGAA

PHS2-pET151     AATTTCTTTCTTTTTGGTGCAACGGCCGATCAGGTCCCTCGACTGCGTAAAGAACGAGAA    8022
ATPHS2_SEQ_F3   AATTTCTTTCTTTTTGGCGCTACGGCCGATAAAGTCTCTCGACTGCCTAAAGAACGAGAA
ATPHS2_SEQ_F4   AATTTCTTTCTTTTTGGTGCAACGGCCGATCAGGTCCCTCGACTGCGTAAAGAACGAGAA

PHS2-pET151     GACGGACTGTTCAAACCCGATCCTCGGTTCGAAGAGGCAAAGCAGTTTGTCAAAAGTGGA    8082
ATPHS2_SEQ_F3   GACGGACTGTT
ATPHS2_SEQ_F4   GACGGACTGTTCAAACCCGATCCTCGGTTCGAAGAGGCAAAGCAGTTTGTCAAAAGTGGA

PHS2-pET151     GGACGTGGTGATTACTTCCTGGTTGGGTATGACTTCCCCAGCTACATGGACGCTCAGGCC    8142
ATPHS2_SEQ_F4   GGACGTGGTGATTACTTCCTGGTTGGGTATGACTTCCCCAGCTACATGGACGCTCAGGCC

PHS2-pET151     AAAGTTGACGAAGCTTATAAGGACCGGAAGGGGTGGCTGAAAATGTCGATATTGAGCACA    8202
ATPHS2_SEQ_F4   AAAGTTGACGAAGCTTATAAGGACCGGAAGGGGTGGCTGAAAATGTCGATATTGAGCACA

PHS2-pET151     GCCGGGTCAGGAAAGTTCAGCAGTGACCGTACAATAGCTCAGTATGCCAAAGAGATTTGG    8262
ATPHS2_SEQ_F4   GCCGGGTCAGGAAAGTTCAGCAGTGACCGTACAATAGCTCAGTATGCCAAAGAGATTTGG

PHS2-pET151     AACATTGAGGCTTGTCCTGTTCCCTAAAAGGGCGAGCTCAGATCCGGCTGCTAACAAAGC    35
ATPHS2_SEQ_F4   AACATTGAGGCTTGTCCTGTTCCCTAAAAGGGCGAGCTCAGATCCGGCTGCTAACAAAGC

PHS2-pET151     CCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGG    95
ATPHS2_SEQ_F4   CCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGG

PHS2-pET151     GGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCC    155
ATPHS2_SEQ_F4   GGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCC

PHS2-pET151     CGCAAGAGGCCCGGCAGTACCGGCATAACCAAGCCTATGCCTACAGCATCCAGGGTGACG    215
ATPHS2_SEQ_F4   CGCAAGAGGCCCGGCAGTACCGGCATAACCAAGCCTATGCCTACAGCATCCAGGGTGACG
```

Appendix 2 – Synthetic gene for CDP

App2.1 Nucleotide sequence of the *E. coli* codon optimised CDP

The gene for CDP from *Clostridium thermocellum* was codon optimised for expression in *E. coli* and synthesised and cloned into pET15b to form pET15b-CDP (BamHI sites underlined).

GGATCCGATGATTACCAAAGTGACGGCACGCAACAATAAAATTACCCCGGTGGAACTGCTGAACCAAAAATTCGGCAATA
AAATCAACCTGGGCAACTTTGCAGACGCTGTGTTCACCGATGCGGCCTTTAAAAATGTTGCGGGCATTGCCAACCTGCCG
ATGAAAGCGCCGGTGATGCAGGTTCTGATGGAAAATTGCATCGTTTCGAAATATCTGAAACAATTTGTCCCGGATCGTAG
CGTCTGTTTCGTGGAAGAAGGCCAGAAATTTTACATTGTGCTGGAAGACGGTCAAAAAATCGAAGTGCCGGAAGATGTTA
ACAAAGCACTGAAAGCTACCGTTTCAGATGTCAAACATTGGGCAGGCTATCTGACGGAAGACGGTGAACACGTTATTGAT
CTGCTGAAACCGGCTCCGGGTCCGCATTTCTACGTCAATCTGCTGATCGGCAACCGTCTGGGTTTTAAACGCACCCTGCA
GACCACGCCGAAATCTGTGGTTGATCGTTTCGGCCGCGGTTCTTTTCGTAGTCACGCAGCTACCCAGGTTCTGGCAACGC
GTTTTGACATGCGCCAAGAAGAAAATGGCTTCCCGGCTAACCGCCAGTTTTATCTGTACGAAGATGGTAAACAAATCTTT
TATAGCGCACTGATTGATGACAATATCGTGGAAGCTACCTGCAAACACTCTTGTAACCGTACCGTTATCAAATACAAAAC
GGCATGCAATCTGGAAATTACCCGCACGATCTTCCTGGTGCCGCACAAAAAAGGCTTTCCGCTGGCGACCGAACTGCAGC
GTATTGAAATCAAAAATGCGTCCGATAAAGCCCGCAACCTGTCAATTACCTATACGGGCATGTTCGGCACGGGTGCGGTT
CATGCCATTTTTGAAGACGTGACCTATACGAATGTTATCATGCAGTCTGCGGCCCTGTACAACGATAAAGGCGAATTTAT
TGGTATCACCCCGGACTATTACCCGGAAGAATTTAAACAGGATACCCGTTTTGTCACGATGATTGTGCGCAACGGCGATG
AAAAATCCTTTCCGCAATCATTCTGCACCGATTATAATGACTTTGTGGGCACCGGTACGCTGGAACACCCGGCGGGCGGT
TGTAATCTGAACAATAAACTGAACCGTAAAGGTCCGGGCTTTTTCGCACTGGGTGCTCCGTTCACCGTCGAACCGGGTAA
AACGGTGATTATCGATACCTTTACGGGCCTGAGCTCTAGTAAAGACAACGAAAATTATAGCGATGCCGTGATGCTGCGTG
AACTGGATAATCTGCTGCGCTACTTCGAAAAATCGGAAAGCGTTGAAGAAACCCTGAACGAAATCATCAACTTCCATGAA
AACTACGGTAAATACTTCCAGTTCAACACGGGCAATAAACTGTTTGATAGTGGTTTCAACCGTAATCTGGCATTCCAGGT
GCTGTATCAAACCTTCATGTCTCGCAGTTTTGGCCAGACGCAAAAAGGTTACCGTGAAATTGGCTTTCGCGAAATCCAGG
ACCTGTTCGCCTCCATGTACTACTTCATCAACATCGGCTATCAAGATTTCGTTAAAGAACTGCTGTTTGAATGGACCGCA
AATGTCTACAAAATGGGTTACGCTAACCATAACTTCTATTGGGTGGGCAAACAGCCGGGTCTGTACTCCGATGACTCACT
GTGGCTGCTGCAAGCGTATTACCGCTATATTATCTACACCAAAGACACGAGCGTTCTGAACGAAGAAGTCCCGGTGGCGG
ATGGCAACAATGAAAAACGTGCCGTCCGCGAAACCCTGAAAGCAATTATCCAGTATTCGGCTTGCATTAGCGTGGGTGAC
CACGGTCTGCCGCTGCTGGATCTGGCCGACTGGAATGATTGTCTGAAAATTGCGTCCAACTCAATCGATGGTGCCACCAA
AGAAAAACTGTATTACGAACAGCTGAAGAAAACCAACGGCAAATATGGTGACCGTTTCATGAGCGATTACTCGGAAAGCG
TGATGAATGCGTTTCTGCTGAAACTGGCCATTGATCATCTGGCGGAAATCGCCACCCTGGATAATGACACGCAGCTGGCG
CAGCAAATGTCGGAACTGAGCAAAGAAGTTACCGATCGCATTCAAAAACACGCATGGAAAGAAAACTTTTTCGCTCGTGT
CCTGATCAACCGCTATAAAGACGGCTCTTATACCTACCTGGGCGCGAAAGGTGATAAACTGAGTGCCGACCCGAATATTG
ATGGTGTGTATTTTCTGAACAGTTTCGCGTGGTCTGTTCTGAGTGACGTCGCAACCGATGAACAGATTGCTATCATGGTT
GATGTCATCAAAAAACATCTGCTGACCCCGTATGGTCTGCGTCTGGTTACGCCGGCCGACCTGAATAAAATCGCAAACGA
TACCGCTACGGGTCATTACTTTTTCGGCGATCGCGAAAACGGTGCAGTCTTTAAACACGCGTCCATGATGGCCGTGGCAG
CTCTGATTAAAGCGGCCAAAAAAGTGAAAGACAATGAACTGGCGAAAGAAATGGCACGTATTGCTTATTTCATGATCGAT
CTGGTCCTGCCGTACAAAAACCTGGAAAATCCGTTTCAGGTGGCCGGCAACCCGCGCATTTGCACCCAATATATCAATAC
CGATACGGGCGAAAACATTGGTCCGCTGCTGTCCGGCACCGCAACGTGGCTGAACCTGAATCTGATTTCACTGGCCGGCA
TCGAATATACCCGTGATGGTATTTCGTTCAACCCGATCCTGCGCGAAGAAGAAACCCAGCTGAATTTTACGCTGAAAGCG
CCGAAATGTTCTTACAAATTCAGTATCACCAAACCGGTGGGCTTTGCCCGCATGGAATCCTCAGAATACGAACTGTTTGT
TGACGGTCAGAAAATCGATAATACCGTTATTCCGATGTACACCGACGAAAAAGAACATATTGTCACCCTGAAATTCAAAT
GAGGATCC

App.2.2. Amino acid sequence encoded in the CDP open reading frame of pET15b-CDP

(Purification tag highlighted in blue).

MGSSHHHHHHSSGLVPRGSHMLEDPMITKVTARNNKITPVELLNQKFGNKINLGNFADAVFTDAAFKNVAGIANL
PMKAPVMQVLMENCIVSKYLKQFVPDRSVCFVEEGQKFYIVLEDGQKIEVPEDVNKALKATVSDVKHWAGYLTED
GEHVIDLLKPAPGPHFYVNLLIGNRLGFKRTLQTTPKSVVDRFGRGSFRSHAATQVLATRFDMRQEENGFPANRQ
FYLYEDGKQIFYSALIDDNIVEATCKHSCNRTVIKYKTACNLEITRTIFLVPHKKGFPLATELQRIEIKNASDKA
RNLSITYTGMFGTGAVHAIFEDVTYTNVIMQSAALYNDKGEFIGITPDYYPEEFKQDTRFVTMIVRNGDEKSFPQ
SFCTDYNDFVGTGTLEHPAGGCNLNNKLNRKGPGFFALGAPFTVEPGKTVIIDTFTGLSSSKDNENYSDAVMLRE
LDNLLRYFEKSESVEETLNEIINFHENYGKYFQFNTGNKLFDSGFNRNLAFQVLYQTFMSRSFGQTQKGYREIGF
REIQDLFASMYYFINIGYQDFVKELLFEWTANVYKMGYANHNFYWVGKQPGLYSDDSLWLLQAYYRYIIYTKDTS
VLNEEVPVADGNNEKRAVRETLKAIIQYSACISVGDHGLPLLDLADWNDCLKIASNSIDGATKEKLYYEQLKKTN
GKYGDRFMSDYSESVMNAFLLKLAIDHLAEIATLDNDTQLAQQMSELSKEVTDRIQKHAWKENFFARVLINRYKD
GSYTYLGAKGDKLSADPNIDGVYFLNSFAWSVLSDVATDEQIAIMVDVIKKHLLTPYGLRLVTPADLNKIANDTA
TGHYFFGDRENGAVFKHASMMAVAALIKAAKKVKDNELAKEMARIAYFMIDLVLPYKNLENPFQVAGNPRICTQY
INTDTGENIGPLLSGTATWLNLNLISLAGIEYTRDGISFNPILREEETQLNFTLKAPKCSYKFSITKPVGFARME
SSEYELFVDGQKIDNTVIPMYTDEKEHIVTLKFK*

App 2.3. Alignment of CDP and CBP from *Clostridium thermocellum*

Alignment was performed using ClustalW (EBI). * = identical, := identity score of 8, . = identity score of 7.

```
CDP     MITKVTARNNKITPVELLNQKFGNKINLGNFADAVFTDAAFKNVAGIANLPMKAPVMQVL 60
CBP     ------------------------------------------------------------

CDP     MENCIVSKYLKQFVPDRSVCFVEEGQKFYIVLEDGQKIEVPEDVNKALKATVSDVKH-WA 119
CBP     ----------------------------------MKFGFFDDANKEYVITVPRTPYPWI 25
                                            *: . :*.**    **. . : *
CDP     GYLTEDGEHVIDLLKPAPGPHFYVNLLIGNRLGFKRTLQTTPKSVVDRFGRGSFRSHAAT 179
CBP     NYLGTEN--------------FFS--LISNTAG----------------GYCFYRDARLR 53
        .**  :.               *:   **.*  *                *   :*.
CDP     QVLATRFDMRQEENGFPANRQFYLYEDGKQIFYS---ALIDDNIVEATCKHSCNRTVIKY 236
CBP     RITRYRYNNVPIDMGG---RYFYIYDNGD--FWSPGWSPVKRELESYECRHGLGYTKIAG 108
        ::   *::    : *    * **:*::*.  *:*   : :. :: .  *:*. . * *
CDP     KTACNLEITRTIFLVPHKKGFPLATELQRIEIKNASDKARNLSITYTGMFGTGAVHAIFE 296
CBP     KRNG-IKAEVTFFVPLNYNG-----EVQKLILKNEGQDKKKITLFSFIEFCLWNAYDDMT 162
        *    ::   *:*:  : :*     *:*:: :** .:. :::::     *    .:  :

CDP     DVTYTNVIMQSAALYNDKGEFIGITPDYYPEEFKQDTRFVTMIVRNGDEKSFPQSFCTDY 356
CBP     NFQRN----FSTGEVEIEGSVIYHKTEYR--ERRNHYAFYSVNAKISGFDSDRDSFIGLY 216
        :.  .     *:.  : :*..*  ..:*   * ::.  * :: .: .. .*  :**   *
CDP     NDFVGTGTLEHPAGGCNLNNKLNRKGPGFFALGAPFTVEPGKTVIIDTFTGLSSSKDNEN 416
CBP     NGFDAPQAVVNGKS----NNSVADGWAPIASHSIEIELNPGEQKEYVFIIGYVENKDEEK 272
        *.* .. :: :  .    **.:    . : : .  : ::**:      : *  ..**:*:
CDP     YS--DAVMLRELDNLLRYFEKSESVEETLNEIINFHENYGKYFQFNTGNKLFDSGFNRNL 474
CBP     WESKGVINKKKAYEMIEQFNTVEKVDKAFEELKSYWNALLSKYFLESHDEKLNRMVNIWN 332
        :.  ..:  ::  :::. *:. *.*:::::*: .: :   . : ::: :: ::  .*
CDP     AFQVLYQTFMSRSFGQTQKGY-REIGFREIQDLFASMYYFINIGYQDFVKELLFEWTANV 533
CBP     QYQCMVTFNMSRSASYFESGIGRGMGFRDSNQDLLGFVHQI----PERARERLLDLAATQ 388
         :* :    **** .  :.*  * :***: :: : .: : *     : .:* *:: :*.
CDP     YKMGYANHNFYWVGKQP-----GLYSDDSLWLLQAYYRYIIYTKDTSVLNEEVPVADGNN 588
CBP     LEDGSAYHQYQPLTKKGNNEIGSNFNDDPLWLILATAAYIKETGDYSILKEQVPFNNDPS 448
         : * * *::  : *:      . :.**.***: *   **  * * *:*:*:**. :. .
CDP     EKRAVRETLKAIIQYSACISVGDHGLPLLDLADWNDCLKIASNSIDGATKEKLYYEQLKK 648
CBP     KADTMFEHLTRSFYHVVNN-LGPHGLPLIGRADWNDCLNLNCFSTVPD-------ESFQT 500
        :  :: * *.  : : .   :* *****:. *******:: . *           *.::.
CDP     TNGKYGDRFMSDYSESVMNAFLLKLAIDHLAEIATLDNDTQLAQQMSELSKEVTDRIQKH 708
CBP     TTSKDGK-----VAESVMIAGMFVFIGKDYVKLCEYMGLEEEARKAQQHIDAMKEAILKY 555
        *..* *.      :**** * :: :  .. .::.   .  : *:: .:  . :.: * *:
CDP     AWKENFFARVLINRYKDGSYTYLGAKGDKLSADPNIDGVYFLNSFAWSVLSDVAT-DEQI 767
CBP     GYDGEWFLRAYDD------------FGRKVGSKENEEGKIFIESQGFCVMAEIGLEDGKA 603
        .:. ::* *.  :              * *:.:. * :*  *::* .:.*::::.  * :
CDP     AIMVDVIKKHLLTPYGLRLVTPADLNKIANDTATGHYFFGDRENGAVFKHASMMAVAALI 827
CBP     LKALDSVKKYLDTPYGLVLQNPAFTRYYIEYGEISTYPPGYKENAGIFCHNNAWIICAET 663
          :* :**:* ***** * .**  .   :    . *  * :**..:* * .   :.*
CDP     KAAKKVKDNELAKEMARIAYFMIDLVLPYKNLENPFQVAGNPRICTQYINTDTGENIGPL 887
CBP     VVGRGDMAFDYYRKIAPAYIEDVSDIHKLEPYVYAQMVAGKD-------AKRHGEAKNSW 716
        ..:     :  :::*      :. :   :    .  ***:         .  **  ..
CDP     LSGTATWLNLNLIS-LAGIEYTRDGISFNPILREEETQLNFTLKAPKCSYKFSITKPVGF 946
CBP     LTGTAAWNFVAISQWILGVKPDYDGLKIDPCIPKAWDGYKVTRYFRGSTYEITVKNPN-- 774
        *:***:*  : : . : *::   **:.::* : :     :.*     .:*:::::.:*
CDP     ARMESSEYELFVDGQKIDNTVIPMYTDEKEHIVTLKFK 984
CBP     -HVSKGVAKITVDGNEISGNILPVFNDGKTHKVEVIMG 811
         ::...  :: ***::*...::*::.* * * * : :
```

Appendix 3 – *Euglena* proteins co-purified with laminarin phosphorylase activity

The fractions from the laminarin purification protocols were subject to proteomic analysis. Proteins were identified by peptide mass-fingerprinting against the translated transcriptome and searched against the NCBI protein database using BLASTP. The closest homologue found for each transcript is listed. Highlighted in yellow are GH81 family members.

Proteins from first purification protocol		
Transcript number	Closest homologue	Score
App.3.1 *Euglena* proteins identified in Eg2.1 gel slice		
dm.6752	pyruvate carboxylase, gene 1 [Xenopus laevis]	3312
dm.14567	conserved hypothetical protein [Vibrio harveyi]	3188
dm.12520	pyridoxal-dependent decarboxylase domain-containing protein 1 [Gallus gallus]	3066
dm.14	malate synthase-isocitrate lyase [Euglena gracilis]	2888
dm.11270	putative alanyl-tRNA synthetase [Leishmania infantum]	2694
dm.1790	predicted protein [Physcomitrella patens]	2078
dm.12625	cullin-associated NEDD8-dissociated protein 1-like [Glycine max]	2062
dm.44758	ankyrin repeat protein [Paracoccidioides brasiliensis]	1947
dm.15323	coatomer subunit beta'-1 [Arabidopsis thaliana]	1701
dm.3707	cell division cycle protein 48-like protein, expressed [Micromonas sp]	1655
lm.62373	insulin-degrading enzyme [Danio rerio]	1604
lm.59664	exosome complex exonuclease RRP44-like [Nasonia vitripennis]	1481
dm.10697	inversin protein alternative isoform [Trichomonas vaginalis]	1431
dm.16810	WD40 repeat-containing protein [Polysphondylium pallidum]	1385
dm.501	pyruvate, phosphate dikinase [Caldithrix abyssi]	1372
dm.9721	coatomer subunit beta-1 [Arabidopsis lyrata]	1306
dm.33057	tryptophan synthase subunit beta [Planococcus donghaensis] See Appendix 10	1289
lm.40065	no hits	1288
lm.41899	ubiquitin-activating enzyme e1, putative [Leishmania donovani]	1090
lm.55348	alpha-2-macroglobulin domain-containing [Ilumatobacter coccineum	1066
dm.27248	hypothetical protein Thini_1924 [Thiothrix nivea]	1032
dm.35794	no hits	1027
dm.11750	coatomer subunit gamma-2-like [Vitis vinifera]	1010
dm.8434	hypothetical protein [Rhodanobacter spathiphylli]	922
dm.22743	hypothetical protein, conserved [Trypanosoma brucei]	902
dm.17299	ubiquitin activating enzyme [Naegleria gruberi]	899
dm.468	aconitate hydratase 2 [Methylomicrobium album]	820
dm.3530	hypothetical protein [Monosiga brevicollis]	788
dm.12763	hypothetical protein [Monosiga brevicollis]	387
lm.74157	regulator of chromosome condensation (RCC1)like protein putative [Albugo laibachii]	730
dm.27489	UV-damaged DNA-binding protein 1A [Arabidopsis lyrata]	678

dm.49218	cell division cycle and apoptosis regulator protein 1 [Clonorchis sinensis]	677
dm.54710	no hits	667
lm.58565	no hits	539
dm.7256	heat shock protein, putative [Trypanosoma cruzi]	645
dm.186	pyruvate: NADP+ oxidoreductase [Euglena gracilis]	579
lm.73638	no hits	573
dm.71637	no hits	304
dm.65801	ubiquitinyl C19 peptidase [Naegleria gruberi]	536
dm.46	heat shock protein 90 [Euglena gracilis]	471
dm.39195	hypothetical protein [Trypanosoma cruzi]	465
dm.10321	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Danio rerio]	455
dm.7071	similar to conserved hypothetical protein [Tribolium castaneum]	450
dm.7625	cell division cycle and apoptosis regulator protein 1 [Clonorchis sinensis]	449
dm.28974	Phosphoenolpyruvate carboxylase 1 [Chlamydomonas reinhardtii]	424
dm.13136	intraflagellar transport protein 122 homolog [Equus caballus]	417
dm.34328	no hits	414
dm.30978	ubiquitin-activating enzyme E1 [Salpingoeca sp.]	379
lm.57406	ring finger protein 123 [Mustela putorius furo]	376
lm.3583	kinesin [Trypanosoma cruzi]	375
dm.78696	sec24-like transport protein [Arabidopsis thaliana]	375
dm.11453	carnosine synthase 1-like [Cavia porcellus]	362
lm.84438	beta-galactosidase [Psychromonas marina]	343
lm.8430	hypothetical protein [Dysgonomonas mossii]	343
lm.9447	hypothetical protein [Dysgonomonas mossii]	60
dm.322	heat shock protein 60 [Euglena gracilis]	316
dm.38431	no hits	310
dm.31243	flagellar associated protein [Chlamydomonas reinhardtii]	309
dm.7327	carbamoyl-phosphate synthase [ammonia], putative [Phytophora infestans]	268
dm.3395	heat shock protein [Trypanosoma cruzi]	264
dm.41656	ubiquitin carboxyl-terminal hydrolase 4 isoform b [Rattus norvegicus]	244
lm.61306	kinesin K39, putative, partial [Leishmania infantum]	239
lm.76211	ankyrin repeat protein [Trichomonas vaginalis]	232
lm.15628	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	211
dm.13155	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	141
lm.15629	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	46
dm.138	elongation factor 1 alpha [Euglena gracilis]	206
dm.22458	serine/threonine protein kinase [Tetrahymena thermophile]	203
dm.20564	L-aminoadipate-semialdehyde dehydrogenase [Capsaspora owczarzaki]	196
dm.24693	no hits	190
lm.47385	calpain-15 [Gallus gallus]	170
lm.43993	paraflagellar rod protein 2C [Leishmania mexicana]	170
lm.58785	conserved hypothetical protein [Leishmania mexicana]	164
dm.17280	putative acetyl-CoA carboxylase [Leishmania infantum]	156
lm.27495	hypothetical protein [Phytophthora sojae]	155
lm.49833	von Willebrand domain-containing protein [Capsaspora owczarzaki]	155

dm.21534	no hits	155
dm.46328	basal body protein [Chlamydomonas reinhardtii]	155
dm.78615	putative ATP-dependent RNA helicase [Arabidopsis thaliana]	155
lm.84453	hypothetical protein, variant [Salpingoeca sp. ATCC 5081]	155
dm.2759	elongation factor EF-3 [Coccomyxa subellipsoidea]	153
dm.32024	5-oxoprolinase [Azospirillum brasilense]	152
dm.31760	putative adenylosuccinate synthetase [Leishmania braziliensis]	147
dm.7644	predicted protein [Physcomitrella patens]	142
lm.403	unnamed protein product [Tetraodon nigroviridis]	137
dm.57883	putative protein serine/threonine kinase [Polysphondylium pallidum]	136
dm.38076	splicing factor 3B subunit 3-like [Oreochromis niloticus]	135
dm.37350	putative Tyrosine kinase-like (TKL) protein [Neospora caninum]	130
dm.35236	putative Tyrosine kinase-like (TKL) protein [Neospora caninum]	37
dm.6466	eukaryotic initiation factor [Coccomyxa subellipsoidea]	128
dm.51164	predicted protein [Physcomitrella patens]	126
dm.51828	ubiquitin fusion degradation protein UfdB, putative [Penicilllium sp.]	122
dm.20604	no hits	118
dm.6861	GH65 family protein [Thermoanaerobacter wiegelii]	115
lm.26838	adenylosuccinate synthetase [Trypanosoma cruzi]	109
lm.80211	dynactin [Coprinopsis cinerea]	106
dm.30499	gamma-glutamylcysteine synthetase [Laeonereis acuta]	105
lm.42878	tetratricopeptide repeat protein 21B-like [Xenopus tropicalis]	103
dm.25799	hypothetical protein [Trypanosoma cruzi]	96
dm.13040	acetyl-CoA acetyltransferase [Reinekea sp. MED297]	91
lm.8338	hypothetical protein [Salpingoeca sp. ATCC 5081]	89
lm.86223	ankyrin 2,3/unc44 [Aedes aegypti]	89
dm.11294	conserved hypothetical protein [Leishmania Mexicana]	88
lm.77499	hypothetical protein [Trypanosoma cruzi]	86
dm.15786	acetyl-CoA carboxylase [Nannochloropsis gaditana]	85
dm.41655	WD40 repeat-containing protein [Dictyostelium discoideum]	67
dm.278	DNA-directed RNA polymerase II 135 kDa polypeptide, putative [Oryza sativa]	64
lm.21412	clathrin interactor 1 [Mustela putorius]	62
dm.446	RNA-binding region RNP-1 domain-containing protein [Polysphondylium pallidum]	60
lm.55849	hypothetical protein [Branchiostoma floridae]	59
lm.24990	flagellar motor protein [Oceanicola granulosus]	58
dm.21288	predicted protein [Phaeodactylum tricornutum]	58
lm.57893	ankyrin 2,3/unc44 [Aedes aegypti]	57
dm.30785	predicted protein [Physcomitrella patens subsp. patens]	55
dm.61506	Molecular co-chaperone STI1 (ISS) [Ostreococcus tauri]	53
dm.6921	hypothetical protein [Daphnia pulex]	53
lm.44496	ankyrin repeat domain-containing protein 50 -like [Amphimedon queenslandica]	50
lm.41918	maternal embryonic leucine zipper kinase, putative [Ixodes scapularis]	49
lm.17228	maternal embryonic leucine zipper kinase, putative [Ixodes scapularis]	29
dm.51076	ccar1, putative [Ixodes scapularis]	46
lm.67640	putative protein serine/threonine kinase	44
lm.72440	polyribonucleotide nucleotidyltransferase [Nannochloropsis	43

	gaditana]	
dm.27702	no hits	41
lm.51516	proline iminopeptidase [Aeromonas caviae]	40
lm.66536	epidermal growth factor receptor substrate 15-like 1-like [Acyrthosiphon pisum]	39
dm.24752	WD repeat-containing protein 19-like [Strongyloides stercoralis]	36
lm.65316	peptidase C19 family ubiquitinyl hydrolase [Naegleria gruberi]	34
dm.17922	hypothetical protein [Trypanosoma cruzi]	34
lm.18544	kinesin family member 12 [Salpingoeca sp.]	34
lm.50851	HAD superfamily hydrolase [Chlorobium tepidum TLS]	34
dm.435	ATP-dependent RNA helicase dhh1 [Zea mays]	34
dm.9754	protein phosphatase 1, regulatory subunit 8 [Xenopus laevis]	34
lm.25046	hypothetical protein [gamma proteobacterium	34
lm.73994	viral A-type inclusion protein [Trichomonas vaginalis]	34
lm.85920	R-SNARE protein, tomsyn-like family [Chlamydomonas reinhardtii]	34
dm.57166	no hits	34
dm.11945	adaptor-related protein complex 2 [Dictyostelium fasciculatum]	34
lm.84772	ubiquitin carboxyl-terminal hydrolase 12 [Vitis vinifera]	34
dm.30548	hypothetical protein [Gibberella zeae]	33
dm.11777	ubiquinone/menaquinone biosynthesis methyltransferase UbiE/COQ5 [endosymbiont of Tevnia jerichonana]	32
dm.3232	vitamin B6 biosynthesis protein [Wallemia sebi]	32
lm.66673	DEP domain containing 5 isoform 2 [Capsaspora owczarza]	30
dm.16358	kinesin-like protein KLP6-like [Strongylocentrotus purpuratus]	28
lm.77341	putative kinesin K39, partial [Leishmania braziliensis]	28
lm.11279	conserved Plasmodium protein, unknown function [Plasmodium falciparum]	28
dm.168	RNA-binding region RNP-1 domain-containing protein	28
lm.32501	hypothetical protein [Trypanosoma cruzi]	28
dm.1638	heat shock protein 60 [Euglena gracilis]	28
dm.32489	predicted protein [Physcomitrella patens subsp. patens	28
lm.83814	PCMT-domain-containing protein [Coccomyxa subellipsoidea]	28
lm.87063	no hits	28
lm.71419	ATP-dependent DEAD/H RNA helicase, putative [Trypanosoma cruzi]	28
dm.14940	putative kinesin [Leishmania braziliensis]	28
dm.4783	acyl-CoA dehydrogenase FadE [Moraxella catarrhalis]	28
lm.69342	Rho-associated kinase, putative [Ixodes scapularis]	28
dm.25565	no hits	28
dm.70524	Regulator of nonsense transcripts-like protein [Medicago trunculata]	28
lm.53	beta-tubulin [Euglena gracilis]	27
dm.4000	UDP-glucose dehydrogenase [Capsaspora owczarzaki]	26
lm.56904	RNA-binding region RNP-1 domain-containing protein [Polysphondylium pallidum]	25
dm.3761	DNA polymerase I alpha catalytic subunit, putative [Trypanosoma cruzi]	25
dm.17952	thymidylate kinase [Modestobacter marinus]	23
lm.68357	conserved hypothetical protein [Trypanosoma congolense]	22
dm.47148	presequence protease 2, chloroplastic/mitochondrial-like [Glycine max]	22
dm.11926	protein transport protein Sec24C, putative [Trypanosoma cruzi]	22

lm.64702	no hits	22
dm.58933	no hits	22
dm.12424	predicted protein [Naegleria gruberi]	22
lm.92865	no hits	22
lm.77127	SAPS-domain-containing protein [Coccomyxa subellipsoidea	16
dm.3270	caseinolytic peptidase B protein homolog [Pan troglodytes]	14
dm.484	phosphofructokinase, putative [Toxoplasma gondii]	14
lm.74092	2,3 cyclic-nucleotide 2-phosphodiesterase [Thermus thermophiles]	14
lm.85256	RNA polymeraseII elongator complex subunit [Naegleria gruberi]	13
App.3.2 *Euglena* proteins identified in Eg2.2 gel slice		
dm.14567	conserved hypothetical protein [Vibrio harveyi]	4833
dm.14561	hypothetical protein [Shewanella woodyi]	1131
lm.14	malate synthase-isocitrate lyase [Euglena gracilis]	4127
dm.42166	malate synthase-isocitrate lyase [Euglena gracilis]	445
lm.50762	malate synthase-isocitrate lyase [Euglena gracilis]	153
lm.55348	hypothetical protein [Ilumatobacter coccineum]	2563
dm.13947	hypothetical protein [Ilumatobacter coccineum]	1318
dm.6752	pyruvate carboxylase, gene 1 [Xenopus laevis]	2334
dm.10697	inversin protein alternative isoform [Trichomonas vaginalis]	2088
dm.10321	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Danio rerio]	1710
dm.17280	putative acetyl-CoA carboxylase [Leishmania infantum]	1642
dm.8434	hypothetical protein [Rhodanobacter spathiphyllum]	1624
dm.8064	pyruvate decarboxylase, putative [Perkinsus marinus]	1561
dm.27248	hypothetical protein Thini_1924 [Thiothrix nivea]	1419
dm.32024	5-oxoprolinase [Azospirillum brasilense]	1406
dm.33057	Tryptophan synthase beta chain [Planococcus halocryophilus] See Appendix 10	1198
lm.20526	phosphoribosylanthranilate isomerase [Thermobaculum terrenum]	202
dm.20564	protein synthetase, putative [Acanthamoeba castellanii]	993
dm.25026	aminoadipate-semialdehyde dehydrogenase [Capsaspora owczarzaki	808
lm.21957	aminoadipate-semialdehyde dehydrogenase [Capsaspora owczarzaki	694
lm.21099	protein translocase subunit secA, chloroplastic [Coccomyxa subellipsoidea C-169]	775
dm.13277	Endo-1,3-beta-glucanase, C-terminal fragment, family GH81 [Ectocarpus siliculosus]	758
dm.32489	zinc metalloprotease (insulinase family) [Arabidopsis thaliana]	718
lm.9201	Hydroxylamine reductase [Opitutaceae bacterium]	678
dm.7327	carbamoyl-phosphate synthase [ammonia], putative [Phytophora infestans]	669
lm.8430	von Willebrand factor A [Zunongwangia profunda]	635
dm.34328	no hits	586
lm.62373	novel protein similar to H.sapiens IDE, insulin-degrading enzyme (IDE, zgc:162603) [Danio rerio]	555
dm.17299	ubiquitin activating enzyme [Naegleria gruberi]	512
dm.322	heat shock protein 60 [Euglena gracilis]	485
dm.1638	heat shock protein 60 [Euglena gracilis]	65
dm.11294	conserved hypothetical protein [Leishmania Mexicana]	481

lm.13055	hypothetical protein [Mycobacterium sp.]	415
dm.9720	hypothetical protein [Mycobacterium sp.]	337
dm.199	acetyl-coa synthetase [Volvox carteri]	395
dm.22743	hypothetical protein, conserved [Trypanosoma brucei]	361
dm.17211	exportin-1 [Gallus gallus]	359
lm.55849	hypothetical protein [Branchiostoma floridae]	340
dm.54710	predicted protein [Naegleria gruberi]	335
lm.58565	predicted protein [Naegleria gruberi]	272
lm.3719	phosphoenolpyruvate carboxykinase [GTP] [Waddlia chondrophila]	305
dm.1649	phosphoenolpyruvate carboxykinase [GTP] [Waddlia chondrophila]	142
dm.28607	ubiquitin carboxyl-terminal hydrolase 12 [Arabidopsis thaliana]	301
dm.37621	hydrogenobyrinic acid a,c-diamide cobaltochelatase, partial [Coccomyxa subellipsoidea	288
dm.12520	pyridoxal-dependent decarboxylase domain-containing protein 1 [Gallus gallus]	277
lm.52434	dihydropyrimidine dehydrogenase [Polysphondylium pallidum]	274
lm.19930	fatty acid synthase isoform 2 [Amylomyces rouxii]	259
lm.15627	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	232
lm.15621	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	163
lm.15630	membrane alanyl aminopeptidase [Ectocarpus siliculosus]	128
dm.462	glutamate synthase-like protein [Chaetomium thermophilum]	232
dm.3530	hypothetical protein [Monosiga brevicollis]	227
dm.12763	hypothetical protein [Monosiga brevicollis]	82
dm.33277	protein transporter SEC23 [Rhizopus oryzae]	226
dm.51828	ubiquitin fusion degradation protein UfdB, putative [Penicilllium sp.]	200
dm.11270	putative alanyl-tRNA synthetase [Leishmania infantum]	175
dm.78696	sec24-like transport protein [Arabidopsis thaliana]	173
dm.27489	UV-damaged DNA-binding protein 1A [Arabidopsis lyrata]	169
dm.186	pyruvate: NADP+ oxidoreductase [Euglena gracilis]	159
lm.90921	ubiquitin activating enzyme [Naegleria gruberi]	159
lm.29460	hypothetical protein [Phytophthora sojae]	156
dm.150	glutamine synthetase III [Guillardia theta]	149
dm.12625	predicted protein [Naegleria gruberi]	145
lm.95239	putative exoribonuclease 1 [Trypanosoma congolensis]	144
dm.51164	sec24-like transport protein [Arabidopsis thaliana]	144
lm.57893	ankyrin 2,3/unc44 [Aedes aegypti]	144
dm.484	phosphofructokinase, putative [Toxoplasma gondii]	137
dm.31760	putative adenylosuccinate synthetase [Leishmania braziliensis]	124
dm.11295	choline/carnitine O-acetyltransferase [Trypanosoma cruzi]	110
lm.31707	UDP-glucose ceramide glucosyltransferase-like 1, isoform CRA_d [Rattus norvegicus]	95
dm.33444	dihydropyrimidinase [Dictyostelium purpureum]	91
lm.23873	WD repeat protein [Methanosaeta harundinacea]	88
lm.68214	anthranilate/para-aminobenzoate synthase component I [uncultured marine group II euryarchaeote]	87
dm.13143	isocitrate dehydrogenase [Salpingoeca sp]	86
lm.59166	ADP-ribosyltransferase 1-like protein [Adineta vaga]	84
dm.4323	cytosolic fructose-1,6-bisphosphatase [Euglena gracilis]	84
lm.84438	beta-galactosidase [Psychromonas marina]	81
dm.96795	hypothetical protein [Volvox carteri]	76

dm.13156	aminopeptidase with a membrane alanine aminopeptidase domain [Thalassiosira pseudonana]	75
lm.53309	ubiquitinyl hydrolase [Coccomyxa subellipsoidea]	73
lm.91991	NFX1-type zinc finger-containing protein 1-like [Danio rerio]	71
dm.7515	heat shock protein 70 cytosolic isoform [Rhynchobodo]	70
lm.32739	rubrerythrin [Geobacter daltonii]	69
lm.27495	dymeclin hypothetical protein [Phytophthora sojae]	67
lm.62160	predicted protein [Nematostella vectensis]	64
dm.21288	predicted protein [Phaeodactylum tricornutum]	61
lm.87294	proteasome activator complex subunit 4 [Mus musculus]	60
lm.8338	hypothetical protein [Salpingoeca sp]	56
lm.47385	similar to predicted protein [Hydra magnipapillata]	52
dm.21714	short-chain dehydrogenase/reductase SDR [Burkholderia oklahomensis]	49
lm.51633	putative DNA helicase [uncultured marine group II euryarchaeote]	48
lm.8553	propionyl-CoA carboxylase beta chain, mitochondrial precursor [Phytophthora infestans	47
dm.12	aldehyde dehydrogenase [Trichodesmium erythraeum]	47
dm.1790	predicted protein [Physcomitrella patens]	43
lm.59664	exosome complex exonuclease RRP44-like [Crassostrea gigas]	43
dm.30978	ubiquitin-activating enzyme E1 [Salpingoeca sp.]	41
dm.73610	hypothetical protein [Batrachochytrium dendrobatidis]	40
dm.3784	Hsc70-interacting protein [Entamoeba histolytica]	38
dm.61167	peptidase C19 family ubiquitinyl hydrolase [Naegleria gruberi]	35
lm.84453	hypothetical protein, variant [Salpingoeca sp.]	35
dm.12386	isoleucyl-tRNA synthetase, cytoplasmic [Equus caballus]	35
lm.75584	ubiquitin carboxyl-terminal hydrolase [Capsaspora owczarzaki	31
dm.9303	glucose-6-phosphate 1-dehydrogenase [Condylura cristata]	30
dm.25799	hypothetical protein [Trypanosoma cruzi]	30
dm.80676	tetratricopeptide repeat-containing protein [Arabidopsis thaliana]	29
dm.314	biotin carboxylase [Clostridium sp. BNL1100]	29
dm.120	heat shock protein HSP70 [Trypanosoma cruzi]	28
dm.25565	No hits	28
dm.4304	D-3-phosphoglycerate dehydrogenase [Novosphingobium nitrogenifigens]	27
dm.24109	kinesin K39, putative, partial [Leishmania infantum]	27
dm.37941	flavine-containing monoxygenase [Populus trichocarpa]	25
lm.38812	L-serine dehydratase, iron-sulfur-dependent, single chain form [Pseudomonas stutzeri]	24
lm.49430	ankyrin repeat protein [Paracoccidioides brasiliensis]	23
lm.11898	prolyl-tRNA synthetase [Trypanosoma cruzi]	22
dm.37504	ubiquitin carboxyl-terminal hydrolase 48-like, partial [Anolis carolinensis]	21
lm.58393	conserved hypothetical protein [Pediculus humanus]	20
dm.28974	hypothetical protein [Guillardia theta]	19
lm.11603	magnesium chelatase subunit H [Synechococcus	19
dm.37470	hypothetical protein [Sulfuricella denitrificans]	19
dm.16810	WD40 repeat-containing protein [Polysphondylium]	18
dm.31837	ubiquitin-activating enzyme e1, putative [Leishmania donovani]	18
dm.49306	No hits	18
lm.51040	protein serine/threonine kinase [Polysphondylium pallidum]	17
dm.11778	ubiquinone/menaquinone biosynthesis methyltransferase	17

dm.46928	Cytoplasmic dynein 2 heavy chain 1[Tripneustes gratilla]	17
lm.67640	alpha-type protein kinase [Euglena gracilis]	17
dm.446	RNA-binding region RNP-1 domain-containing protein [Polysphondylium pallidum]	17
dm.3232	vitamin B6 biosynthesis protein [Wallemia sebi]	17
dm.46	heat shock protein 90 [Euglena gracilis]	17
dm.8303	Saccharopine dehydrogenase [Caldithrix abyssi]	17
lm.3583	kinesin [Trypanosoma cruzi]	16
dm.93772	Protein GBF-1, isoform a [Caenorhabditis elegans]	16
dm.28650	No hits	16
lm.92865	No hits	15
lm.50421	Tetratricopeptide TPR_2 repeat protein [Arthrospira maxima]	14
dm.65397	predicted protein [Micromonas pusilla]	14
dm.17873	6-phosphogluconate dehydrogenase [Euglena gracilis]	13

Proteins from second purification protocol		
Transcript number	Closest homologue	Score
App.3.3 *Euglena* proteins identified in EgC		
dm.14570	hypothetical protein [marine gamma proteobacteria]	557
lm.36680	hypothetical protein [Kordia algicida]	378
lm.14	malate synthase-isocitrate lyase [Euglena gracilis]	530
dm.322	heat shock protein 60 [Euglena gracilis]	146
dm.3530	hypothetical protein [Monosiga brevicollis]	122
dm.138	elongation factor 1 alpha [Euglena gracilis]	92
dm.8145	propionyl-CoA carboxylase beta chain, mitochondrial precursor [Phytophthora infestans]	83
lm.8553	propionyl-CoA carboxylase beta chain, mitochondrial precursor [Phytophthora infestans]	65
dm.1598	pyruvate kinase [Takifugu rubripes]	79
dm.186	pyruvate: NADP+ oxidoreductase [Euglena gracilis]	54
dm.27598	vacuolar protein 8 [Coniophora puteana	40
lm.8430	von Willebrand factor A [Zunongwangia profunda]	39
dm.25642	hypothetical protein [Sorangium cellulosum]	36
lm.15620	puromycin-sensitive aminopeptidase-like protein [Salpingoeca sp.]	27
App.3.4 *Euglena* proteins identified in EgD		
dm.14570	hypothetical protein [marine gamma proteobacterium]	1117
lm.36680	hypothetical [Kordia algicida]	663
lm.14	malate synthase-isocitrate lyase [Euglena gracilis]	951
dm.138	elongation factor 1 alpha [Euglena gracilis]	202
dm.3530	hypothetical protein [Monosiga brevicollis]	165
lm.8550	propionyl-CoA carboxylase beta chain, mitochondrial precursor [Phytophthora infestans]	153
dm.186	pyruvate: NADP+ oxidoreductase [Euglena gracilis]	83
dm.322	heat shock protein 60 [Euglena gracilis]	50
dm.1598	pyruvate kinase [Takifugu rubripes]	33
dm.27598	vacuolar protein 8 [Coniophora puteana]	17
lm.15620	puromycin-sensitive aminopeptidase-like protein [Salpingoeca sp.]	16

Appendix 4 – Carbohydrate-active enzymes in the *Euglena* transcriptome

Glycosyl Hydrolases		Glycosyl Transferases		Accessory modules	
Euglena transcript #	CAZy module	*Euglena* transcript #	CAZy module	*Euglena* transcript #	CAZy module
lm.80490	GH1	dm.12422	GT1	lm.25689	CBM48
lm.93530	GH1	dm.19420	GT1	lm.101413	CBM57
dm.51553	GH2	dm.19717	GT1	dm.80173	EXPN-CBM63
dm.56224	GH2	dm.33982	GT1	lm.68827	CE3
lm.44961	GH2	dm.80517	GT1	dm.66677	CE13
lm.84438	GH2	dm.90384	GT1	dm.71771	CE13
lm.98570	GH2	lm.22765	GT1	lm.102226	CE13
lm.98571	GH2	lm.22766	GT1	lm.102297	CE13
dm.104047	GH3	lm.34813	GT1	lm.48774	CE13
dm.80650	GH3	lm.53934	GT1	lm.60242	CE13
dm.89679	GH3	lm.75841	GT1	lm.90170	CE13
lm.101242	GH3	lm.75842	GT1	lm.90173	CE13
lm.72409	GH3	lm.79780	GT1		
lm.81977	GH3	lm.86108	GT1	Distantly related Glycosyl Transferases	
lm.89053	GH3	lm.87840	GT1		
lm.89054	GH3	dm.47703	GT1-GH78	*Euglena* transcript #	CAZy module(s)
lm.89055	GH3	dm.27722	GT2	lm.72487	GT23
lm.93078	GH3	dm.27724	GT2	lm.91630	GT23
lm.95673	GH3	dm.27726	GT2	lm.55165	GT25
lm.95674	GH3	dm.27728	GT2	lm.94749	GT25
dm.35835	GH5	dm.27873	GT2	dm.32272	GT32
dm.35839	GH5	dm.27874	GT2	dm.37769	GT32
dm.35840	GH5	dm.27876	GT2	dm.39718	GT32
dm.44807	GH5	dm.42731	GT2	dm.48218	GT32
dm.69508	GH5	dm.58009	GT2	dm.81916	GT32
dm.83397	GH5	dm.73971	GT2	lm.13695	GT32
lm.40866	GH5	dm.96021	GT2	lm.32837	GT32
lm.40874	GH5	lm.108044	GT2	lm.68834	GT32
lm.40875	GH5	lm.31276	GT2	dm.30790	GT34
lm.61020	GH5	lm.31364	GT2	dm.37781	GT34
lm.71927	GH5	lm.32222	GT2	dm.37782	GT34
dm.91911	GH117	lm.32223	GT2	dm.37784	GT34
lm.76046	GH117	lm.32225	GT2	dm.68760	GT34
lm.76047	GH117	lm.32226	GT2	lm.52552	GT34
dm.440	GH17	lm.32229	GT2	lm.68129	GT34
dm.56861	GH17	lm.48352	GT2	lm.83427	GT34
lm.47728	GH17	lm.56291	GT2	lm.96814	GT4

dm.47800	GH18	lm.68894	GT2	dm.87760	GT47
lm.103637	GH18	lm.81887	GT2	dm.22087	GT61
dm.34783	GH18-CBM18	lm.98495	GT2	dm.33771	GT61
lm.29686	GH18-CBM18	dm.15135	GT4	lm.24531	GT61
lm.102325	GH19	dm.79942	GT4	lm.24533	GT61
dm.89574	GH20	dm.81869	GT4	lm.24541	GT61
lm.88319	GH20	dm.90804	GT4	lm.32128	GT61
lm.88320	GH20	lm.20804	GT4	lm.32131	GT61
dm.56632	GH27	lm.21766	GT4	lm.32134	GT61
dm.63503	GH30	lm.33890	GT4	lm.32135	GT61
dm.63509	GH30	lm.58009	GT4	lm.70292	GT61
lm.17784	GH30	lm.62122	GT4	lm.70293	GT61
lm.17787	GH30	lm.67740	GT4	lm.84274	GT61
lm.17797	GH30	lm.68525	GT4	lm.90176	GT61
lm.17800	GH30	lm.68818	GT4	lm.99017	GT61
lm.17814	GH30	lm.75595	GT4	lm.97192	GT69
lm.17815	GH30	lm.75596	GT4	dm.12391	GT75
lm.17825	GH30	lm.76927	GT4	dm.37464	GT90
lm.17826	GH30	lm.79157	GT4	dm.67893	GT90
lm.17841	GH30	lm.84277	GT4		
lm.44824	GH30	lm.93157	GT4		
lm.66301	GH30	lm.93215	GT4		
dm.22178	GH31	lm.95961	GT4		
dm.53023	GH31	lm.97615	GT4		
dm.62947	GH31	dm.57564	GT8		
dm.65434	GH31	dm.86901	GT8		
lm.75791	GH31	dm.86904	GT8		
lm.75792	GH31	dm.86905	GT8		
lm.75793	GH31	lm.54268	GT8		
lm.81297	GH31	lm.62538	GT8		
dm.25013	GH36	lm.89703	GT8		
lm.23614	GH36	lm.93668	GT8		
lm.52294	GH36	lm.93674	GT8		
lm.52298	GH36	lm.93676	GT8		
lm.52303	GH36	lm.93679	GT8		
lm.52304	GH36	lm.93682	GT8		
dm.79592	GH38	lm.93685	GT8		
lm.77431	GH38	lm.93697	GT8		
lm.89296	GH38	lm.93704	GT8		
lm.89298	GH38	lm.93705	GT8		
lm.71907	GH42	lm.98440	GT8		
dm.47656	GH43	dm.25919	GT10		
lm.50076	GH43	dm.61080	GT10		

lm.93323	GH43	dm.72664	GT10
dm.48562	GH47	dm.78084	GT10
dm.48567	GH47	dm.78087	GT10
dm.59995	GH47	dm.78088	GT10
dm.59996	GH47	lm.67669	GT10
dm.70793	GH47	lm.78120	GT10
dm.70795	GH47	lm.89155	GT10
dm.70799	GH47	lm.90278	GT10
dm.72005	GH47	lm.91263	GT10
lm.82143	GH47	lm.91264	GT10
lm.82528	GH47	lm.92122	GT10
lm.89589	GH47	lm.71174	GT11-GT15
dm.30807	GH5	lm.97566	GT13
dm.22013	GH55	dm.49632	GT15
lm.96365	GH63	lm.70358	GT15
lm.97052	GH63	lm.83960	GT15
lm.63754	GH64	dm.31521	GT17
dm.92069	GH65	dm.88372	GT17
lm.12469	GH65	lm.88149	GT17
lm.41777	GH65	lm.92197	GT17
lm.50851	GH65	lm.29618	GT20
lm.95741	GH65	dm.60521	GT22
lm.71986	GH78	dm.60522	GT22
dm.20767	GH81	dm.85690	GT22
dm.25642	GH81	dm.85692	GT22
dm.31657	GH81	dm.85694	GT22
dm.31843	GH81	lm.71029	GT22
dm.31844	GH81	dm.67541	GT23
dm.66774	GH81	dm.67542	GT23
dm.82797	GH81	dm.83534	GT23
lm.26257	GH81	dm.96497	GT23
lm.28976	GH81	lm.100539	GT23
lm.28977	GH81	lm.26965	GT23
lm.39241	GH81	lm.69542	GT23
lm.39306	GH81	lm.81778	GT23
lm.39307	GH81	lm.87230	GT23
lm.4691	GH81	lm.89315	GT23
lm.80430	GH81	lm.89316	GT23
lm.84039	GH81	lm.94751	GT23
lm.90769	GH81	lm.31707	GT24
lm.91687	GH81	dm.47345	GT25
lm.94770	GH85	lm.69918	GT25
lm.94771	GH85	dm.42760	GT28
lm.94772	GH85	dm.26818	GT31

lm.94775	GH85
lm.102467	GH88
dm.99395	GH99

dm.36833	GT31
dm.39545	GT31
dm.39554	GT31
dm.39555	GT31
dm.39557	GT31
dm.39558	GT31
dm.83370	GT31
lm.27958	GT31
lm.48152	GT31
lm.76406	GT31
lm.95852	GT31
lm.97727	GT31
dm.27659	GT32
dm.31546	GT32
dm.56081	GT32
dm.71099	GT32
lm.26881	GT32
lm.29288	GT32
lm.46212	GT32
lm.60500	GT32
lm.68996	GT32
lm.85050	GT32
lm.93810	GT32
lm.92144	GT33
lm.93348	GT34
dm.27721	GT40
lm.68449	GT40
dm.35031	GT41
lm.52466	GT41
lm.92993	GT41
dm.39277	GT47
dm.39390	GT47
dm.39998	GT47
dm.40001	GT47
dm.66841	GT47
dm.76900	GT47
lm.68145	GT47
lm.85331	GT47
lm.87611	GT47
lm.88157	GT47
lm.89017	GT47
lm.89707	GT47
lm.90869	GT47
lm.92839	GT47

lm.92840	GT47
dm.22210	GT48
lm.13240	GT48
lm.13249	GT48
lm.13336	GT48
dm.58200	GT49
dm.64067	GT49
lm.94033	GT50
dm.96080	GT57
lm.99883	GT57
lm.68532	GT58
lm.100691	GT59
dm.33780	GT61
dm.33783	GT61
dm.33784	GT61
dm.47644	GT61
dm.56369	GT61
dm.74359	GT61
lm.32130	GT61
lm.32132	GT61
lm.32133	GT61
lm.32136	GT61
lm.47661	GT61
lm.70294	GT61
lm.70295	GT61
lm.79988	GT61
lm.86157	GT61
lm.95124	GT61
dm.17756	GT66
dm.41728	GT66
lm.68366	GT66
dm.60061	GT69
dm.60062	GT69
lm.47975	GT75
lm.47982	GT75
dm.100639	GT76
lm.98056	GT76
dm.20225	GT77
dm.26143	GT77
dm.26148	GT77
dm.46237	GT77
dm.46238	GT77
dm.48743	GT77
dm.57100	GT77

dm.57112	GT77
dm.64325	GT77
dm.70621	GT77
lm.28466	GT77
lm.43281	GT77
lm.43282	GT77
lm.43283	GT77
lm.43284	GT77
lm.43910	GT77
lm.52104	GT77
lm.56292	GT77
lm.57291	GT77
lm.90263	GT77
lm.90264	GT77
lm.90266	GT77
lm.90271	GT77
lm.90272	GT77
lm.92804	GT77

Appendix 5 – Numbers of CAZys annotated in the database for various species.[1-3]

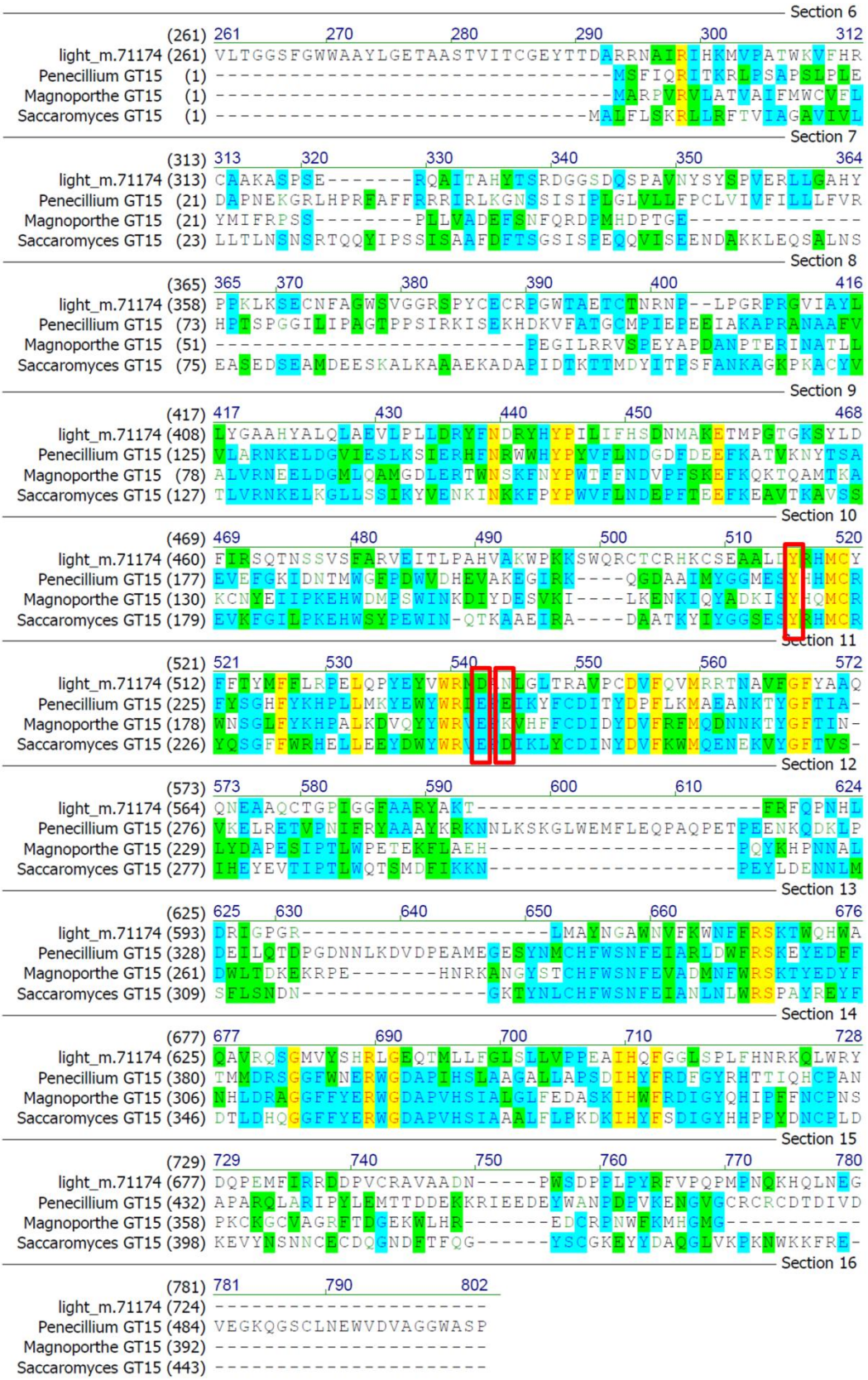
	#GH	#GT	#CBM
Fungi			
Ascomycota			
Aspergillus	251	92	43
Botryotinia	242	106	65
Leptosphaeria	216	100	54
Magnaporthe	232	94	71
Penicillium	219	103	50
Myceliophthora	194	88	46
Neurospora	174	78	27
Basidiomycota			
Cryptococcus	85	68	1
Malassezia	25	41	1
Piriformospora	200	73	101
Sporisorium	103	66	7
Yeasts			
Ashbya	45	58	9
Candida	56	69	10
Kluyveromyces	48	63	11
Saccharomyces	51	66	12
Yarrowia	54	69	6
Oomycetes			
Albugo	88	68	10
Phytophora	279	157	40
Plant			
Arabidopsis	400	562	126
Oryza	428	607	119
Bathycoccus	49	171	25
Micromonas	41	86	30
Osterococcus	30	70	22
Red algae			
Cyanidioschyzon	21	54	13
Porphyridium	31	83	40
Glaucophytes			
Cyanophora	84	128	24
Cryptophytes			
Guillardaria	53	257	29
Chlorarachinophytes			
Bigelowiella	112	109	7
Animals			
Homo	98	235	40
Caenorhabditis	111	270	80
Drosophila	103	150	279
Euglenozoa			
Euglena	127	228	4
Leishmania	12	41	1
Trypanosoma	19	36	1

Appendix 6 – Alignments of didomain CAZys (lm.71174 and dm.47703) with well characterised single domain proteins

Alignments were performed using AlignX (Invitrogen) and confirmed by modelling proteins using SWISS-MODEL in automated mode.[4] The key catalytic residues are highlighted in red boxes.

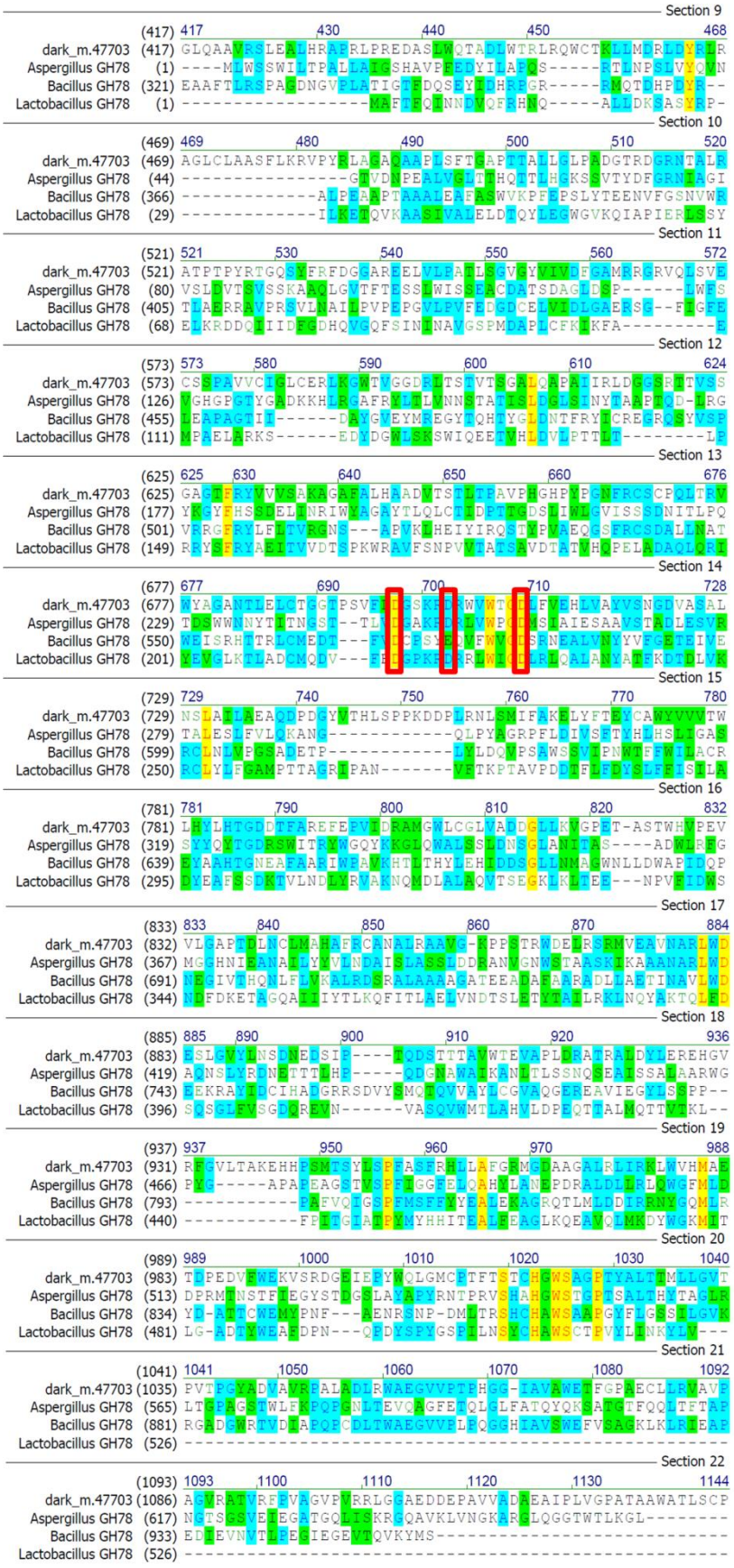
App.6.1 Alignment of lm.71174 with GT11s[5]

App.6.2 Alignment of lm.71174 with GT15s[6]



App.6.3 Alignment of dm.47703 with GT1s[7]

App.6.4 Alignment of dm.47703 with GH78s[8]



Appendix 7 – Rhamnosidase from *Euglena* transcript dm.44703 C-terminus (EgRase)

App.7.1 Synthetic gene for EgRase

The synthetic gene for the C-terminus of dm.44703 was ordered from Genescript (www.genescript.com), codon optimised for expression in *Nicotiana tabacum* (Fig App.7.1), in case *E. coli* expression was no possible, sub-cloned into expression vector pET15b, to form pET15b-dm.47703-CTerm (Fig App.7.2).

A:

CATATGCCCGGGAAGCTTCTCATGGATAGATTAGATTACAGGTTAAGAGCAGGACTTTGTCTCGCAGCATCATTTTTGAAGAGGG
TGCCTTACAGATTAGCAGGTGCTCAAGCTGCACCACTCTCATTTACTGGAGCACCTACTACAGCTCTTTTGGGTTTACCAGCTGA
TGGAACTAGAGATGGTAGGAATACAGCATTGAGAGCTACACCAACTCCTTACAGGACTGGACAATCATATTTTAGATTCGATGGA
GGTGCAAGGGAAGAGTTAGTGCTCCCAGCTACACTTAGTGGAGTTGGTTACGTTATCGTGGATTTCGGAGCTATGAGAAGGGGTA
GAGTTCAGTTGAGTGTGGAATGTTCTTCACCTGCTGTTGTGTGTATTGGACTCTGCGAGAGACTTAAGGGTTGGACTGTGGGAGG
TGATAGGCTTACATCCACTGTTACATCTGGAGCTTTGCAAGCTCCTGCAATTATAAGATTAGATGGAGGTTCTAGGACTACAGTT
AGTTCCGGAGCTGGTACATTTAGATATGTTGTGGTTTCAGCTAAAGCAGGTGCTTTCGCACTCCATGCTGCAGATGTGACTAGTA
CACTTACTCCAGCTGTTCCTCATGGACACCCATACCCTGGTAATTTTAGATGTTCATGCCCACAGCTTACTAGGGTTTGGTATGC
TGGAGCAAACACACTTGAATTGTGCACTGGAGGTACACCTTCAGTTTTTCTTGATGGTAGTAAGAGAGATAGGTGGGTGTGGACT
GGTGATCTCTTCGTTGAACATCTTGTTGCATACGTGTCTAATGGAGATGTTGCTTCAGCACTTAACAGTTTGGCTATTTTAGCTG
AGGCACAGGATCCAGATGGTTATGTTACACACCTTTCCCCACCTAAGGATGATCCTCTCAGAAACCTTTCTATGATCTTTGCAAA
AGAATTGTATTTCACTGAGTACTGTGCTTGGTATGTTGTTGTGACATGGTTGCATTACTTACACACTGGTGATGATACATTTGCT
AGAGAATTCGAGCCAGTTATTGATAGGGCAATGGGATGGCTCTGCGGTCTTGTGGCTGATGATGGATTACTCAAGGTTGGTCCTG
AAACTGCATCTACATGGCATGTTCCAGAGGTTGTGCTTGGAGCTCCTACTGATTTGAATTGTTTAATGGCTCACGCATTTAGATG
CGCTAACGCATTGAGGGCTGCAGTTGGTAAACCACCTTCAACAAGATGGGATGAATTGAGAAGTAGGATGGTTGAGGCAGTGAAC
GCTAGGTTATGGGATGAATCCTTGGGAGTTTACCTCAACTCCGATAACGAGGATTCTATTCCAACTCAAGATTCTACTACAACTG
CAGTTTGGACTGAAGTGGCTCCTCTTGATAGAGCAACAAGGGCTCTCGATTACCTTGAAAGAGAGCATGGAGTTAGGTTTGGTGT
GCTTACTGCTAAAGAGCATCACCCATCTATGACATCATATTTGAGTCCTTTTGCATCCTTCAGACATCTTTTGGCTTTCGGAAGG
ATGGGTGATGCTGCAGGTGCATTAAGACTCATAAGGAAGTTGTGGGTTCACATGGCTGAAACTGATCCAGAGGATGTGTTTTGGG
AAAAAGTTTCAAGAGATGGAGAAATCGAGCCATATTGGCAGCTTGGAATGTGTCCTACATTCACTTCCACATGTCATGGATGGTC
TGCAGGTCCAACATACGCTTTGACAACTATGTTACTCGGAGTTACTCCAGTGACACCTGGTTATGCAGATGTTGCTGTGAGACCT
GCTTTGGCAGATTTAAGATGGGCTGAAGGAGTTGTGCCAACTCCTCACGGAGGTATCGCTGTTGCATGGGAAACATTTGGTCCAG
CTGAGTGTCTTTTGAGAGTTGCAGTGCCTGCTGGAGTTAGAGCAACTGTGAGGTTCCCAGTTGCTGGTGTTCCTGTGAGAAGGCT
TGGAGGTGCAGAAGATGATGAGCCAGCTGTTGTGGCTGATGCAGAGGCTATACCTTTGGTTGGTCCTGCTACTGCTGCTTGGGCA
ACACTTTCCTGTCCTTCTGATTTCTCCTTTGAATGACTCGAG

B:

MGSSHHHHHHSSVLPPRGSHMPGKLLMDRLDYRLRAGLCLAASFLKRVPYRLAGAQAAPLSFTGAPTTALLGLPADGTRDGRNTA
LRATPTPYRTGQSYFRFDGGAREELVLPATLSGVGYVIVDFGAMRRGRVQLSVECSSPAVVCIGLCERLKGWTVGGDRLTSTVTS
GALQAPAIIRLDGGSRTTVSSGAGTFRYVVVSAKAGAFALHAADVTSTLTPAVPHGHPYPGNFRCSCPQLTRVWYAGANTLELCT
GGTPSVFLDGSKRDRWVWTGDLFVEHLVAYVSNGDVASALNSLAILAEAQDPDGYVTHLSPPKDDPLRNLSMIFAKELYFTEYCA
WYVVVTWLHYLHTGDDTFAREFEPVIDRAMGWLCGLVADDGLLKVGPETASTWHVPEVVLGAPTDLNCLMAHAFRCANALRAAVG
KPPSTRWDELRSRMVEAVNARLWDESLGVYLNSDNEDSIPTQDSTTTAVWTEVAPLDRATRALDYLEREHGVRFGVLTAKEHHPS
MTSYLSPFASFRHLLAFGRMGDAAGALRLIRKLWVHMAETDPEDVFWEKVSRDGEIEPYWQLGMCPTFTSTCHGWSAGPTYALTT
MLLGVTPVTPGYADVAVRPALADLRWAEGVVPTPHGGIAVAWETFGPAECLLRVAVPAGVRATVRFPVAGVPVRRLGGAEDDEPA
VVADAEAIPLVGPATAAWATLSCPSDFSFE*

Figure App.7.1: The sequence of EgRase

A: The DNA sequence of the ordered EgRase gene, including the tags (blue) added, encoding NdeI, XmaI and HindIII restriction sites at the start and XhoI at the end, to facilitate cloning. **B:** The protein sequence of the expressed EgRase with the tag (blue) added during cloning and from the pET15b, encoding a hexahistidine tag to facilitate purification.

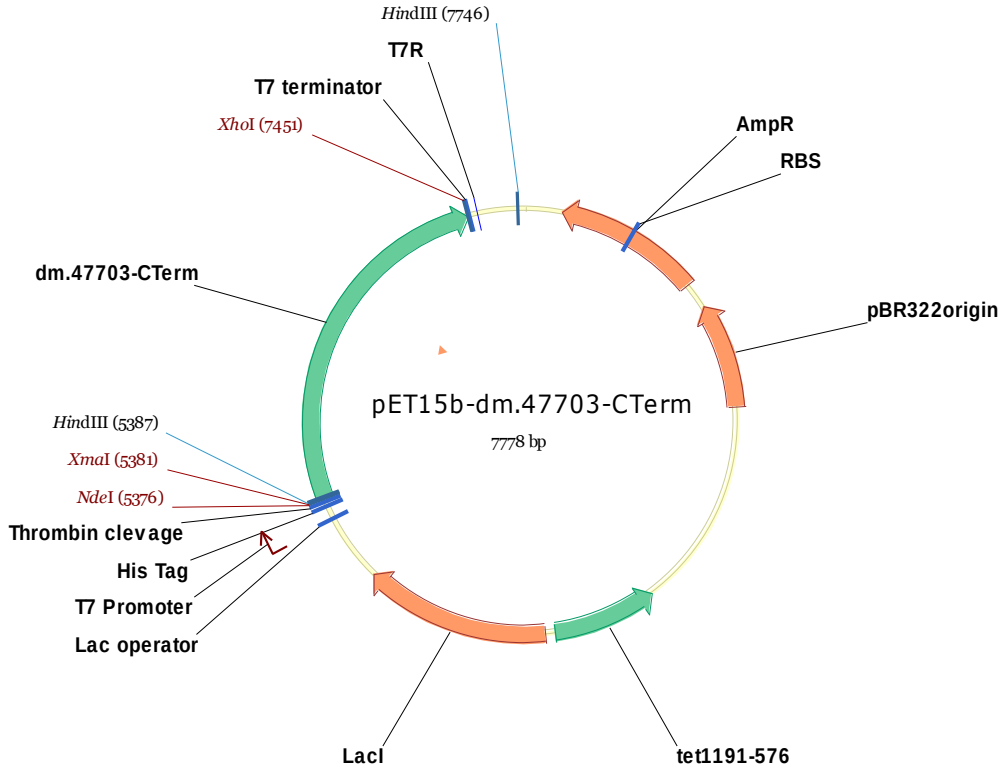


Figure App.7.2: Plasmid map of pET15b-dm.47703-CTerm

App.7.2 Expression and purification of EgRase

pET15b-dm.47703-CTerm was transformed into *E. coli* strain BL21 and selected on carbenicillin. EgRase was expressed in these cells, grown in AIM at 37 °C until OD_{600} reached 0.1 when the temperature was reduced to 18 °C. After culturing overnight the cells were collected, lysed and the protein purified using the normal protocol (see Section 8.1.2) to yield approximately 0.5 mg per litre cell culture of >95% pure protein, as judged by SDS-PAGE (Fig App.7.3).

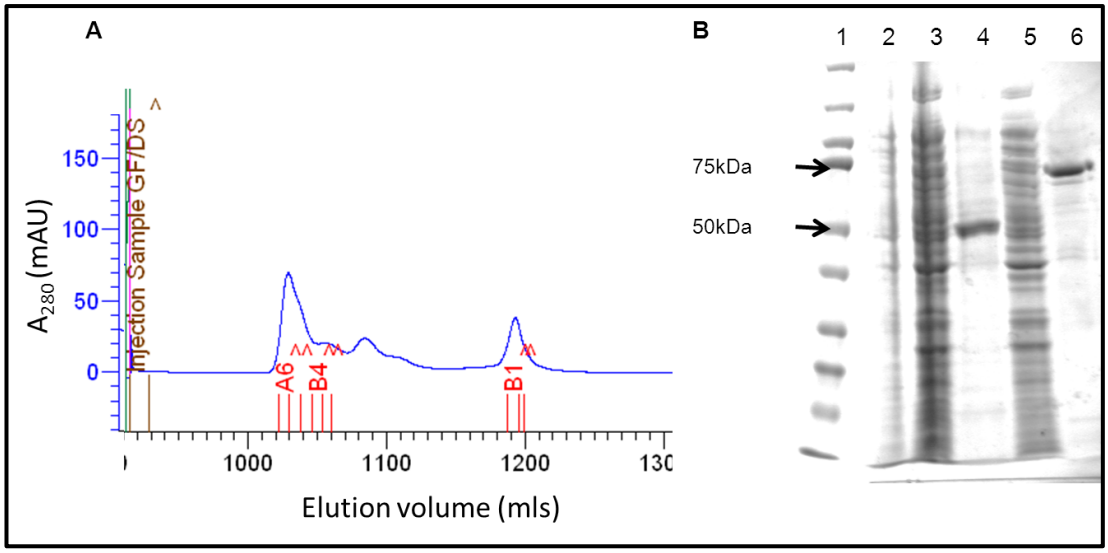


Figure App.7.3: Purification of EgRase

EgRase was purified from *E. coli* using Ni-IMAC and GF (**A**) to give one peak of pure protein (fractions A6, B6, B5), some partially degraded protein fractions (B4-B2) and one peak of UV absorbent small molecules (B1). **B.** The pure protein were combined and run on SDS-PAGE. Lane 1: Kaleidoscope ladder (BioRad). Lane 2: cells containing pET15b-dm.47703-CTerm, but grown in LB. Lane 3: cells expressing the protein. Lane 4: insoluble cell fraction. Lane 5: soluble cell fraction. Lane 6: combined fractions A6, B6, B5.

App.7.3 Activity of EgRase

The hydrolysis of PNP-α-L-Rhamnose (PNP-Rha) by the EgRase was measured under standard reaction conditions (20 mM MES, pH 7.5, 10% DMSO, 1.75 µg/ml EgRase containing 10 % GF buffer) in the range 0-10 mM substrate and incubated at 37 °C. After 3 hours, to allow significant turnover at this very low enzyme concentration, an equal volume (100 µl) of Na_2CO_3 (2 M) was added and the released PNP was measured by absorption at 415 nm, with comparison to standards (Fig App.8.4). This gave a K_M of 400 ± 65 µM and a specific activity of 106 ± 4 µmol/min/mg protein, giving a turnover of 136 per second, in good agreement with the well characterised, and crystallised *Bacillus* homologue (K_M = 280 µM, k_{cat} = 140 per second).[8]

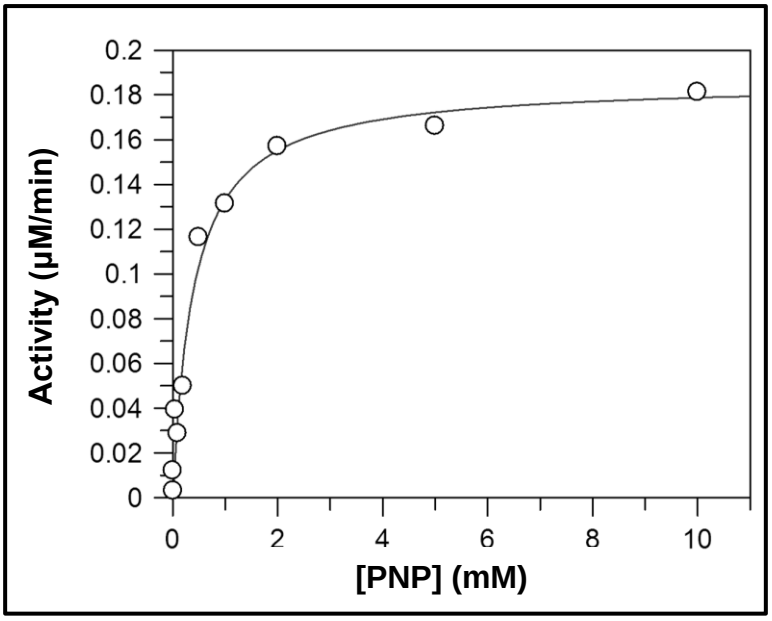


Figure App.7.4: Activity assay of EgRase

The activity assay was performed in duplicate and the average shown. The data was processed using GraFit (Erithacus Software Ltd.) to calculate the V_{max} and K_M.

Appendix 8 – Euglena transcripts for core metabolic enzymes

(Only the first *Euglena* homologue for each enzyme is detailed)

E.C.	Activity	Transcript	Nearest homologue	Evalue
App.8.1 Glycolysis/gluconeogenesis				
5.4.2.2	phosphoglucomutase	lm.27422	phosphoglucomutase [Phytophthora infestans T30-4]	6.00E-120
5.3.1.9	Glucose-6-phosphate isomerase	lm.56040	Glucose-6-phosphate isomerase [Crassostrea gigas]	0.00E+00
3.1.3.11	fructose-bisphosphatase	dm.3989	fructose-1,6-bisphosphatase [Guillardia theta]	5.00E-165
2.7.1.11	6-phosphofructokinase	lm.19925	putative 6-phospho-1-fructokinase [Leishmania mexicana]	9.00E-145
2.7.1.90	diphosphate-fructose-6-phosphate 1-phosphotransferase	lm.13041	Pyrophosphate--fructose 6-phosphate 1-phosphotransferase subunit beta [Ectocarpus siliculosus]	0.00E+00
4.1.2.13	Fructose-bisphosphate aldolase	dm.182	fructose-bisphosphate aldolase, class I [Galdieria sulphuraria]	7.00E-153
5.3.1.1	Triose-phosphate isomerase	dm.3382	triosephosphate isomerase (TIM) [Galdieria sulphuraria]	3.00E-107
1.2.1.12	glyceraldehyde-3-phosphate dehydrogenase	lm.52	glyceraldehyde 3-phosphate dehydrogenase [Bodo saltans]	0.00E+00
2.7.2.3	phosphoglycerate kinase	light 6737	hypothetical protein [Guillardia theta]	0.00E+00
5.4.2.1	phosphoglycerate mutase	lm.11257	phosphoglycerate mutase [Thalassiosira pseudonana]]	9.00E-108
4.2.1.11	phosphopyruvate hydratase	dm.588	phosphopyruvate hydratase [Candidatus Accumulibacter phosphatis]	0.00E+00
2.7.1.40	pyruvate kinase	dm.39974	pyruvate kinase [Galdieria sulphuraria]	3.00E-162

App.8.2 Citrate cycle				
2.3.3.1	citrate synthase	lm.23523	citrate synthase I [Rhodospirillum centenum SW]	5.00E-180
4.2.1.3	aconitate hydratase	lm.27090	hypothetical protein [Selaginella moellendorffii]	0.00E+00
1.1.1.41	isocitrate dehydrogenase	lm.56820	isocitrate dehydrogenase [Salpingoeca sp.]	2.00E-166
6.2.1.4	succinate-CoA ligase	lm.9621	succinate-CoA ligase [Polysphondylium pallidum]	3.00E-137
1.3.5.1	succinate dehydrogenase	dm.59812	predicted protein [Ostreococcus lucimarinus]	0.00E+00
4.2.1.2	fumarate hydratase	lm.18575	hypothetical protein [Saccoglossus kowalevskii]	0.00E+00
1.1.1.37	malate dehydrogenase	lm.15162	malate dehydrogenase, mitochondrial precursor [Phytophthora infestans]	2.00E-147
App.8.3 Pentose phosphate pathway				
1.1.1.47	glucose 1-dehydrogenase	dm.50276	GDH/6PGL endoplasmic bifunctional protein isoform X1 [Ficedula albicollis]	0.00E+00
3.1.1.17	gluconolactonase	dm.27330	hypothetical protein [Paenibacillus sp.]	1.00E-50
2.7.1.12	gluconokinase	lm.34185	gluconokinase [Rhizobium leguminosarum]	1.00E-57
1.1.1.44	phosphogluconate dehydrogenase	lm.22792	6-phosphogluconate dehydrogenase putative [Albugo laibachii]	0.00E+00
5.3.1.6	Ribose-5-phosphate isomerase	lm.10968	predicted protein [Bathycoccus prasinos]	4.00E-82
2.7.6.1	ribose-phosphate diphosphokinase	lm.72623	ribose-phosphate pyrophosphokinase 2 isoform X1 [Echinops telfairi]	3.00E-84
5.1.3.1	ribulose-phosphate 3-epimerase	dm.6906	ribulosephosphate 3epimerase putative [Albugo laibachii]	8.00E-106
2.2.1.1	transketolase	lm.12767	transketolase [Chthoniobacter flavus]	0.00E+00
2.2.1.2	transaldolase	lm.15850	transaldolase [Rattus norvegicus]	2.00E-145
4.1.2.13	Fructose-bisphosphate aldolase	lm.11373	hypothetical protein [Monosiga brevicollis]	6.00E-176
2.7.1.11	6-phosphofructokinase	lm.15159	putative 6-phospho-1-fructokinase [Leishmania mexicana]	2.00E-139
3.1.3.11	fructose-bisphosphatase	lm.38478	inositol phosphatase/fructose-1,6-bisphosphatase [Nitrosococcus	3.00E-163

			halophilus]	
2.7.6.1	ribose-phosphate diphosphokinase	lm.72623	PREDICTED: ribose-phosphate pyrophosphokinase 2 isoform X1 [Echinops telfairi]	3.00E-84
3.1.1.31	6-phosphogluconolactonase	lm.67599	6-phosphogluconolactonase, putative [Phytophthora infestans]	2.00E-78
App.8.4 Fatty acid biosynthesis				
6.4.1.2	acetyl-CoA carboxylase	lm.45716	putative acetyl-CoA carboxylase [Leishmania infantum]	0.00E+00
2.3.1.86	fatty-acyl-CoA synthase	lm.76569	fatty acid synthase [Wickerhamomyces ciferrii]	6.00E-54
3.1.2.14	oleoyl-[acyl-carrier-protein] hydrolase	lm.47166	hypothetical protein [Streptomyces scabrisporus]	2.00E-24
Carotenoid biosynthesis.				
2.5.1.32	phytoene synthase	lm.32898	predicted protein [Thalassiosira pseudonana]	4.00E-108
1.3.5.5	phytoene desaturase	lm.43440	phytoene dehydrogenase [Phaeodactylum tricornutum]	0.00E+00
5.2.1.12	carotene isomerase	dm.32619	hypothetical protein [Guillardia theta]	9.00E-140
App.8.5 Calvin cycle				
4.1.1.39	Ribulose-bisphosphate carboxylase	lm.3071	ribulose-bisphosphate carboxylase small chain precursor [Chlamydomonas sp.]	1.00E-60
1.2.1.13	glyceraldehyde-3-phosphate dehydrogenase	lm.3559	glyceraldehyde-3-phosphate dehydrogenase [Pyrocystis lunula]	0.00E+00
3.1.3.37	sedoheptulose-bisphosphatase	lm.14721	sedoheptulose-1,7-biphosphatase [Volvox carteri f. nagariensis]	2.00E-156
2.7.1.19	phosphoribulokinase	dm.49223	predicted protein [Micromonas sp.]	5.00E-166

Appendix 9 – Euglena transcripts for biosynthetic enzymes

(Only the first two *Euglena* homologues for each enzyme are detailed)

E.C.	Activity	Transcript	Nearest homologue	Evalue	Transcript	Nearest homologue	Evalue
App 9.1 Isoprenoid biosynthesis							
2.3.1.9	acetyl-CoA C-acetyltransferase	lm.23285	acetyl-CoA acetyltransferase [Reinekea blandensis]	4.00E-150	lm.46227	hypothetical protein[uncultured bacterium]	4.00E-155
2.3.3.10	hydroxymethylglutaryl-CoA synthase	lm.98584	hydroxymethylglutaryl-CoA synthase 1-like [Acyrthosiphon pisum]	4.00E-83	lm.20443	hypothetical protein [Trypanosoma cruzi]	3.00E-174
1.1.1.34	hydroxymethylglutaryl-CoA reductase (NADPH)	lm.36658	putative 3-hydroxy-3-methylglutaryl-CoA reductase [Leishmania braziliensis]	4.00E-170			
2.7.1.36	mevalonate kinase	lm.47203	mevalonate kinase [Pyrococcus sp.]	5.00E-39			
2.7.4.2	phosphomevalonate kinase	dm.76890	predicted protein [Naegleria gruberi]	2.00E-82			
4.1.1.33	diphosphomevalonate decarboxylase	none					
2.2.1.7	1-deoxy-D-xylulose-5-phosphate synthase	dm.12516	1-deoxy-D-xylulose 5-phosphate synthase, chloroplast precursor [Ectocarpus siliculosus]	0.00E+00			
1.1.1.267	1-deoxy-D-xylulose-5-	lm.41205	putative plastid 1-deoxy-D-xylulose 5-phosphate	0.00E+00			

	phosphate reductoisomerase		reductoisomerase precursor [Pyropia yezoensis]				
2.7.7.60	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	lm.51784	4-diphosphocytidyl-2C-methyl-D-erythritol synthase [Chlamydomonas reinhardtii]	3.00E-104			
2.7.1.148	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	lm.83522	4-diphosphocytidyl-2c-methyl-d-erythritol kinase [Thalassiosira pseudonana]	1.00E-100			
4.6.1.12	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	dm.50045	putative plastid 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase precursor [Guillardia theta]	5.00E-77			
1.17.7.1	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	lm.27792	hypothetical protein [Volvox carteri]	0.00E+00			
1.17.1.2	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	dm.13897	hypothetical protein [Chlorella variabilis]	0.00E+00			
2.5.1.1	geranyl diphosphate synthase	lm.32576	geranylgeranyl pyrophosphate synthase [Oscillatoria acuminate]	1.00E-122	lm.90469	geranylgeranyl diphosphate synthase [Synechococcus sp.]	2.00E-71
2.5.1.10	farnesyl diphosphate	lm.19578	predicted protein [Phaeodactylum tricornutum]	2.00E-71	dm.26131	hypothetical protein [Phytophthora sojae]	4.00E-79

	synthase						
2.5.1.29	geranylgeranyl diphosphate synthase	dm.48870	predicted protein [Nematostella vectensis]	9.00E-45	dm.10389	predicted protein [Physcomitrella patens]	1.00E-143
2.5.1.32	Phytoene synthase	lm.32898	predicted protein [Thalassiosira pseudonana]	3.00E-108			
2.5.1.21	Squalene synthase	lm.22898	farnesyltransferase [Trypanosoma cruzi]	1.00E-112			
5.4.9.88	cycloartenol synthase	dm.6181	cycloartenol synthase;-2,3-epoxysqualene mutase-like protein [Thalassiosira pseudonana]	0			
App.9.2 Vitamin E biosynthesis							
1.13.11.27	4-hydroxyphenyl pyruvate dioxygenase	lm.13577	4-hydroxyphenylpyruvate dioxygenase [Oxytricha trifallax]	3.00E-157	dm.13964	4-hydroxyphenylpyruvic acid dioxgenase [Capsaspora owczarzaki]	1.00E-155
2.5.1.-	homogenitisate phytyltransferase	lm.60507	hypothetical protein [Selaginella moellendorffii]	8.00E-122	lm.48543	predicted protein [Phaeodactylum tricornutum]	3.00E-91
2.1.1.-	tocopherol cyclase	dm.38942	predicted protein [Thalassiosira pseudonana]	9.00E-64			
2.1.1.95	tocopherol O-methyltransferase	lm.59814	hypothetical protein [Selaginella moellendorffii]	7.00E-86	dm.3700	hypothetical protein [Volvox carteri]	1.00E-153

App.9.3 Lysine biosynthesis							
4.1.3.21	homocitrate synthase	dm.16450	2-isopropylmalate synthase [Phytophthora infestans]	0.00E+00	lm.48249	2-isopropylmalate synthase/homocitrate synthase [Phycisphaera mikurensis]	2.00E-178
4.2.1.26	homoaconitase	lm.36101	homoaconitase [Melioribacter roseus]	0.00E+00	lm.11199	Isopropylmalate dehydratase [Guillardia theta]	0.00E+00
1.1.1.87	homoisocitrate dehydrogenase	lm.15833	isocitrate/isopropylmalate dehydrogenase [Aciduliprofundum boonei]	8.00E-170	lm.80792	isocitrate/isopropylmalate dehydrogenase family protein [Caldilinea aerophila]	1.00E-90
2.6.1.39	aminoadipate aminotransferase	dm.43146	kynurenine/alpha-aminoadipate aminotransferase [acanthomeba castellani]	6.00E-144	lm.82932	aminotransferase [Deinococcus radiodurans]	4.00E-87
1.2.1.31	aminoadipate reductase	lm.21957	aminoadipate-semialdehyde dehydrogenase [Capsaspora owczarzaki]	0	lm.23118	protein synthetase, putative [Acanthamoeba castellanii]	0.00E+00
1.5.1.10	saccharopine dehydrogenase	lm.12750	saccharopine dehydrogenase [Acanthamoeba castellanii]	1.00E-133			
1.5.1.17	saccharopine dehydrogenase	lm.14473	lysine ketoglutarate reductase [Caldithrix abyssi]				
App.9.4 Arginine and proline biosynthesis							
2.7.2.11	Glutamate kinase	lm.12713	predicted protein [Nematostella vectensis]	3.00E-71	dm.21434	glutamate 5-kinase [Halobacteroides halobius]	7.00E-75
1.2.1.41	Gamma-glutamyl phosphate reductase	lm.12713	predicted protein [Nematostella vectensis]	3.00E-71	dm.21435	glutamate 5-kinase [Halobacteroides halobius]	7.00E-75
2.6.1.13	ornithine-Δ-	dm.2778	Ornithine aminotransferase,	0	lm.13761	predicted protein [Naegleria	4.00E-

	aminotransferase		mitochondrial [Dicentrarchus labrax]			gruberi]	159
2.1.3.3	ornithine carbamoyltransferase	lm.57420	predicted protein [Physcomitrella patens]	9.00E-108	dm.3271	ornithine carbamoyltransferase family protein [Trichomonas vaginalis]	3.00E-134
6.3.4.5	argininosuccinate synthase	dm.3269	Argininosuccinate synthase [Plesiocystis pacifica]	0			
4.3.2.1	argininosuccinate lyase	dm.26272	argininosuccinate lyase [Thermobaculum terrenum]	1.00E-119			
3.5.3.6	arginine deiminase	lm.85339	probable arginine deiminase [Cyanidioschyzon merolae]	7.00E-93			
4.3.1.12	ornithine cyclodeaminase	dm.78705	ornithine cyclodeaminase [Candidatus Korarchaeum cryptofilum]	5.00E-100			
1.5.1.2	pyrroline-5-carboxylate reductase	dm.52519	pyrroline-5-carboxylate reductase [Anaerolinea thermophile]	4.00E-94			
1.5.99.8	proline dehydrogenase	dm.3408	proline oxidase [Capsaspora owczarzaki]	5.00E-128	lm.30651	proline oxidase, putative [Phytophthora infestans]	2.00E-59
App.9.5 Threonine biosynthesis							
2.7.2.4	aspartokinase	lm.48603	aspartate kinase [Fibrisoma limi]	7.00E-173			
1.2.1.11	aspartate-semialdehyde dehydrogenase	lm.25814	aspartate-semialdehyde dehydrogenase [Bermanella marisrubri]	2.00E-175			
1.1.1.3	homoserine dehydrogenase	lm.48604	aspartate kinase [Fibrisoma limi]	7.00E-173	lm.20768	hypothetical protein [Selaginella moellendorffii]	2.00E-104

2.7.1.39	homoserine kinase	none					
4.2.3.1	threonine synthase	lm.36284	threonine synthase [Geobacter metallireducens]	2.00E-172			
App.9.6 Branched-chain amino acid biosynthesis							
4.3.1.19	Threonine dehydratase	lm.62724	threonine dehydratase[Batrachochytrium dendrobatidis]	7.00E-164	lm.15790	threonine dehydratase, medium form [Polaromonas sp.CF318]	0.00E+00
2.2.1.6	acetolactate synthase I/III	lm.11204	pyruvate decarboxylase, putative [Perkinsus marinus]	0.00E+00			
1.1.1.86	ketol-acid reductoisomerase	dm.442	ketol-acid reductoisomerase [Gemmata obscuriglobus]	2.00E-164	lm.36809	ketol-acid reductoisomerase [SAR324 cluster bacterium]	4.00E-54
4.2.1.9	dihydroxy-acid dehydratase	lm.22604	dihydroxy-acid dehydratase [Beggiatoa alba]	0.00E+00	lm.48541	dihydroxy-acid dehydratase 2-like [Nematostella vectensis]	0.00E+00
2.6.1.42	Branched chain amino acid amino transferase	lm.13579	Glu/Leu/Phe/Val dehydrogenase family protein [Microscilla marina]	0.00E+00			
2.3.3.13	2-isopropylmalate synthase	lm.48249	2-isopropylmalate synthase/homocitrate synthase [Phycisphaera mikurensis]	2.00E-178	dark_m.16450	hypothetical protein [Phytophthora sojae]	0
4.2.1.33	3-isopropylmalate dehydratase	lm.11199	Isopropylmalate dehydratase [Guillardia theta]	0	lm.36101	homoaconitase [Melioribacter roseus]	0
1.1.1.85	3-isopropylmalate dehydrogenase	lm.13576	3-isopropylmalate dehydrogenase [Planctomyces limnophilus]	2.00E-136	lm.15833	isocitrate/isopropylmalate dehydrogenase [Thermoplasmatales archaeon]	3.00E-174

App.9.7 Aromatic amino acid biosynthesis							
2.5.1.54	3-deoxy-7-phosphoheptulonate synthase	lm.61463	probable 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase isoenzyme [Ustilago hordei]	8.00E-152	lm.32049	phospho-2-dehydro-3-deoxyheptonate aldolase [Dickeya zea]	6.00E-149
4.2.3.4	3-dehydroquinate synthase	lm.52098	hypothetical protein [Batrachochytrium dendrobatidis]	0.00E+00			
4.2.1.10	3-dehydroquinate dehydratase	lm.52099	hypothetical protein [Batrachochytrium dendrobatidis]	0.00E+00			
1.1.1.25	Shikimate dehydrogenase	lm.52100	hypothetical protein [Batrachochytrium dendrobatidis JAM81]	0.00E+00	lm.97034	shikimate 5-dehydrogenase [Achromobacter xylosoxidans]	2.00E-70
2.7.1.71	Shikimate kinase	lm.52101	hypothetical protein [Batrachochytrium dendrobatidis]	0.00E+00	lm.64084	Shikimate kinase (ISS) [Ostreococcus tauri]	1.00E-24
2.5.1.19	EPSP synthase	lm.52102	hypothetical protein [Batrachochytrium dendrobatidis]	0.00E+00			
4.2.3.5	Chorismate synthase	dm.12955	chorismate synthase [Desulfotalea psychrophila]	2.00E-151			
4.1.3.27	anthralinate synthase	lm.68214	anthranilate/para-aminobenzoate synthase component I [uncultured euryarchaeote]	2.00E-136	dm.24194	phenazine biosynthesis protein PhzE [Streptomyces sviceus]	8.00E-139
5.4.99.5	chorismate mutase	lm.67331	hypothetical protein [Phytophthora sojae]	6.00E-98	lm.43560	tyrosine biosynthesis bifunctional enzyme [Phytophthora sojae]	1.00E-161
4.2.1.51	prephenate dehydratase	lm.67331	hypothetical protein [Phytophthora sojae]	6.00E-98	lm.43561	tyrosine biosynthesis bifunctional enzyme	1.00E-161

						[Phytophthora sojae]	
1.3.1.12	prephenate dehydrogenase	lm.67331	hypothetical protein [Phytophthora sojae]	6.00E-98			
2.6.1.57	aromatic amino acid transaminase	lm.50263	aromatic amino acid aminotransferase [Rubrivivax benzoatilyticus]	7.00E-19			
2.6.1.85	aminodeoxychorismate synthetase	dm.24194	phenazine biosynthesis protein PhzE [Streptomyces sviceus]	8.00E-139			
4.1.3.38	aminodeoxychorismate lyase	lm.32069	hypothetical protein [Selaginella moellendorffii]	7.00E-167			
2.4.2.18	Anthralinate phosphoribosyltransferase	dm.33057	tryptophan synthase subunit beta [Planococcus antarcticus]	2.00E-163			
5.3.1.24	Phosphoribosylanthralinate isomerase	dm.33057	tryptophan synthase subunit beta [Planococcus antarcticus]	2.00E-163			
4.1.1.48	Indoyl-3-glycerol phosphate synthase	dm.33057	tryptophan synthase subunit beta [Planococcus antarcticus DSM 14505]	2.00E-163			
4.2.1.20	Tryptophan synthase B subunit	dm.33057	tryptophan synthase subunit beta [Planococcus antarcticus]	2.00E-163			
4.2.1.20	Tryptophan synthase A subunit	lm.34856	Tryptophan synthase (alpha / beta chains) [Ectocarpus siliculosus]	8.00E-42			

App.9.8 Tetrapyrrole biosynthesis							
2.3.1.37	5-aminolevulinate synthase	lm.11190	5-aminolevulinate synthase [Oceanibaculum indicum]	7.00E-160			
1.2.1.7	glutamyl-tRNA reductase	lm.12397	Glutamyl tRNA reductase [Volvox carteri]	7.00E-175			
5.4.3.8	glutamate-1-semialdehyde 2,1-aminomutase	dm.2724	glutamate-1-semialdehyde aminotransferase/glutamate-1-semialdehyde 21-aminomutase [Nannochloropsis gaditana]	0.00E+00	lm.7162	predicted protein [Micromonas sp.]	3.00E-174
4.2.1.24	aminolevulinate dehydratase	lm.37481	MGC84775 protein-like [Saccoglossus kowalevskii]	5.00E-122	dm.7076	hypothetical protein Thalassiosira oceanica]	3.00E-153
2.5.1.61	hydroxymethylbilane synthase	lm.59777	hydroxymethylbilane synthase [Oscillatoria nigro-viridis]	6.00E-95			
App.9.9 Ascorbate biosynthesis							
1.1.1.22	UDP-Glc dehydrogenase	lm.8509	UDP-glucose dehydrogenase [Capsaspora owczarzak]	0.00E+00			
5.1.3.2	UDP-Glc 4-epimerase	dm.42120	UDP-glucose 4-epimerase [Nannochloropsis gaditana]	3.00E-144	dm.18834	UDP-glucose 4-epimerase [Nannochloropsis gaditana]	4.00E-149
2.7.7.64	UDP-sugar pyrophosphorylase	lm.84866	predicted protein [Micromonas sp.]	8.00E-163	lm.77285	predicted protein [Micromonas sp.]	1.00E-148
2.7.1.43	glucuronokinase	lm.74877	predicted protein [Nematostella vectensis]	6.00E-160			
1.1.1.19	GlcUA reductase	lm.70407	Mannitol dehydrogenase rossman domain family [Ectocarpus siliculosus]	1.00E-140	dm.24690	mannitol-1-phosphate 5-dehydrogenase [Providencia sneebia]	2.00E-123

3.1.1.17	aldonolactonase	dm.27330	Gluconolactonase [Crocosphaera watsonii]	8.00E-48	Already known		
1.1.3.8/1.3.2.3	L-gulonolactone oxidase	dm.53576	FAD/FMN-containing dehydrogenase [Moorea producens]	2.00E-13	lm.29571	Putative actin interacting protein [Polysphondylium pallidum]	0.00E+00
2.7.7.69	GDP-L-galactose phosphorylase	lm.98124	Hypothetical protein, partial [Rhipicephalus pulchellus]	3.00E-62			
3.1.3.-	Aldoase	lm.49313	hypothetical protein [Trichoplax adhaerens]	1.00E-60	lm.96714	inositol monophosphatase [Burkholderia ambifaria]	4.00E-74
1.1.1.122	D-threo-aldose 1-dehydrogenase	dm.6726	aldo/keto reductase [Alicyclobacillus acidocaldarius subsp. acidocaldarius]	5.00E-75	lm.28331	predicted protein [Micromonas sp.]	7.00E-81
1.3.2.3	L-galactono-1,4-lactone dehydrogenase	dm.53576	FAD/FMN-containing dehydrogenase [Moorea producens]	2.00E-13			
5.1.3.18	GDP-D-mannose 3', 5'-epimerase	lm.30913	NAD-dependent epimerase/dehydratase [Planococcus donghaensis]	1.00E-146	lm.21407	NAD-dependent epimerase/dehydratase [Pedosphaera parvula]	1.00E-121
2.7.1.6	galactokinase	lm.57669	hypothetical protein [Monosiga brevicollis]	5.00E-158			
1.1.1.-	Galuronic acid reductase	lm.15162	malate dehydrogenase, mitochondrial precursor [Phytophthora infestans]	2.00E-147	lm.86647	malate dehydrogenase, mitochondrial precursor [Phytophthora infestans]	5.00E-41
2.7.7.10	UDP-hexose pyrophosphatase	lm.20758	galactose-1-phosphate uridylyltransferase [Planctomycete limnophilus]	1.00E-36	light only		

App.9.10 Trypanothione biosynthesis							
4.1.1.18	Lysine decarboxylase	lm.45128	decarboxylase [Microlunatus phosphovorus]	0	lm.75103	putative arginine/lysine/ornithine decarboxylase [Sulfitobacter sp.]	2.00E-166
4.1.1.17	Ornithine decarboxylase	none					
2.6.1.76	Diaminobutyrate-2-oxoglutarate transaminase	lm.13761	predicted protein [Naegleria gruberi]	4.00E-159	lm.27919	aminotransferase class-III [Dictyostelium purpureum]	7.00E-166
4.1.1.86	L-2,4-diaminobutyrate decarboxylase	lm.51975	cysteine synthetase/pyridoxal dependent decarboxylase [Micromonas pusilla]	1.00E-140			
2.5.1.16	Spermidine synthase	lm.8151	spermidine synthase [Rhizopus delemar]	8.00E-106	dm.16659	predicted protein [Micromonas sp.]	7.00E-99
6.3.2.3	Glutathione synthase	lm.34786	hypothetical protein [Dictyostelium purpureum]	3.00E-97			
6.3.2.2	Glutathamate cysteine ligase	lm.43917	glutamate-cysteine ligase, catalytic subunit [Taeniopygia guttata]	0.00E+00			
4.1.1.50	S-adenosylmethionine decarboxylase	dm.44803	S-adenosylmethionine decarboxylase [Leishmania donovani]	3.00E-91			
6.3.1.9	Glutathionylspermidine synthase	lm.36171	bifunctional glutathionylspermidine amidase/glutathionylspermidine synthetase [Simiduia agarivorans]	4.00E-98	lm.20283	hypothetical protein [Plasmodiophora brassicae]	3.00E-48
3.5.1.78	Glutathionylspermidine	lm.36172	bifunctional glutathionylspermidine	1.00E+00	lm.20284	hypothetical protein [Plasmodiophora brassicae]	3.00E-48

	amidase		amidase/glutathionylspermidine synthetase [Simiduia agarivorans]				
1.8.1.12	Trypanothione reductase	dm.46314	trypanothione reductase, putative [Trypanosoma cruzi marinkellei]	0.00E+00			
App.9.11 Ovothiol biosynthesis							
NA	5-histidylcysteine sulfoxide synthase	dm.8268	generic methyltransferase [Salpingoeca sp.]	0.00E+00	lm.33343	hypothetical protein [Rhizopus delemar]	6.00E-171
NA	5-histidylcysteine sulfoxide N-mehtyl transferase	dm.8269	generic methyltransferase [Salpingoeca sp.]	0.00E+00	lm.51716	hypothetical protein [Trichoplax adhaerens]	2.00E-89

Appendix 10 – Proteomic analysis of dm.33057

The protein encoded on the *Euglena* transcript dm.33057 was co-purified with the laminarin phosphorylase activity and identified by proteomic analysis. This transcript encodes four domains, which are normally found on separate polypetides, for the biosynthesis of tryptophan. Peptides were found that match 31% of its sequence, evenly spread throughout, except in the region before the first domain. This sample had been purified by AIEX, HIC and has a mass in the range 100-150 kDa on denaturing PAGE indicating that all four domains are translated as a single polypeptide (Fig App.10.1).

1	GALAPPLATL	SGGPAGMGLV	AFDGAAAVVT	DLTADRQALM	AGINALAVGD
51	GGPTHLAAAL	DAGVAAFGDS	EPRRARFLEV	LTDRLGDDGR	WALERAAAHK
101	AAGVVVLVVG	VGPAPEAGPL	QVLASDPAEE	HWMRLGALPA	VPAVADIVGS
151	GGDGQNTWNV	STPAAIVAAG	AGIRMAKHGN	RSASSNSGSA	DLLEQFGAYL
201	PVSPTVVPEL	VDRTNFSFLY	APAFHPALRN	VAKVRQDLGT	KTVFNFLGPL
251	LNPAAPQYNV	VGVSDRAMAE	VMARCLSRRP	GVQALVVHSE	DGMDKISPVR
301	NTHSWRVEGG	KEPVYQLLRP	EDFGLVSDPA	KDCGAGGGSP	AHNAAALLRV
351	LQGTSPDPLL	ADFVLLQAAA	LAHLAGKAAS	YAEGMALMRE	TVASGRALAA
401	LNEYVRLSTR	LSRDSNRAIL	QTTVRRRQVD	VALAQQLVPE	AELRALPHFS
451	RPCVDLAAAL	KGSTTPVAVM	AEIKRASPSE	GDINPSLDVA	AIAGQYTEGG
501	AVALSVLTEP	TWFKGSLQDL	TTAREVVEKA	GAQRPAVLRK	EFVFTKYQIT
551	EAKAHGADTV	LLIAAIEPAM	REAGESLQSL	LDYSRSLGME	PLVEVVTEDE
601	VDAAIGCGAK	VIGINNRDLT	TFRVDLGRTE	TLTRYAEFAY	PEKAGEVVWV
651	SLSGVNDNAD	IHILQKLSPR	VRTVLVGTSL	MRAADPAALL	RTFTSGVVSP
701	ADLAKIKAPP	KARPLVKICG	LKTPEQVISA	LRDGVDFVGL	MFYPKSHRYI
751	ADPATARRMV	ELTHQCALPG	HPVWGQLPLA	VGVFMQESLE	RINELSRAVG
801	FDAVQLYGYR	PGDIDLAQLK	KADGSPLPVI	WACSVTTAED	VAAVAYPPGI
851	AALCLDRKAG	GALGGTGQAF	DWGLLKGFAA	PVPVWLAGGI	DEGNAPQAAE
901	VPGIATIDLS	SGVETDKVKD	FWKIRTVTRL	IKNRLPAYFG	PFGGQFVPEI
951	LMPALIDLEK	QYLRVFRNPD	FWREVKWYWA	NYAGRPTLLY	KATKQTEDIR
1001	KTCAPGMGAQ	IWYKREDLLH	TGAHKVNNAL	GQALMAKWLG	KRRMIAETGA
1051	GQHGVATATV	CALFGLECIV	YMGALDTVRQ	KLNVLRMEAL	GAKVVAVEAG
1101	QKTLNDAVNE	ALRDWATTSH	YSHYMIGSCV	GPHPFPTMVR	DFQSIISREA
1151	RVQCLEQLGR	LPDKVFACTG	GGSNAIGTFD	AFVQDAGVQI	FGVEAGGDGH
1201	DLNSATLSKG	SIGFVHGART	YLLQDEESGQ	VKPTHSVSAG	LDYPGVGPEH
1251	SYLKSMGRVQ	YINASDKAAL	DAFQHLARTE	GILAALESSH	AVAATIEASR
1301	NSPQEMNLVC	CMSGRGDKDM	NTIEEWLRGD	GVPHH	

Figure App.10.1: Proteomic analysis of dm.33057

The peptides identified with mascot (red) matched throughout the sequence, including the TrpD (amino acids 130-408, yellow), TrpC (419-693, green), TrpF (713-933, cyan) and TrpB (935-1332, grey) domains.

Appendix 11 – Natural product synthases encoded in the *Euglena* transcriptome

Transcripts were identified using BLASTP to search for keto synthase (for PKSs), condensation, and adenylation (for NRPSs) domains. There is a high degree of uncertainty using this technique.

A – Amino acid adenylation. AA-syn – Amino acid synthesis. ANK – Ankyrin domain. AT – Acyl transfer. C – Condensation. CoAL – Acyl-CoA ligase. DH – Dehydratase. AmT – Aspartate amino transferase. EH – Enoyl CoA hydratase ER – Enoyl reductase. HCS – HMGCoA synthase. KR – keto reductase. KS – Polyketide synthase. TE – thioesterase.

App.11.1 Polyketide synthases

	Transcript	Role	Domains
1	lm.8157	PKS	KS-AT-DH-ER-KR-ACP-EH-EH-TE-HCS
2	lm.60697	PKS	DH-KS-ACP-AmT
3	lm.53854	PKS	KS-AT-KR-ACP-KS
4	lm.82030	PKS	KR-ACP-KS
5	lm.42557	PKS	KR-ACP-KS-DH-KR
6	lm.91532	PKS	A-ACP-KS
7	lm.95952	PKS	ACP-KS
8	lm.88225	PKS	KS-AT
9	lm.23151	PKS	KS
10	lm.94376	PKS	C-ACP
11	lm.88941	PKS	KS
12	lm.110121	PKS	KS
13	lm.102218	PKS	KS
14	lm.97081	PKS	ACP-KS

App.11.2 Nonribosomal peptide synthases

	Transcript	Activity	
1	lm.66007	NRPS	C-A-ACP
2	lm.9669	NRPS	C-C-A-ACP-TE
3	lm.32232	NRPS	C-C-A-ACP-C-A
4	lm.96272	NRPS	C-A-A
5	lm.21957	AA-syn	C-A-ACP-TE
6	lm.23118	AA-syn	A-ACP-TE
7	lm.78138	A	A-ANK
8	lm.54590	A	A-ANK
9	lm.87820	A	A-ACP-TE
10	lm.77877	A	A-ACP-ANK
11	lm.98982	A	A
12	lm.89785	A	A-ACP
13	lm.47668	A	A
14	lm.94698	A	A
15	lm.44795	A	A
16	lm.44327	A	CoAL
17	lm.11010	A	CoAL
18	lm.81072	A	CoAL
19	lm.97175	C	CoAL
20	lm.26470	C	CoAL
21	lm.22119	C	CoAL
22	lm.37810	C	CoAL
23	lm.3119	C	CoAL
24	lm.28443	C	CoAL
25	lm.9400	C	CoAL
26	lm.26875	C	CoAL
27	lm.12406	C	CoAL
28	lm.3346	C	CoAL
29	lm.44112	C	CoAL
30	lm.17087	C	CoAL

References

1 Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-D238 (2009).

2 Curtis, B. A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59-65 (2012).

3 Bhattacharya, D. *et al.* Genome of the red alga *Porphyridium purpureum*. *Nat. Commun.* **4**, 1941-1941 (2013).

4 Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201 (2006).

5 Takahashi, T. *et al.* A sequence motif involved in the donor substrate binding by α-1,6-fucosyltransferase: the role of the conserved arginine residues. *Glycobiology* **10**, 503-510 (2000).

6 Lobsanov, Y. D. *et al.* Structure of Kre2p/Mnt1p: a yeast α-1,2-mannosyltransferase involved in mannoprotein biosynthesis. *J. Biol. Chem.* **279**, 17921-17931 (2004).

7 Mulichak, A. M., Lu, W., Losey, H. C., Walsh, C. T. & Garavito, R. M. Crystal atructure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: interactions with acceptor and nucleotide ligands. *Biochemistry* **43**, 5170-5180 (2004).

8 Cui, Z., Maruyama, Y., Mikami, B., Hashimoto, W. & Murata, K. Crystal structure of glycoside hydrolase family 78 α-L-rhamnosidase from *Bacillus* sp. GL1. *J. Mol. Biol.* **374**, 384-398 (2007).