8-27-2014

Substrate Adaptability of Periplasmic Substrate-binding Proteins as a Function of their Evolutionary Histories

Nathalie Boucher
University of Connecticut - Storrs, nathalie.bou@gmail.com

Substrate Adaptability of Periplasmic Substrate-binding Proteins as a Function of their Evolutionary Histories

Nathalie Boucher, Ph.D.

University of Connecticut, 2014

The ATP-binding cassette (ABC) transporters play an important role in the uptake of nutrients in bacterial and archaeal cells. The bacterium *Thermotoga maritima* has an unusually high number of ABC transporters. However, only a few of them have been characterized due to the variety of substrates they can transport and the time consuming efforts needed to thoroughly characterize each ABC transporter. In this work, a technique has been optimized to determine the sugars that interact with the substrate-binding proteins (SBP) of the ABC transporter complexes in *Thermotoga maritima*. Using this high throughput assay, the ligands of trehalose-binding protein (TreE) and xylose-binding protein (XylE2) in *Thermotoga maritima* as well as representatives of the mannoside-binding proteins (Man) in the *Thermotoga* species were identified. The phylogeny of the mannose-binding proteins is interesting because of the presence of two paralogous proteins in this family (ManD and ManE). Ten representatives of the mannoside-binding proteins (ManD and ManE) encoded in the Thermotogales were characterized. At 37°C and 60°C, the ManD and ManE homologs bind cellobiose, cellotriose, cellotetraose, β-mannotriose, and β-mannotetraose. However, the ManE homologs have additional function, as they are able to bind β-mannobiose, laminaribiose, laminaritriose and sophorose.

Nathalie Boucher – University of Connecticut, 2014

An examination of the selective pressure that acted on *manD* and *manE* and the identification of the residues located in the binding-site of their encoded proteins suggest that early in the evolution of *manD* and *manE,* these genes had different codon sites under positive selection which encode important regions of their binding site. Taken together, these analysis suggest that these paralogs likely evolved under different evolutionary constrains.

Substrate Adaptability of Periplasmic Substrate-binding Proteins as a Function of their Evolutionary Histories

Nathalie Boucher

B.S. Laval University, Quebec, Canada, 2000

M.Sc. Laval University, Quebec, Canada, 2002

Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

At the

University of Connecticut

2014

2014

APPROVAL PAGE

Doctor of Philosophy Dissertation

Substrate Adaptability of Periplasmic Substrate-binding Proteins as a Function of their Evolutionary Histories

Presented by

Nathalie Boucher, B.S., M.Sc.

Major Advisor_____

Kenneth M. Noll

Associate Advisor_____

Daniel J. Gage

Associate Advisor_____

J. Peter Gogarten

University of Connecticut

2014

Acknowledgements

I would like to thank my advisor, Dr. Kenneth Noll, for his mentorship during my graduate studies. Dr. Noll welcomed me in his laboratory and helped me find the best project that matched my skills. I greatly appreciate his willingness to incorporate my scientific ideas and I am grateful for all his professional advice.

I thank my labmates, especially Nicholas and Chaman. Together, we had numerous scientific and non-scientific debates that taught me other points of view on different topics. I thank Kunica for all her help while I was in Albany and Kristen for her helpful comments and edits on my thesis. I also thank my past laboratory members for their friendship during those years.

I thank my family, especially Pascal for his unconditional support through my graduate studies. I am grateful to my son, at first, for all those toothless smiles, and later, for all those great smiles that lighted up my days.

I dedicate this thesis to my aunt, who was a mother to me and who saw the potential in me. She encouraged me to achieve my dreams and she taught me to stand up for my ideas.

Table of Contents

Chapter 1 1

General introduction 1

Evolutionary history of the ABC transporters and the SBP 4

The Thermotogales 6

Carbohydrate ABC transporters in *Thermotoga maritima* 7

Carbohydrate uptake: catabolism and regulation 8

The SBP and the substrate translocation 11

Objective statement 14

Chapter overviews 15

Contribution from other researchers 17

Chapter 2 18

Identification of genomic variations of different genomovars of *Thermotoga maritima* 18

Introduction 18

Materials and Methods 21

Strains 21

PCR amplification and DNA sequencing 21

Sequence depositions 23

Prediction software 23

Data acquisition and phylogenetic analysis 23

Results 27

Identification of two genomovars of *Thermotoga maritima* MSB8 27

Sequence of the genomic DNA deleted from the genomovar TIGR 28

Evolution of Mal and Tre transporters 33

Genetic variations of *malF2* 34

Discussion 36

Recent up-dates: sequencing project of *T. maritima* MSB8 genomovar DSM3109 36

Conclusion 39

Chapter 3 41

Development of a screening assay to identify the ligands that interact with thermophilic substrate-binding proteins (SBPs) 41

Introduction 41

DSF assay as a screening method to detect ligand-binding interactions 42

Principle of the DSF assay 43

Materials and Methods 46

Differential scanning fluorimetry (DSF) screening of ligand binding 46

Sugar purity 47

Results 48

Effect of the change of pH and ligand concentration on the unfolding temperature 48

DSF analysis of ligand binding of MalE1 and MalE2 54

Discussion 58

Adaptation of the DSF assay for use with thermophilic SBPs 58

Conclusion 60

Chapter 4 63

Characterization of the ABC-transporters located in the newly sequenced 10 kb region of *Thermotoga maritima* MSB8 genomovar DSM3109 63

Introduction 63

Materials and Methods 65
Cloning, expression and purification of TreE and XylE2 65
Differential scanning fluorimetry (DSF) screening of ligand binding 66
Intrinsic fluorescence spectroscopy 66
Phylogenetic tree for XylE2 67
Results 69
Ligand stabilization of TreE and XylE2 69
Binding affinities of TreE and XylE2 70
Discussion 77
Transcriptional regulation 80
Conclusion 81

Chapter 5 83
Reexamination of the binding properties of the SBPs encoded by TM0595, TM1150, TM1199 (*lptE*) and TM0418 (*inoE*). 83
Introduction 83
Materials and Methods 85
Cloning, expression and purification 85
Intrinsic fluorescence spectroscopy of InoE 87
Growth of *T. maritima* on myo-inositol-1-phosphate 87
Results 89
InoE 89
LptE (TM1199) 97
TM1150 and TM0595 97
Discussion 100

InoE: Inositol binding protein 100
LptE: Galactose and lactose-binding protein 101
TM1150 and TM0595 104
Conclusion 106

Chapter 6 109
Characterization of the mannoside-binding proteins in the Thermotogales 109
Introduction 109
Materials and Methods 111
Strains 111
Cloning 111
Protein expression and purification 112
Differential scanning fluorimetry to measure ligand-protein interactions and protein thermostabilities 115
Sugar purity 115
Intrinsic fluorescence spectroscopy 116
Results 118
Synteny 118
The thermostabilities of ManE and ManD are consistent with the OGTs of their hosts 121
The ManE and ManD orthologs interact with sugars composed of β-D-glucose and β-D-mannose 123
ManE orthologs bind similar sugars 130
The ManD orthologs do not bind laminaribiose, laminaritriose and sophorose 137
Discussion 139
Conclusion 142

Chapter 7 144

Substrate adaptability of mannoside-binding proteins as a function of their evolutionary histories 144

Introduction 144

Material and Methods 146

Data acquisition and phylogenetic analysis 146

Crystal structure superposition and structural alignments 146

dN/dS ratio: branch-site model analysis 147

Results 149

Phylogenetic analysis 149

Detection of the codon sites under neutral evolution and positive selection 151

Residues involved in the change of function 154

Discussion 159

Why are ManD orthologs still maintained? 164

Conclusion 166

Appendix 1: List of genomic variations between *T. maritima* MSB8 genomovars TIGR, DSM3109 and ATCC. 171

Appendix 2: Details of the differential scanning fluorimetry (DSF) protocol. 175

Appendix 3: Thermostabilities of MalE1, MalE2, TreE and XylE2 measured by DSF and represented by ΔTm values. 176

Appendix 4: Glycosidic bonds of disaccharides and oligosaccharides 178

Appendix 5: Sequence alignment using $BglE_{Tmar}$ (pdb:3i5o), $ManE_{Tmar}$ (pdb: 1vr5) and $ManD_{Tmar}$ from structures superposition using CHIMERA 179

Appendix 6: *manD* and *manE* codon alignment for the PAML analysis 181

Appendix 7: Newick trees with each *manD*/*manE* branch flagged for PAML dN/dS ratio analysis..... 195

Appendix 8: PAML control file for *manD* under hypothesis testing branch-model H_0 196

Appendix 9: PAML control file for *manD* under hypothesis testing branch-model H_1 198

Appendix 10: PAML control file for *manE* under hypothesis testing branch-model H_0 200

Appendix 11: PAML control file for *manE* under hypothesis testing branch-model H_1 202

Appendix 12: PAML branch-site model output: the *manD* branch set as foreground . 204

Appendix 13: PAML branch-site model output: the *manE* branch set as foreground.. 219

References..... 234

List of Figures

Figure 1. Operon organization of different carbohydrate ABC transporters. 3
Figure 2. PCR amplifications using primers to the loci TM1848 and TM1847.. 29
Figure 3. Organization of the ORFs between the TM1848 and TM1847 orthologs in *Thermotoga* species.. 30
Figure 4. An unrooted Mal3-type ABC transporters tree 31
Figure 5. Schematic representation of the *mal2* operon as described in the 1999 annotation of the *T. maritima* MSB8 genome showing genetic variations found in this study in the two genomovars. 35
Figure 6. Illustration of the steps involved in the differential scanning fluorimetry (DSF) assay. 45
Figure 7. Effect of pH on the unfolding temperature curve of *T. maritima* MalE1 using DSF. 49
Figure 8. Effect of pH on the unfolding temperature (Tm) using DSF 50
Figure 9. Effect of the pH on the ΔTm using DSF on MalE1 and MalE2. 55
Figure 10. ΔTm of *T. maritima* TreE and XylE2 determined by DSF.. 72
Figure 11. Emission spectra of *T. maritima* TreE with and without trehalose. 75
Figure 12. Emission spectra of *T. maritima* TreE. 76
Figure 13. Unrooted maximum likelihood tree of substrate-binding proteins homologous to TM0114 depicting the relationships among the xylose-binding proteins..... 79
Figure 14. Emission spectra of InoE with and without *myo*-inositol 92
Figure 15. Emission spectra of InoE with and without the addition of MI-1-phosphate and a titration curve with MI-1-phosphate..... 93
Figure 16. Galactose catabolism and utilization pathway..... 103
Figure 17. Syntenic regions of the mannoside ABC transporter operon *manEFGKL* and *manD* and in the Thermotogales. 120
Figure 18. Summary of the interactions of the ManE and ManD orthologs with sugars determined by DSF.. 129
Figure 19. Competitive assay of the $ManE_{Mpri}$ and $ManE_{Fnod}$..... 134
Figure 20. Competitive assay of the $ManE_{Tlet}$ 135

Figure 21. Rooted phylogenetic tree depicting the relationships among mannoside-binding proteins and related SBPs.. 150
Figure 22. LIGPLOT of $BglE_{Tmar}$ bound to sugars. 157
Figure 23. Multi-sequence alignment of $ManE_{Tmar}$, $ManD_{Tmar}$ and $BglE_{Tmar}$.. 158
Figure 24. Map of the codon sites under neutral selection and positive selection on $ManE_{Tmar}$ structure superposition 162
Figure 25. Map of the codon sites under neutral selection and positive selection on $ManD_{Tmar}$ structure superposition 163

List of Tables

Table 1. Primers used for sequencing the 10 kb region by primer walking..... 25
Table 2. Effect of ligand concentration on ΔTm of MalE1 using DSF..... 53
Table 3. Effect of ligand concentration on ΔTm of MalE2 using DSF..... 53
Table 4. Summary of the binding properties and ligand induced thermostabilities of MalE1 and MalE2..... 56
Table 5. ΔTm values of TreE using a titration with trehalose and maltose..... 73
Table 6. ΔTm values of XylE2 using a titration with glucose and xylose..... 73
Table 7. Apparent binding affinities (K_d) of TreE and XylE2 measured at 60°C..... 74
Table 8. Summary of the proteins and loci discussed in Chapter 5. 86
Table 9. Thermostabilities InoE and the SBPs encoded by TM1150 and TM0595 measured by DSF. 90
Table 10. Apparent binding affinities (K_d values) of InoE measured at 20°C and 60°C. 94
Table 11. Growth of *T. maritima* grown in defined media with different carbon sources. 96
Table 12. Thermostabilities LptE measured by DSF and represented by ΔTm values... 98
Table 13. Designations of the ManE and ManD orthologs used in this study. 113
Table 14. Primers used to amplify the ManD- and ManE-encoding genes. 114
Table 15. Thermal stabilities of recombinant ManD and ManE proteins determined by differential scanning fluorimetry (DSF) 122
Table 16. Thermostabilities of ManE orthologs measured by DSF..... 125
Table 17. Thermostabilities of ManD orthologs measured by DSF 127
Table 18. Summary of the binding properties and apparent binding constants (K_d) for the ManD and ManE orthologs..... 131
Table 19. Maximal fluorescence change of the ManD and ManE orthologs after addition of gentiobiose (10 μM, Gen) or Konjac glucomannan (0.12 mg, GM). 136
Table 20. Competition assay of the ManD orthologs..... 138
Table 21. List of the codon sites for class 1 and 2a using the branch-site model calculated by Bayes Empirical Bayes (BEB) in PAML 153

Chapter 1

General introduction

Throughout all three domains of life, cells are protected from their environment by a cell membrane. Transporters are essential to facilitate the movement of charged ions and molecules across the cell membrane. They are involved in the uptake of nutrients, export of toxic compounds, formation of ion gradients and excretion of toxins. For over a decade, efforts have been made to categorize all transporters in prokaryotes and eukaryotes. Among all classes of transporters, the ATP-binding cassette (ABC) transporter system is one of the largest superfamilies, containing more than eighty distinct families (1). In prokaryotes, these systems are composed of three components: a substrate-binding protein (SBP), two transmembrane proteins (TMP) and two ATP-binding proteins (ABP). The SBP binds a substrate (ligand), which is translocated by the transmembrane proteins, while the ATP-binding proteins provide the energy-coupled reaction required for the translocation. Although ABC transporter systems are found in all domains of life, the SBPs are only present in Archaea and Bacteria.

The ABC transporters are characterized by their high binding affinities for specific substrates and can bind a wide range of substrates, such as carbohydrates, amino acids, oligopeptides, metals and other compounds. Among the ABC transporter systems, three families transport carbohydrates: two carbohydrate uptake transporters (CUT1 and CUT2) and the peptide/opine/nickel uptake transporter (PepT). In general, the CUT1,

CUT2 and PepT families differ by the type of substrates that they transport and by the organization of their operons (Figure 1). The TMP and the ABP can be homodimeric or heterodimeric (Figure 1) (2–4). The members of the CUT1 family can transport disaccharides, oligosaccharides, glycerol-phosphate, and polyols, while the members of the CUT2 family generally transport monosaccharides (5). In contrast, members of the PepT family primarily transport oligopeptides and metals, although some members were found to transport carbohydrates as well (6–8).

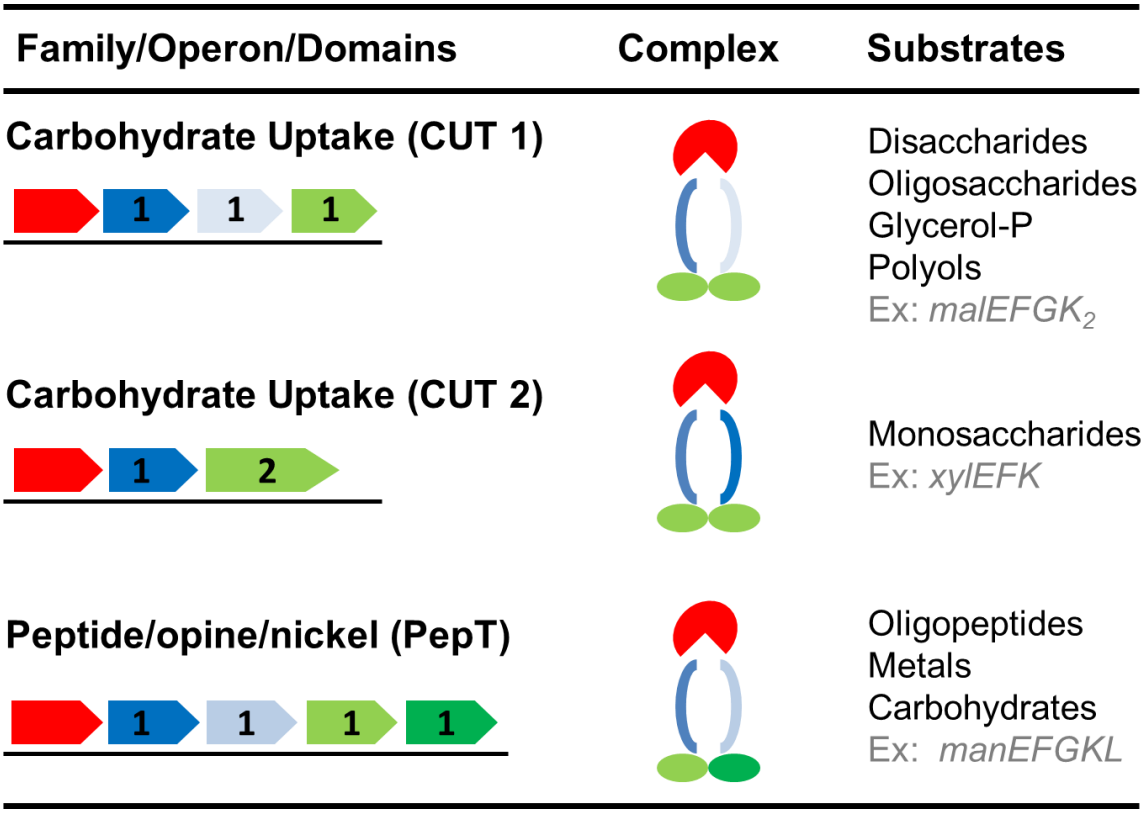


Figure 1. Operon organization of different carbohydrate ABC transporters. The ABC transporter genes and protein components are representated as follows: SBP (red), TMP (light and dark blue) and ABP (light and dark green). The number in the gene box indicates the number of domain encoded by the corresponding gene.

Evolutionary history of the ABC transporters and the SBP

The classification of transporters is primarily established using sequence homologies and phylogenies of the permeases, the proteins that contain the transmembrane domain (1, 5). Since these proteins are present in all transporters, they are well suited for classification purposes. Most evolutionary studies of ABC transporters were performed more than a decade ago, at a time when very few sequences were available in the databases. Phylogenetic analyses that include all ABC transporter systems, independent of their functions, suggest that they diverged early into three different classes before the separation of the three domains of life (9). Members of Class 1 transporters are mainly exporters with fused ABPs. Class 2 transporters are involved in functions other than transport. Class 3 transporters are mainly importers. The bacterial and archaeal ABC importers or substrate binding protein-dependent transporter transporters belong to Class 3.

As previously mentioned, the ABC transporters are composed of at least two components (ABP and TMP) while the ABC transporter importers have an additional subunit, the SBP. Some phylogenetic studies of the operons, as well as the genes that encode each component, have been published. The sequences of the ABP are the most conserved (10, 11) while the sequences of the TMP and the SBP are more divergent (3, 12). Usually the TMP, ABP and the SBP within a given transporter have the same evolutionary history (3, 11). However, some exceptions were noted in *T. maritima* and *T. litoralis* (13–16). One of these exeption is the putative ABC transporter encoded by TM1149-TM1153 in *T. maritima* (discussed in Chapter 5) which was suggested that one or more of its components have been replaced during its evolution (15). Since the SBP,

TMP and ABP-encoding genes tend to be organized in the same order within operons (2–4) and these genes have usually the same evolutionary history, it was proposed that the importers evolve from the duplication of an ancestral operon encoding all the components (3, 11, 12).

Although the SBP is the most divergent of the ABC components (3, 12), their overall structures remain similar (17, 18). They are composed of two globular domains connected by a hinge (17, 19) which were speculated to have arisen from gene duplication followed by fusion of an ancestral gene encoding a single domain (12, 20). The SBPs are divided into distinct categories, based either on their structural β-sheet topologies (21) or folding (22, 23). A major structural event described as "domain dislocation" by Fkukami-Kobayashi might have taken place in the course of their evolutionary history (20). The domain dislocation affects the rearrangement of the secondary structures by the intramolecular rearrangement and exchange of strands as opposed to domain swapping which does not affect the topology of the protein. According to the authors, this structural change might have occurred before the divergence of the bacteria and archaea. However, since most SBP-encoding genes are found in Bacteria and Archaea, it is unclear if they were present in the eukaryotic lineage and then lost in this domain (9).

The evolutionary history of particular ABC transporter family is often difficult to establish. This is mainly due to the relatively high rates of gene transfers (4, 16, 24), low sequence identities and sequence divergence of some portions of the SBPs (20). Therefore, the functions of the ABC transporters are difficult to infer from sequence information and phylogenetic histories alone. However, it was previously noted that

SBPs with chemically similar ligands tend to cluster with each other in phylogenetic trees (12, 20). This was also observed through this work in phylogenetic examinations of the genes encoding the trehalose-, xylose- and mannoside-binding proteins in *T. maritima*.

The Thermotogales

The most studied and the first member isolated of the Thermotogales is *Thermotoga maritima* MSB8 (25). The Thermotogales are gram-negative anaerobes commonly found in geothermal environments (26), heated marine sediments (25, 27, 28), oil reservoirs (29, 30) and bioreactors (31). Most members are thermophilic organisms that grow at temperatures above 70°C. Members of the *Thermotoga* genus such as *Thermotoga maritima* and *Thermotoga* species RQ2 appear to be the most thermophilic members, able to grow at temperatures as high as 90°C (25). Other Thermotogales, such as *Fervidobacterium nodosum* and *Thermotoga lettingae*, which were isolated from a hot spring in New-Zeland (26) and sulfate-reducing bioreactor (31), respectively, have lower optimum growth temperature ranging from 65-70°C. In recent years, efforts have been made to find mesophilic Thermotogales (32) and currently, two mesophilic species, *Mesotoga prima* and *Mesotoga infera*, have been successfully isolated (28, 33).

All the known members of the Thermotogales possess a common feature, a loose outer envelope named a “toga” (25), from which their name is derived. Although the toga obviously acts as a physical barrier and serves as anchor to some glycoside hydrolases (34, 35), its exact composition and function remain unclear (36). However, two toga-associated structural proteins were identified, the anchor protein OmpA1 and the porin OmpB (36–39).

The Thermotogales are often model organisms to study thermostable glycoside hydrolases because of their potential for industrial applications. Given the ability of the proteins from these organisms to withstand high temperatures, the Thermotogales are attractive model organisms to study protein thermostability in general. In addition, the Thermotogales are attractive organisms to study the early evolution of life because they are believed to be a deep branching bacterial lineage which might have retained characteristics reflecting the lifestyle of the last common ancestor of bacteria (40, 41).

Carbohydrate ABC transporters in *Thermotoga maritima*

Compared to other prokaryotes, *T. maritima* encodes a high proportion of ABC transporters per millions of base pairs (Mb) (42) and these transpotersare often clustered in the vicinity of genes encoding hydrolases and transcription factors. The reason for this unusually high number of ABC transporters is unclear but it could provide a selective advantage in environments enriched with polysaccharides and biomass.

The sequence analysis of the genome of *T. maritima* released in 1999 suggested that many genes were horizontally acquired from members of the Archaea (40). Although later analysis demonstrated that the number of these genes was overestimated, (41) some SBPs were later shown to be closely related to archaeal homologs (16) and were apparently shared with archaea.

After the release of the genome sequence of *T. maritima* in 1999, many researchers attempted to elucidate the sugar specificities of the annotated ABC transporters. Two notable studies were able to demystify the functions of many ABC transporters encoded by *T. maritima* by using different experimental approaches. Conners et *al.* predicted the

functions of the ABC transporter encoding-genes by using expression data from microarray experiments (43), while Nanavati *et al.* identified the functions of various substrate-binding proteins by measuring their binding affinities using intrinsic fluorescence (8). Nanavati *et al.* found ABC transporter proteins able to bind a variety of sugars: cellobiose, laminaribiose, xylose, xylobiose, xylotriose, xyloglucan, mannobiose, mannotriose, mannotetraose, *myo*-inositol, ribose, maltose, maltotriose, mannotetraose and trehalose.

In the high temperature ecosystems that the Thermotogales inhabit, their carbon sources might sustain thermochemical changes. At 80°C, a major known thermochemical change is the Maillard reaction that affects reducing sugars and amino compounds (44). The effect of thermochemical changes on carbon source utilization has been studied (45, 46). One study examined the effect of the thermochemical changes on the growth of *P. furiosus* (45). It was found that a fed-batch culture gave a larger cell yield as compared to a batch culture, most significantly when the cells were grown on β-1,4-cellooligosaccharides, β-1,3-laminarioligosaccharides, maltose and cellobiose. The author concluded that the cell yield increase was because the thermochemical change on the sugar was limited when the carbon source was not present in the culture for an extended length of time. Another study examined the effect of the products formed by the Maillard reaction on the growth of *Aeropyrum pernix*. Their findings suggested that these products are toxic and inhibit cell growth (46).

Carbohydrate uptake: catabolism and regulation

Every organism requires energy to perform its cellular-based activities to grow and divide. *Thermotoga maritima* is a heterotrophic organism that ferments organic carbon to produce acetate, lactate, CO_2 and H_2 (25, 47). The molecules of glucose are converted into pyruvate through the Embden-Meyerhof (EM) and Entner-Doudoroff (ED) glycolytic pathways (47, 48) to meet the energy requirements of the organism. The EM and ED operate simultaneously and their contributions to the production of lactate are approximately 87% and 13%, respectively (48). *T. maritima*, *Thermotoga neapolitana*, *Thermotoga naphthophila*, *Thermotoga petrophila T. lettingae* and *F. nodosum* can grow a variety of carbohydrates such as mono-, di- and polysaccharides (25–27, 30, 31). Most can grow on fructose, galactose, glucose, xylose, ribose, lactose, maltose, sucrose and starch at the exception of *F. nodosum* which is unable to grow on xylose or ribose and *T. naphthophila* and *T. petrophila* which are unable to grow on xylose.

Unlike many other prokaryotes, many *Thermotoga* species do not contain genes that encode phosphotransferase systems (PTS) for substrate transport. So far, only the genomes of *Thermotoga* species RQ2 and *T. naphthophila* reveal the presence of fructose PTS encoding-genes (49). As discussed previously, carbon uptake appears to be primarily done by ABC transporter systems. Recently, Rodionov's research group demonstrated that the transcription factor ROK (repressor, open reading frame, kinase) controls the transcription of many ABC transporters (50, 51). It is suggested that the ROK proteins binds upstream of the ABC transporter operon and prevents its transcription. The presence of the effector, usually a carbohydrate capable of being transported by the specific ABC transporter system, can derepress the expression of the genes controlled by ROK. A similar mode of repression has been identified for the

trehalose/maltose ABC transporter and the maltodextrin ABC transporter of the archaeon *Pyrococcus furiosus* (52, 53). The *P. furiosus* maltose ABC transporter operon encodes the maltose ABC transporter proteins (MalEFGK), a trehalose synthase, and the transcriptional repressor TrmB. Studies demonstrate that TrmB binds the promoter and prevents the transcription of the operon. In the presence of maltose and trehalose, TrmB binds these sugars and relieves the inhibition of transcription. TrmB also regulates the transcription of the maltodextrin ABC transporter operon, which encodes the maltodextrin ABC transporter proteins (MdxEFGK) and an amylopullulanase by using a mechanism similar to that it uses to control the maltose ABC transporter. TrmB binds upstream of *mdxE* and only maltotriose and sucrose can relieve the inhibition of transcription of this operon. The repression of the transcription of these two ABC transporters is higher in the presence of glucose (54).

The SBP and the substrate translocation

The SBPs of gram-negative bacteria contain a signal peptide (Class I) but they are soluble and located in the periplasm while SBPs of gram-positive bacteria and archaea are attached to the cytoplasmic membrane. The SBPs of gram-positive organisms have a lipobox motif containing a conserved cysteine (Class II). The SBPs are attached to the membranes by lipid modification of this cysteine by a thioether linkage to a diglyceride from the membrane (55). Like the gram-positive organisms, archaea have their SBPs anchored in their membranes (56, 57). Some SBPs, mainly members of the CUT family, have a mode of attachment that is similar to that found in gram-positive bacteria. For example, the archaeal trehalose/maltose-binding protein (TMBP) in *T. litoralis* and *P. furiosus* contain a cysteine in their N-termini like that found in the lipobox motif (56, 57). Other archaeal SBPs, mainly members of the PepT family, are cleaved in their N-termini and are anchored in the membrane through a C-terminal hydrophobic domain that is integrated into the membrane (56, 57). A third mode of attachment was described in *Sulfolobus solfataricus*. These SBPs contain a signal sequence in the N-terminus, similar to class III signal peptides, that anchors them by their N-terminus (57–59). Interestingly, although the genome of *T. maritima* has exchanged SBP encoding genes with the archaea through horizontal gene transfer (HGT) (14–16, 40), none of the SBP amino acid sequences in *T. maritima* are predicted to contain a lipobox motif (Class II). The SBPs encoded by *T. maritima* contain a signal peptide and are soluble (Class I) as the SBP found in other gram-negative organism. It is possible that the signal peptide sequence has evolved rapidly by relaxed selection following their acquisition as proposed by Li *et al.* (60).

The translocation of the substrate to the inside of the cell by the bacterial ABC transporter is usually performed by a complex of proteins that includes an SBP, two permeases containing the TMPs and the two ATPases containing the ABPs. A variety of models have been proposed to explain the different steps of substrate translocation to the inside of a cell. For simplicity, most models are derived from the structures and experiments using the maltose ABC transporter in *Escherichia coli* ($MalEFGK_2$) will be discussed. The $MalEFGK_2$ complex is a well characterized importer, which was crystallized in different conformational states (61–65). As implied in the name, the ABC transporter complex hydrolyzes ATP to provide the energy required to translocate the molecule across the membrane. The two ABPs, which contain the Walker-A, Walker-B and several other motifs, are located in the cytoplasm. The Walker-A motif binds ATP by interacting with its phosphoryl groups (P-loop) while the Walker-B motif contains an acidic amino acid residue involved in the binding of magnesium (66, 67). The C-motif is a signature sequence (LSGGQ) found in all members of the ABC superfamily. This motif interacts with the γ phosphate of the ATP (68, 69) and mutations in this motif (70–72) suggest that it might be involved in ATP hydrolysis. The D-loop, Q-loop and H-loop participate in the hydrolysis of the ATP (73).

MalE (the SBP) initiates the uptake of the carbohydrate in the periplasm. The sugar binds to the MalE binding site located in a cleft between its two globular domains that are connected by a hinge (19, 21). When liganded, MalE undergoes a conformational change that brings together the globular domains (19), often described as a "Venus fly trap" mechanism (21, 74). It is accepted that the substrate translocation to the inside of the cell occurs according to the alternating access model proposed by Jardetzky (75). In the

absence of a ligand, the membrane-embedded permeases (MalF and MalG) form a cavity exposed to the cytoplasm, a conformation called "inward-facing". The ligand is translocated in an alternate access mode in which the cavity is opened to the periplasmic face, a conformation called "outward-facing" promoted by the binding of ATP to allow the ligand to cross the membrane (76). The $MalEFGK_2$ crystal structures suggest that the ATP hydrolysis cycle is performed according the "switch model". In the inward-facing conformation of the transporter, the ABPs are separated and require ATP and MalE to stimulate their closure and to reorient the cavity formed by the permeases in the direction of the periplasm, the outward-facing conformation that promotes substrate translocation (62, 64, 77, 78). Surprisingly, the unliganded MalE and ATP are able to promote the closure of the ABPs (77). When bound to ATP, the ABPs form a dimer in a "head-to-tail" configuration (79), each nucleotide is bound at the interface of the dimer between the P-loop of one ABP and the C-loop of the other ABP (64). After hydrolysis of the 2 ATPs, one ATP per ABP, the complex opens to release the ADP and P_i and the transporter returns to its nucleotide-free form. The absence of ATP drives the transporter to its inward-facing conformation suggesting that ATP hydrolysis controls the conformational changes from the inward-facing to the outward-facing state (78, 80).

Since the discovery of the substrate binding protein-dependent transporters, it was assumed that the SBP directs the ligand specificity of the ABC transporter. However, recent crystal structures showed that the *E. coli* MalF permease binds the 3 glucosyl units at the non-reducing end of the maltoheptaose (65). This new finding raises the possibility that the substrate specificity of a transporter may be dictated by its permease as well as its substrate binding protein.

Objective statement

The objective of this study is to determine the evolutionary mechanisms leading to substrate preferences of the mannoside-binding periplasmic substrate-binding proteins in the Thermotogales. To accomplish this objective, this thesis is divided in two aims:

1) **Development of a pipeline to discover the binding properties of SBPs of unknown function.** Due to the challenges to determine the functions of thermophilic SBPs, a technique of differential scanning fluorimetry (DSF) will be optimized to screen ligands that interact with the SBPs and the putative interactions will be confirmed by spectroscopy. The *T. maritima* MalE1 and MalE2, two SBPs of known function will be used to validate the optimized DSF assay by comparing the sugars identified with this technique to the known binding properties previously determined by spectroscopy. The pipeline will be used to identify the binding properties of two SBPs of unknown function. This step will demonstrate that the pipeline using DSF assay as a screening technique is effective to discover unknown binding properties.
2) **Establish the relationship between the binding properties of SBPs and their phylogenetic histories.** The mannoside-binding proteins encoded by many Thermotogales will be characterized to determine their respective binding properties using the pipeline previously developed. Then, the residues involved in the binding site will be predicted using structure superimposition. The selective pressures on these proteins will be examined using a branch-site model to determine the sites that have undergone neutral evolution and positive selection. This will provide insight on the type of selective pressures that have

shaped their sequences and to determine if these selective pressures have acted on sites that encode important regions of the binding sites.

Chapter overviews

Chapter 2 describes the identification of different genomovars of *T. maritima* and the analysis of their respective genomic differences. After the release of the genome sequence of *T. maritima* strain MSB8 in 1999, discrepancies were noted between DNA sequences in the genome sequence and published cloned *T. maritima* genes. In this research, I explained these discrepancies by finding that two genomovars of *T. maritima* MSB8 were circulating in laboratories. The genomovars were designated as DSM3109 and TIGR, the latter is the genomovar sequenced in 1999 by The Institute of Genomic Research (TIGR). The major genomic difference between the genomovars is an additional 8,870 bp of genomic DNA in the genomovar DSM3109 that is missing from genomovar TIGR. Sequencing revealed that the region contains operons for putative trehalose and xylose ABC transporters. In addition to this genetic event, other genomic variations between genomovar TIGR and DSM3109 were found.

Chapter 3 describes a novel application of differential scanning fluorimetry (DSF) that is used to determine ligand-binding interactions. The original experimental applications of DSF are not capable of detecting ligand interactions with the thermophilic proteins found in *T. maritima*. By reducing the pH of the assay, I was able to identify the putative ligands that interact with the previously characterized substrate-binding proteins (SPBs) of *T. maritima*. The chapter includes a description of important parameters and their effects on the identification and detection of ligand interactions.

Chapter 4 describes the functional characterization of the two SBPs encoded by each operon found in the 8,870 bp of additional genomic DNA of the genomovar DSM3109 (see Chapter 1). By analyzing their respective binding properties using the DSF assay and spectroscopy, the two operons were demonstrated to encode a trehalose ABC transporter (TreEFG) and a xylose ABC transporter (XylE2F2K2). At 60°C, TreE binds trehalose, sucrose and glucose while XylE2 binds xylose, glucose, and fucose (81).

Chapter 5 contains the analysis of the binding properties of SBPs previously characterized and encoded by TM1199, TM1150, TM0595 and TM0418 (*inoE*) in *T. maritima*. The substrate of the SBPs encoded by TM1199, TM1150, and TM0595 remained undetermined and their binding properties were assessed using the DSF assay. These results combined with comparative genomics analyses suggest that the SBP encoded by TM1199 is a putative galactose and lactose-binding protein. Also, a spectroscopic measurement demonstrated that InoE encoded by TM0418 binds inositol-6-phosphate at 20°C and 60°C but not *myo*-inositol at 60°C (82).

Chapter 6 presents the results from the characterization of the ten representatives of the Thermotogales mannoside-binding proteins that show evidence of a complex evolutionary history including vertical inheritance, horizontal gene transfers and possibly a gene duplication. Many Thermotogales encode two mannoside-binding proteins, ManD and ManE. Their binding properties were measured at 37°C and 60°C by DSF and spectroscopy and their thermal stabilities were measured by DSF. These analyses revealed that at 37°C and 60°C, the ManD and ManE homologs bind cellobiose, -triose, -tetraose, β-mannotriose, and β-mannotetraose. However, ManE binds β-mannobiose, laminaribiose, laminaritriose and sophorose while ManD does not bind these sugars.

Chapter 7 presents an analysis to determine which residues are located in the binding sites of the ManD and ManE homologs using the solved structure of $ManE_{Tmar}$ and the modeled structure of $ManD_{Tmar}$ superimposed on the structure of $BglE_{Tmar}$ bound to cellobiose and laminaribiose as a reference. A branch-site analysis was performed to identify which sites in the branches leading to *manD* and *manE* were under selective pressure. These analyses suggest that both *manD* and *manE* contain codon sites that were under positive selection and these sites encode an important region of the binding site.

Contribution from other researchers

Dr. Pascal Lapierre at the University of Connecticut Biotechnology Center made major contributions on the bioinformatic analyses of the genomic variations of the genomovars of *Thermotoga maritima* MSB8. Dr. Lapierre advised and contributed to the construction of phylogenetic trees and the dN/dS ratio analysis. Dr. Kristen Swithers and Chaman Ranjit performed preliminary examinations of the 10 kb region. In addition, Dr. Duval Nanavati engineered the recombinant MalE1, MalE2 and InoE as well as the SBPs encoded by TM1199, TM1150 and TM0595.

Chapter 2

Identification of genomic variations of different genomovars of *Thermotoga maritima*

Portions of this chapter were published in Boucher, N. and K. M. Noll. 2011. Ligand screening by differential scanning fluorimetry of thermophilic transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8. Appl. Environ. Microbiol. 77:6395-6399.

Introduction

In the course of examining the evolution of maltose ABC transporters among the Thermotogales, it was noticed that most of the sequenced species have three operons that are members of this family (*mal1*, *mal2* and *mal3*) (16). The genome of the first sequenced species, *Thermotoga maritima* strain MSB8 was unique in having only two *mal* operons, *mal1* and *mal2*. The *T. maritima* operons appeared to have arisen by duplication and they belong to the family that includes the *mal* transporter in *Escherichia coli* (16). The Thermotogales *mal3* family is related to the archaeal trehalose/maltose transporters from *Thermococcus litoralis* and *Pyrococcus furiosus* (16, 83, 84). Based on earlier examination, it was speculated that an ancestral Thermococcales may have acquired this transporter via horizontal gene transfer from a member of the Thermotogales (16). The arrangement of the genes adjacent to the *mal3* orthologs in

several *Thermotoga* species is similar to a region of the *T. maritima* genome, but there was no *mal* transporter in that region in the annotated genome sequence.

In 1994, Liebl, *et al.* published the sequence of a portion of a *malE* gene adjacent to a β-glucosidase gene (*bglA*) that was cloned and characterized (85). When Liebl's MalE protein sequence was queried against the *Thermotoga* spp., the sequence was found to have a strong match to the MalE3 orthologs, more so than the *T. maritima* MalE1 and MalE2. Since a *bglA* gene is upstream of the *malE3* gene in all the other *Thermotoga* species, the region containing the *T. maritima mal3* ortholog was suspected to have been deleted during cultivation of that strain (16). The hypothesis was confirmed by reexamining the isolate that was sequenced in 1999 and the strain used by Liebl, *T. maritima* MSB8 deposited at the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ, DSM 3109). Sequencing that region in the strain from the DSMZ showed that there is a third complete *mal* transporter operon in *T. maritima* MSB8 and the sequence published in 1999 is from a laboratory variant missing a portion of chromosomal DNA containing the *mal3* genes. Furthermore, another ABC transporter operon was found, a putative xylose ABC transporter, as well as Liebl's *bglA* gene that was also missing from the genome sequence for *T. maritima* MSB8 published in 1999. Besides their genetic variation, no discernable phenotypic differences are observable between the genome-sequenced strain and that from the DSMZ. I propose to call these two strains genomovars and to designate the strain deposited at the DSMZ collection as *Thermotoga maritima* MSB8 genomovar DSM3109 and the genome-sequenced variant *Thermotoga maritima* MSB8 genomovar TIGR. These results expand our knowledge of

the physiological capabilities of this organism and also provide a cautionary tale for genome sequencing projects.

Materials and Methods

Strains

Thermotoga maritima MSB8 genomovar DSM3109 (DSM 3109), *Thermotoga petrophila* RKU-1 (DSM No. 13995), and *Thermotoga naphthophila* RKU-10 (DSM No. 13996) were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). *Thermotoga maritima* MSB8 genomovar TIGR and *Thermotoga* sp. RQ2 were kindly provided by Karl O. Stetter. *Thermotoga neapolitana* NS-E was provided by the late Holger W. Jannasch.

PCR amplification and DNA sequencing

The region between the *cbpA* and ROK orthologs (TM1848 and TM1847, respectively in *T. maritima*) was amplified by PCR using the IDUltra Taq DNA polymerase enzyme kit (ID Labs Biotechnology, Inc.) containing 20 µl of the following mix: 1X PCR buffer, 0.2 mM of each dNTP, 2.5 U IDUltra Taq DNA polymerase, 1 mM of each primer (Table 1) and 200 ng of genomic DNA. The products were amplified using an MJ Mini thermocycler using an initial denaturation step at 94°C for 2 min 45 s and 35 cycles of 94°C for 15 s, 51°C for 30 s and 72°C for 12.5 min and a final extension at 72°C for 10 min. The PCR products were resolved by 0.9% agarose gel electrophoresis in 1X TAE buffer and visualized by staining with ethidium bromide.

The 10 kb PCR product from *T. maritima* MSB8 genomovar DSM3109 was sequenced using a primer walking method. After amplification using the protocol described above, the product was precipitated using 2 volumes of 95% ethanol and 1/10

volume of 3 M sodium acetate (pH 5.2) and washed twice with 70% ethanol. The purified PCR product was sequenced using the primers listed in the Table 1. All the primers in this work were designed using Primer3 v0.4.0 or Primer-BLAST (86). To increase the reliability of the sequences, at least two identical sequences were obtained and sequencing was performed on both DNA strands.

The 16S rRNA genes from both *T. maritima* MSB8 genomovars were sequenced using two sets of primers. A fragment of 1478 bp was amplified from each using the primers 5´-TAACACATGCAAGTCGAGCG-3´ and 5´-GGCTACCTTGTTACGACTT-3´ using the same PCR conditions described above but with IDproof Taq DNA polymerase (ID Labs Biotechnology, Inc.) and the following PCR program: 94°C for 2 min 45 s and 35 cycles at 94°C for 15 s, 50°C for 30 s and 72°C for 90 s.

The purified PCR products (QiaQuick PCR purification kit, Qiagen) were used as templates in Sanger sequencing reactions containing a single primer (either 5´-TAACACATGCAAGTCGAGCG, 5'-CGGGTATCTAATCCGGTTTG-3´, 5´-TGGGGAAGCCGGTCTCCTGG-3´, or 5´-GGCTACCTTGTTACGACTT-3´) and BigDye Terminator v3.1 (Applied Biosystem). Each reaction was analyzed by capillary electrophoresis on a 3130xl Genetic Analyzer (Applied Biosystem).

The *malF2* genes from both *T. maritima* MSB8 genomovars were sequenced using three sets of primers that hybridized with TM1836 and TM1839 (gi|12057205:1812428-1814385). Three fragments of 697 bp (5´-AAACGATGGGGAAGAGAACC-3´, 5´-GAACGTGTACAGGACGTTACTCA-3´), 809 bp (5´-AGAACGGATCGTTGAACC AC-3´, 5´-TGGGAAAGAGGTCCTTTTGA-3´) and 745 bp (5´-GAATAGACGGCG

TTCACCTC-3´, 5´-GACAGGCCAGTG TCGAAGAT-3´) were amplified using the same PCR conditions as described for the 16S rRNA gene with the following PCR program: 94°C for 2 min 45 s and 35 cycles at 94°C for 15 s, 55°C for 30 s and 72°C for 45 s. The PCR purification and sequencing reactions were carried out as above.

Sequence depositions

The missing genomic DNA between TM1848 and TM1847 from genomovar DSM3109 and *malF2* from genomovars DSM3109 and TIGR were deposited in GenBank, as JF907620, JF907621 and JF907622, respectively.

Prediction software

ORFs were predicted using GeneMark (87) and ORF Finder (88). The operons were predicted by FGENESB. The transcription and translation terminators were predicted by FindTerm and RibEX (89). Protein transmembrane helices were predicted by the TMHMM Server v. 2.0 (90).

Data acquisition and phylogenetic analysis

Protein sequences were retrieved by BLASTP searches of the sequenced microbial genomes database at NCBI (91) using as queries amino acid sequences of two *T. maritima* MalE, MalF, or MalG homologs. The top 50 sequences retrieved by BLASTP searches were assembled into datasets and repeated sequence entries and extremely closely related sequences (determined using neighbor-joining trees of aligned sequences) were removed. Those sequences from three-gene operons (homologs of MalE, MalF, and

MalG) were retained and aligned using ClustalX v1.83.1 (92). The genes in operons were individually aligned and, then concatenated in a single alignment. The authentic frameshift in TM1837 was corrected as described (93). The concatenate dataset was used to construct a consensus tree using PHYML v2.4.4 performed with 1,000 bootstrap resamplings, the JTT substitution model, a fixed proportion of invariable sites, one category of substitution rate, and the BIONJ input tree.

Table 1. Primers used for sequencing the 10 kb region by primer walking.

Sequence (5'-3')	Location		Strand
CAGTCTTTCACTTATCATCATGG	*1827017	*1827039	+
ACGATGTTCTTTCTATCACCGAAGTGC	446	472	+
CCTTGCGGCTCTCGCGGAAA	486	505	-
TGTTTCGAGTAACTCCAGGCTC	957	978	+
ACTCGCTGCTGGGTTGGAGT	1080	1099	-
CATGATCTCGTCAGAACTTCCTGAAGT	1193	1219	-
TGCCCCTACGTCTATTCCTC	1276	1295	+
GGTGAGCGTGCTGAGAGATGGG	1912	1933	-
ACCTTCGCCTTCTTGTAGACAGC	2078	2100	+
GATTACAAGAAGATGTACAGGGAAGCGG	2199	2226	-
CCTGTCGGTTGCGGAGAACA	2265	2284	-
TGCGCTACTCCAAGCCACGC	2800	2819	+
TCCGTTTATCGTGTATACCAGAAT	2853	2876	+
TGGAAGCGTCGAAGCTGGTG	2901	2920	-
AGTGCCGGTACCAGTGATCTACATG	3000	3024	-
TTCTGTTACAATATCGACACC	3647	3667	+
TATCTCTCTGTTCTCGTTCTCTAC	3758	3781	-
ATGCGGCGAAGCAGGCAGAG	4415	4434	-
CGGAATGAAAGATGCGGCAGAGAAAC	4474	4499	-
TCCGTTCTCCACCGCAAACCA	4496	4516	+
TTGGGAGGTGTTTTTCATGAG	4620	4640	-
CGCAGCGATGAACGTCAAAAGCC	5110	5132	+
TCGTGCTTCCCATGTCTGCTCCT	5149	5171	-

CCCGATCAGGCTTTATCGACGAAAT	5613	5637	+
TTCACCGGTGCGTGGTTCGG	5831	5850	-
CACGCACCGGTGAAAGCCCT	5837	5856	+
CTCGATGCACTGCGCGTGTT	5936	5955	-
CGCGCAGTGCATCGAGTGTC	5940	5959	+
TTGATAACGCCAACGATAGGAGTGGC	5972	5997	-
AAAGAGGCGGGGGAGAGCCC	6649	6668	+
TTCAGAGGTACGTGCACGCTGC	6755	6776	-
CGAGGACCCACCCACCGAGT	7020	7039	+
AGGAACTGGCCGTACGCCTG	7130	7149	-
TTCAAGGAACGGAGCAAACT	7624	7643	+
AGTGGAAATCATTCCGATGC	7743	7762	-
AGGAGCGACCAGACGAAGTA	8271	8290	+
AAGGCCATACAGGAGGGAGT	8304	8323	-
CACGGTTTCCCTGAACACTT	8837	8856	+
CACAATCTCTTGAGGGCACA	8871	8890	-
CCCGTGTCACCGTTCTTTAC	9321	9340	+
GTTGTGATGGGCTGACCTTT	9581	9600	-
GGAGACAGAGCCTTTGAGAT	*1828254	*1828235	-

* Positions on the GenBank sequence (AE000512.1) from the genome project of *T. maritima* genomovar TIGR.

Results

Identification of two genomovars of *Thermotoga maritima* MSB8

PCR amplifications with oligonucleotides that hybridized to the loci TM1848 and TM1847 were performed using DNA isolated from *T. maritima* from a culture of the strain that was used to prepare DNA for the TIGR sequencing project (kindly provided by Dr. Karl Stetter) (40) and from a culture of the strain obtained from the DMSZ collection (DSM 3109). The amplification from the inoculum used for the TIGR sequencing project gave a fragment of 1,216 bp, as predicted from the genome sequence data in GeneBank, while the amplification from the DMSZ collection gave a fragment of approximately 10 kb (Figure 2). Other Thermotogales species gave fragments predicted by their reference genome sequences (Figure 2). The amplification from *T. neapolitana* gave a second amplicon of approximately 3.5 kb. The sequence of this fragment showed that it contained the alpha-glucan phosphorylase gene located at CTN_1407 indicating that this product resulted from a non-specific amplification.

To ensure that both cultures were genuinely *T. maritima* MSB8, the 16S ribosomal RNA gene of each culture was sequenced. Both sequences were identical and matched with 100% identity to the sequence of the 16S rRNA gene deposited in 1987 as *Thermotoga maritima* MSB8 DSM 3109 (M21774.1). This demonstrated that these two cultures represent genomic variants or genomovars of *T. maritima* MSB8. I propose to designate the strain deposited at the DSMZ collection as *Thermotoga maritima* MSB8 genomovar DSM3109 while the variant used by TIGR for its genome sequencing project will be referred as to *Thermotoga maritima* MSB8 genomovar TIGR.

Sequence of the genomic DNA deleted from the genomovar TIGR

The sequence of the deleted genomic DNA was completed by primer walking using the 10 kb PCR product from *T. maritima* MSB8 genomovar DSM3109 as template. This genomovar was found to have an additional 8,870 bp of genomic DNA between the loci TM1848 and TM1847. This region is missing from the genome of the genomovar TIGR between the positions 1,827,804 and 1,827,805. The new sequence has been deposited in GenBank as JF907620. This additional genomic DNA contains seven ORFs as predicted by GeneMark.hmm 2.4. The sequence comparison of this region with the syntenic regions of other *Thermotoga* species shows that they all have similar genes at this location (Figure 3) suggesting that genomovar DSM3109 lost 8,870 bp of genomic DNA during laboratory cultivation giving rise to genomovar TIGR.

The first ORF encodes a beta-glucosidase (BglA) that was first characterized in 1994 before publication of the genome sequence (85). The sequence matches with 100% identity to the sequence in GenBank (X74163) deposited by Liebl.

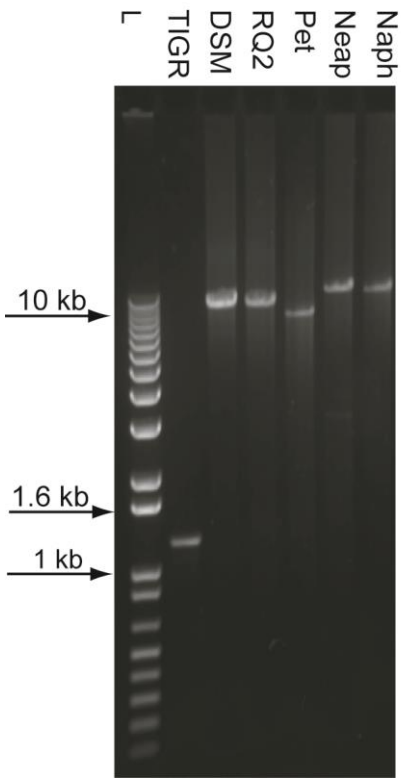


Figure 2. PCR amplifications using primers to the loci TM1848 and TM1847. L, DNA ladder 1 kb plus (Invitrogen); TIGR, *T. maritima* MSB8 genomovar TIGR; DSM, *T. maritima* MSB8 genomovar DSM3109; RQ2, *Thermotoga* sp. RQ2; Pet, *T. petrophila* RKU-1; Neap, *T. neapolitana* NS-E; and Naph, *T. naphthophila* RKU-10.

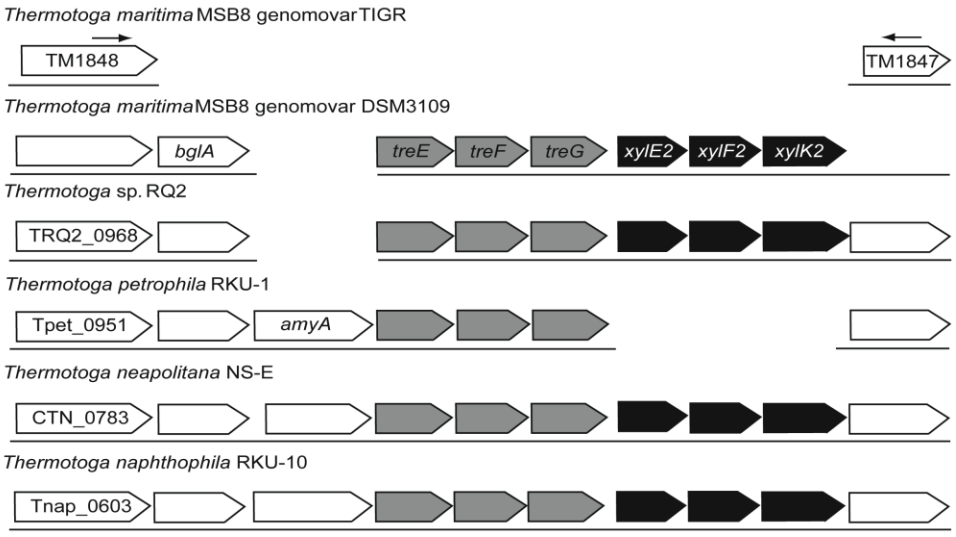


Figure 3. Organization of the ORFs between the TM1848 and TM1847 orthologs in *Thermotoga* species. The arrows represent the primer annealing sites used for PCR amplification of the 10 kb region. The *tre* operon is represented by grey boxes while the *xyl2* operon is represented by black boxes.

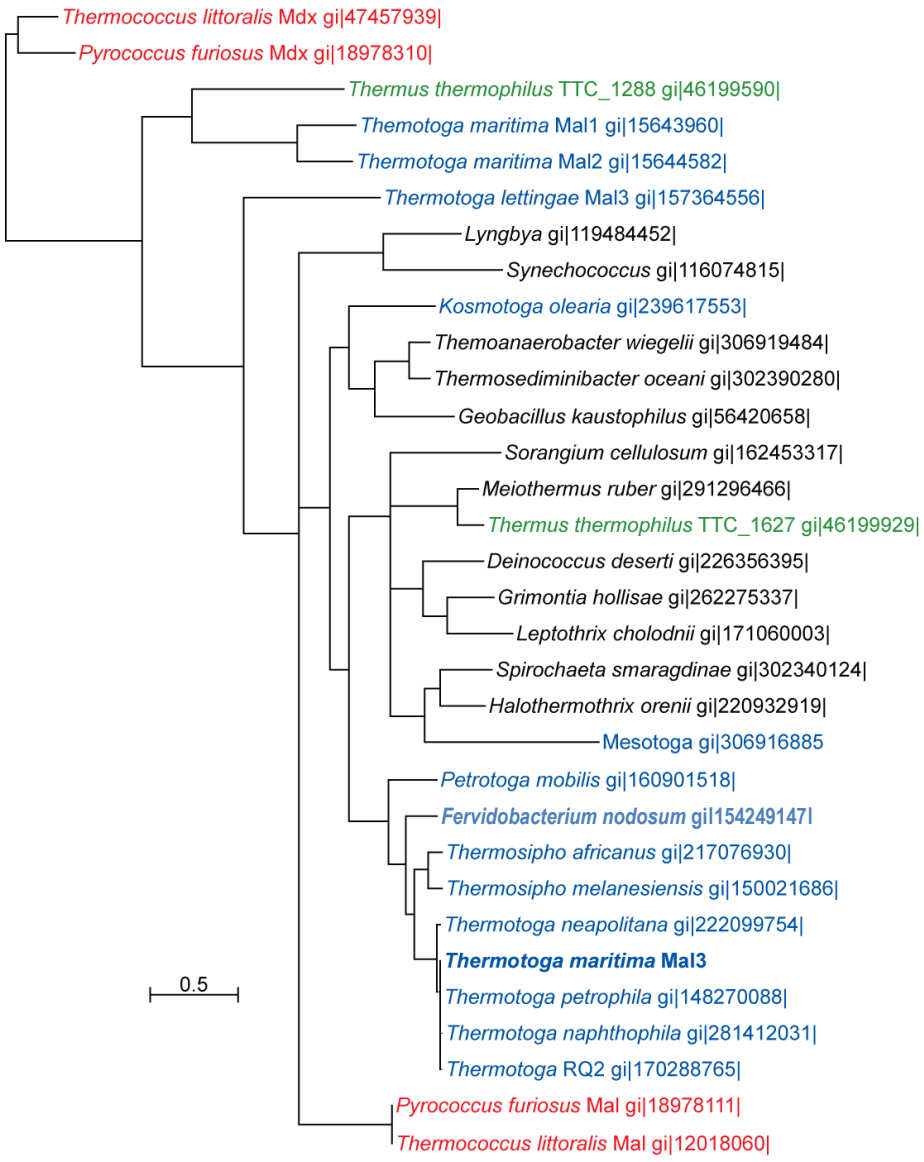


Figure 4. An unrooted Mal3-type ABC transporters tree. Concatenated aligned sequences of MalEFG orthologs were used to construct a maximum likelihood tree using PhyML. Only relationships with bootstrap support values ≥70% are shown. The archaeal Mdx transporters were used as the outgroup. Thermotogales transporters are shown in blue, archaeal in red, *Thermus* in green, and other bacteria in black. The *T. maritima* Mal3 transporter is in boldface. The gi numbers of the corresponding MalE orthologs are indicated.

The next three ORFs encode a putative member of the maltose and trehalose ABC transporter family. *T. maritima* has two known maltose operons, *mal1* and *mal2*, that were previously characterized (8, 93, 94). It was speculated that the *mal3* operon discovered in other Thermotogales was missing from the *T. maritima* genome (16), but I have found that the strain used for the genome sequence, genomovar TIGR, lost the region after the strain was deposited in the DSMZ collection. The newly sequenced operon will be designated as a putative trehalose and maltose ABC transporter system (*tre*) because the substrate-binding protein is phylogenetically closer to the trehalose and maltose-binding protein (TMBP) encoded in *Thermococcus litoralis* than MalE1 and MalE2 found in *T. maritima* (Figure 4). The operon encodes a periplasmic substrate-binding protein (TreE) and two transmembrane proteins (TreF and TreG) (Figure 3). Both TreF and TreG proteins have six predicted transmembrane helices. The region downstream of *bglA* contains a palindromic sequence that suggests the presence of a rho-independent transcription terminator. This suggests that *bglA* is not transcribed in the same operon as *tre*.

The fourth, fifth and sixth ORFs encode a second putative ABC transporter system containing a periplasmic substrate-binding protein, a transmembrane protein with eight predicted transmembrane helices, and an ATP-binding protein with two nucleotide-binding domains (Figure 3). The substrate-binding protein of this ABC transporter system is homologous to XylE1 (TM0114) from *T. maritima*, a substrate-binding protein that binds both glucose and xylose (8, 95). This suggests that this ABC transporter operon is a putative *xyl2*.

TM1847 encodes a putative transcription factor that belongs to the ROK (Repressor, Open reading frame, Kinase) family. However, the 5´ portion of the gene is deleted in genomovar TIGR. As in other *Thermotoga* species, genomovar DSM3109 contains a full ORF likely to encode a functional protein (Figure 3).

Evolution of Mal and Tre transporters

The *Thermotoga mal1* and *mal2* operons are related to the *E. coli mal* operon and arose from a gene duplication in the ancestor of the *Thermotoga* species (16). The new *tre* operon found in genomovar DSM3109 is only distantly related to *mal1* and *mal2* and is more closely related to the trehalose/maltose ABC transporter systems of *Thermococcus litoralis* and *Pyrococcus furiosus* as well as putative operons previously designated as *mal3* found in most other Thermotogales species (Figure 4) (16). The orthologous *tre* (*mal3*) operons in other *Thermotoga* species are sister groups and the evolution of the Tre transporter resembles the evolution of the Thermotogales lineage as revealed through 16S rRNA gene sequence comparisons. This suggests that most Thermotogales species inherited this operon *via* vertical gene transfer. The presence of bacterial lineages interspersed within the Thermotogales can be explained by subsequent transfer of genes event between the Thermotogales other bacterial species. The only exception is the Tre transporter in *T. lettingae* that is basal to the clade containing the other Thermotogales. *T. lettingae* seems to have acquired this transporter from another bacterial donor. Interestingly, the archaeal trehalose/maltose transporters are also of bacterial origin. There is no evidence for a specific relationship between these archaeal

transporters and the Thermotogales Tre transporters. Their inheritances were apparently independent of one another.

Genetic variations of *malF2*

In an effort to determine if other genetic variations exist in the other maltose ABC transporters (*mal1* and *mal2*) of the two genomovars, the region encoding the defective *malF2* gene was sequenced in both genomovars. According to the published genome sequence for genomovar TIGR (40), an ancestral *malF2* suffered nonsense mutations that created two ORFs, TM1837 and TM1838, neither of which encode a functional MalF2. However, when this region in both genomovars was sequenced, no frameshifts were found, suggesting both genomovars encode a functional MalF2. Together, the loci TM1838 and TM1837 encode a putative full-length MalF2 protein of 577 amino acids (Figure 5).

Other, minor differences were found in the DNA sequence of the *mal2* regions in the two genomovars. In genomovar DSM3109, *malF2* contains a single nucleotide difference compared to *malF2* in genomovar TIGR that results in a valine at position 528 in genomovar TIGR instead of an isoleucine as in genomovar DSM3109 (Figure 5, Appendix 1). Also, genomovar TIGR has an insertion of 4 nucleotides (AGGG) upstream of *malF*, in the intergenic region between *malE* and *malF* that creates the DNA repeat AGGGAGGG (from +51 to +44) (Figure 5).

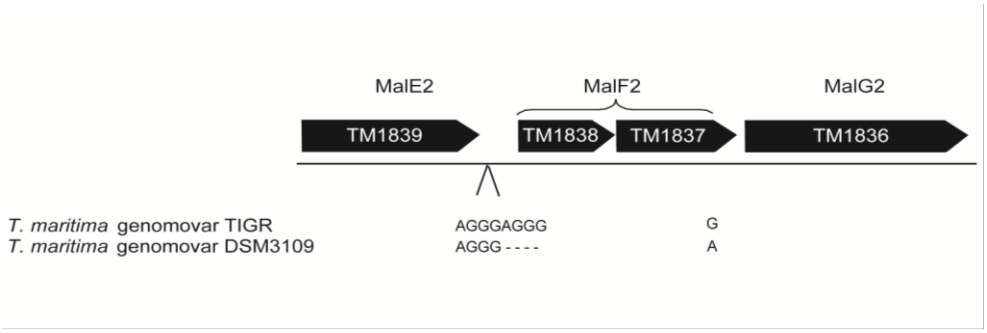


Figure 5. Schematic representation of the *mal2* operon as described in the 1999 annotation of the *T. maritima* MSB8 genome showing genetic variations found in this study in the two genomovars. The substrate-binding protein-encoding gene (*malE2,* TM1839) followed by the two genes coding for the transmembrane proteins (*malF2,* TM1838 and TM1837; and *malG2,* TM1836) are shown. Differences in the sequences between genomovars TIGR and DSM3109 were determined in this study using the sequences JF907621 and JF907622.

Discussion

The term genomovar was introduced by Rosselló in 1991 to describe different variants, or genomic groups, of *Pseudomonas stutzeri* (96). The term genomovar can be used for those strains exhibiting genomic variations with no detectable phenotypic differences that would allow them to be classified as separate species (97, 98). I have demonstrated that the *T. maritima* isolate used for the genome sequencing project published in 1999 (40) is a variant that lost 8,870 bp of genomic DNA between the loci TM1848 and TM1847 when compared to the isolate deposited in the DSMZ strain collection in 1986 (25). Both genomovars of *T. maritima* MSB8 have the same 16S ribosomal RNA gene sequence and no reported phenotypic differences. I propose to designate the type strain (DSM 3109) as *T. maritima* MSB8 genomovar DSM3109 and the sequenced variant as *T. maritima* MSB8 genomovar TIGR. The 8,870 bp region is largely syntenic with regions from other *Thermotoga* species that also have a *bglA* gene and a putative *tre* operon suggesting that *T. maritima* MSB8 genomovar TIGR lost these genes.

Recent up-dates: sequencing project of *T. maritima* MSB8 genomovar DSM3109

A sequencing project for *T. maritima* genomovar DSM3109 was initiated by the Joint Genome Institute (JGI) and was completed in 2011 (GenBank:AGIJ00000000.1). The deletion of the 10 kb region is the largest genetic event in genomovar TIGR. Other genetic variations were identified: mainly insertions, deletions and substitutions of a single nucleotide (Appendix 1).

In 2013, Latif *et al.* sequenced and annotated *T. maritima* MSB8 deposited at the American Culture Collection (ATCC) which was renamed as genomovar ATCC (GenBank: CP004077.1) (99). Both genomovars DSM3109 and ATCC contain the 10 kb region. This finding suggests that the deletion occurred while the strain was cultivated in Prof. Stetter's laboratory, likely after its deposition in the DSMZ and ATCC. The growth and storage condition of the genomovar TIGR are unknown and therefore it is difficult to speculate on the selection that could have led to the loss of the 10 kb region. The ABC transporters Mal1, Mal2 and Xyl1 in *T. maritima* have similar functions as the two ABC transporters (Tre and Xyl2) missing in the genomovar TIGR. The absence of the trehalose and xylose ABC transporters (Tre and Xyl2) is not detrimental and the redundancy likely allowed the genomovar TIGR to survive in laboratory conditions.

Other than the 10 kb region, there are 99 variations at the single-nucleotide level including many indels (Appendix 1). The variations are found among 58 genes and 11 intergenic regions (Appendix 1). The deletion of 4 nucleotides in the intergenic region between *malE2* and *malF2* has been validated in the genomovar DSM3109 (Figure 5) and is unique to the genomovar since the sequence of the genomovar ATCC published by Latif *et al.* does not contain the deletion. This is evidence of how rapidly small genetic differences can appear in laboratory strains. While one authentic single-nucleotide polymorphism and sequencing errors in the *mal2* operon of genomovar TIGR (Figure 5) and 16S rRNA gene (Appendix 1) were validated by me using Sanger sequencing, the other areas of variation were not resequenced in the genomovar TIGR and therefore it is difficult to determine if the variations are polymorphisms or sequencing errors that occurred in those genome sequencing projects.

The genomovar TIGR was sequenced using shotgun cloning and Sanger sequencing while the genomovars DSM3109 and ATCC were sequenced using pyrosequencing. The genomovar DSM3109 was sequenced using 454 and Illumina technologies while the genomovar ATCC was sequenced only using Illumina technology. The sequence coverage of the genomovars TIGR, DSM3109 and ATCC are 7-fold, 30-fold and 1700-fold, respectively. Coverage depth, sequencing and cloning biases might lead to different types of sequencing errors (100–102). Also, the fewer number of genome sequences in the database in 1999 might have affected the assembly, closure, and annotation processes. After the release of the sequence of the genomovar TIGR, approximately 31 genes were annotated by NCBI curators as containing an authentic frameshift or a point of mutation leading to early protein termination. The *malF2* is an example of one of those genes annotated as containing an authentic frameshift. However, the resequencing of the *malF2* from the genomovar TIGR and DSM3109 demonstrated that the frameshift was a sequencing error and that no frameshift is present in either genomovar.

Conclusion

The identification of the genomovar TIGR explains the discrepancies found between sequences from *Thermotoga maritima* MSB8. Previously, it was unclear why some variations were found between some sequences and sequencing errors alone could not explain all the discrepancies. The release of the sequence of the 10 kb region of the genomovar DSM3109 prompted Rodinov's research team to examine the function of the ROK transcription factor encoded by the locus TM1847. The protein was not initially investigated because the TIGR genomovar genome annotation suggested that the protein was truncated. Rodinov's laboratory found the binding sites of the transcription factor encoded by TM1847 (GluR) as well its transcriptional control mechanism. The results show that GluR binds two DNA sequences that are located upstream of *treE* and *xylE2* where it acts as a repressor. Glucose, an effector of GluR, prevents the repressor from binding to the binding sequences which allows the transcription of the operons (50). Without the sequence of the 10 kb region, those discoveries would have never been made.

The complete sequence of the genomovar DSM3109 is now available to the research community. Since then, the isolate deposited at the ATCC has been sequenced as well (99) and the sequence is also available through NCBI. Approximately a hundred variations were found between the genomovars TIGR, DSM3109 and ATCC. Some have been authenticated in this study. Additional sequencing is necessary to confirm the genetic variations observed between the genomovars to make sure that they are not from sequencing errors. The genomovar ATCC was sequenced using Illumina technology which can have an error-rate bias in homopolymer regions and in high- and low-GC

regions (103–105). The genomovar DSM3109 was sequenced using 454 and Illumina technologies. The use of two technologies should reduce the error-rate biases.

Since there are no genetic tools to manipulate *Thermotoga maritima*, these natural variants might be relevant for future studies if the variations are genuine and they are affecting the proteins' functions. For microbiologists, it is a reminder that strains cultivated in laboratories over time do not always maintain the genotype of the original isolate. As strains are cultivated and undergo multiple passages under different growth conditions, new populations arise, carrying new genotypes that can differ from one laboratory to another.

Chapter 3

Development of a screening assay to identify the ligands that interact with thermophilic substrate-binding proteins (SBPs)

Portions of this chapter were published in Boucher, N. and K. M. Noll. 2011. Ligand screening by differential scanning fluorimetry of thermophilic transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8. Appl. Environ. Microbiol. 77:6395-6399.

Introduction

Ligand-binding assays are useful to determine a protein's ligand specificity, function, and binding properties. A variety of techniques are available to identify protein-ligand interactions that are based on different modes of detection such as radioactivity, fluorescence or antibody binding. Some methods require the use of bioanalytical equipment to measure the changes of specific intrinsic properties of the protein upon ligand binding. Generally, ligand-binding assays require a large concentration of protein and are rather time consuming. Most of these techniques only detect the binding of one ligand at a time, which limits their use in annotating putative binding proteins in genome sequencing projects. This impedes our knowledge of their functions in different physiological pathways. In addition, the thermophilic nature of the proteins from organisms like the Thermotogales adds experimental challenges to the determination of their binding properties. To improve our knowledge of the functions of the substrate-

binding proteins (SBPs) encoded by the genomes of thermophiles, new techniques need to be developed to meet those challenges. I implemented a fast, high-throughput method using differential scanning fluorimetry (DSF) optimized to analyze thermophilic SBPs.

DSF assay as a screening method to detect ligand-binding interactions

DSF is based on the property that a protein bound to its ligand is more thermostable than the ligand-free protein (106, 107). The protein-ligand interaction can be detected by measuring the difference between the unfolding temperatures of the ligand-free and ligand-bound forms. In some respects, the technique can be compared to differential scanning calorimetry (DSC) which uses the same protein thermodynamic principle to detect ligand-binding. However, DSF measures the unfolding temperature curve using fluorescence from an extrinsic dye rather than measuring the release of energy from the protein as ligand is bound. Unlike DSC, the DSF assay is not based on an intrinsic property of the protein, therefore the technique is defined as an indirect ligand-binding assay. Thus the assay cannot be used to determine binding constants. DSF is suited to detect ligand-binding interactions qualitatively and therefore allows one to screen many potential ligands for those that can subsequently be used in methods employing measures of intrinsic binding properties.

DSF was developed by the pharmaceutical industry to identify targets for drugs (108). It is often referred to as thermal shift, thermal melt or ThermoFluor®. However, over the years, its application widened mainly because of recent improvements in fluorescence detection devices. DSF is a fast, easy and cost-effective method that requires low

amounts of proteins (~2 μg), which makes this assay very attractive as a screening technique.

Principle of the DSF assay

Amphiphilic dyes are necessary in DSF to detect protein unfolding since their fluorescence is quenched in water and increases when the dye binds to hydrophobic protein regions (109). The most commonly used dyes for DSF are 1,8-ANS (1-anilino-8-naphthalene sulfonate), bis-ANS (4,4'-dianilino-1,1'-binaphthyl-5,5'-disulfonate) and Sypro Orange. In recent years, Sypro Orange has been commonly used as extrinsic fluorescent dye in DSF assays because of its high signal-to-noise ratio (109) and because its fluorescence can be measured using a real-time PCR thermocycler. The optical properties of Sypro Orange are similar to SYBR green whose fluorescence is compatible with a filter (ex/em: 497/520 nm) commonly used with real-time PCR thermocyclers. Sypro Orange has a maximal excitation at 470 nm and maximal emission at 570 nm (110), with approximately 40-50% of its relative fluorescence at 520 nm.

In a DSF measurement, the protein sample (2-2.5 μg) and the fluorescent dye are heated to 98°C. As the temperature increases, the protein starts to unfold allowing the dye to bind to the exposed hydrophobic sites which causes an increase in the fluorescence intensity (108, 109, 111, 112) (Figure 6). At the unfolding temperature (Tm), the concentration of native protein and unfolded protein are equal ($[Protein_{native}]=[Protein_{unfold}]$).

After the protein unfolding curve is recorded, the protein melting temperature is calculated using a curve fitting to a Boltzmann function (109) (see Appendix 4). To fit

the Boltzmann equation, only the curve from the lowest point to the highest point from which the slope is derived from the fitted formula is used. A value, called ΔTm is used to express the ligand-induced protein thermostability. The ΔTm is the difference between the Tm in the presence of the ligand and in the absence of the ligand ($Tm_{ligand} - Tm_{free\ protein}$). A ΔTm close to zero indicates no thermal stability change in the presence of a ligand while a positive value indicates an interaction between the ligand and the protein. The ΔTm values do not necessarily reflect the strengths of ligand binding.

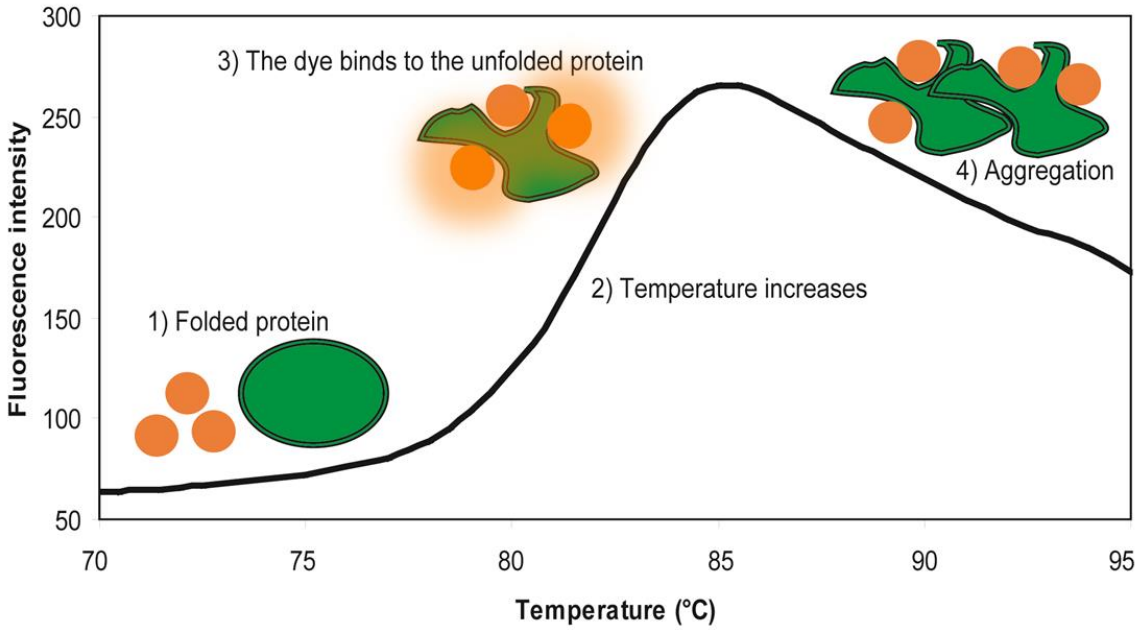


Figure 6. Illustration of the steps involved in the differential scanning fluorimetry (DSF) assay. 1) Protein sample and dye is mixed together at room temperature. 2) The sample is gradually heated up to 98°C. 3) When the dye binds to the hydrophobic regions of the unfolded protein there is an increase of fluorescence. 4) The protein sample aggregates leading to a decrease of fluorescence. Sypro Orange dye is represented as orange circles while the protein is represented in green.

Materials and Methods

Differential scanning fluorimetry (DSF) screening of ligand binding

The samples were mixed to a final volume of 20 µl at the following amounts unless otherwise indicated: 2.5 µg protein, 4 µl 5X citric acid buffer (pH 3.50-3.75, composition described in Appendix 2), 0.032 µl 5000X Sypro Orange (Invitrogen), 1.2 µl 2.5 M NaCl and 2 µl ligand, typically with a final concentration of 10 mM ligand. The final concentrations of cellobiose, mannan, pullulan and sorbitol were 25 µM, 0.225% (w/v), 0.5% (w/v) and 0.1% (w/v), respectively. The fluorescence intensities were measured using an iCycler IQ real time detection system (Bio-Rad) with excitation at 490 nm and emission at 530 nm. The samples were heated from 25ºC to 94ºC with a heating rate of 0.5ºC per min. All the assays were done in triplicate.

The midpoint temperature of the unfolding transition (Tm) was obtained with the program Gnuplot from curve fitting to the Boltzmann equation (109).

$$y = \mathrm{L} + \frac{U - L}{1 + \exp^{(\frac{\mathrm{Tm} - x}{a})}}$$

The value of *y* represents the fluorescence at temperature *x*. Where L, U and *a* are the minimum and maximum fluorescence intensities and the slope of the curve, respectively. To fit the Boltzmann equation, only the curve from the lowest point of inflexion to the highest point of inflexion from which the slope is derived from the fitted formula was used as described in Figure 6. The delta Tm (ΔTm) of the protein for a ligand was

calculated as the difference between the Tm of the ligand-bound and ligand-free protein. The complete results from the DSF are provided in Appendix 3.

Sugar purity

The purity of the sugars was at least 98% unless noted otherwise. The sugar molecules and the linkage of the disaccharides and oligosaccharides are summarized in the Appendix 4. Arabinose, melibiose, raffinose, cellobiose, sorbitol, tagatose, lactose, ribose, trehalose, maltose, L-fucose, xylose, fructose, galactose, L-rhamnose, myo-inositol, glucose, sucrose, mannan, mannose, α-1,4-maltotetraose (95%), and pullulan were obtained from Sigma-Aldrich. The xyloglucan oligosaccharide (95%), β-1,4-mannotriose (95%), and β-1,4-mannotetraose (95%) were supplied from Megazyme and the α-1,4-maltotriose (97%) from ICN.

Results

Effect of the change of pH and ligand concentration on the unfolding temperature

Preliminary testing of the DSF method was performed on thermostable proteins whose binding affinities were known. The recombinant maltose-binding proteins MalE1 and MalE2 from *T. maritima* that were previously characterized in our laboratory were chosen to test this assay (8, 93). Figure 7 shows the effect of the pH on the unfolding temperature curve of *T. maritima* MalE1. At pH values above 4 the unfolding temperatures of MalE1 could not be recorded with the thermocycler, but when the assays were carried out in citric acid buffers at pH values between 2 and 4, the unfolding temperature curves of were detectable. Figure 8 depicts the effect of acidic pH on the Tm of a thermophilic (*T. maritima*, MalE1) and a mesophilic protein (ManE, *Mesotoga prima*). For both mesophilic and thermophilic proteins, the effect of the pH is linear between pH 2 and pH 4. However, the effect of the pH plateaued at pH values above 5. pH values below 2 were not tested because it has been previously shown that proteins might be able to refold at such lower pH values (113).

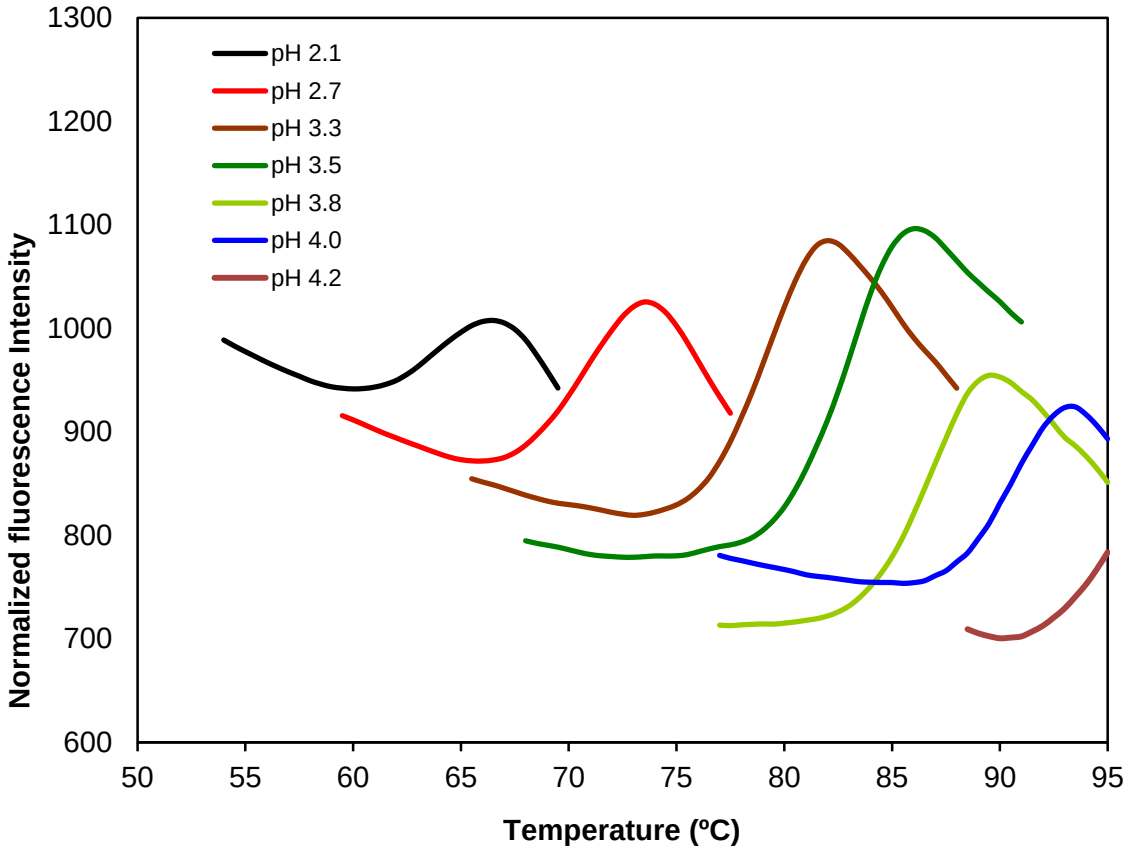


Figure 7. Effect of pH on the unfolding temperature curve of *T. maritima* MalE1 using DSF.

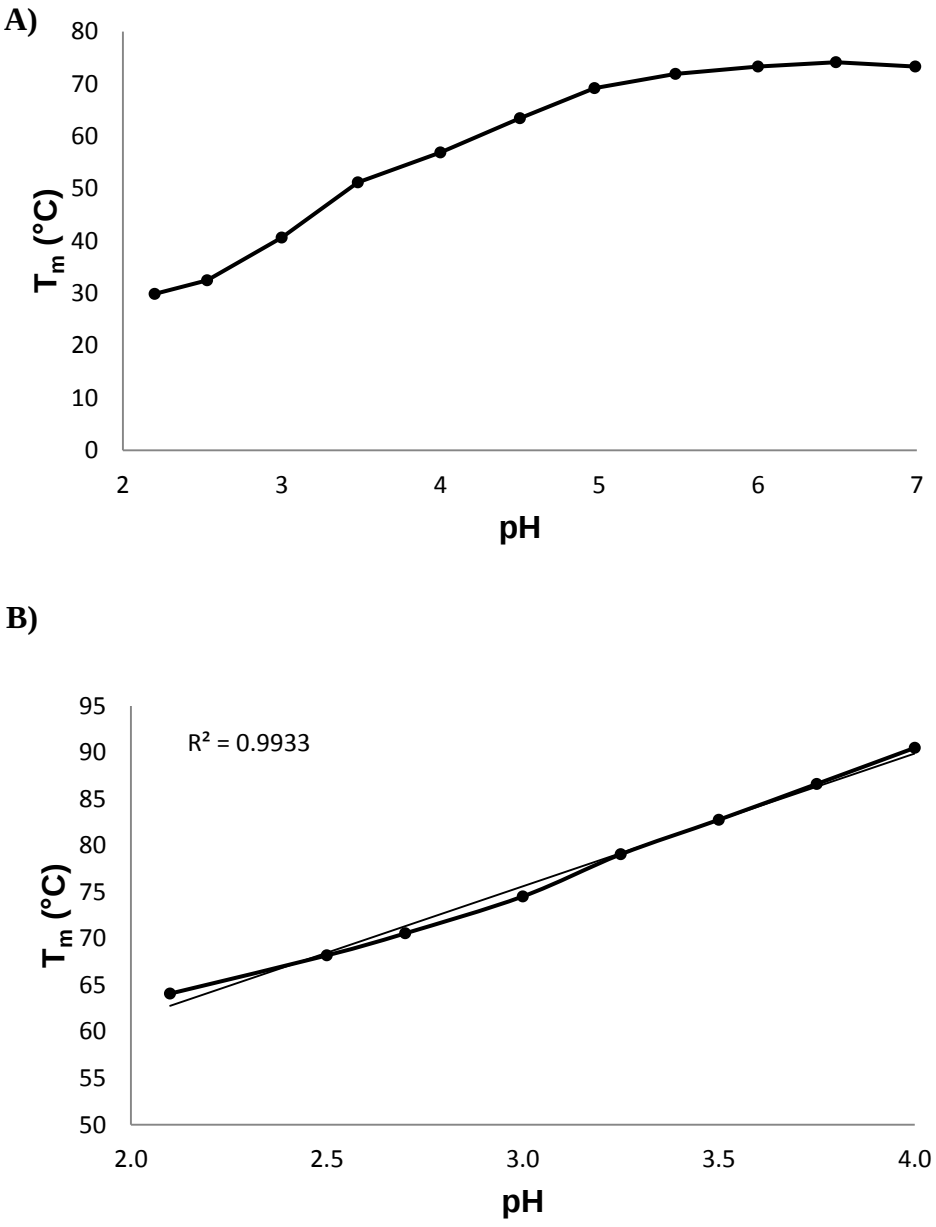


Figure 8. Effect of pH on the unfolding temperature (Tm) using DSF. A) A thermophilic protein (MalE1, *T. maritima*) (pH 2 to 4). B) A mesophilic protein (ManE, *Mesotoga prima*) (pH 2 to 7).

Under physiological conditions, the SBPs bind with high affinity to their ligands at low concentration (114). Giuliani *et al.* used a DSF assay to determine putative ligand interactions with mesophilic SBPs (115). Their ligand concentration optimization demonstrated that ligand concentrations of 100-fold higher than physiological concentrations were needed (500 to 1000 μM, with a ligand to protein ratio of 100) to detect ligand-protein interactions (115). Their DSF assay was carried out using ligand concentrations between 500-1000 μM at pH 7.5 (115). However, a DSF assay of thermophilic SBPs would need to be performed at lower pH values to detect the unfolding at temperatures below 98°C, so the optimal concentration of ligand under these conditions had to be assessed. Ligand titrations of maltose, maltotriose and maltotetraose were performed to determine the effect of the ligand concentration on the ΔTm of MalE1 and MalE2 under acidic conditions (Table 2 and Table 3). The Table 2 and Table 3 show that higher concentrations of ligand elicit higher ΔTm values (Table 2 and Table 3). Often concentrations of ligands above 2 mM were necessary to observe a ΔTm value above 2°C. This could suggest that at lower pH, the sugars have less contact with the proteins. Sugars that were shown in a previous study using intrinsic fluorescence not to bind to MalE1 and MalE2 do not increase the protein ΔTm significantly at 10 mM (Appendix 3) (93).

The effect of subtle pH changes on the ΔTm values were also examined at pH 3.0 and pH 3.5 (Figure 9). MalE1 thermostability apparently responds more dramatically to pH change than does MalE2 (Figure 9). A pH reduction from 3.5 to 3.0 led to a smaller ΔTm, which is more significant in the presence of maltose, showing a decrease from 1.8

to 1.1 (Figure 9). This highlights the importance of finding the optimal pH by performing a pH titration prior conducting the ligand screening.

Table 2. Effect of ligand concentration on ΔTm of MalE1 using DSF. The assay was performed in triplicate (standard deviations are shown).

	ΔTm (°C)		
Concentration (μM)	**Maltose**	**Maltotriose**	**Maltotetraose**
no ligand	0.0 ± 0.1	0.0 ± 0.1	0.0 ± 0.1
3.2	0.1 ± 0.1	0.3 ± 0.2	0.8 ± 0.1
16	0.2 ± 0.2	1.2 ± 0.1	23.0 ± 0.1
80	0.2 ± 0.1	3.0 ± 0.2	5.6 ± 0.1
400	0.7 ± 0.1	5.0 ± 0.1	7.8 ± 0.2
2000	2.3 ± 0.1	7.3 ± 0.1	-[a]
10000	4.7 ± 0.3	9.9 ± 0.1	-[a]

[a] No unfolding curve was observed

Table 3. Effect of ligand concentration on ΔTm of MalE2 using DSF. The assay was performed in triplicate (standard deviations are shown).

	ΔTm (°C)		
Concentration (μM)	**Maltose**	**Maltotriose**	**Maltotetraose**
no ligand	0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.2
3.2	0.2 ± 0.4	0.0 ± 0.1	0.2 ± 0.4
16	0.4 ± 0.4	0.2 ± 0.2	1.1 ± 0.9
80	0.2 ± 0.1	0.5 ± 0.2	2.1 ± 0.8
400	1.1 ± 0.2	1.2 ± 0.1	3.6 ± 0.4
2000	1.8 ± 0.6	3.0 ± 0.1	5.5 ± 0.5
10000	2.4 ± 0.8	5.5 ± 0.2	-[a]

[a] No unfolding curve was observed

DSF analysis of ligand binding of MalE1 and MalE2

MalE1 binds maltotriose, maltose, and mannotetraose with respective K_d values of 0.008 μM, 24 μM, and 38 μM (8, 93). At 10 mM sugar, maltotriose, maltose, and mannotetraose increased the thermal stability of MalE1 with ΔTm values of 10.8°C, 4.5°C, and 4.2°C, respectively.

MalE2 binds maltose and maltotriose with respective K_d values of 8.4 μM and 11 μM, but not mannotetraose (8, 93). At 10 mM sugar maltose and maltotriose increased the thermal stability of MalE2 with ΔTm values of 1.0°C and 6.9°C, respectively while the ΔTm value in the presence of 10 mM mannotetraose was calculated to be 0.2°C. However, trehalose did not increase the thermal stability of MalE2 (ΔTm of -0.1°C), though the protein was reported to bind this sugar (K_d, 9.5 μM) (8, 93, 94).

Interestingly, maltotetraose stabilized both MalE1 and MalE2 with ΔTm values of 11.8°C and 6.9°C, respectively. This sugar was not previously tested by spectroscopy. No change in the thermal stability of MalE1 and MalE2 was recorded in the presence of other sugars. These results are summarized in Table 4 and the complete data set is located in Appendix 3.

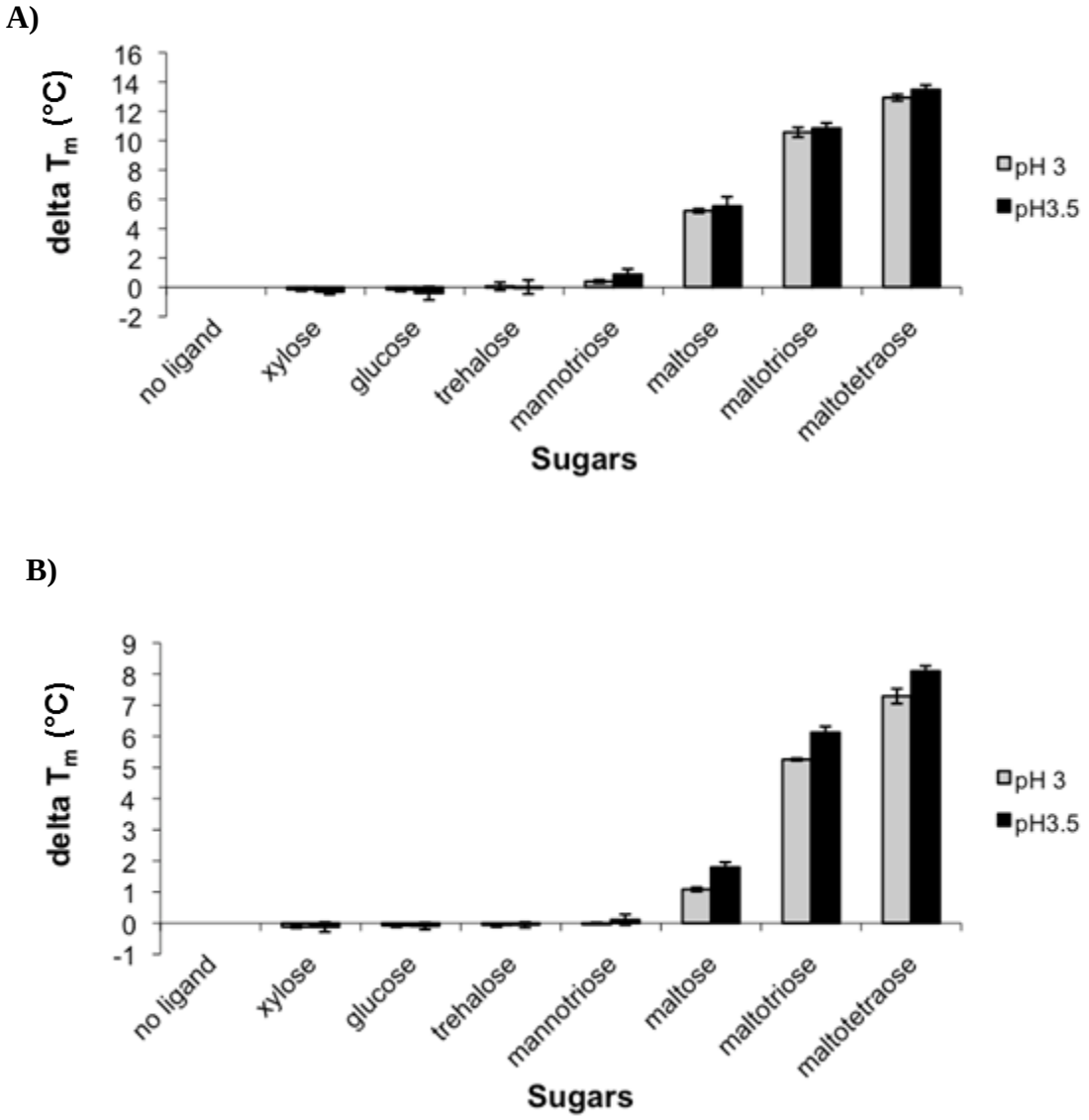


Figure 9. Effect of the pH on the ΔTm using DSF on MalE1 and MalE2. A) MalE1 and B) MalE2. The assay was performed triplicate and the error bars represent the standard deviation.

Table 4. Summary of the binding properties and ligand induced thermostabilities of MalE1 and MalE2. The binding constants (K_d) determined by intrinsic fluorescence are from Nanavati *et al.* (8, 93) while the DSF results are from this study. The DSF assay was performed in triplicate and the standard deviations are shown.

Sugars	**MalE1**		**MalE2**	
	K_d (μM)	DSF (ΔTm)	K_d (μM)	DSF (ΔTm)
maltose	24	4.5 ± 0.1	8.4	1.0 ± 0.3
maltotriose	0.008	10.8 ± 0.5	11	6.9 ± 0.2
maltotetraose	n.d.	11.8 ± 0.2	n.d.	8.7 ± 0.4
mannotetraose	38	4.2 ± 0.1	-	0.2 ± 0.3
trehalose	-	0.2 ± 0.1	9.5	0.3 ± 0.2
glucose	-	0.0 ± 0.1	-	-0.1 ± 0.2
xylose	-	0.2 ± 0.3	-	-0.1 ± 0.2

(+) protein-ligand interaction, (-) no interaction, (n.d.) not determined in the study

Optimal conditions

The optimal conditions were determined based on the effect of the pH and the ligand concentration on the unfolding temperature curve. For thermophilic proteins, the assay can becarried out in a volume of 20 µl containing 2.0-2.5 µg of protein, 8X Sypro Orange, 150 mM NaCl, 1X citric acid buffer (made as described in Appendix 2 at the appropriate pH and 10 mM ligands. The effect of 150 mM, 500 mM and 1000 mM NaCl was tested. The salt concentration is indirectly proportional to the Tm value (MalE2 at pH 4.4; 150 mM: 91.8 °C, 500 mM: 89.1 °C, 1000 mM: 86.7 °C). Although high salt concentration reduces the Tm value, 150 mM NaCl was used because this concentration is typical for many ligand-binding assays. As described previously, 10 mM ligand does not appear to increase the ΔTm of a ligand known to not bind to a protein while that concentration allows detection of protein interaction with a known ligand in an acidic environment. For maximal ΔTm, the pH should be as close as possible to a physiological pH (pH 7). For a ligand-free protein, I sought to use a Tm between 80-85 °C. Within this range, the protein's unfolding temperature curve in the presence of a ligand that causes an interaction will likely be detected.

Discussion

Adaptation of the DSF assay for use with thermophilic SBPs

DSF was previously used to identify ligands of SBPs from mesophilic organisms (115, 116). However, protein unfolding temperatures (Tm) from hyperthermophilic organisms, such as *Thermotoga maritima*, can be above 100°C (95, 117, 118) which is outside the heating capability of a real-time PCR thermocycler.

The challenge was to find a protocol that mildly denatures the protein to reduce its Tm without 1) compromising the interaction of the protein with its ligand and 2) impairing the fluorescence of the dye or its detection. Mild denaturing agents such as urea and guanidine hydrochloride were tested. Guanidine hydrochloride is commonly used in DSC to reduce the Tm of a protein. However, these denaturants were found to be incompatible with either the reagents or with the equipment. In the presence of guanidine hydrochloride, the Sypro Orange dye did not fluoresce while with urea, ammonia fumes can be produced and damage the thermocycler.

Low pH places the protein into a molten globule state, a third conformational state between the native and unfolded states (119–121). When a protein is in a molten globule state, the protein retains its secondary structures, but is not as tight as the native state. Low pH treatment was previously used to analyze different conformational protein states and the unfolding dynamics of the maltose binding protein of *Escherichia coli* (122, 123). That study showed that a molten globule state produced by low pH treatment reduces the unfolding temperature (Tm). The molten globule state offers an additional advantage over a partially denatured protein since the secondary structures are maintained in the

molten globule state as they are in the native protein, which likely maximizes the points of contact between the putative ligand and the protein.

Some sugars are known to prevent protein aggregation and can increase the stability of a protein without necessarily interacting with the binding pocket (124, 125). However, our data using MalE1 and MalE2 recombinant proteins indicate that there is no significant increase in the protein stability in the presence of a sugar that is not a ligand as demonstrated by measures of changes in intrinsic fluorescence. The data shown in Table 4 demonstrates that sugars that do not bind to MalE1 and MalE2 have a ΔTm less than 0.2 °C (for example MalE2 and mannotetraose). The stability of MalE2 was not increased in the presence of trehalose although this sugar was shown to bind to the protein using fluorescence spectroscopy at 20°C and pH7. This is the only false negative result I observed. Trehalose is stable at high temperatures and low pH (126), so hydrolysis does not account for our observation. It is likely that the tertiary structure of MalE2 might not offer sufficient contacts between the sugar and the binding pocket to increase its stability in the presence of trehalose.

Conclusion

Despite the one false negative result, the ΔTm values determined by DSF generally demonstrate an interaction between the protein and its ligand. Overall, DSF measurements performed under acidic conditions provide a robust method to screen sugars that find those that interact with thermophilic SBPs. The assay can be used to screen hundreds of putative ligands in less than a day. This assay can dramatically improve our knowledge of the function of the SBPs encoded in the genomes of thermophilic organisms.

The assay was optimized to screen carbohydrates and sugar alcohols. However, our laboratory successfully used this assay to determine the putative interactions between *Thermotoga lettingae* and *Fervidobacterium nodosum* BtuF proteins with B12 and cobinamide (127). Thus this can be a generally applied method to screen the ligand specificities of thermostable proteins.

In addition to screening for putative ligand interactions, DSF can used to measure the unfolding temperature (Tm) of thermophilic proteins. Traditionally, ITC is the method of choice to determine the Tm of thermophilic proteins using guanidine hydrochloride as a mild denaturant. Our laboratory successfully performed DSF to determine the Tm of *myo*-inositol phosphate synthase (MIPS) from *Pyrococcus furiosus*, which grows at temperatures up to 103°C (128, 129).

DSF was successfully optimized to determine the interactions between thermophilic proteins and carbohydrates and sugar alcohols. However, the ABC system transporters can take up a broad variety of ligands such as metals, vitamins, amino acids, and dipeptides. Giuliani *et al*., included in their ligand library compounds other than

carbohydrates such as dicarboxylic acids, aromatic acids, medium and long chain of fatty acids, amino acids, di-tri-tetrapeptides, polyamines, vitamins and metals in order to determine their interactions with mesophilic proteins (115, 116). Butzin *et al.* performed the DSF assay on BtuF using B12 and cobinamide as ligands using acidic environment (127). However, other types of ligand were not tested using DSF in acidic environment. In addition, other than substrate-binding proteins, other types of proteins such as enzymes and transcription factors have not been tested. The DSF assay could be useful to study these types of proteins and other types of ligands. However, additional work is necessary to determine if acidic environments can be used with each type of protein and ligand.

This assay can easily be used for high-throughput screens. The Protein Structure Initiative (PSI) is a collaborative project to determine protein crystal structures from representatives of the different protein families (130). *T. maritima* proteins were targeted for the initiative at the Joint Center for Structural Genomics (131). A protein library containing clones that express products from many *T. maritima* genes were constructed (personal communication, Dr. Kenneth M. Noll). A DSF assay was attempted to determine the putative interactions with these gene products, but the results were not conclusive (personal communication from Dr. Kenneth Noll). That screen may have failed if the proteins were not sufficiently denatured to allow ligand binding to be observed. The use of low pH conditions as described here might allow such high-throughput screens to succeed. Also, any protein libraries constructed from other thermophile organisms can be screened by DSF to determine the putative ligand interactions. With the new optimized conditions, the DSF assay opens the door to high-throughput screening which might unveil not only unknown protein-ligand interactions

but provide an overall understanding of the physiology and metabolism of thermophilic organisms.

Chapter 4

Characterization of the ABC-transporters located in the newly sequenced 10 kb region of *Thermotoga maritima* MSB8 genomovar DSM3109

Portions of this chapter were published in Boucher, N. and K. M. Noll. 2011. Ligand screening by differential scanning fluorimetry of thermophilic transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8. Appl. Environ. Microbiol. 77:6395-6399.

Introduction

Previously, a newly discovered region of 8870 bp in the genome of *Thermotoga maritima* MSB8 genomovar DSM3109 was sequenced revealing two operons encoding two ABC transporter systems. Phylogenetic analyses indicated that the region encodes a putative third maltose ABC transporter and a putative second xylose ABC transporter. The genome of *T. maritima* MSB8 was annotated to contain two maltose ABC transporters (*mal1* and *mal2*), both of which were later shown to encode maltose binding proteins (8, 40, 93). These two proteins (MalE1 and MalE2) are more related to the bacterial maltose-binding protein while the putative third maltose ABC transporter is more related to the archaeal trehalose/maltose ABC transporter (TMBP) from *Thermococcus litoralis* and *Pyrococcus furiosus* (16). Consequently, to reflect the phylogenetic relation between the putative Mal3 and the archaeal homologs, the putative third maltose ABC transporter system is renamed as the trehalose ABC transporter (Tre)

throughout this thesis. The second putative xylose ABC transporter (Xyl2) is distantly related to the characterized Xyl1 which binds xylose and glucose (8, 132). Interestingly, the gene organization in the vicinity of these two ABC transporter operons in the *Thermotoga* species shows that those regions have sustained differential gene loss. *T. maritima* MSB8 genomovar TIGR, a laboratory variant of genomovar DSM3109 that was deposited in the Deutsche Sammlung von Mikroorganismen und Zellkulturen, lost both operons. However, these non-essential genes were retained in *T. maritima* MSB8 genomovar DSM3109.

To understand the function of these operons, the genes encoding the substrate-binding proteins (SBPs) of each operon were cloned and expressed in *E. coli*. The binding properties of the SBPs were examined using differential scanning fluorimetry (DSF) to determine their binding specificities and fluorescence spectroscopy to determine their binding affinities based on the intrinsic fluorescence changes of their aromatic amino acids. The study of the function of these newly discovered putative maltose ABC transporters in *T. maritima* might improve our understanding of why some *Thermotoga* species retained these ABC transporters.

Materials and Methods

Cloning, expression and purification of TreE and XylE2

The putative SBP-encoding genes *treE* and *xylE2* from *T. maritima* MSB8 genomovar DSM3109 were amplified using IDProof Taq DNA polymerase enzyme kit (ID Labs Biotechnology, Inc.). Primers containing an NdeI restriction site were used for *treE* amplification (5´-CATATGAAAATCACTATGACATCTGGAGGGGT-3´, 5´-CATATGTTACTGTCCAAGCAGGAATTTGAGC-3´) and a primer containing NdeI and BamHI restriction sites were used for *xylE2* amplification (5´-CATATGGAGGA CATGACAATACTTTTGGCACCG-3´, 5´-GGATCCTCAGCGCTTGGAATAGATCT TCTTTGC-3´). The primers were designed to exclude the first 60 nucleotides of the ORFs that encode the predicted signal peptides (SignalP 3.0) (133). The PCR products were ligated into the pGEMeasy vector (Promega) and transformed in *E. coli* JM109 competent cells (Promega). The plasmids were digested with the appropriate restriction enzymes and the digested fragments were ligated into a pET15b vector digested with the appropriate restriction enzymes. The resulting pET15b-*treE* and pET15b-*xylE2* plasmids were transformed into *E. coli* BL21-CodonPlus (DE3)-RIPL competent cells (Statagene). The clones were validated by sequencing.

The proteins were extracted and purified according to the following protocol for ligand screening using DSF and fluorescence spectroscopy. 500 µl from a 5 ml of a culture grown in LB medium containing ampicillin (50 µg/ml) was inoculated into 50 ml of the same medium, grown with shaking at 37°C for 4-5 h, and then induced with 1 mM isopropyl-ß-D-thiogalactopyranoside (IPTG) for 16 h at 24ºC. The cells were harvested

by centrifugation at 2,400 x g and lysed using the B-PER bacterial protein extraction kit (Pierce). The overexpressed His-tagged proteins were purified with a nickel-Sepharose high performance column (GE) according to the manufacturer's protocol. The recombinant TreE tended to aggregate, especially at 4ºC. This phenomenon was also observed for the related *Thermococcus litoralis* TMBP (134). To prevent this, the cell lysate containing TreE was treated with 8 µl of RNAse A (Qiagen) and DNAse 1 (1 unit/µl, Pierce) per gram of cell paste (134). The proteins were dialyzed 3 times against 667 ml buffer (20 mM sodium phosphate, pH 7.4) for at least 3 hours at room temperature.

For the pH titration experiment, the proteins were expressed with a fast induction protocol. The cells were grown as described above to an optical density (OD_{600}) between 0.3 and 0.6, and then induced with 1 mM IPTG for 2 h at 37ºC. The cells were pelleted and lysed using the protocol described above. The proteins were concentrated using a 30,000 NMWL centrifugal filter unit (Millipore) at room temperature and washed 3 times with 5 mL of buffer (20 mM sodium phosphate pH 7.4, and 500 mM NaCl).

Differential scanning fluorimetry (DSF) screening of ligand binding

The DSF assays were carried out as described in Chapter 3 and Appendix 2.

Intrinsic fluorescence spectroscopy

All fluorescence measurements were performed using an SLM Aminco-Bowman 2 spectrofluorimeter. The protein samples were incubated at 60°C and the temperature of the cuvette with the sample was equilibrated at 60°C for 5 min. Fluorescence emission

spectra were measured at an excitation wavelength of 280 nm for XylE2 and 295 nm for TreE and the emission intensities were measured over the wavelength range of 300 to 400 nm. The dissociation constants were measured by adding increasing amounts of selected carbohydrates into a stirred cuvette at 60°C. In a typical experiment, 3 µl of different stock solutions of the carbohydrate were added to 1.5 ml of 0.45 µM or 0.08 µM protein solution (20 mM sodium phosphate, pH 7.4). After the addition of the sugar, the sample was stirred for 2 min to reach equilibrium and then the fluorescence intensity at the predetermined emission wavelength was recorded for 45 s.

The binding affinities (K_d) of TreE for trehalose and XylE2 for glucose and xylose were obtained from curve fitting using KaleidaGraph to the equation accounting for ligand depletion:

$$F = F_0 + \Delta F/2[Pt] ((K_d + [Pt] + [Lt]) – ((K_d + [Pt] + [Lt])^2 – 4[Pt] [Lt])^{1/2}$$

where F is the measured fluorescence at ligand concentration [Lt] and [Pt] is the total protein concentration (135, 136). This formula was used since the protein concentration was greater than the measured K_d. If the protein concentration was less than the measured K_d, K_d values were obtained from curve fitting using GraphPad Prism to the equation:

$$F= (Fmax * [L]) / (K_d + [L])$$

where F is the measured fluorescence at ligand concentration [L].

Phylogenetic tree for XylE2

The data set was assembled by searching 575 prokaryotic genomes using (TblastP) with TM0114 as a query and retrieving all hits below an E value of $10e^{-20}$. Sequence alignment was done using Muscle v4.0.170 (137, 138) and the maximum likelihood tree

was calculated with RAxML version 7.0.4 (139), gamma with four categories and WAG substitution model.

Results

Ligand stabilization of TreE and XylE2

To determine the protein-ligand interactions of the newly identified substrate-binding proteins, the genes encoding TreE and XylE2 were expressed in *E. coli*. For each protein 26 sugars were screened using DSF. The complete data are listed in Appendix 3. The thermal stability of TreE was increased the most in the presence of 10 mM trehalose and maltose while 10 mM glucose and xylose increased the thermal stability of XylE2 the best (Figure 10). However, the upper values of the unfolding transition curves were above 94°C and so the exact ΔTm could not be determined. Sugar titrations of maltose and trehalose were made for TreE as well as glucose and xylose for XylE2 (Table 6) to determine the exact ΔTm using a known concentration of sugar. The sugar titrations suggest that TreE interacts the most with maltose while XylE2 interacts the most with glucose.

In addition to trehalose and maltose, TreE was stabilized by 10 mM sucrose, maltotriose, maltotetraose, and glucose with ΔTm values of 11.4°C, 7.6°C, 4.7°C, and 3.7°C, respectively (Figure 10, Appendix 3). In addition to xylose and glucose, XylE2 was stabilized by 10 mM sucrose, 10 mM L−fucose, 2.5 mM cellobiose, and 10 mM myo-inositol with ΔTm values of 10.9°C, 8.8°C, 6.0°C, and 5.4°C, respectively (Figure 10, Appendix 3).

Binding affinities of TreE and XylE2

The binding affinities of TreE and XylE2 were determined by fluorescence spectroscopy. To determine which sugars to use in these measurements, the ΔTm values determined by DSF were ranked from highest to lowest and the K_d values of sugars were measured starting from the top of the list until the values were greater than 10 μM. Fluorescence emission spectra were measured at an excitation wavelength of 280 nm and 295 nm to monitor aromatic residues and tryptophan, respectively. The K_d values were determined at 60°C, the highest temperature that the instrument could reach.

TreE binds trehalose, sucrose, and glucose with K_d values of 0.024 μM, 0.300 μM, and 56.78 μM, respectively (Table 7). Trehalose elicits an increase of fluorescence of approximately 9% (Figure 11). No fluorescence changes were observed with maltose (Figure 12) or maltotriose. However, after the addition of 1 mM maltose, subsequent addition of 1 mM trehalose did not elicit any change of fluorescence suggesting that TreE was already saturated with maltose (Figure 12).

XylE2 is a member of the CUT2 family, which generally transports only monosaccharides (5, 140). This protein binds glucose, xylose, and L-fucose with K_d values of 0.059 μM, 0.042 μM, and 1.436 μM, respectively (Table 7). Although sucrose, myo-inositol, and cellobiose stabilized XylE2 with ΔTm values of 10.9°C, 5.4°C, and 6.0°C, respectively (Appendix 3), the K_d values were very high at 116.4 μM, 56.47 μM, and 3.955 μM, respectively. It is likely that these high K_d values are due to binding by other sugars that were present as impurities in these sugars. The purity of the sucrose, myo-inositol and cellobiose used in these experiments is 98%, which can translate to 1 nM of impurities for every 50 nM of the sugar. The fact that XylE2 binds glucose at high

affinity (K_d: 90 nM) means that only 4.5 μM of sugar containing 2% impurities would be yield the observed binding curve.

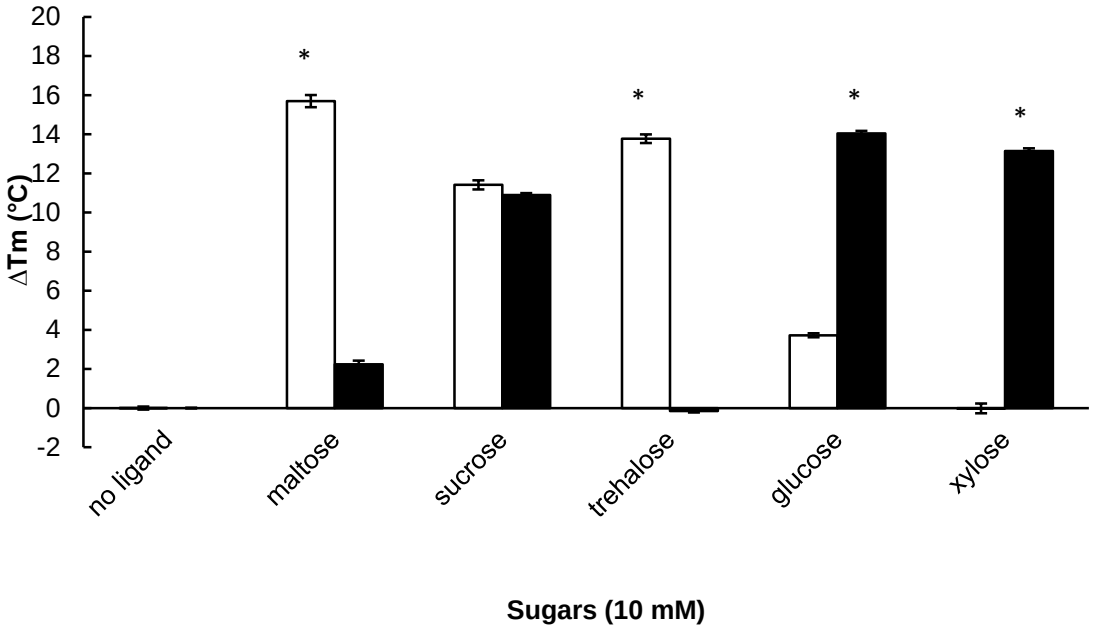


Figure 10. ΔTm of *T. maritima* TreE and XylE2 determined by DSF. TreE (white), XylE2 (black). (*) Since the upper limits of the melting curve exceeded the temperature limits of the thermocycler, the actual ΔTm is greater than or equal to the value indicated. The assay was performed triplicate and the error bars represent the standard deviations.

Table 5. ΔTm values of TreE using a titration with trehalose and maltose. The DSF assay was performed in triplicate (standard deviations are shown).

	ΔT_m (°C)	
Concentration (μM)	**Trehalose**	**Maltose**
no ligand	0.0 ± 0.2	0.0 ± 0.1
3.2	4.5 ± 0.3	5.8 ± 0.2
16	7.0 ± 0.2	8.6 ± 0.2
80	9.3 ± 0.2	11.2 ± 0.2
400	11.6 ± 0.2	13.7 ± 0.2
2000	14.0 ± 0.2	-[a]
10000	-[a]	-[a]

[a] No unfolding curve was observed

Table 6. ΔTm values of XylE2 using a titration with glucose and xylose. The DSF assay was performed in triplicate (standard deviations are shown).

	ΔT_m (°C)	
Concentration (μM)	**Glucose**	**Xylose**
no ligand	0.0 ± 0.2	0.0 ± 0.2
3.2	0.9 ± 0.1	0.5 ± 0.2
16	2.9 ± 0.3	1.8 ± 0.2
80	5.9 ± 0.2	3.8 ± 0.2
400	9.3 ± 0.5	6.8 ± 0.2
2000	12.4 ± 0.4	10.1 ± 0.2
10000	-[a]	13.5 ± 0.2

[a] No unfolding curve was observed

Table 7. Apparent binding affinities (K_d) of TreE and XylE2 measured at 60°C. K_d values were measured by changes in intrinsic fluorescence upon ligand binding.

Protein	**Sugar**	**K_d values at 60°C (µM)**
TreE	trehalose	0.034
	sucrose	0.300
	glucose	56.78
XylE2	glucose	0.059
	xylose	0.042
	L-fucose	1.436

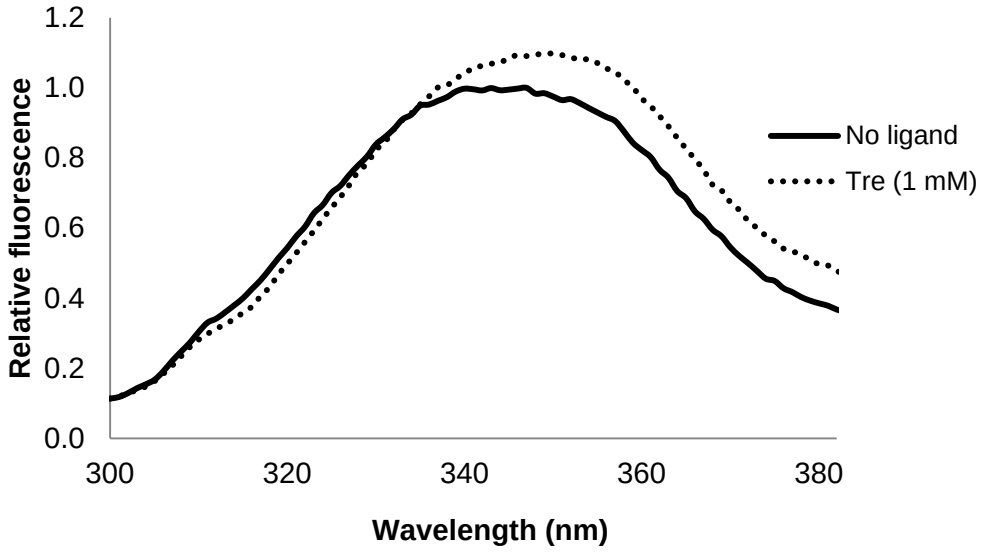


Figure 11. Emission spectra of *T. maritima* TreE with and without trehalose at 60°C.

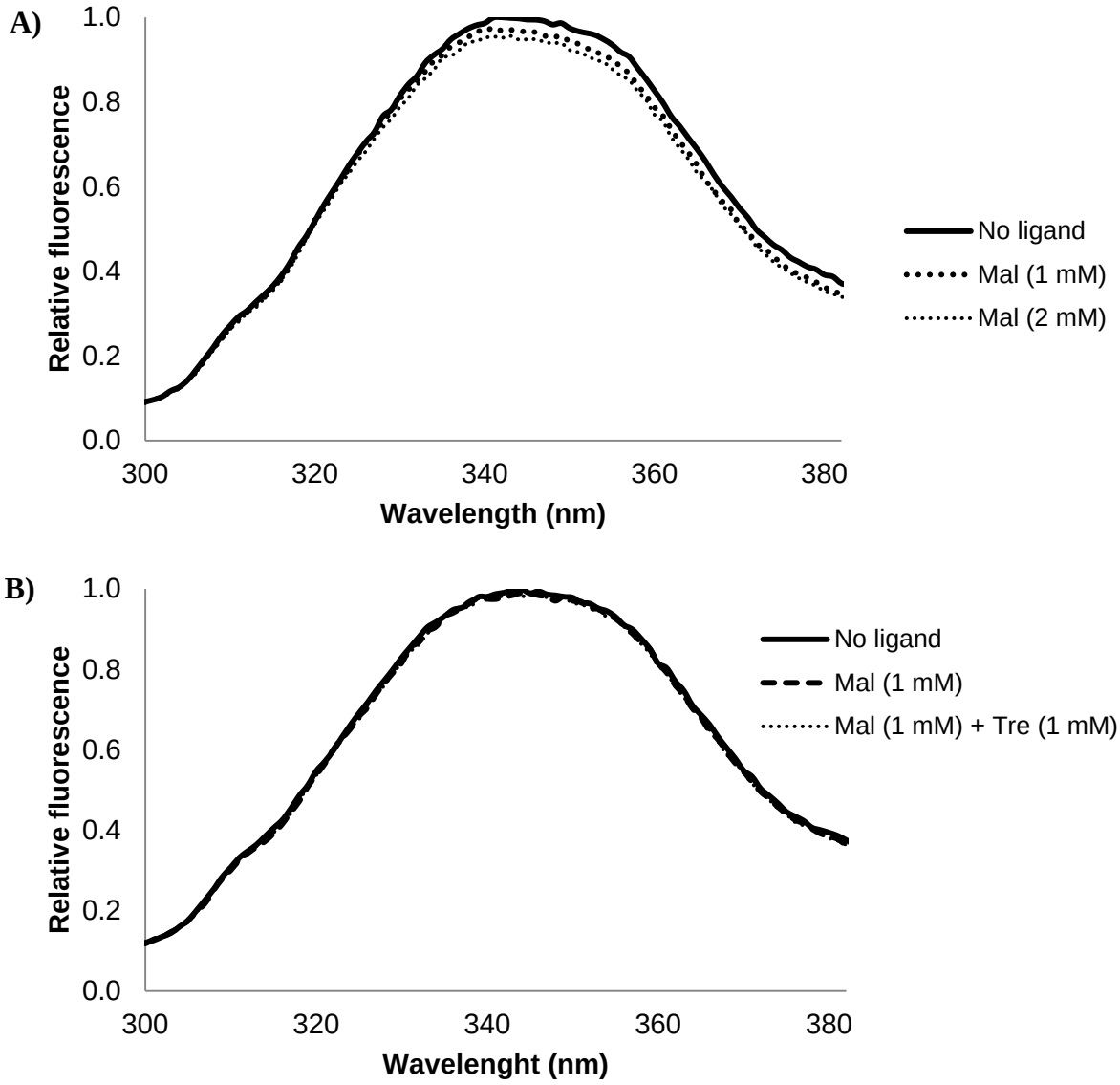


Figure 12. Emission spectra of *T. maritima* TreE at 60°C. A) With and without the addition of maltose (1 mM and 2 mM). B) With and without the addition of maltose (1 mM) and subsequent addition of trehalose (1 mM).

Discussion

The 8.9 kb region of *T. maritima* MSB8 genomovar DSM3109 contains two putative operons, *tre* and *xyl2*, encoding two carbohydrate ABC transporter systems. The *tre* operon encodes a periplasmic substrate-binding protein (TreE) and two transmembrane proteins (TreF and TreG). The *tre* operon is distantly related to the *T. maritima mal1* and *mal2* operons (16). At 20°C, MalE1 has the highest binding affinity for maltotriose (0.008 μM) while MalE2 binds maltose the best (8.4 μM) (8, 93). TreE was demonstrated to have the highest binding affinity for trehalose (0.024 μM). Both MalE2 and TreE bind trehalose; however, their relative binding affinities cannot be compared because their K_d values were determined at different temperatures (20°C and 60°C). The family of thermostable Mal proteins that includes the *T. maritima* Tre proteins is closely related the trehalose/maltose ABC transporter systems of *Thermococcus litoralis* and *Pyrococcus furiosus* (Figure 3) that bind maltose and trehalose (83, 134). *T. litoralis* TMBP binds trehalose and maltose with a binding affinity of 160 nM at 85°C (134). Although the binding affinities for these sugars were never determined for the *P. furiosus* TMBP, its sequence is identical to that of the *T. litoralis* TMBP (141), suggesting they have similar binding affinities. Unlike MalE1 and MalE2 (8, 93, 94), TreE recognizes sucrose and glucose, though it binds the latter with very low affinity (Table 7). Some maltose binding proteins in other organisms also have affinities for either sucrose or glucose. For example *Thermus thermophilus* MalE1 (TTC_1627) binds sucrose, maltose, and trehalose with K_d values of 401 nM, 103 nM, and 67 nM, respectively, at 70°C (its optimal growth temperature) (142). At 60°C, the binding affinities for sucrose and trehalose were calculated for *T. maritima* TreE and they were 300 nM and 24 nM,

respectively (Table 7). Although *T. thermophilus* MalE1 is a member of the thermostable Mal3 and Tre clade (Figure 4), it does not bind glucose (142) like *T. maritima* TreE does (K_d 56.78 μM at 60°C). Also, Herman *et al.* showed that glucose binding by the *T. litoralis* TMBP is modulated by temperature. The K_d values for glucose were 20 μM at 25°C and 40 μM at 60°C, while no binding was observed at 85°C (143). Taken together, these results suggest that glucose is not a physiologically relevant ligand of the trehalose and maltose binding proteins.

The second putative operon found in *T. maritima* MSB8 genomovar DSM3109 was designated as *xyl2* as it codes for a periplasmic substrate-binding protein (XylE2), a transmembrane protein (XylF2), and an ATP-binding protein (XylK2). The substratebinding protein XylE2 is phylogenetically related to TM0114 (XylE1), but their precise evolutionary relationship to one another cannot be resolved with confidence (Figure 13). It is interesting that both proteins bind glucose and xylose (8, 132), but XylE2 binds L−fucose at 60°C. This could illustrate another example of evolutionary divergence of a periplasmic substrate binding protein. However, to compare the functional divergence of XylE1 and XylE2 their respective binding affinities for L−fucose have to be compared at the same temperature and therefore, more analysis is required.

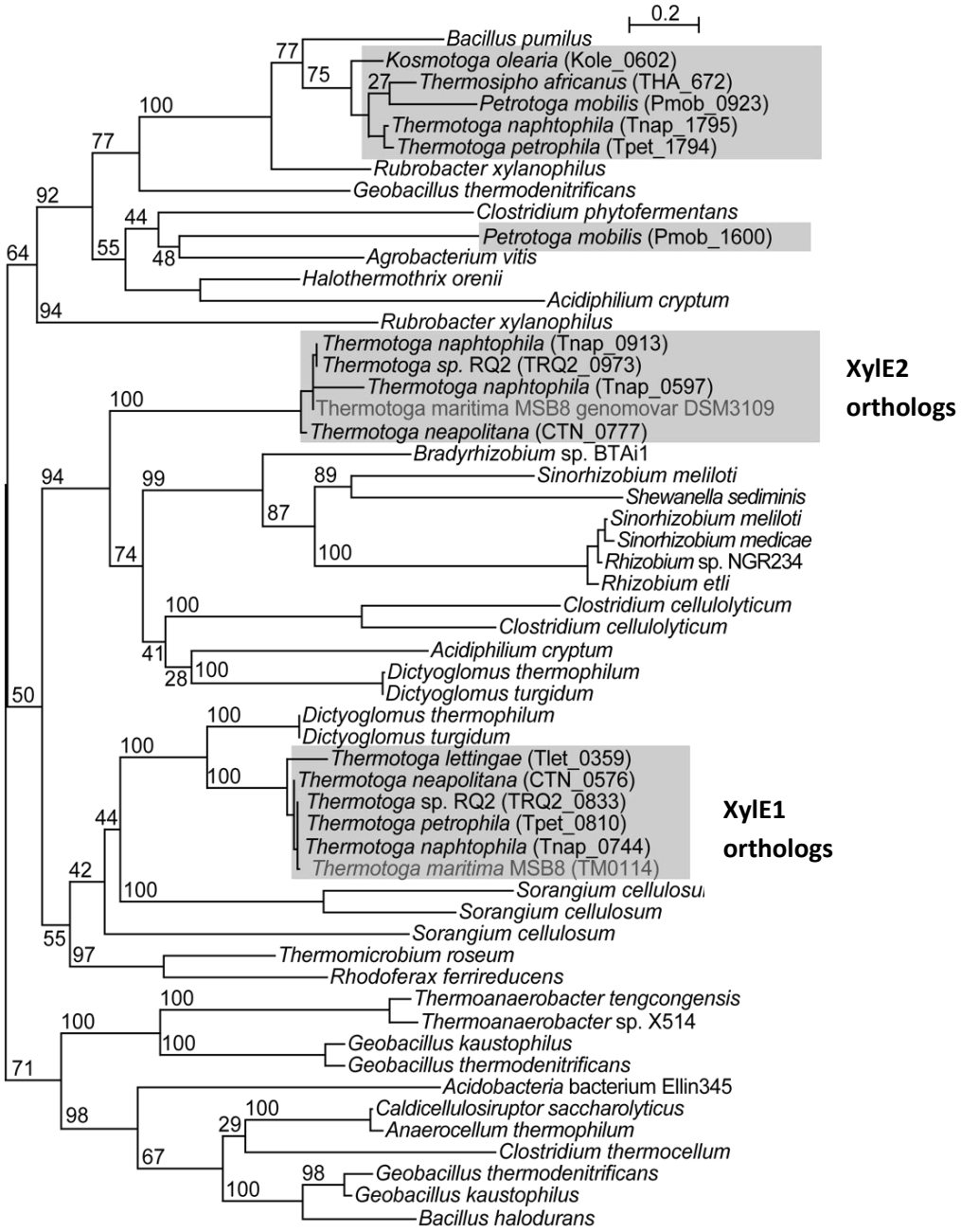


Figure 13. Unrooted maximum likelihood tree of substrate-binding proteins homologous to TM0114 depicting the relationships among the xylose-binding proteins. The scale bar indicates an evolutionary distance of 0.2 amino acid substitution per position. The numbers on the branches indicate the bootstrap support for each node.

Maltose binding by TreE could not be detected through changes in its intrinsic fluorescence. Its unfolding temperature (T_m) was dependent on maltose concentration (Table 5) and phylogenetic analyses show that its closest relatives transport maltose (Figure 4). Upon addition of trehalose, the fluorescence of TreE increases (Figure 11), however, the addition of trehalose to a protein sample containing maltose did not elicit any change of fluorescence suggesting that TreE was saturated with maltose (Figure 12). This experiment suggests that TreE binds maltose. Changes in intrinsic fluorescence cannot always be observed during binding because it reflects the average contribution of the changes of the intrinsic fluorescence of all tryptophans (144, 145). Similarly, XylE1 binds glucose but its binding was not detected by intrinsic fluorescence changes (8). However, Tian *et al.* solved the crystal structure of the XylE1 in complex with glucose and observed an increase in protein thermostability in the presence of glucose using circular dichroism (132).

Transcriptional regulation

Since the characterization of the trehalose and xylose-binding proteins (TreE and XylE2), further studies have been done on the transcriptional regulation of the trehalose and xylose ABC transporter operons. This region was interesting, especially since the gene for the transcription factor ROK is encoded in this region in *T. maritima* genomovar DSM3109. This gene was truncated in the genome sequence of *T. maritima* genomovar TIGR published in 1999, so its function was not studied prior to the discovery of the 10 kb region. The operon *xylEFK2* was renamed *gluEFK* by those who studied the regulation of operons in this regions since its transcription appears to be under the control

of GluR (TM1847), the transcriptional regulator located downstream of *gluEFK* (50). GluR can bind upstream of *gluEFK* repressing its expression and glucose is thought to elicit derepression of the operon (50). In addition to *treEFG*, *T. maritima* has another operon (*treTR*) involved in trehalose metabolism. It encodes a trehalose synthase (TreT), which is involved in the synthesis of trehalose and other trehalose analogs (146), and the transcriptional regulator TreR (50). TreR was experimentally demonstrated to bind upstream of *treTR* and *treEFG* using an electrophoretic mobility shift assay (EMSA). In the presence of trehalose TreE could potentially derepress the transcription of *treEFG* and *TreTR.* Interestingly, the EMSA experiments demonstrated that there is an additional binding site for GluR upstream of *treEFG* suggesting that GluR not only controls the transcription of *gluEFK* but *treEFG* as well (50). Real-time PCR also demonstrated that GluE and TreE were upregulated on cells grown on glucose while only TreE was upregulated in cells grown on trehalose (147).

Conclusion

In this study, the functions and the binding properties of two previously unknown ABC transporters in *T. maritima* MSB8 genomovar DSM3109 were determined. Based on these properties the two operons were named *treEFG* and *xylEFK2*. TreE binds trehalose, sucrose and glucose (with very low affinity) while the second SBP XylE2, binds xylose, glucose and L-fucose. TreE also binds maltose, though its binding affinity for maltose could not be determined. The addition of trehalose to a protein sample containing maltose did not elicit any change of fluorescence, which indicates that TreE was already saturated with maltose. To complete the characterization of TreE,

competitive-binding experiments should be used to measure its binding affinity for maltose.

The SBPs TreE and XylE2 share similar functions with MalE1, MalE2 and XylE but the results of this study suggest these transporters might have distinct functions. Since the binding affinities of XylE1 and MalE2 have not been measured at 60°C their respective binding properties cannot be directly compared with those of XylE2 and TreE. If the binding affinities of MalE1, MalE2 and XylE1 were measured at 60°C this would provide insight on how these genes were maintained in the genome of the genomovar DSM3109.

Chapter 5

Reexamination of the binding properties of the SBPs encoded by TM0595, TM1150, TM1199 (*lptE*) and TM0418 (*inoE*).

Some of the results from this chapter were published in Rodionova, I.A., Leyn, S.A., Burkart, M.D., Boucher, N., Noll, K.M., Osterman, A.L., and Rodionov, D.A. 2013. Novel inositol catabolic pathway in *Thermotoga maritima*. Environmental Microbiology. 15(8):2254-66.

Introduction

Previously, Nanavati *et al*. demonstrated that many annotated oligopeptide-binding proteins in *T. maritima* bind carbohydrates. However, no binding activity could be determined for the SBPs encoded by TM1199 (*lptE*), TM1150, and TM0595 (8). Characterization was attempted by the use of the intrinsic fluorescence of the aromatic amino acids. However, the unusually high number of tryptophans present in these proteins (25W in TM1199, 11W in TM1150, 13W in TM0595) may have hindered detection of the change in fluorescence elicited by the binding. In this study, the binding properties of these proteins were analyzed using the DSF assay. This method is suitable for these high-tryptophan proteins because its mechanism of detection is independent of their tryptophan content (see Chapter 3 for a detailed description).

In addition, the binding properties of the inositol-binding protein (InoE) encoded by TM0418 was reexamined because new information indicates its known ligand may not be its preferred ligand. InoE was shown to bind *myo*-inositol (MI) at 20°C with a binding affinity of 24.0 ± 1 μM (8). However, the low binding affinity suggests that the MI is not the primary ligand of InoE and the protein might bind another sugar or inositol derivative with higher affinity. InoE is encoded downstream of *iolRMNGKO* and bioinformatics and experimental evidence demonstrated that this operon encodes proteins of a novel inositol catabolic pathway (82). The low binding affinity of InoE and the gene organization around *inoE* suggest that the SBP may bind *myo*-inositol-1-phosphate (MI-1-phosphate) better than MI. To determine if InoE binds MI-1-phosphate, its binding properties were examined using DSF and intrinsic fluorescence. Additionally, growth experiments were performed to determine if the MI-1-phosphate could support the growth of *T. maritima*. This study revealed that InoE binds MI-I-phosphate, but that growth is not supported by this inositol derivative suggesting that MI-I-phosphate serves another function for *T. maritima*.

Materials and Methods

Cloning, expression and purification

The SBPs encoded by TM0418 (*inoE*), TM0595, TM1150 and TM1199 (*lptE*) were cloned as previously described (8). Table 8 summarizes information about these proteins and the encoded genes in the vicinity of these loci. Protein expression and purification were performed as describe in Chapter 4. InoE, LptE and the SBPs encoded byTM0595 and TM1150 were stored in phosphate buffer (20 mM sodium phosphate pH 7.4, and 100 mM NaCl). If mentioned in the text, dithiothreitol (DTT) was added to the storage buffer of LptE at a final concentration of 0.5 mM.

Table 8. Summary of the proteins and loci discussed in Chapter 5.

Substrate-binding protein		**Annotation of neighboring gene**		
Name and/or locus	Ligand	Locus	Predicted or known function	Ref
InoE (TM0418)	*myo*-inositol[1] and *myo*-inositol-1-phosphate[2]	TM0411	transcriptional regulator (IolR)	2, 3
		TM0412	2-keto-MI dehydrogenase (IolM)	3
		TM0413	di-keto-inositol hydrolase (IolN)	3
		TM0414	*myo*-inositol dehydrogenase (IolG)	3
		TM0416	5-keto-L-gluconate-D-tagaturonate epimerase (IolO)	3
LptE (TM1199)	galactose and lactose	TM1190	galactokinase (GalK)	
		TM1191	galactose-1-phosphate uridylyltransferase (GalT)	
		TM1192	α-galactosidase (GalA)	4, 5
		TM1193	β-galactosidase (LacZ1)	6
		TM1195	galactosidase (LacZ2)	
		TM1200	transcriptional regulator (GalR)	
		TM1201	endo-1,4-β-galactanase (GanB)	7
TM0595	unknown	none		
TM1150	glucose-6-phosphate	TM1154	6-phosphogluconolactonase	8, 9
		TM1155	glucose-6-phosphate dehydrogenase	

[1](8)
[2](82)
[3](50)
[4](148)
[5](149)
[6](150)
[7](151)
[8](152)
[9](153)

Intrinsic fluorescence spectroscopy of InoE

All fluorescence measurements were performed using an SLM Aminco-Bowman 2 spectrofluorometer. The temperature of the cuvette with the sample was equilibrated at 23°C or 60°C for 2 min. Fluorescence emission spectra were measured at an excitation wavelength of 295 nm and the emission intensities were measured over the wavelength range of 305 to 400 nm. The dissociation constants were measured by adding increasing amounts of selected carbohydrates into a stirred cuvette at the desired temperature. In a typical experiment, 3 µl of different stock solutions of the carbohydrate was added to 1 ml of 0.5 µM protein solution (20 mM sodium phosphate, 100 mM NaCl, pH 7.4). After the addition of the sugar, the sample was stirred for 2 min to reach equilibrium and then the fluorescence intensity at the predetermined emission wavelength was recorded. The K_d values were obtained from curve fitting using KaleidaGraph to the equation:

$$F = F_0 + \Delta F/2[Pt]\ ((K_d + [Pt] + [Lt]) - ((K_d + [Pt] + [Lt])^2 - 4[Pt]\ [Lt])^{1/2}$$

where F is the measured fluorescence at ligand concentration [L] and [Pt] is the total protein concentration.

Growth of *T. maritima* on myo-inositol-1-phosphate

A volume of 3 ml of defined medium (154), containing 0.25% of carbon source (maltose, dextrose or MI-1-phosphate) and a control without carbon source were inoculated with *T. maritima* grown on TTM medium containing 0.25% maltose as carbon source. The TTM medium contained (l^{-1}): 20 g NaCl, 4.8 g HEPES, 0.5 g cysteine, 1.0 g yeast extract, 2.5 g $Na_2S_2O_3$, 0.0033 g Na_2O_4W, and 500 ml of mineral solution, 10 ml of trace minerals solution and 10 ml vitamins solution. The mineral solution contained (l^{-1}):

6 g K_2HPO_4, 6g NH_4SO_4, 12 g NaCl, 2.6 g $MgSO_4 \cdot 7H_2O$, 0, 0.16 g $CaCl_2 \cdot 2H_20$. The trace mineral solution contained (l^{-1}): 3 g $MgSO_4 \cdot 7H_2O$, 0.5 g $MnSO_4 \cdot 2H_2O$, 1 g NaCl, 0.1 g $FeSO_4 \cdot 7H_2O$, 0.1 g $CoSO_4$, 0.1 g $CaCl_2 \cdot 2H_2O$, 0.01 g $ZnSO_4$, 0.01 g $CuSO_4 \cdot 5H_2O$, 0.01 g $AlK(SO_4)_2$, 0.01 g H_3BO_3, 0.01 g $Na_2MoO_4 \cdot 2H_2O$, 1.5 g nitrilotriacetic acid ($C_6H_9NO_6$) with pH adjusted at 6.5 with KOH. The vitamin solution contained (l^{-1}): 10 mg pyridoxine/HCl, 5 mg calcium pantothenate, 5 mg nicotinic acid, 5mg aminobenzoic acid (ABA), 5 mg riboflavin, 5 mg thiamin/HCl, 5 mg thioctic acid, 2 mg biotin, 2 mg folic acid and 0.1 mg cyanocobalamin. The sugars, the defined and the TTM media were prepared as previously described (155). The MI-1-phosphate (Sigma Aldrich, ≥96%) was treated with $MgSO_4$ with at least 6X the weight of MI-1-phosphate to precipitate the barium and create a magnesium salt. The barium sulfate precipitate was removed by filtration. As a positive control, *T. maritima* was grown in a defined medium containing 0.25% MI-1-phosphate and 0.25% maltose indicating that no potential inhibitor was left from the MI-1-phosphate preparation. All the growth experiments were performed at 80°C under anaerobic conditions.

Results

InoE

The thermostability of InoE in the presence of potential ligands was examined by DSF. The stability of the SBP increased in the presence of 10 mM MI (ΔTm, 2.4°C), 1.68 mM MI-1-phosphate (ΔTm, 9.5°C), 10 mM fructose-6-phosphate (ΔTm, 14.1°C), and 10 mM glucose-6-phosphate (ΔTm, ≥15°C) (Table 9). Although MI-1-phosphate, fructose-6-phosphate and glucose-6-phosphate increased the thermostability of the InoE, these results are not necessarily indicative of ligand binding in the binding pocket because all these compounds contain phosphate that can increase a protein's thermostability nonspecifically.

Nanavati *et al.* measured the binding affinity at 20°C of InoE with MI but the binding affinity was not tested at a temperature closer to the physiological growth temperature of *T. maritima* (8). Although a small increase of fluorescence (1.02%) was observed in the presence of MI at 23°C, no change of fluorescence was detected at 60°C with concentrations of MI up to 1.5 mM (Figure 14) (82). This result suggests that InoE is unable to participate in the transport of MI at temperatures close to the *T. maritima* optimal growth temperature.

In the presence of MI-1-phosphate, a decrease of relative fluorescence was observed from 1.00 to 0.80 at 23°C and 1.00 to 0.82 at 60°C. InoE binds MI-1-phosphate with K_d values of 7.7 ± 0.4 μM at 20°C and 19.8 ± 5.3 μM at 60°C (Figure 15, Table 10) (82). Glucose-6-phosphate was tested at 20°C but saturation was not reached at concentrations up to 50 uM.

Table 9. Thermostabilities InoE and the SBPs encoded by TM1150 and TM0595 measured by DSF represented by ΔTm values. The assay was performed in triplicate and standard deviations are shown. If the Tm_{ligand} could not be calculated precisely because of incomplete denaturation or inability to observe a complete unfolding curve, Tm_{ligand} was estimated using the maximum fluorescence intensities.

Ligand	**InoE**	**TM1150**	**TM0595**
arabino-galactose	0.1 ± 0.1	0.0 ± 1.3	-1.4 ± 0.5
arabinose	0.1 ± 0.2	1.2 ± 0.7	-0.1 ± 0.2
cellobiose	-0.1 ± 0.1	-0.2 ± 0.8	-0.1 ± 0.2
fructose	-0.1 ± 0.2	0.9 ± 0.7	-0.1 ± 0.2
fructose-6-P	14.1 ± 0.2	N/A	N/A
L-fucose	0.0 ± 0.2	1.1 ± 0.8	-0.2 ± 0.2
galactose	-0.3 ± 0.1	0.3 ± 1.0	-0.3 ± 0.2
glucose	0.0 ± 0.1	0.7 ± 0.6	0.0 ± 0.2
glucose-6-P	>15	N/A	N/A
MI-1-phosphate	9.5 ± 0.2	N/A	N/A
lactose	-0.3 ± 0.1	1.0 ± 0.6	-0.1 ± 0.2
maltose	0.0 ± 0.2	1.0 ± 0.7	-0.2 ± 0.2
maltotetraose	-0.1 ± 0.1	0.5 ± 0.8	0.7 ± 1.3
maltotriose	0.1 ± 0.2	-0.1 ± 0.7	-0.1 ± 0.2
mannose	-0.4 ± 0.2	0.5 ± 0.6	0.1 ± 0.2

mannotetraose	0.5 ± 0.3	0.6 ± 0.6	-0.8 ± 0.8
melibiose	0.0 ± 0.1	0.1 ± 1.0	-0.1 ± 0.2
myo-inositol	2.4 ± 0.1	1.1 ± 1.0	0.1 ± 0.2
pullulan	0.2 ± 0.3	0.1 ± 0.7	-0.5 ± 0.2
raffinose	0.2 ± 0.2	0.4 ± 0.7	0.0 ± 0.2
rhamnose	-0.2 ± 0.1	-1.1 ± 0.7	0.1 ± 0.2
salicin	0.1 ± 0.2	0.1 ± 1.0	0.0 ± 0.2
sorbitol	0.3 ± 0.1	0.3 ± 0.7	0.3 ± 0.8
sucrose	0.1 ± 0.2	1.1 ± 0.7	-0.1 ± 0.2
tagatose	0.1 ± 0.1	0.9 ± 0.6	-0.1 ± 0.2
trehalose	0.2 ± 0.2	0.7 ± 0.6	0.0 ± 0.2
xyloglucan	-0.1 ± 0.1	0.8 ± 1.0	-0.1 ± 0.2
xylose	-0.2 ± 0.1	0.8 ± 0.7	-0.1 ± 0.2
no ligand	0.0 ± 0.2	0.0 ± 0.9	0.0 ± 0.2

(N/A) sugar not tested

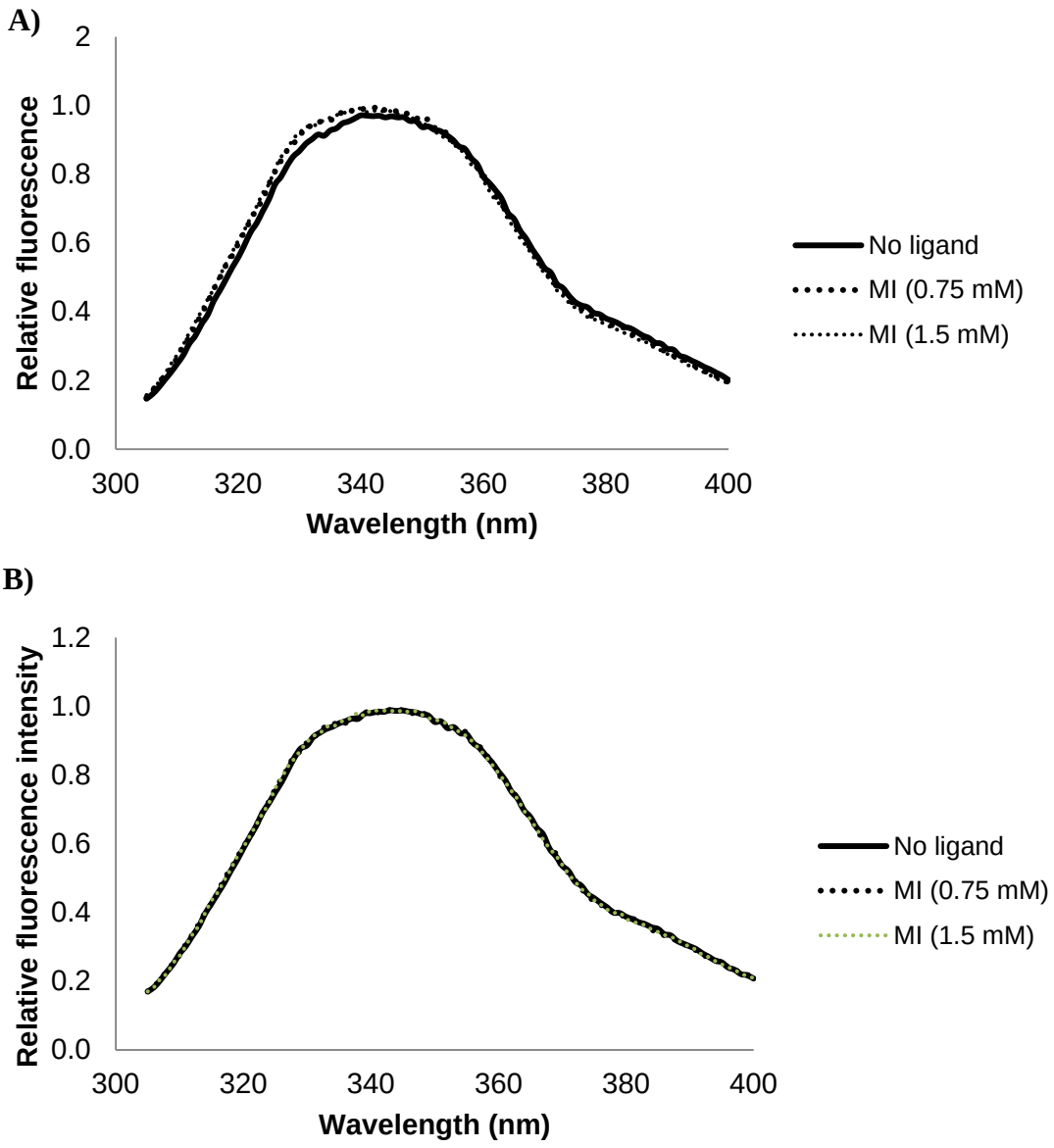


Figure 14. Emission spectra of InoE with and without *myo*-inositol at (A) 20°C and (B) 60°C.

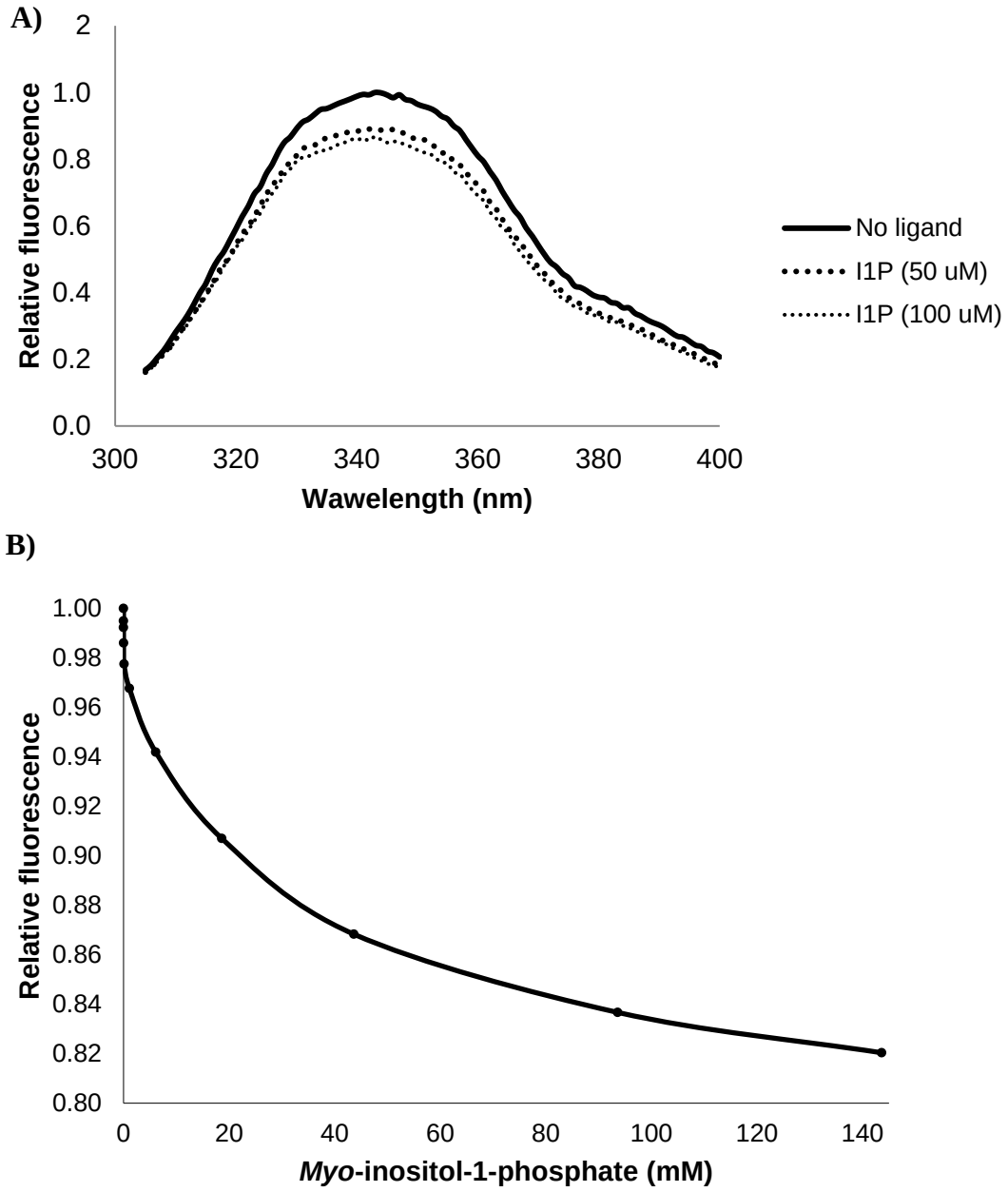


Figure 15. (A) Emission spectra of InoE with and without the addition of MI-1-phosphate and (B) a titration curve with MI-1-phosphate at 60°C.

Table 10. Apparent binding affinities (K_d values) of InoE measured at 20°C and 60°C. The K_d values were measured in duplicate by intrinsic fluorescence change upon ligand binding and standard deviations are shown.

Ligand	**K_d (μM)**	
	20°C	**60°C**
myo-inositol	24 ± 1[*]	no binding[**]
myo-inositol-1-phosphate	7.7 ± 0.4	19.8 ± 5.3

[*] from (8)
[**] at 1.5 mM of ligand

MI-1-phosphate as a carbon source for *T. maritima*

Growth experiments were performed to determine if the growth of *T. maritima* is supported on MI-1-phosphate (Table 11). The cells were grown on TTM medium containing 0.25% maltose and then, inoculated into a defined medium (154) containing 0.25% (w/v) carbon source (maltose, dextrose or MI-1-phosphate) and a control without carbon source. To ensure that the MI-1-phosphate did not contain any inhibitor, an initial growth experiment was carried out with defined medium containing 0.25% maltose and 0.25% MI-1-phophate and a positive control containing 0.25% maltose. Both cultures had similar final optical density (Table 11) suggesting that no inhibitor was present in the MI-1-phosphate solution. The growth on single carbon source (dextrose, maltose or MI-1-phosphate) was evaluated. Growth was observed only on 0.25% maltose (Table 11). No growth was observed on dextrose, MI-1-phosphate and on the control without carbon source (Table 11).

Table 11. Growth of *T. maritima* grown in defined media with different carbon sources. The optical densities of cultures were measured after 40 hours of growth. The growth assays performed on a single carbon source were performed twice and their standard deviations are shown.

Carbon source (0.25%)	**Optical density ($O.D_{660}$)**
none	0.04 ± 0.03
maltose	0.25 ± 0.04
dextrose	0.06 ± 0.02
MI-1-phosphate	0.07 ± 0.03
maltose and MI-1-phosphate	0.25

LptE (TM1199)

The thermostability of LptE without the addition of dithiothreitol (DTT) was increased the most in the presence of 10 mM mannan (ΔTm, 12.0 ± 0.4°C), 10 mM sucrose (ΔTm, 7.2 ± 0.1°C) and 10 mM xyloglucan (ΔTm, 6.7 ± 0.4°C) (Table 12). The interaction of the protein with the polysaccharide mannan was surprising, especially since mannose and mannotriose could not induce an increase in the thermostability of the protein (ΔTm, 0.4 ± 0.1°C and 0.1 ± 0.1°C, respectively) (Table 12). The protein thermostability was also increased in the presence of sucrose, which often stabilizes proteins without interacting with the binding pockets, suggesting that the interaction was non-specific.

LptE contains four cysteines. These residues can form disulfide bonds leading to a misfolded protein, perhaps forming non-specific interactions with sugars. The DSF assay was performed using protein stored in a phosphate buffer (pH 7.4) containing 0.5 mM DTT to reduce potential disulfide bonds. The thermostability of LptE stored with DTT had highest fluorescence increases in the presence of 10 mM arabinogalactose (ΔTm, 13.1 ± 0.4°C), 10 mM lactose (ΔTm, 6.8 ± 0.2°C), 10 mM mannotetraose (ΔTm, 6.4 ± 0.2°C) and 10 mM galactose (ΔTm, 3.6 ± 0.2°C) (Table 12). These results suggest that LptE interacts the most with arabinogalactose, lactose, mannotetraose and galactose.

TM1150 and TM0595

The thermostabilities of the SBPs encoded by TM1150 and TM0595 remained unchanged in presence of any of the tested sugars. The delta values were close to zero after taking in account the standard deviations (Table 9).

Table 12. Thermostabilities LptE measured by DSF and represented by ΔTm values. When mentioned, dithiothreitol (DTT) was added at 0.5 mM final concentration. The assay was performed in triplicate and standard deviations are shown. If the Tm_{ligand} could not be calculated precisely because of incomplete denaturation or inability to observe a complete unfolding curve, Tm_{ligand} was estimated using the maximum fluorescence intensities.

Ligand	**ΔTm**	
	No DTT	**+DTT**
arabino-galactose	N/A	13.1 ± 0.4
arabinose	2.3 ± 0.2	0.1 ± 0.2
cellobiose	3.9 ± 0.0	0.2 ± 0.3
fructose	-0.1 ± 0.1	-0.1 ± 0.2
L-fucose	0.3 ± 0.2	0.2 ± 0.2
galactose	0.4 ± 0.1	3.6 ± 0.2
glucose	0.0 ± 0.1	0.0 ± 0.2
lactose	0.2 ± 0.1	6.8 ± 0.2
maltose	1.8 ± 0.1	0.1 ± 0.2
maltotetraose	-0.1 ± 0.1	0.2 ± 0.2
maltotriose	0.2 ± 0.1	0.2 ± 0.2
mannose	0.4 ± 0.1	0.0 ± 0.2
mannan	12.0 ± 0.4	0.3 ± 0.3
mannotetraose	N/A	6.4 ± 0.2
mannotriose	0.1 ± 0.1	N/A

melibiose	0.2 ± 0.1	0.3 ± 0.2
myo-inositol	0.4 ± 0.4	0.3 ± 0.2
pullulan	1.7 ± 0.1	0.3 ± 0.3
raffinose	0.1 ± 0.2	0.1 ± 0.2
rhamnose	0.2 ± 0.2	0.2 ± 0.2
ribose	0.3 ± 0.2	0.3 ± 0.3
salicin	-0.1 ± 0.1	0.1 ± 0.2
sorbitol	-0.1 ± 0.1	0.0 ± 0.2
sucrose	7.2 ± 0.1	0.0 ± 0.2
tagatose	0.1 ± 0.2	0.4 ± 0.2
trehalose	3.0 ± 0.5	0.3 ± 0.2
xyloglucan	6.7 ± 0.2	0.2 ± 0.2
xylose	0.3 ± 0.1	0.1 ± 0.2
no ligand	0.0 ± 0.0	0.0 ± 0.3

(N/A) sugar not tested

Discussion

InoE: Inositol binding protein

The *T. maritima* inositol pathway converts *myo*-inositol to D-tagaturonate which then can enter into the tagaturonate utilization pathway (82, 156). The inositol pathway was elucidated using bioinformatics analysis and biochemical characterization of the recombinant enzymes encoded within the *iol* operon. Biochemical analysis revealed that the operon encodes a *myo*-inositol (MI) dehydrogenase (IolG), a 2-keto-*myo*-inositol dehydrogenase (IolM), a diketo-inositol hydrolase (IolN) and a 5-keto-L-gluconate-D-tagaturonate epimerase (IolO) (82). The *iol* operon is located upstream of the *inoEFGK* operon which encodes an ABC transporter system. The substrate-binding protein InoE was previously demonstrated to bind *myo*-inositol (MI) at 20°C with a K_d of 24.0 ± 1 μM (8). The low binding affinity suggested that MI is not to the primary ligand for InoE so the SBP might bind another inositol derivative with higher affinity. Because of the proximity of the *iolRMNGKO* and *inoEFGK* operon, it was speculated that SBP encoded by *inoE* might bind phosphorylated *myo*-inositol with higher affinity. To demonstrate the involvement of the inositol ABC transporter in the inositol catabolic pathway, the binding properties of InoE were examined with MI-1-phosphate at 20°C and 60°C. InoE binds MI-1-phosphate with K_d values of 7.7 ± 0.4 μM at 20°C and 19.8 ± 5.3 μM at 60°C (Figure 15, Table 10). Additionally, InoE does not bind MI at 60°C. These results show that InoE binds with better affinity MI-1-phosphate than MI. Previously, it was demonstrated that *T. maritima* growth could not be supported by inositol (157). Inositol is most commonly found in a phosphorylated form (158). To test if MI-1-phosphate

could support *T. maritima* growth, cell growth was measured at 80°C on a defined medium containing 0.25% MI-1-phosphate. Growth could not be supported in the presence of MI-1-phosphate alone while growth was measured on a medium containing 0.25% maltose. Growth was also possible on a medium containing 0.25% maltose and 0.25% MI-1-phosphate suggesting that the MI-1-phosphate preparation did not contain an inhibitor. It possible that the defined medium was limiting because the cells did not grow to high densities, the maximal absorbance reached of cells grown on maltose was only 0.25 ± 0.04, and *T. maritima* did not grow on dextrose. Although InoE can bind MI-1-phosphate at 60°C, it is possible that the SBP is unable to bind the ligand *in vivo* or that the ABC transporter is not functional at 80°C. MI, MI-1-phosphate, *scully*-inositol, 2-keto-*myo*-inositol and glucose did not inhibit the shift of the predicted DNA target bound to IolR in an electrophoretic mobility shift assay (50) suggesting that none of the sugars is an effector of the transcriptional repressor IolR. Thus, it is unclear what the natural ligand of InoE is. Another explanation could be that MI-1-phosphate is not used as carbon source, but instead as a precursor to the biosynthesis of di-*myo*-inositol-phosphate (DIP). DIP is a known compatible solute in *T. maritima* (159–161) and the genome encodes a complete biosynthetic pathway for DIP (162).

LptE: Galactose and lactose-binding protein

The gene *lptE* is located near genes that encode a hydrolase and a regulator involved in galactoside catabolism. The putative operon encodes a putative galactokinase (GalK), a putative UDP-glucose-hexose-1-phosphate uridylyltransferase (GalT), an α-galactosidase (GalA) (148, 149), a β-galactosidase (LacZ1) (150), a β-galactosidase

(LacZ2), an endo-1,4-β-galactosidase (GanB) (Yang et al., 2006) and a putative galactose repressor (157) (Figure 16). LacZ1 hydrolyzes lactose to glucose and galactose and the galactose is transferred to a galacto-oligosaccharide by transgalactosylation (150, 163). GanB hydrolyzes pectic galactans to galactose and beta-1,4-D-galactobio-, -trio and -tetraose (151). GalK and GalT convert α-D-galactose to glucose-6-phosphate by the Leloir pathway (Figure 16). The putative galactokinase GalK is a homolog of the characterized GalK protein in *Pyrococcus furiosus* (164, 165) and *Pyrococcus horikoshii* (166) that share 42% and 44% identity, respectively. GalK also has a strong preference for galactose and less so for D-galactosamine (51).

The gene *lptE* (TM1199) was previously annotated as a component of a putative galactoside ABC transporter operon (*ltpEFGKL*) (43, 51, 157). Nanavati *et al.* cloned and expressed *lptE* and attempted to determine the protein's binding properties, but none of the sugars tested were able to bind to the recombinant protein at 20°C (8). LptE contains 25 tryptophans, which might have hindered detecting changes in fluorescence elicited by binding of the sugar. Additionally, its binding properties were also measured at 20°C, a temperature below the growth temperature of *T. maritima*.

A)

B)

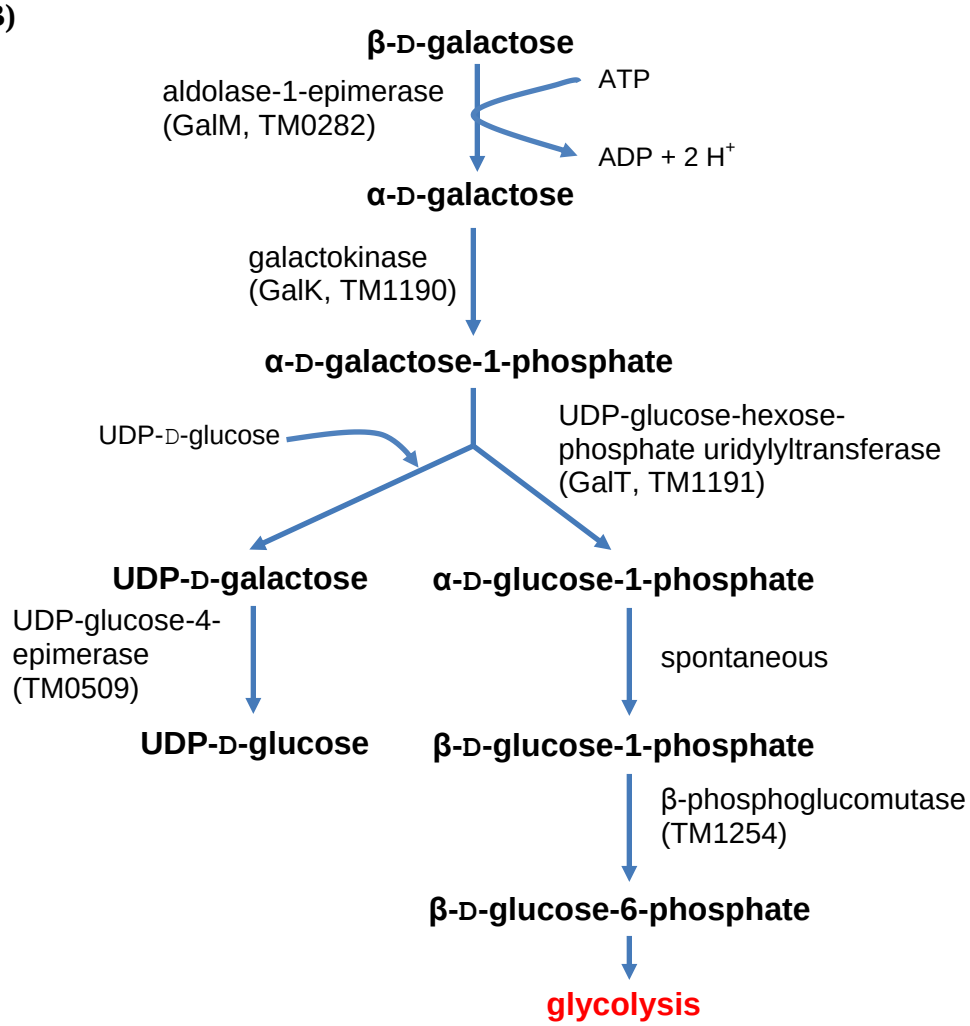


Figure 16. Galactose catabolism and utilization pathway. A) Schematic of the region between TM1190 and TM1201 (represented from 3' to 5'). B) Leloir pathway

The DSF assays performed here indicate that disaccharides containing a molecule of galactose such as arabinogalactose and lactose increased the thermal stability of LptE. Previously, transcript analysis studies had shown an increase of the transcripts of TM1190-TM1201 in cells grown on lactose versus glucose (167) and on galactose versus ribose (157). These results suggest that the LptEFGKL transporter system transports lactose and galactose.

TM1150 and TM0595

Typically, the genes encoding the carbohydrate ABC transporters in *T. maritima* are mostly clustered near genes encoding hydrolases and transcription regulators that respond to their respective substrate. However, TM1150 and TM0595 are not found near genes encoding hydrolases or transcription regulators. The genes proximal to TM1150 encode all the components of a transporter from the peptide/opine/nickel family (PepT). The genes encoding the transporter are located downstream from the genes encoding a glucose-6-phosphate dehydrogenase (152, 153) and a putative 6-phosphogluconolactonase which are involved in the synthesis of 6-phospho-D-gluconate from β-D-glucose-6-phosphate. This suggests that the SBP encoded by TM1150 may be involved in the transport of glucose-6-phosphate. It is possible that the ligand of the SBP encoded by TM1150 was not tested by the DSF assay.

A bioinformatics analysis of the SBP encoded by TM0595 reveals that the protein is not complete which suggests that it is not functional. The conserved domain tool from NCBI showed that the middle of the substrate-binding domain has suffered an insertion of 150 amino acids while the Pfam database from the Sanger institute showed that the

domain is truncated. The loci TM0591, TM0592 and TM0593 encode an ATPase, a permease and a SBP, respectively and are annotated as an amino acid ABC transporter. These genes that encode the ABC transporter components belong to a polar amino acid transporter family (PAAT) while the SBP encoded by TM0595 seems to belong to a different family. The gene organization near TM0595 is not typical of the CUT1, CUT2 or the PepT family. These findings suggest that TM0595 is not functional.

Conclusion

Because MI-1-phosphate is unable to sustain growth of *T. maritima*, it is possible that another derivative of inositol is the substrate of InoE. Further screens can be performed using other untested ligands and inositol derivatives. Since compounds that contain a phosphate can increase the thermal stability of the SBP without interacting with the binding pocket, screening techniques that are independent of the protein thermostability such as the intrinsic fluorescence of the aromatic residues might be better suited. Alternatively, since the InoE transcript should be up-regulated if the cells are grown on the substrate for this ABC transporter, the InoE transcript could be analyzed by real-time PCR during growth on possible InoE ligands. A similar approach has been successful to determine the function of other importers (43, 167, 168). However, an increase of the transcript level of InoE would not be direct proof that the carbon source is utilized by the inositol transporter and the results would need to be confirmed using ligand binding assays. Finally, it is possible that MI-1-phosphate is used not as a carbon source but as a precursor to synthetize DIP, a common compatible solute.

LptE contains a high number of tryptophans that appear to hinder the use of the intrinsic fluorescence method. Other techniques such as equilibrium dialysis could be performed to confirm the interactions of LptE with arabinogalactose, galactose, lactose and mannotetraose that were demonstrated by DSF.

The region around TM1150 encodes all the components of a PepT family ABC transporter. However, the DSF assay and measures of the intrinsic fluorescence of the protein were unsuccessful to determine its binding properties. It is possible that the protein binds a ligand that was not tested. Although the analysis of the binding properties

of many annotated dipeptide and oligopeptides ABC transporters in *T. maritima* demonstrated that they bind carbohydrates, it is possible that this SBP has a stronger affinity for dipeptides or small oligopeptides. The protein shares approximately 30% identity with a known oligopeptide-binding protein that binds small peptides (169, 170). The binding specificity of the oligopeptide-binding protein with small peptides can be examined using competitive binding assays, isothermal titration calorimetry or intrinsic fluorescence.

The DSF assays performed here indicate that disaccharides containing a molecule of galactose such as arabinogalactose and lactose increased the thermal stability of LptE. Previously, transcript analysis studies had shown an increase of the transcripts of TM1190-TM1201 in cells grown on lactose versus glucose (167) and on galactose versus ribose (157). These results suggest that the LptEFGKL transporter system transports lactose and galactose.

Typically, the genes encoding the carbohydrate ABC transporters in *T. maritima* are mostly clustered near genes encoding hydrolases and transcription regulators that respond to their respective substrate. However, TM1150 and TM0595 are not found near genes encoding hydrolases or transcription regulators. The genes proximal to TM1150 encode all the components of a transporter from the peptide/opine/nickel family (PepT). The genes encoding the transporter are located downstream from the genes encoding a glucose-6-phosphate dehydrogenase (152, 153) and a putative 6-phosphogluconolactonase which are involved in the synthesis of 6-phospho-D-gluconate from β-D-glucose-6-phosphate. This suggests that the SBP encoded by TM1150 may be

involved in the transport of glucose-6-phosphate. It is possible that the ligand of the SBP encoded by TM1150 was not tested by the DSF assay.

A bioinformatic analysis of the SBP encoded by TM0595 reveals that the protein is not complete which suggests that it is not functional. The conserved domain tool from NCBI showed that the middle of the substrate-binding domain has suffered an insertion of 150 amino acids while the Pfam database from the Sanger institute showed that the domain is truncated. The loci TM0591, TM0592 and TM0593 encode an ATPase, a permease and a SBP, respectively and are annotated as an amino acid ABC transporter. These genes that encode the ABC transporter components belong to a polar amino acid transporter family (PAAT) while the SBP encoded by TM0595 seems to belong to a different family. However, the gene organization near TM0595 is not typical of the CUT1, CUT2 or the PepT family. These findings suggest that TM0595 is not functional.

Chapter 6

Characterization of the mannoside-binding proteins in the Thermotogales

Introduction

The mannoside-binding protein is a component of the mannoside ABC transporter system that belongs to the PepT family. Many SBPs from the PepT family have been previously characterized, including the *Pyrococcus furiosus* cellobiose binding protein ($CbtA_{Pfur}$), the *T. maritima* mannoside-binding proteins ($ManD_{Tmar}$, $ManE_{Tmar}$), and the *T. maritima* β-glucoside-binding protein ($BglE_{Tmar}$) (7, 8, 118). The binding properties of these proteins overlap as they are able to bind saccharides containing either β-D-glucose or β-D-mannose or both. *T. maritima* encodes two mannoside-binding proteins ($ManD_{Tmar}$ and $ManE_{Tmar}$) one of which is suspected to have arisen from a gene duplication. At 20°C, the $ManD_{Tmar}$ binds β-mannobiose, -triose, and –tetraose as well as galactosyl mannobiose and cellobiose while $ManE_{Tmar}$ binds only β-mannobiose (8). $BglE_{Tmar}$ is a close relative of $ManD_{Tmar}$ and $ManE_{Tmar}$ that binds laminaribiose, laminaripentaose, cellobiose and cellopentaose (8, 118). $CbtA_{Pfur}$ binds various β-D-glucose disaccharides and β-glucosides such as cellobiose, -triose, -tetraose, and -pentaose as well as laminaribiose and laminaritriose (7). Its binding properties were not examined with sugars composed of β-D-mannose. However, *P. furiosus* does not grow on glucomannan or ivory nut mannan that contains β-D-mannose saccharides (Driskill et al. 1999) even though a β-mannosidase was detected (171, 172).

The different of functions of $ManD_{Tmar}$ and $ManE_{Tmar}$ are interesting because, based on the functions of other relatives, it suggests that either the ManE or the ManD ortholog has diverged from its ancestral state. The $ManE_{Tmar}$ might have lost many of its functions or $ManD_{Tmar}$ diverged andd novel functions that were essential to the fitness of *T. maritima*. To understand how these SBPs have evolved, seven Thermotogales ManE orthologs and three ManD orthologs were chosen for study. To determine how the SBPs adapted in their new hosts, their protein thermostabilities were measured using differential scanning fluorimetry (DSF). The ligand interactions with the recombinant proteins were screened using DSF and their binding affinities were measured by intrinsic fluorescence. Their binding affinities were measured at 37°C and 60°C to determine if the proteins' functions are temperature-dependent and to understand how their functions have changed through their evolutionary history.

Materials and Methods

Strains

Thermotoga maritima MSB8 genomovar DSM 3109, *Thermotoga petrophila* RKU-1 (DSM 13995), *Thermotoga naphthophila* RKU-10 (DSM 13996), *Thermotoga lettingae* TMO (DSM 14385) and *Fervidobacterium nodosum* Rt17-B1 (DSM 3606) were obtained from the German Collection of microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). *Thermotoga sp.* strain RQ2 and *Mesotoga prima* were provided by Karl O. Stetter and Camilla L. Nesbø, respectively. *Thermotoga neapolitana* NS-E was provided by the late Holger W. Jannasch.

Cloning

The designations of the putative mannoside-binding proteins studied here are provided in Table 13. The putative mannoside-binding protein encoding genes were amplified using the FailSafe^{TM} PCR system (Epicentre) with primers containing the restriction enzyme recognition sequence of BamH1 for $ManE_{Tlet}$ and of Nde1 for all the other ManD and ManE orthologs (listed in Table 14). To prevent protein insolubility, the sequence encoding the signal peptide was excluded from the PCR product. The signal peptide cleavage sequences were determined using SignalP 4.0 (173). The genes were sub-cloned into pGEM®-T vector (Promega) and then digested with the appropriate restriction enzyme. The digested products were ligated into pET15b vector and the plasmids were transformed in Rosetta^{TM}(DE3) competent cells (EMD Millipore) for protein expression. All clones were validated by sequencing and were 100% identical to

the sequence found in the NCBI database with the exception of the clone from *T. neapolitana* which had a nucleotide discrepancy with the NCBI reference sequence (NC_011978.1) leading to amino acid change from isoleucine (Ile) to threonine (Thr) at position 167 (I167T). Based on the crystal structure of a close relative BglE (TM0031, pdb: 2o7i), the amino acid is not located in the binding pocket of the protein.

Protein expression and purification

For the protein expression, the cells were grown in 50 ml LB medium containing ampicillin (50 μg/ml) for 4 to 5 h at 37°C and then induced with 1 mM IPTG (isopropyl-D-thiogalactopyranoside) for 16 h at 18°C. In a typical experiment, the cells were harvested by centrifugation at 2,400 x g and washed with PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4) and lysed using the B-PER protein extraction reagents (Thermo Scientific Pierce) containing 0.4 mg of DNAse, 0.4 mg of lysozyme and 0.8X of Halt Protease Inhibitor Cocktail (Thermo Scientific Pierce) per ml of cell paste. For high yield (>1 mg), the overexpressed His-tagged proteins were purified with a high-performance nickel-Sepharose® (GE Healthcare Life Sciences) using a polyethylene filter ~30 μm (Thermo Scientific Pierce) according to the manufacturer's protocol and followed by dialysis against 2 liters of buffer (20 mM sodium phosphate, pH 7.4). If a lower yield was sufficient (0.6 to 1 mg), the proteins were purified using His SpinTrap Kit (GE Healthcare Life Sciences) followed by filtration on an Amicon® Ultra 30,000 NMWL (EMD Millipore) filter and washed 3 times using 5 mL of 20 mM sodium phosphate, pH 7.4.

Table 13. Designations of the ManE and ManD orthologs used in this study. The symbol (-) indicates that the organism does not encode an orthologous protein.

Organism	**Designation**	
	ManE	**ManD**
T. maritima	$ManE_{Tmar}$	$ManD_{Tmar}$
Thermotoga RQ2	$ManE_{TRQ2}$	$ManD_{TRQ2}$
T. naphthophila	$ManE_{Tnap}$	$ManD_{Tnap}$
T. petrophila	$ManE_{Tpet}$	-
T. neapolitana	$ManE_{Tnea}$	-
T. lettingae	$ManE_{Tlet}$	-
F. nodosum	$ManE_{Fnod}$	-
M. prima	$ManE_{Mpri}$	-

Table 14. Primers used to amplify the ManD- and ManE-encoding genes. The restriction enzyme recognition sequences are underlined.

Protein/Locus	**Forward (top) and Reverse (bottom) Primer Sequences (5´-3´)**
$ManE_{Tmar}$	CTCGAGATGCAGACTTTTGAGAGAAACAAA
TM1223	CTCGAGTTACTTTGCTTCAATACCAAAGA
$ManE_{TRQ2}$	CTCGAGATGCAGACTTTTGAGAGAAACAAA
TRQ2_1595	CTCGAGTTACTTTGCTTCAATACCAAAGA
$ManE_{Tpet}$	CTCGAGATGCGGACTTTTGAGAGAAACAAA
Tpet_1545	CTCGAGTTACTTTGCTTCAATACCGAAGAGTGT
$ManE_{Tnea}$	CTCGAGATGCAAACTTTCGAAAGGAACAAG
CTN_1348	CTCGAGTTATTTTGCTTCGATACCAAACAGTGTC
$ManE_{Tlet}$	GGATCCGGAAACATTTGAAAGAAGCAAAACA
Tlet_1438	GGATCCCTATTTTGCTTTTATGTTGAAAAGCA
$ManE_{Mpri}$	CTCGAGATG CAA GTT TAC GAT CGA AAA GAA ACT
YP_006346550	CTCGAGTTACTTAGCCTCCAGGGCTGTCAG
$ManE_{Fnod}$	CTCGAGATG GAC GTA GTT TAC AAA AGA GAT GAG
Fnod_1553	CTCGAGTTATTTATTGGTAATTCCTATAAGCAT
$ManD_{Tmar}$	CTCGAGATGCAAGTTTTAGAACGAAACGAAACT
TM1226	CTCGAGTTATTTTGCCGGTTTGAGATTAAATAG
$ManD_{TRQ2}$	CTCGAGATGTTAGAACGAAACGAAACTATGTAC
TRQ2_1592	CTCGAGTTATTTTGCCGGTTTGAGATTAAATAG
$ManD_{Tnap}$	CTCGAGATGCAAGTTTTAGAACGAAACGAAACT
Tnap_1566	CTCGAGTTATTTTGCCGGTTTGAGATTAAATAG

Differential scanning fluorimetry to measure ligand-protein interactions and protein thermostabilities

The samples were mixed to final volumes of 20 µl, according to the following preparation: 2.0-2.5 µg protein, 4 µl 5X citric acid/Na_2HPO_4 buffer (as described in Appendix 2; pH 7 unless specified in the text), 8X SYPRO® Orange (Life Technologies), 150 mM NaCl, and 2 µl ligand (or mili-Q water), typically with a final concentration of 10 mM. The final concentrations of cellobiose, mannan, pullulan, glucomannan and sorbitol were 25 mM, 0.225% (wt/vol), 0.5% (wt/vol), 0.05% (wt/vol) and 0.1% (wt/vol), respectively. The fluorescence intensities were measured using a CFX Connect™ Real-Time PCR Detection System (Bio-Rad) with excitation at 490 nm and emission at 530 nm. The samples were heated from 25 to 98°C with a heating rate of 0.5°C per min. All the assays were done in triplicate. The midpoint temperature of the unfolding transition (Tm) was obtained with the program Gnuplot from curve fitting to a Boltzmann equation as previously described (81) (Chapter 3 and Appendix 2).

For the ligand screening of $ManD_{Tmar}$, $ManD_{Tnap}$, $ManD_{TRQ2}$, $ManE_{Tlet}$, $ManE_{Mpri}$, and $ManE_{Fnod}$ the assays were carried out at pH 7. The assays were carried out at pH between 3 and 4 for $ManE_{Tmar}$, $ManE_{TRQ2}$, $ManE_{Tnap}$, and $ManE_{Tpet}$ because their respective unfolding temperatures could not be observed at pH 7.

Sugar purity

The purity of the sugars was at least 98% and sugars were of D configuration unless noted otherwise. Arabinose, melibiose, raffinose, β-1,4-cellobiose, sorbitol, tagatose, lactose, ribose, α-1,1-trehalose, α-1,4-maltose, L-fucose, salicin, β-1,6-gentiobiose

(85%), xylose, fructose, galactose, L-rhamnose, *myo*-inositol, mannosamine hydrochloride, glucose, sucrose, mannan, mannose, β-1,2-sophorose, pullulan, β-1,4-cellotriose (93%), β-1,4-cellotetraose (85%), α-1,3-mannobiose (95%) and β-1,4-maltotetraose (95%) were obtained from Sigma-Aldrich. Xyloglucan oligosaccharide (95%), β-1,4-mannobiose (95%), β-1,4-mannotriose (95%), and β-1,4-mannotetraose (95%), β-1,3-laminaribiose (95%), β-1,3-laminaritriose (95%) and glucomannan were supplied from Megazyme, and β-1,4-maltotriose (97%) and arabinogalactose (3-O-β-D-galacto-pyranosyl-D-arabinose) (98%) were from ICN. Mannitol was from Calbiochem.

Intrinsic fluorescence spectroscopy

The fluorescence measurements were performed using either a fluoroMax-3 graciously provided by the Wadsworth Center (NY) or fluoroMax-4 spectrofluorometer (Horiba Scientific) graciously provided by Dr. Carol Teschke. The protein samples were incubated at the appropriate temperature of 37°C or 60°C. Fluorescence emission spectra were measured at an excitation wavelength of 295 nm or 280 nm, and the emission intensities were measured over the wavelength range of 310 to 370 nm. The dissociation constants were measured by adding increasing amounts of carbohydrate into a stirred cuvette. In a typical experiment, 2.4 µl of different stock solutions of the carbohydrate were added to 1.2 ml of 0.3-0.5 µM protein solution (20 mM sodium phosphate, pH 7.4). For the measurement using glucomannan, a solution of 0.05% (wt/vol) was used. After the addition of the sugar, the sample was stirred for 1 min to reach equilibrium and then the fluorescence intensity at the predetermined emission wavelength was recorded. The

K_d values were obtained from curve fitting using KaleidaGraph to the equation accounting for ligand depletion as follows:

$$F = Fo \pm \frac{\Delta F}{2[Pt]} ((Kd + [P] + [L]) - \sqrt{(Kd + [P] + (L))^2 - 4[P][L]})$$

where F is the measured fluorescence at ligand concentration [L] and [P] is the total protein concentration. This formula was used since the protein concentration was typically greater than the measured K_d (135, 136).

Results

Synteny

The syntenic regions of the mannoside ABC transporter operon *manEFGKL* and *manD* and in the Thermotogales are shown in Figure 17. In all the organisms studied, the mannoside ABC transporter operon (*manEFGKL*) has a typical organization and encodes an SBP (ManE), two transmembrane proteins (ManF and ManG) and two ATP-binding proteins (ManK and ManL). A putative LacI transcription factor is encoded downstream of the transporter operon (TM1218 in *T. maritima*). *T. maritima*, *Thermotoga* species RQ2 and *T. naphthophila* have a putative operon that encodes a ManE paralog (ManD) (93), a mannan endo-1,4-β-mannosidase (ManB) (174–177), a putative hydrolase (encoded by TM1225 in *T. maritima*), and a ROK transcription factor (ManR) (50). However, the operon does not contain the genes encoding the other ABC transporter components, so the encoded ManD is called an orphan SBP (178). The *T. petrophila*, *T. neapolitana* and *T. lettingae* genomes do not encode ManD. In addition, *T. lettingae* does not encode ManR. Since most *Thermotoga* species have *manD* and *manR*, it is likely that these genes were lost in *T. petrophila*, *T. neapolitana* and *T. lettingae*. The *F. nodosum* and *M. prima* genomes do not have any of the genes found in this operon. The distribution and location of the hydrolase-encoding genes are variable. A putative hydrolase-encoding gene found in the proximity of *manD* is also found in *M. prima* downstream of the *manEFGKL* operon. The hydrolase-encoding gene is found in all studied organism except *F. nodosum*. Downsteam of *manEFGKL* operon, *F. nodosum*

encodes an endoglucanase (Cel5A) that can hydrolyze carboxylmethyl cellulose, regenerated amorphous cellulose, barley β-D-glucan and galactomannan (179)

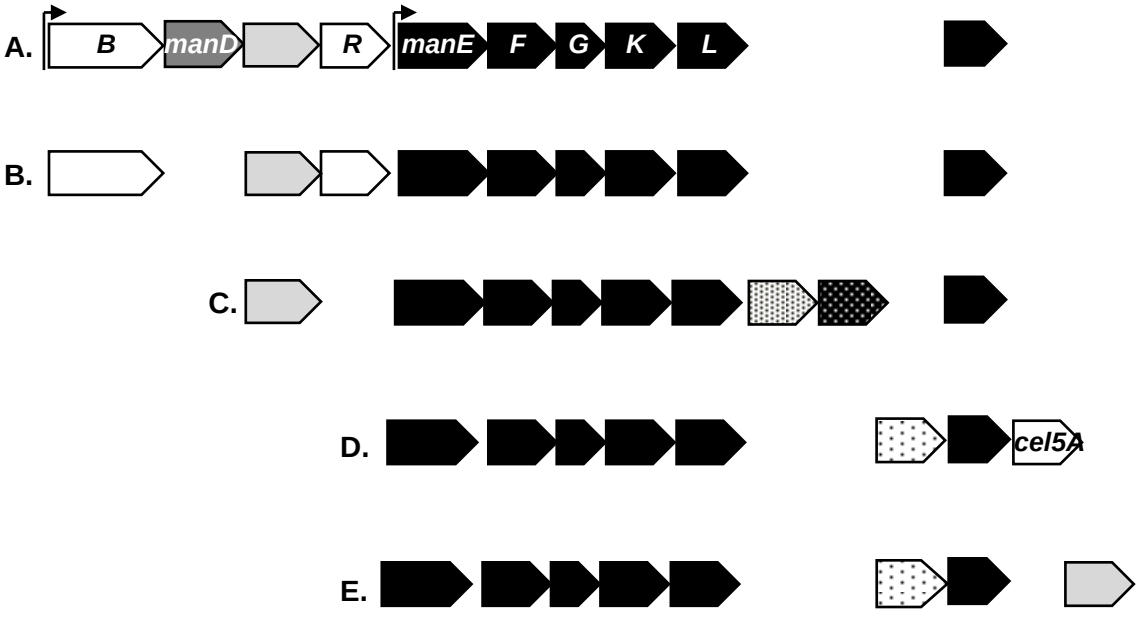


Figure 17. Syntenic regions of the mannoside ABC transporter operon *manEFGKL* and *manD* and in the Thermotogales. (A) *T. maritima*, *Thermotoga* species RQ2 and *T. naphthophila*; (B) *T. petrophila* and *T. neapolitana*; (C*) T. lettingae*; (D) *F. nodosum*; and (E) *M. prima*. The *manD* and *manEFGKL* genes are represented in dark grey and black, respectively. Genes for a putative glycoside hydrolase (light grey, encoded by TM1225 in *T. maritima*), various putative hydrolases (different stippled pattern boxes), endo-1,4-β-glucanase (*cel5A*), endo-1,4-β-mannosidase (*manB*), ROK transcription factor (*manR*) and putative transcription factor (hatched boxes) are represented. The promoters for *T. maritima* were drawn upstream of *manB* and *manE* according to Kazanov et *al.*

The thermostabilities of ManE and ManD are consistent with the OGTs of their hosts

Thermal stabilities of the recombinant proteins were examined by differential scanning fluorimetry (DSF) to estimate their unfolding temperatures (T_m) (Table 15). Although DSF is not routinely performed to determine unfolding temperatue (T_m), this assay was previously used to compare the T_m of recombinant proteins from different organisms (128). *T. maritima*, *Thermotoga sp.* RQ2, *T. petrophila* and *T. neapolitana* are the organisms from this study that can grow at the highest optimum growth temperatures (OGT), all above 77°C (25, 27, 30). At pH 7, $ManE_{Tmar}$, $ManE_{TRQ2}$, $ManE_{Tpet}$ and $ManE_{Tnea}$ proteins have the highest estimated T_m values (≥ 95°C) (Table 15). The $ManE_{Tnap}$ was not characterized in this study since its sequence is identical to that of $ManE_{Tmar}$. *T. lettingae* and *F. nodosum* have lower OGTs, ranging from 65-70°C (26, 31). At pH 7, ManE proteins from these two species have lower T_m values of 89.43 and 80.22°C, respectively (Table 15). *M. prima* is a mesophile with an OGT of 37°C, the lowest OGT of all the organisms characterized in this study (28) (Table 15). At pH 7, $ManE_{Mpri}$ was found to have the lowest unfolding temperature, 73.32°C (Table 15). Overall, the T_m values of the ManE orthologs are consistent with the OGTs of their hosts.

Only *T. maritima*, *Thermotoga sp.* RQ2 and *T. naphthophila* encode the ManE paralog, ManD (Figure 17). At pH 7, the T_m of their ManD proteins are 86.13, 89.27 and 81.55 °C, respectively, lower than those of their ManE proteins (Table 15).

Table 15. Thermal stabilities of recombinant ManD and ManE proteins determined by differential scanning fluorimetry (DSF). The unfolding temperatures (Tm) of substrate-binding proteins are compared to the optimum growth temperatures (OGTs) of their hosts. The assay was performed in triplicate and standard deviations are shown.

Organism	OGT (ºC)	Growth temperature range (ºC)	Tm at pH 7 (ºC)	
			ManE	ManD
T. maritima	80	50-90	>98[a]	86.13 ± 0.36
Thermotoga RQ2	80	50-90	>97 [a]	89.27 ± 0.13
T. naphthophila	80	48-86	-[b]	81.55 ± 0.09
T. petrophila	80	47-88	>95[b]	-[c]
T. neapolitana	77	55-90	>97[b]	-[c]
T. lettingae	65	50-75	89.43 ± 0.04	-[c]
F. nodosum	65-70	40-80	80.22 ± 0.25	-[c]
M. prima	37	20-50	73.32 ± 0.07	-[c]

[a] Partial or no unfolding temperature was detected within the maximum limit of detection of the equipment (98 ºC)

[b] The protein sequence is 100% identical to $ManE_{Tmar}$

[c] The *manD* gene is missing from the genome

The ManE and ManD orthologs interact with sugars composed of β-D-glucose and β-D-mannose

ManE and ManD were subjected to a DSF screen with different sugars to determine which sugars should be used for measurements of binding affinities (81). The complete results from these DSF measurements are listed in Table 16 and Table 17 and a summary is provided in Figure 18.

All the ManE orthologs show high thermal stabilization by β-mannobiose, β-mannotriose, cellotriose, cellotetraose, laminaribiose, laminaritriose, sophorose, salicin, β-mannotetraose, cellobiose and gentiobiose ($\Delta Tm \geq 5.3°C$) (Table 16). In most cases, their thermal stabilities in the presence of the ligands were so high that the complete unfolding temperature curves could not be observed. All ManE orthologs showed a strong interaction with glucomannan ($\Delta Tm > 6.8°C$) except for $ManE_{Mpri}$, which only interacted weakly with glucomannan (ΔTm: $3.3 \pm 3.9°C$) (Table 16 and Figure 18).

All the ManD orthologs interacted strongly with cellobiose, cellotriose and β-mannotetraose ($\Delta Tm > 7.6°C$). To a lesser extent, ManD orthologs interacted with β-mannotriose ($\Delta Tm \geq 3.8°C$), β-mannobiose ($\Delta Tm \geq 3.4°C$), laminaritriose ($\Delta Tm \geq 2.7°C$), cellotetraose ($\Delta Tm \geq 2.5°C$), laminaribiose ($\Delta Tm \geq 1.5°C$), and glucomannan ($\Delta Tm \geq 1.3°C$). Unlike the ManE orthologs, the ManD orthologs interacted only weakly with sophorose ($\Delta Tm \leq 1.1°C$) and gentiobiose ($\Delta Tm \leq 0.5°C$) (Table 17 and Figure 18).

The thermal stabilities of all recombinant ManD and ManE proteins were increased in the presence of carbohydrates containing beta linked mannose and glucose. In contrast, the ManE and orthologs interacted only weakly with α-1,3-mannobiose (ΔTm: 0.2 to

3.8°C) and the ΔTm values for the ManD orthologs for this sugar were near zero (ΔTm: -1.7 to - 0.1°C) (Table 16 and Table 17).

Table 16. Thermostabilities of ManE orthologs measured by DSF and represented by ΔTm values. The assay was performed in triplicate and standard deviations are shown. If the Tm_{ligand} could not be calculated precisely because of incomplete denaturation or inability to observe a complete unfolding curve, Tm_{ligand} was estimated using the maximum fluorescence intensities.

Ligand	**ΔTm (°C) of ManE ortholog**						
	Tmar	**TRQ2**	**Tpet**	**Tnea**	**Fnod**	**Mpri**	**Tlet**
arabinogalactose	-0.1±0.9	-0.3±0.5	0.2±0.1	2.7	2.1±0.7	0.4±0.2	2.7±0.4
α-mannobiose	3.3	0.9±0.1	3.8±0.1	0.2±0.9	1.7±0.6	1.3	1.2±0.4
arabinose	-0.2±0.1	-0.1±0.5	-0.2±0.1	0.1±0.2	0.6±0.7	0.0±0.2	-0.1±0.3
cellobiose	>5.4	>11.2	>11.5	>7.0	>20.0	>24	>8.2
cellotetraose	>12.5	>11.2	>10.0	>6.8	>20.0	24.8±0.3	>8.2
cellotriose	>12.5	>11.2	>10.0	>6.8	>20.0	24.7±0.3	>8.2
fructose	-0.1±0.1	-0.1±0.5	-0.2±0.2	0.1±0.8	-0.1±0.5	0.0±0.2	0.0±0.3
fucose	0.1±0.1	-0.1±0.5	-0.1±0.1	0.2±0.5	0.4±0.5	0.0±0.2	-0.1±0.4
galactose	1.2±0.1	1.1±0.5	0.7±0.1	0.9±0.4	0.5±0.7	3.4±0.2	1.5±0.5
gentiobiose	>5.4	9.6±0.5	9.9±0.2	5.3±0.2	12.4±0.7	10.1±0.2	>8.2
glucomannan	>12.5	10.8±1.8	>10.0	>6.8	>20.0	3.3±3.9	>7.9
glucose	3.7±0.1	4.1±0.5	5.0±1.3	4.6±0.3	2.8±0.7	4.6±0.2	3.1±0.3
kanamycin	7.4	3.0±0.4	4.9	>6.84	0.4±0.6	-1.0	0.1±0.3
lactose	4.0±0.1	3.6±0.5	3.6±0.2	3.9±0.4	1.6±0.5	7.3±0.2	3.2±0.4
laminaribiose	>12.5	>11.2	>10.0	>6.8	>20.0	21.8±0.2	>8.2
laminatriose	>12.5	>11.2	>10.0	>6.8	>20.0	24.5±0.3	>7.9
maltose	5.2±0.1	5.5±0.6	6.6±0.1	5.1±0.4	4.1±0.5	5.6±0.2	3.8±0.4

maltotetraose	0.4±0.1	1.0±0.5	0.5±0.3	0.6±0.6	1.2±0.6	0.9±0.4	1.22±1.33
maltotriose	4.4±0.1	4.8±0.5	5.5±0.3	4.9±0.3	4.5±0.9	8.6±0.2	4.7±03
mannan	-0.4±0.4	-0.6±1.1	0.0±0.1	-0.1±0.4	0.1±0.5	0.1±0.2	-0.6±0.3
mannitol	0.4±0.2	0.7±0.5	0.0±0.1	0.4±0.3	-0.2±0.5	0.5±0.2	5.3±0.2
β-mannobiose	>12.5	>11.2	>10.0	>6.8	18.4±0.5	23.3±0.3	>8.2
mannosamine	-0.6±0.5	0.0±0.5	-0.1±0.1	-0.5±0.8	0.9±0.5	1.2±0.3	1.13±0.4
mannose	3.5±0.1	3.9±0.5	4.2±0.1	4.2±0.3	4.3±0.5	6.4±0.2	2.8±0.3
β-mannotetraose	>5.4	>11.2	>11.5	>7.0	>20.0	24.5±0.3	>7.9
β-mannotriose	>12.5	>11.2	>10.0	>6.8	>20.0	24.5±0.2	>7.9
melibiose	0.3±0.1	0.2±0.5	0.2±0.1	0.3±0.4	0.3±0.5	0.1±0.2	-0.3±0.5
myo-inositol	2.2	0.3±0.5	0.5±0.3	0.5±0.9	0.1±0.5	0	-0.1±0.3
pullulan	-0.1±0.1	-0.2±0.5	0.0±0.3	-0.2±0.7	0.6±0.6	0.0±0.2	-0.1±0.2
raffinose	0.1±0.1	0.0±0.5	0.2±0.1	0.2±0.3	0.3±0.5	0.1±0.2	-0.3±0.3
rhamanose	0.0±0.1	-0.2±0.5	-0.6±0.2	-0.2±0.4	0.2±0.5	-0.1±0.2	-0.1±0.3
ribose	1.5	-0.2±0.5	0.2±0.2	0.2±0.9	0.1±0.5	-0.1±0.2	-0.1±0.2
salicin	>5.4	11.1±0.5	9.3±0.2	>7.0	17.0±0.6	14.1±0.2	>8.2
sophorose	>12.5	>11.2	>10.0	>6.8	>20.0	19.3±0.3	>7.9
sorbitol	0.7±0.1	1.3±0.5	0.7±0.1	0.9±0.4	1.2±0.5	2.2±0.2	-0.5±0.3
streptomycin	N/A	-2.6±0.3	-2.9±1.0	-3.3±2.4	-0.4±0.5	-0.9	0.2±0.3
sucrose	1.4±0.5	2.4±0.5	0.5±0.1	2.6±0.5	0.4±0.5	0.1±0.2	-0.5±0.6
tagatose	0.1±0.1	-0.1±0.5	0.3±0.2	0.2±0.3	0.3±0.6	0.0±0.2	0.1±0.3
trehalose	0.0±0.1	-0.1±0.5	0.0±0.3	0.2±0.2	0.1±0.6	0.1±0.2	-0.5±0.7
xyloglucan	0.3±0.1	0.3±0.5	0.4±0.2	0.1±0.2	0.3±0.5	0.0±0.2	0.2±0.2
xylose	0.1±0.1	-0.3±0.5	-0.3±0.2	-0.1±0.6	0.1±0.5	0.3±0.2	0.1±0.3

Table 17. Thermostabilities of ManD orthologs measured by DSF and represented by ΔTm values. The assay was performed in triplicate and standard deviations are shown. If the Tm_{ligand} could not be calculated precisely because of incomplete denaturation or inability to observe a complete unfolding curve, Tm_{ligand} was estimated using the maximum fluorescence intensities.

Ligand	**ΔTm (°C) of ManD ortholog**		
	Tmar	**TRQ2**	**Tnap**
arabinogalactose	-0.8 ± 0.4	-0.6 ± 0.4	-0.1 ± 0.1
α-mannobiose	-0.1 ± 0.5	-1.7 ± 0.7	-0.2 ± 0.1
arabinose	-0.3 ± 0.5	-0.1 ± 0.5	0.1 ± 0.1
cellobiose	>8.2	>7.6	>16.4
cellotetraose	>8.2	>8.4	2.5 ± 0.3
cellotriose	8.1 ± 0.3	>7.6	>16.4
fructose	-0.2 ± 0.5	0.2 ± 0.1	0.3 ± 0.1
fucose	-0.1 ± 0.5	-0.2 ± 0.3	0.1 ± 0.1
galactose	0.0 ± 0.5	0.2 ± 0.4	0.1 ± 0.4
gentiobiose	0.3 ± 0.5	0.5 ± 0.4	0.1 ± 0.1
glucomannan	4.3 ± 0.5	1.3	2.4 ± 0.2
glucose	-0.1 ± 0.5	0.1 ± 0.2	0.2 ± 0.1
kanamycin	0.4 ± 0.5	-1.4 ± 0.5	0.7 ± 0.4
lactose	-0.1 ± 0.5	-0.2 ± 0.2	0.0 ± 0.2
laminaribiose	5.2 ± 2.9	1.5 ± 0.2	2.2 ± 0.1
laminatriose	4.1 ± 0.8	>8.4	2.7 ± 0.6
maltose	0.0 ± 0.5	-0.2 ± 0.2	0.2 ± 0.1

maltotetraose	-0.2 ± 0.5	-0.1 ± 0.4	-0.1 ± 0.2
maltotriose	0.0 ± 0.5	0.1 ± 0.1	-0.1 ± 0.1
mannan	-0.2 ± 0.7	-0.9 ± 0.7	0.4 ± 0.1
mannitol	0.0 ± 0.4	0.0 ± 0.6	0.2 ± 0.3
β-mannobiose	8.1 ± 0.3	>7.6	3.4 ± 0.2
mannosamine	-1.5 ± 0.5	-0.3 ± 0.3	0.6 ± 0.1
mannose	-0.2 ± 0.5	-0.2 ± 0.1	-0.2 ± 0.4
β-mannotetraose	>8.1	>7.6	>16.4
β-mannotriose	>8.1	>8.4	3.8 ± 0.2
melibiose	-0.1 ± 0.5	0.4 ± 0.1	0.1 ± 0.2
myo-inositol	-0.1 ± 0.5	-1.2 ± 0.5	-0.2 ± 0.1
pullulan	-0.1 ± 0.5	-0.2 ± 0.1	0.1 ± 0.3
raffinose	0.4 ± 0.5	0.7 ± 0.1	0.0 ± 0.1
rhamanose	-0.1 ± 0.5	0.4 ± 0.4	0.0 ± 0.1
ribose	-0.2 ± 0.5	-0.3 ± 0.4	0.0 ± 0.1
salicin	0.6 ± 0.5	0.3 ± 0.5	0.3 ± 0.1
sophorose	1.1 ± 0.3	-0.4 ± 0.3	0.5 ± 0.1
sorbitol	-0.1 ± 0.5	-0.2 ± 0.2	-0.2 ± 0.3
streptomycin	0.7 ± 0.7	-1.2 ± 0.3	0.4 ± 0.1
sucrose	0.0 ± 0.5	-0.1 ± 0.2	0.2 ± 0.1
tagatose	-0.3 ± 0.5	-0.1 ± 0.2	0.2 ± 0.1
trehalose	-0.0 ± 0.5	-0.4 ± 0.1	-0.2 ± 0.5
xyloglucan	-0.0 ± 0.5	0.0 ± 0.3	0.0 ± 0.1
xylose	-0.0 ± 0.5	0.2 ± 0.1	0.0 ± 0.2

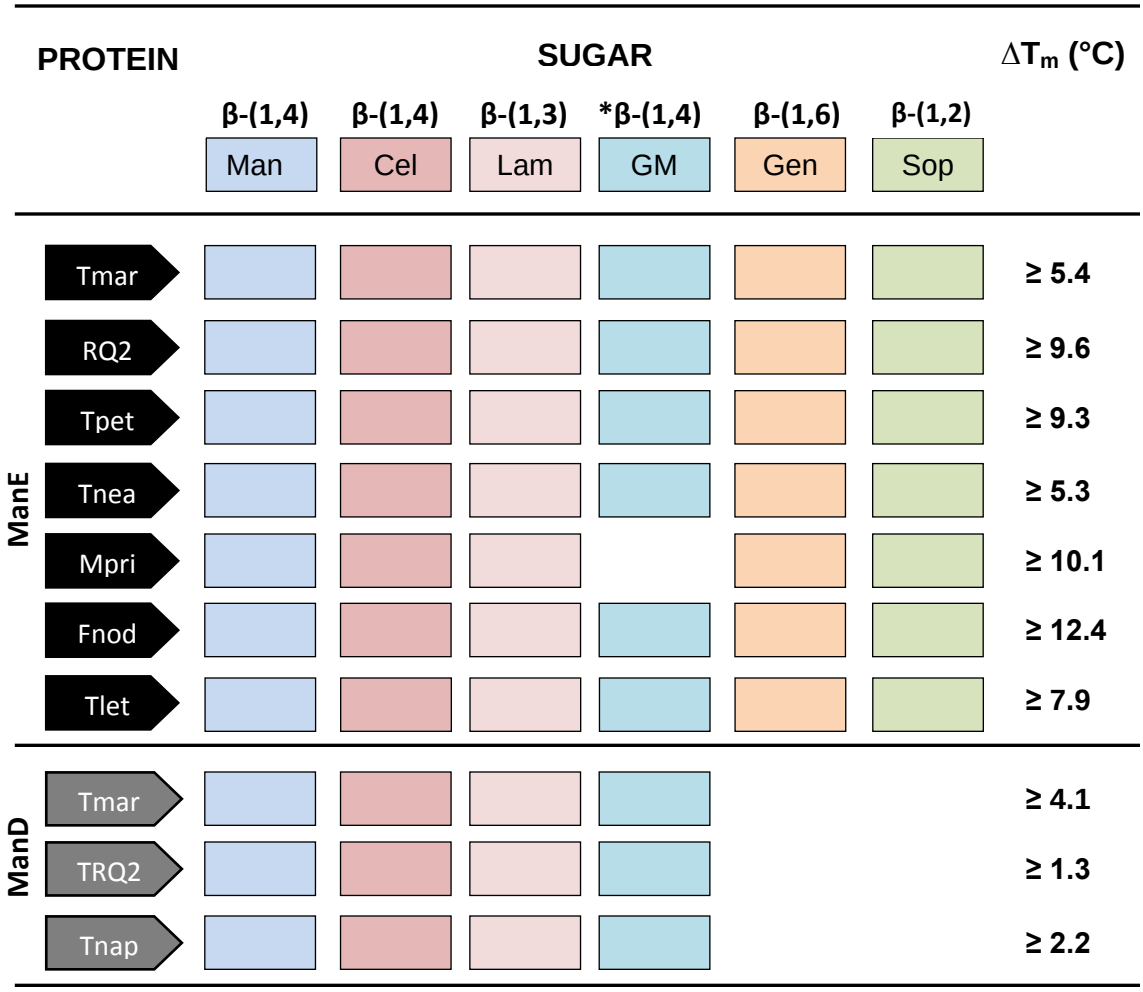


Figure 18. Summary of the interactions of the ManE and ManD orthologs with sugars determined by DSF. The presence of a box indicates that the protein interacts with the indicated sugar. (Man) β-mannobiose, -triose, tetraose; (Cel) cellobiose, -triose, -tetraose; (Lam) laminaribiose, -triose; (GM) glucomannan, (Gen) gentiobiose, and (Sal) salicin. (*) Different chains of β-D-mannose and β-D-glucose that contain mainly β-1,4-glucosyl linkages (92%). The absence of a box indicates that no or only a weak interaction was measured.

ManE orthologs bind similar sugars

To examine how the protein functions of the ManE and ManD orthologs changed in relationship to the temperatures of their hosts, their apparent binding affinities (K_d) were measured at 37°C and 60°C. Their binding properties were measured as the changes of fluorescence of their tryptophan residues upon sugar binding during a ligand titration.

Based on the ΔTm data gathered using DSF, the binding affinities of the ManD and ManE orthologs were examined in the presence of β-mannobiose, β-mannotriose, β-mannotetraose, cellobiose, cellotriose, cellotetraose, laminaribiose, laminaritriose, sophorose, gentiobiose and glucomannan. The binding properties of the ManE and ManD proteins for di-, tri- and tetrasaccharides containing β-D-mannose and β-D-glucose molecules were examined to determine if their binding affinities dramatically change by the addition of pyranose rings. A summary of the ManE and ManD orthologs' binding properties are in Table 18.

All the ManE orthologs characterized in this study bind β-mannobiose, β-mannotriose, β-mannotetraose, cellobiose, cellotriose, cellotetraose, laminaribiose, laminaritriose and sophorose at 37°C and 60°C ($K_d \leq 1027$ nM). Among these ligands, cellobiose had the highest binding affinity at 37°C and 60°C (K_d, 1-21 nM) (Table 18).

Table 18. Summary of the binding properties and apparent binding constants (K_d) for the ManD and ManE orthologs. (M2) β-mannobiose, (M3) β-mannotriose, (M4) β-mannotetraose, (C2) cellobiose, (C3) cellotriose, (C4) cellotetraose, (L2) laminaribiose, (L3) laminaritriose and (Sop) sophorose. A (-) indicates that a competitive assay demonstrated that the protein does not bind to the sugar while a (+) indicates that the protein binds the sugar.

Protein	Assay Temp. (°C)	Apparent K_d (nM)								
		M2	**M3**	**M4**	**C2**	**C3**	**C4**	**L2**	**L3**	**Sop**
$ManE_{Tmar}$	**37**	19	14	325	2	6	146	76	614	325
	60	172	65	251	6	199	69	92	443	415
$ManE_{TRQ2}$	**37**	175	402	323	21	462	190	192	392	1027
	60	183	544	379	4	598	280	191	217	828
$ManE_{Tpet}$	**37**	44	99	81	2	136	2	44	122	709
	60	86	263	116	4	183	70	63	218	972
$ManE_{Tnea}$	**37**	93	351	229	16	667	19	65	212	369
	60	193	349	168	8	443	161	109	201	271
$ManE_{Tlet}$	**37**	74	42	12	(+)	(+)	(+)	(+)	(+)	(+)
	60	26	38	49	1	29	28	(+)	(+)	283
$ManE_{Fnod}$	**37**	50	175	57	3	(+)	(+)	(+)	(+)	(+)
	60	55	110	122	6	(+)	(+)	(+)	(+)	(+)
$ManE_{Mpri}$	**37**	36	41	47	3	201	(+)	(+)	(+)	(+)
	60	82	30	36	2	112	112	26	483	145
$ManD_{Tmar}$	**37**	(+)	190	99	536	12	141	(-)	(-)	(-)
	60	(-)	201	258	1199	246	254	(-)	(-)	(-)
$ManD_{TRQ2}$	**37**	(+)	588	384	346	543	267	(-)	(-)	(-)
	60	(-)	557	429	1095	463	285	(-)	(-)	(-)
$ManD_{Tnap}$	**37**	(+)	97	101	699	117	41	(-)	(-)	(-)
	60	(-)	93	134	2067	163	44	(-)	(-)	(-)

In some instances $ManE_{Mpri}$, $ManE_{Fnod}$ and $ManE_{Tlet}$ had virtually no fluorescence changes after the addition of 300 nM cellobiose, cellotriose, cellotetraose, laminaribiose, laminaritriose and sophorose. Since the screening using DSF showed a large thermostability shift suggesting a ligand-protein interaction with these sugars, a competition assay was performed. If the addition of the test sugar did not elicit a change of fluorescence, then a sugar known to bind, β-mannotetraose (300 nM), was added to the cuvette. If the test sugar is already bound to the protein, the addition of β-mannotetraose should not elicit a change of fluorescence since the protein is already saturated with the first added test ligand. If the test sugar is not bound to the protein, then the addition of β-mannotetraose should elicit a drastic change of fluorescence. The results of such experiments are presented in Figure 19 and Figure 20.

For most of these competition experiments, 300 nM of the test ligand elicited a relative fluorescence change of <0.3 after subsequent addition of the β-mannotetraose (Figure 19 and Figure 20). These results indicate that most SBPs were saturated with 300 nM of the test sugars suggesting that their maximum K_d values are less than 150 nM. For $ManE_{Tlet}$, at 60°C, up to 10 μM laminaribiose was needed to observe <0.2 relative fluorescence change after the addition of β-mannotetraose while a concentration of 1500 nM laminaritriose and 1500 nM of sophorose was needed to observe <0.4 relative change of fluorescence (Figure 20). These results suggests that at 60°C, $ManE_{Tlet}$ maximum K_d value with laminaribiose, laminaritriose and sophorose is 5000 nM, 750 nM and 750 nM, respectively. For $ManE_{Fnod}$, at 37°C, 300 nM of the test ligand sophorose elicited a relative fluorescence change of <0.2 after subsequent addition of the β-mannotetraose, but at 60°C a large relative fluorescence change of approximately 0.78 was measured

(Figure 19). At 1500 nM sophorose, a small or no change was measured at 60°C after subsequent addition of the β-mannotetraose (Figure 19). These results suggest that at 37°C and 60°C, $ManE_{Fnod}$ maximum K_d value with sophorose is 150 nM and 750 nM, respectively.

The binding affinities with glucomannan and gentiobiose were measured by the absence or presence of an intrinsic fluorescence change using a concentration of 0.1 mg/1 ml and 10 μM, respectively. Glucomannan is a polysaccharide containing undefined chains of β-D-mannose and β-D-glucose that contains 92% β-1,4-glucosyl linkages (180). All the ManE orthologs except $ManE_{Tlet}$ had an increase of fluorescence (>5%) in the presence of this polysaccharide (Table 19). Subsequent addition of 300 nM β-mannotetraose shows that glucomannan did not saturate $ManE_{Tlet}$ (Figure 20). The ManD orthologs had a small increase of fluorescence (1.7% to 6.5%) in presence of glucomannan (Table 19). This suggests that ManD is able to bind glucomannan at concentration of 0.1 mg/1 ml.

The fluorescence in the presence of 10 μM gentiobiose varied between 0.2 to 9.8%. $ManE_{Tmar}$ had the largest fluorescence increase in the presence of gentiobiose (3.1% and 9.8% at 37°C and 60°C, respectively) and the subsequent addition of 300 nM β-mannotetraose led to an additional increase of fluorescence of 15.4% and 28.2%, respectively (Table 19). However, gentiobiose is only 85% pure and its impurities are the likely causing the small change of fluorescence (0.2 to 9.8%).

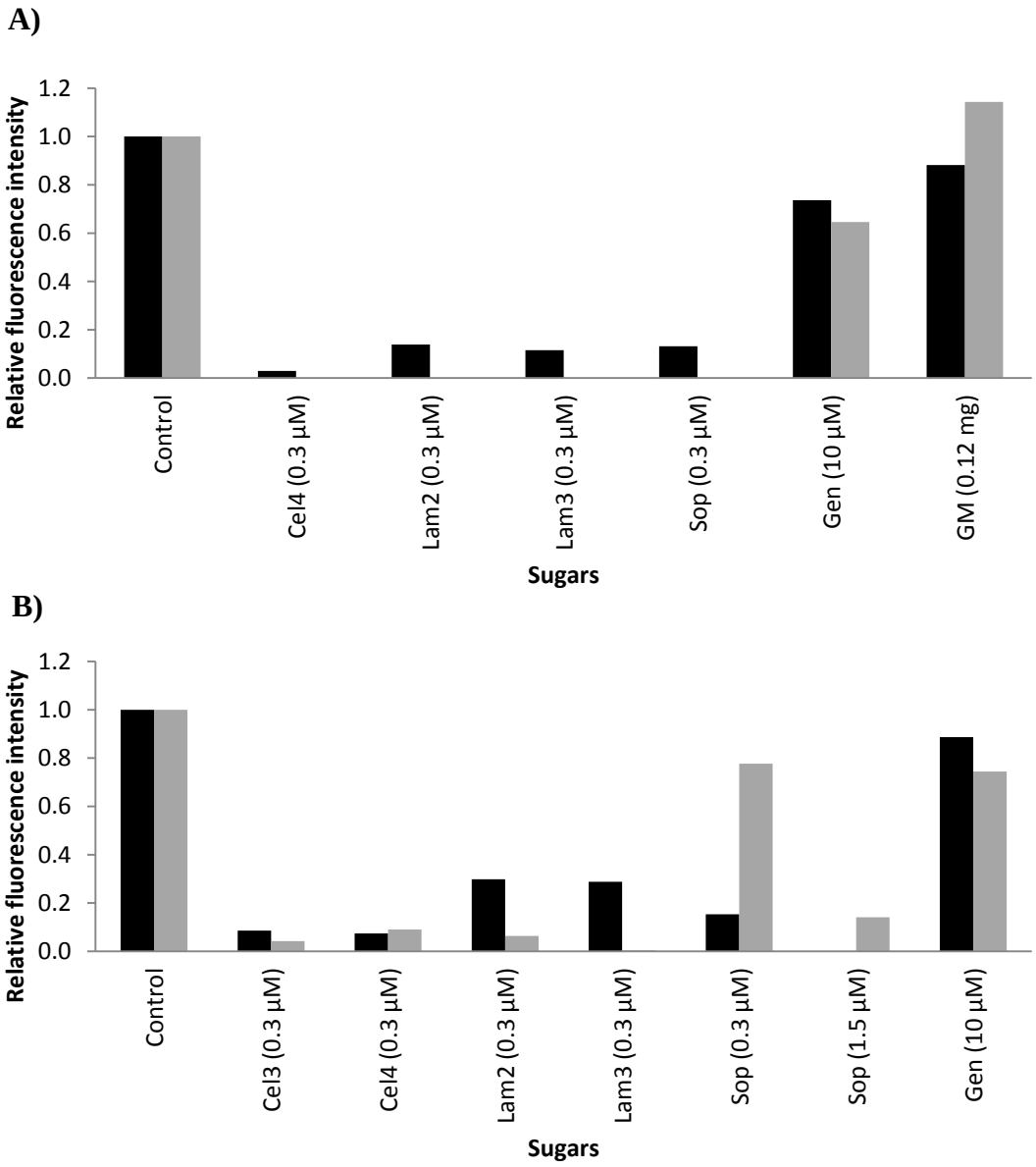


Figure 19. Competitive assay of the $ManE_{Mpri}$ and $ManE_{Fnod}$. Relative change of fluorescence after the addition of β-mannotetraose (300 µM) at 37°C (black) and 60°C (grey). (A) $ManE_{Mpri}$ (B) $ManE_{Fnod}$. (Control) β-mannotetraose, (Cel3) cellotriose, (Cel4) cellotetraose, (Lam2) laminaribiose, (Lam3) laminaritriose, (Sop) sophorose, (Gen) gentiobiose and (GM) glucomannan.

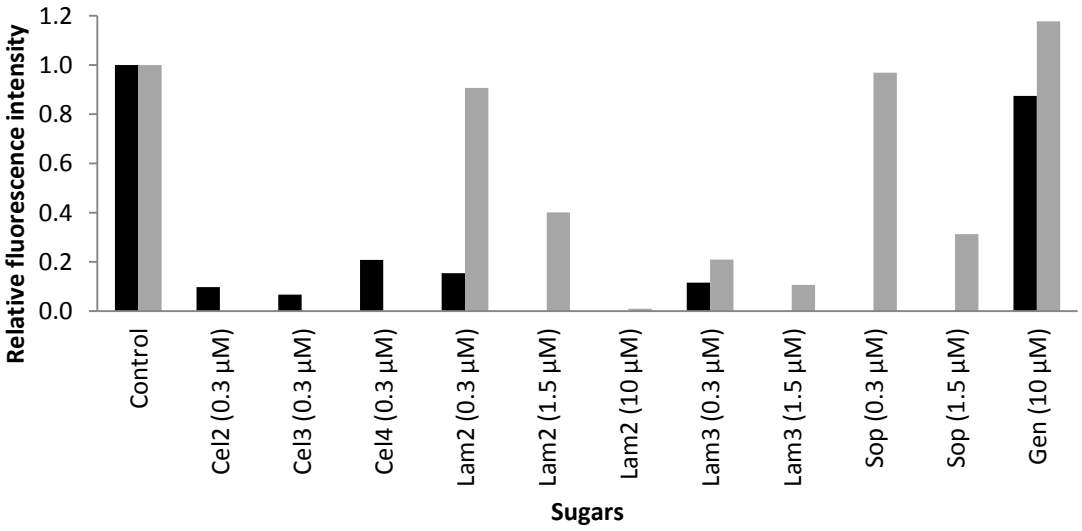


Figure 20. Competitive assay of the $ManE_{Tlet.}$ Relative change of fluorescence of $ManE_{Tlet}$ after the addition of β-mannotetraose (300 µM) at 37°C (black) and 60°C (grey). (Control) β-mannotetraose, (Cel2) cellobiose, (Cel3) cellotriose, (Cel4) cellotetraose, (Lam2) laminaribiose, (Lam3) laminaritriose, (Sop) sophorose, (Gen) gentiobiose and (GM) glucomannan.

Table 19. Maximal fluorescence change of the ManD and ManE orthologs after addition of gentiobiose (10 μM, Gen) or Konjac glucomannan (0.12 mg, GM).

Protein	Fluorescence Change (%)			
	Gen		GM	
	37°C	60°C	37°C	60°C
ManE$_{Tmar}$	3.1	9.8	15.5	28.6
ManE$_{TRQ2}$	0.2	0.4	6.5	5.0
ManE$_{Tpet}$	5.7	1.9	22.4	32.7
ManE$_{Tnea}$	8.2	6.3	25.9	41.3
ManE$_{Tlet}$	0.5	1.4	10.7	9.8
ManE$_{Fnod}$	3.1	2.6	7.0	6.0
ManE$_{Mpri}$	0.2	4.4	2.3	0.5
ManD$_{Tmar}$	0.3	0.3	3.6	1.7
ManD$_{TRQ2}$	0.2	0.4	6.5	5.0
ManD$_{Tnap}$	0.1	0.3	6.0	3.2

The ManD orthologs do not bind laminaribiose, laminaritriose and sophorose

For all the ManD orthologs characterized in this study, at both temperatures, the addition of 10 μM laminiaribiose, 10 μM laminaritriose, and 10 μM sophorose gave little to no change in fluorescence (≤ 3.1%) (Table 20). At 60°C, a similar effect was observed after the addition of 10 μM β-mannobiose (≤ 2.3%) (Table 20). To validate these results, 10 μM β-mannotriose (a ligand shown to bind) was subsequently added to the cuvette. For all the samples, the addition of 10 μM β-mannotriose lead to a change of fluorescence (≥ 6.5%) suggesting that the proteins were not saturated with the sugars (Table 20). For all the ManD orthologs, the addition 10 μM β-mannobiose at 37°C lead to a small change of fluorescence between 3.6% to 11.1% (Table 20). Subsequent addition of 10 μM β-mannotriose to the samples caused an additional change of fluorescence (≥ 5.0%) (Table 20). These results suggest that the proteins were not saturated with the test sugar.

Table 20. Competition assay of the ManD orthologs. Maximal fluorescence change of the ManD orthologs after addition of 10 μM β-mannobiose (M2), laminaribiose (L2), laminaritriose (L3) and sophorose (Sop) followed by subsequent addition of 10 μM β-mannotriose. (+) Ligand determined as able to bind the protein (≥3.7%). (-) Ligand determined as unable to bind the protein with physiological relevance (≤3.1%). The first value indicates the change of fluorescence after the addition of the corresponding sugar. The second value indicates the additional change of fluorescence after the addition of 10 μM β-mannotriose.

Protein	**Assay Temp. (°C)**	**Fluorescence Change (%)**			
		M2	**L2**	**L3**	**Sop**
$ManD_{Tmar}$	**37**	**(+)**	**(-)**	**(-)**	**(-)**
		11.1	0.4	1.6	2.4
		9.6	15.2	14.4	16.4
	60	**(-)**	**(-)**	**(-)**	**(-)**
		0.6	0.1	0.9	3.1
		10.4	13.4	13.2	12.1
$ManD_{TRQ2}$	**37**	**(+)**	**(-)**	**(-)**	**(-)**
		8.3	1.7	1.6	0.3
		9.3	12.7	12.7	13.7
	60	**(-)**	**(-)**	**(-)**	**(-)**
		2.3	0.3	1.6	1.0
		8.7	11.3	12.8	12.1
$ManD_{Tnap}$	**37**	**(+)**	**(-)**	**(-)**	**(-)**
		3.6	1.8	1.4	0.7
		5.0	7.6	9.3	7.0
	60	**(-)**	**(-)**	**(-)**	**(-)**
		1.2	1.2	0.2	1.3
		6.5	6.5	8.2	7.7

Discussion

Nanavati *et al.* characterized the binding properties of $ManE_{Tmar}$ and $ManD_{Tmar}$ at 20°C, a temperature well below the OGT of this organism (80°C, Table 15) (8). That study found that $ManD_{Tmar}$ binds with high affinity β-mannotetraose (K_d: 0.38 μM) and β-mannotriose (K_d: 1.05 μM) and with lower affinity cellobiose (K_d: 9.5 μM), galactosyl mannobiose (K_d: 10 μM) and β-mannobiose (K_d: 15 μM). They found that $ManE_{Tmar}$ only binds β-mannobiose (K_d: 13 μM) and does so with low affinity. These results were interesting because they suggested that the ManE or the ManD orthologs have divergent functions.

To understand how the functions of ManD and ManE orthologs have changed, their binding properties were measured at 37°C and 60°C. These temperatures allow a comparison of their functional temperature ranges to the temperature growth ranges of the corresponding hosts. At 37°C and 60°C, $ManD_{Tmar}$ binds with high affinity β-mannotriose (K_d: 90 nM and 201 nM), β-mannotetraose (K_d: 190 nM and 258 nM), respectively, while it binds weakly cellobiose (K_d: 536 nM and 1200 nM), and not β-mannobiose (Table 18). The binding specificity of $ManD_{Tmar}$ is comparable to the Nanavati study at 20°C although the overall affinity at 37°C and 60°C is higher.

The $ManE_{Tmar}$ binding properties were the most surprising. While Nanavati, et al found that at 20°C $ManE_{Tmar}$ only binds β-mannobiose, I found it had a broader binding specificity with a high affinities (0.002 to 0.614 μM, Table 18), at 37°C and 60°C. $ManE_{Tmar}$ binds β-mannobiose but also β-mannotriose, β-mannotetraose, cellobiose, laminaribiose, laminaribiose, sophorose and also glucomannan (Table 18 and Table 20). All these sugars but sophorose and laminaritriose were tested in the Nanavati study but

they did not elicit a change of the fluorescence by the movement of the protein's tryptophan content (Nanavati et al., 2006). The increase in the sugar specificity and affinity at higher temperatures can be explained by a previous analysis of the thermophilic *T. litoralis* TMBP (181, 182). This SBP has an unfolding temperature higher than 90°C which is similar to those of $ManD_{Tmar}$ and $ManE_{Tmar}$ (Table 15). Herman et *al.* analyzed the structure of TMBP at low (20°C) and high (100°C) temperatures and found that at lower temperature, the structure of the unbound TMBP is compact while at 90°C, the unbound TMBP is partially relaxed. The overall unfolding temperatures (T_m) of the ManE orthologs are higher than those of the ManD orthologs which could be reflected in their respective binding specificities at lower (20°C) and higher (37°C and 60°C) temperatures. It is possible that the ManE orthologs structure is more compact a 20°C than at 37°C and 60°C. This could reduce the flexibility of these SBPs and therefore lead to less ligand specificity at 20°C. The binding specificities of the $ManE_{Mpri}$, $ManE_{Tlet}$, $ManE_{Fnod}$ which are found in organisms with OGTs between 37° and 70°C were similar to those found in organisms with higher OGTs (above 70°C) (Table 15 and Table 18). These results indicate that a temperature of 37°C is sufficient to obtain full binding specificity regardless of the OGT of the host.

The ManE orthologs can bind β-mannobiose, -triose and –tetraose, cellobiose, -triose and –tetraose, laminaribiose and laminaritriose and sophorose with high affinities. The ManD orthologs have similar but fewer functions than the ManE orthologs since they do not bind laminaribiose, laminaritriose sophorose and β-mannobiose. All these sugars are glucose and mannose di- and oligosaccharides containing β-linked glycosidic bonds. The protein thermostabilities of the ManD and ManE homologs were not increased in the

presence of α-1,6-mannobiose, maltose, maltotriose and trehalose suggesting that these SBPs do not interact with mannose and glucose di- and oligosaccharides with α-linkages. The ManE orthologs have broader specificities for glucose di- and oligosaccharides with other glycosidic linkages such as β-1,3-linkages (laminaribiose and -triose) and β-1,2-linkage (sophorose) while the ManD proteins only bind glucose and mannose di- and oligosaccharides with β-1,4-linkages.

Conclusion

The aim of the study was to determine the binding properties of representatives of the mannoside-binding proteins in the Thermotogales to determine if their functions are temperature-dependent and to understand how their functions have changed through time.

This study demonstrated that at 37°C and 60°C, the ManD orthologs have fewer functions than the ManE orthologs. ManD orthologs do not bind β-mannobiose, laminaribiose, laminaritriose and sophorose. Its binding properties indicate that they bind only sugars with β-1,4 glycosidic linkage and they bind disaccharides with less affinity than oligosaccharides. ManE orthologs have a broader sugar specificity that includes sugars in β-1,4, β-1-3 and β-1-2 glycosidic linkages and their binding affinity for di- and oligosaccharides are similar. Although the binding specificities and affinities of the ManE orthologs encoded by different Thermotogales are similar, this study as well as a previous analysis suggests that their binding affinities and specificities are much higher at temperatures above 37°C.

This study demonstrated that the ManD and ManE orthologs have redundant functions but the ManD orthologs have fewer functions than do the ManE proteins. The analysis could not establish any new functions for the ManD orthologs, which could explain the reasons why some *Thermotoga* species maintained this gene. Although common mono, di, tri and tetrasaccharides were tested, the growth substrates they use in their natural environments are unknown. In their extreme environments, their carbon sources might sustain thermochemical changes which can affect their structure and such sugars have not been tested in this study. If such modified sugars were examined, perhaps physiologically relevant differences between the ligand affinities of the ManD and ManE

orthologs would be revealed and explain why the ManD orthologs were retained by some species.

Additional analysis to determine the binding properties of the ManD and ManE with galacto-manno-oligosaccharides and gluco-manno-oligosaccharides might be interesting. At 20°C, $ManD_{Tmar}$ binds β-1,4-galactosyl mannobiose with a K_d of 10 μM (8). *F. nodosum* encodes an endo-1,4-β-glucanase (Cel5A) and many *Thermotoga* species encode an endo-1,4-β-mannosidase (ManB) ortholog. These hydrolases are likely anchored in the toga and they can hydrolyze glucomannan or galactomannan into di- and oligosaccharides. In this study, the binding properties with glucomannan were measured, but because the polysaccharide has an undefined molecular weight no K_d values could be obtained. Since ManD and ManE homologs are likely to bind with better affinities to di- and trisaccharides such as galacto-manno-oligosaccharides and gluco-manno-oligosaccharides than their polysaccharide versions like glucomannan, those oligosaccharides should be tested to perhaps reveal ligand differences between ManE and ManD orthologs.

Additional analyses of dN/dS ratios of the ManD and ManE encoding genes can be performed. These analyses can provide more insights on the type of selective pressures acting on those genes and therefore improve our understanding on how these SBPs have diverged and why some Thermotoga species maintained two copies in their genome.

Chapter 7

Substrate adaptability of mannoside-binding proteins as a function of their evolutionary histories

Introduction

Members of the Thermotogales lineage contain an unusually high proportion of genes encoding ABC transporters (42). Many have been characterized and are often found to have redundant functions (8, 81, 93, 95, 118). Phylogenetic studies show that the mannoside-binding proteins (ManE and ManD) were transferred between archaea and some members of the Thermotogales (8, 14). It is suspected that the ancestral SBP gene underwent a gene duplication event in the *Thermotoga* lineage, which resulted in the presence of paralogous copies (*manE* and *manD*). ManE is present in the *Thermotoga*, *Mesotoga* and *Fervidobacterium* genera and species of these genera have a broad range of optimal growth temperatures, from 37°C to 80°C.

The evolutionary history of the ManD- and ManE-encoding genes can be determined with good confidence and these genes were disseminated through both horizontal and vertical inheritance. Therefore, the mannoside-binding proteins represent an excellent system to conduct different analyses to study the evolutionary mechanisms of the change of function leading to substrate adaptability. The ManD orthologs, the paralogs of the ManE orthologs, were initially suspected to have additional functions based on a previous

study of the ligand specificities of these proteins from *T. maritima* performed at 20°C. In this study the binding properties of these proteins from *T. maritima* and several of its relatives were measured at 37°C and 60°C and revealed that the ManD and ManE homologs have, instead, overlapping functions (see Chapter 6). Surprisingly, no novel functions were found for the ManD paralog encoded by some *Thermotoga* species.

To explain why these genes were maintained in some *Thermotoga* lineages and not in others,, the selective pressure that acted on the ManD and ManE-encoding genes were examined in this chapter. The residues involved in the binding sites of $ManD_{Tmar}$ and $ManE_{Tmar}$ were identified by structural superposition using as a reference the structure of $BglE_{Tmar}$, a relative of the mannoside-binding proteins crystallized with cellobiose or laminaribiose bound to it. A branch-site analysis was performed on nucleotides sequences of homologs of ManE and ManD to determine whichcodons have evolved under neutral evolution and positive selection. Taken together, these analyses suggest that following their divergence from the ancestor of *manD* and *manE* the sequence involved in ligand binding evolved under positive selection, and this selection likely caused the changes observed in ligand binding in the present day ManD and ManE proteins.

Material and Methods

Data acquisition and phylogenetic analysis

The protein alignments were done in MUSCLE v3.8.31 (137, 138). The maximum likelihood tree was calculated using PhyML v3.0 using the WAG substitution model, discrete gamma model under 8 categories, an estimated number of invariable sites, and the best of NNIs and SPRs topology search with 5 random starting trees and 100 bootstrap replicates. The tree was rooted with *T. maritima* TM0071, a previously annotated as an oligopeptide-binding protein that binds xylosides (8), that has orthologs in other Thermotoga species.

Crystal structure superposition and structural alignments

The crystal structure of $ManE_{Tmar}$ (TM1223, pdb: 1vr5) was superimposed on the crystal structure of BglE (TM0031, pdb: 2o7i) in UCSF Chimera 1.8.1 (183) using the default option of the tool MATCHMAKER and MATCH →align. Additional sequences were manually aligned using the default option of the tool MATCH →align. The structural prediction of $ManD_{Tmar}$ (TM1226) was performed using I-TASSER (184). The sequence alignment output generated by CHIMERA is available in Appendix 5 and Figure 23 (without part of the signal peptide and His-tag sequences). The LIGPLOT diagrams were available on the PDBsum database and generated by LIGPLOTv.4.5.3 (185).

dN/dS ratio: branch-site model analysis

To detect episodic selection and to test whether the codon sites were under neutral or positive selection along the *manD* and *manE* branches (Figure 21), an analysis using the branch-site model was performed as implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) software package (186, 187). An alignment of the 19 sequences including the sequences found in the *manE* and *manD* branches of the mannoside-binding protein tree was used to calculate the maximum likelihood (ML) tree (Appendix 6 and 7) (Figure 17, brown dashed box). The protein alignment and ML tree were done as described in the previous section (see data acquisition and phylogenetic analysis). The corresponding codon based alignment was generated using TranslatorX server (188). The branch-site model allows the dN/dS ratio or omega (ω) to vary among codon sites and among branches. The branch tested for positive selection, either the *manD* or *manE* branch (Figure 21) is referred to as the foreground while the other branches are referred to as the background. The modified model A was used and specified in the PAML analysis using the following variables: codeml, model = 2, NSsites = 2 (189, 190). The model estimates three ω values ($0< \omega_0 <1$, $\omega_1 = 1$ and $\omega_2 \geq 1$) and classifies the codon sites into four different categories (class 0, 1, 2a and 2b). Class 0 includes the codon sites under purifying selection ($0< \omega_0 <1$) on all the branches while class 1 contains the codon sites under neutral selection ($\omega_1 = 1$) on all the branches. For class 2a and class 2b, the codon sites on the foreground branch are inferred under positive selection ($\omega_2 \geq 1$). However, in class 2a the codon sites on the background branches are determined under purifying selection ($0< \omega_0 <1$) while in class 2b the codon sites on the background branches are determined under neutral evolution ($\omega_1 = 1$). This

alternate model was compared by likelihood ratio test (LTR) to a null model that fixes the foreground branch at 1 ($\omega_1 = 1$). The Bayes Empirical Bayes (BEB) method was used to calculate posterior probabilities for each codon site classes. An additional LTR was performed which compared the alternate model (H_1: $\omega_2 \geq 1$) to the null model (H_0: $\omega_2 = 1$). The H_1 had a significantly better fit than H_0. All the datasets (trees, codon alignments and control files) used for these analyses are available in Appendix 6 to 13.

The codon sites provided by the PAML output were changed to correspond to the residue positions in the translated *manD*$_{Tmar}$ and *manE*$_{Tmar}$ sequences, so that the codon site and the position residue numbering were the same. Since the sequences of *manD*$_{Tmar}$ and *manE*$_{Tmar}$ are not the same size, corresponding codon sites might translate a residue at a different position in each sequence. These are indicated by the translated position from the *manD*$_{Tmar}$ sequence followed by the symbol (/) and by the translated position from the *manE*$_{Tmar}$ sequence, respectively.

Results

Phylogenetic analysis

The general relationship between the different substrate-binding proteins is difficult to establish due to horizontal transfers and gene duplication events that have occurred throughout the evolutionary history of these proteins. However, the phylogenetic history of these SBPs in the *Thermotoga* lineage mirrors their 16S rRNA gene phylogeny (Figure 21), which indicates that the common ancestor of the *Thermotoga* may have already possessed the mannoside-binding proteins prior to speciation within this group. The branch that contains the $ManE_{Tmar}$ orthologs is designated as the *manE* branch while the *manD* branch that contains the $ManD_{Tmar}$ orthologs is designated as the *manD* branch (Figure 21). These branches also contain other ManE orthologs ($ManE_{TRQ2}$, $ManE_{Tnap}$, $ManE_{Tnea}$, $ManE_{Tpet}$ and Man_{Tlet}) or ManD orthologs ($ManD_{Tmar}$, $ManD_{TRQ2}$ and $ManD_{Tnap}$) (Figure 21). The binding properties of these SBPs as well as $ManE_{Mpri,}$ $ManE_{Fnod}$ were determined as described in Chapter 6 at 37°C and 60°C and by Nanavati et *al.* at 20°C (93). The phylogeny suggests that the gene ancestral to *manE* and *manD* was acquired horizontally by the common ancestor of the Thermotogales from the Archaea (Figure 21). Following this acquisition, either this ancestral gene duplicated to give rise to the modern *manE* and *manD* orthologs or the presence of paralogs is the result of a gene transfer of an orthologous gene from an unsequenced or extinct organism close *Thermotoga* relative prior to its speciation events (Figure 21).

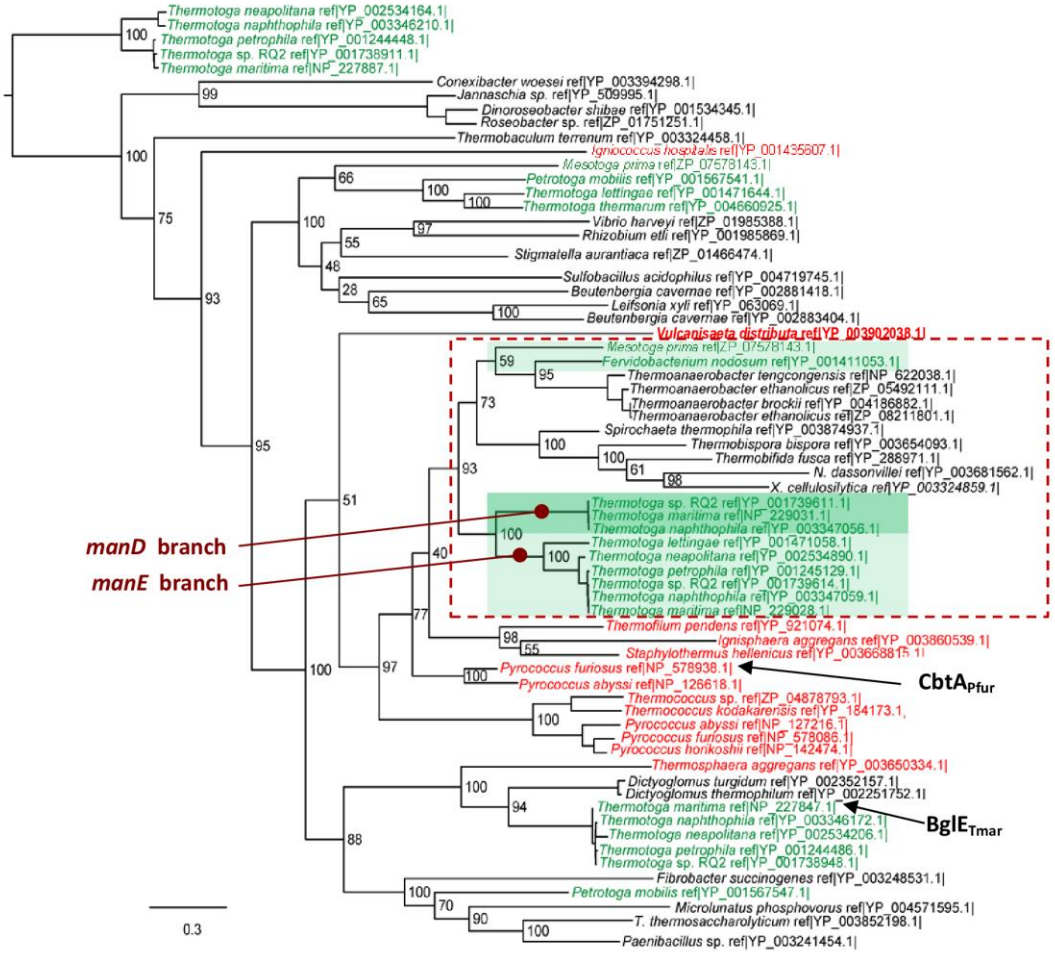


Figure 21. Rooted phylogenetic tree depicting the relationships among mannoside-binding proteins and related SBPs. The Thermotogales are in green text while the archaea are in red and other bacteria are in black. The scale bar indicates an evolutionary distance of 0.3 amino acid substitutions per position. The numbers on the nodes represent the bootstrap support values over 100 replicates. The green boxes represent the mannoside-binding proteins characterized in this study and the ManD orthologs are highlighted in a dark shade while the ManE orthologs are in a light shade. The tree was rooted with representatives of the xyloside-binding protein. The brown dashed box represents the species in the maximum likelihood (ML) subtree use for the PAML branch-site model.

Detection of the codon sites under neutral evolution and positive selection

The results of the branch-site analysis for class 1 and 2a are summarized in Table 21. To keep the same codon sites and the residue position numbering, the codon sites provided by the PAML output were changed to correspond to the residue positions of the translated *manD*$_{Tmar}$ or *manE*$_{Tmar}$ sequence. If the codon sites encode different residue positions, they are designed by the position in the translated *manD*$_{Tmar}$ sequence followed by the symbol (/) and by the translated position in the *manE*$_{Tmar}$ sequence, respectively.

The branch-site model estimates three ω values and classifies the codon sites into class 0, class 1, class 2a and class 2b. Class 1 includes the codon sites that are inferred under neutral selection (ω_1 = 1) on all the branches either on the foreground or the background branch. Four codon sites were found under neutral evolution (4, 20, 539/535 and 546/542) on both branches. However, when the branch leading to *manE* is set as the foreground branch, additional codon sites were determined under neutral evolution (125, 258, 305, 457 and 459).

Class 2a includes codon sites on the foreground branch under positive selection ($\omega_2 \geq$ 1) while the codon sites on the background branch are inferred under purifying selection ($0< \omega_0 <1$). The branch leading to *manD* has 8 codon sites found under positive selection (102, 148, 177, 181, 516, 224, 274 and 521) while the branch leading to *manE* has 3 codon sites found under positive selection (158, 444, and 514). An additional LTR was performed to determine if the model testing for sites under positive selection is significantly better ($\omega_2 \geq 1$) than the model under relaxation of the purifying selection (ω_1 = 1). The alternate model (H_1:$\omega_2 \geq 1$) had a significantly better fit than the null

model H_0 (H_0: $\omega_1 = 1$), which indicates that some codon sites were under positive selection along the foreground branch. No codon site with a posterior probability greater than 0.95 was found for class 2b. This indicates that no codon sites on the foreground branch were under positive selection ($\omega_2 \geq 1$) while codon sites on the background branch were under neutral evolution ($\omega_1 = 1$).

Table 21. List of the codon sites for class 1 and 2a using the branch-site model calculated by Bayes Empirical Bayes (BEB) in PAML with posterior probabilities greater than 0.95. Codon site numbers have been converted to the residue numbers in the amino acid sequences of their corresponding proteins to allow comparisons with residues mentioned in the structural assignments. If the same codon site is identified in both branch but translates a residue at a different position in the translated *manD*$_{Tmar}$ and *manE*$_{Tmar}$ sequences, they are indicated by the same letter in superscript.

Class	**Foreground branch**					
	manD			***manE***		
	Site	**Residue**	**p-value**	**Site**	**Residue**	**p-value**
1	4	F	0.9589	4	F	0.9566
	20	Q	0.9554	20	Q	0.9536
				125	P	0.9555
				258	K	0.9584
				305	E	0.9515
				457	S	0.9535
				459	Y	0.9514
	539[a]	W	0.9607	535[a]	V	0.9568
	546[b]	W	0.9534	542[b]	W	0.9614
2a	102	N	0.9665			
	148	K	0.9687			
				158	I	0.9700
	177	T	0.9756			
	181	S	0.9634			
	224	Y	0.9942			
	274	H	0.9909			
				444	I	0.9760
				514	A	0.9638
	516	G	0.9715			
	521	I	0.9581			

Residues involved in the change of function

Of the ManD and ManE homologs previously characterized (Chapter 6), only the structure of $ManE_{Tmar}$ was solved, but in apo form only (pdb: 1vr5). Cuneo *et al.* solved the structures of $BlgE_{Tmar}$, a close relative of the mannoside-binding proteins, in holo form (118). Its crystal structure was solved bound to cellobiose (pdp: 2o7i) and cellopentaose (pdp: 3i5o), and more recently to laminaribiose (pdp: 4jsd) and laminaripentaose (pdp: 4jso). Figure 22 illustrates the non-covalent interactions between BglE and cellobiose and laminaribiose. $BglE_{Tmar}$ binds the non-reducing end of the sugars through non-covalent interactions such as hydrogen binding and hydrophobic contact Figure 22. Glucose does not bind $BglE_{Tmar}$ (8) and does not increase the Tm significantly (118) suggesting that there are not enough interactions in the binding pocket to promote binding. Figure 22 shows the LigPlot for $BglE_{Tmar}$ bound with cellobiose and laminaribiose. Note that the positions are shifted on the LigPlot and, for clarity, only the positions bound to cellobiose will be used in the text.

Although, the $ManE_{Tmar}$ was solved without a ligand, a structure superposition was performed using $BglE_{Tmar}$ as reference to generate a multi-sequence alignment (Figure 23). Because $BglE_{Tmar}$ has similar functions as $ManD_{Tmar}$ and $ManE_{Tmar}$ and binds cellobiose and laminaribiose (8, 118), the sequence alignment provides additional data on the residues in $ManE_{Tmar}$ $ManD_{Tmar}$ that may interact with glucose di- and oligosaccharides with β-1,3 (laminaribiose) and β-1,4 (cellobiose) linkages (Figure 23 and Appendix 5). These residues might be significant for $ManE_{Tmar}$ since this SBP binds glucose di- and oligosaccharide containing a β-1,3-linkage while $ManD_{Tmar}$ does not and binds cellobiose with a lower affinity than does $ManE_{Tmar}$ (Chapter 6). Figure 23

highlights the residues in $BglE_{Tmar}$ that interact directly only with cellobiose (yellow) and both sugars (aqua) aligned with the corresponding residues in $ManD_{Tmar}$ and $ManE_{Tmar}$ on the sequence alignments.

Most residues in $ManD_{Tmar}$ and $ManE_{Tmar}$ are identical at positions 13, 16, 216, 232, 233, 381, 536 while the residues at positions 14, 234, 427 and 511 have an amino acid substitution. The most notable residue is located at position 14 in the $BglE_{Tmar}$ structure. This residue (Ala) interacts with cellobiose but not with laminaribiose suggesting that a change of this residue can affect the binding specificity and affinity of the protein for cellobiose. When $BglE_{Tmar}$ binds cellobiose, the O6 oxygen at the reducing end of the cellobiose is available to create a hydrogen bond with A14 and forms a hydrophobic contact with W536 while O1 and O2 form hydrophobic contacts with W427. When $BglE_{Tmar}$ is bound to laminaribiose, the O6 of glucose only forms a hydrophobic contact with W427, but O2 forms a hydrogen bond with G13 and a hydrophobic contact with W536. The sequence aligment indicates that the $ManE_{Tmar}$ has an alanine while $ManD_{Tmar}$ has a serine at the position corresponding to the position 14 in the $BglE_{Tmar}$ structure. The residue variation at this position in the $ManD_{Tmar}$ sequence might lead to a decrease of affinity for cellobiose compared to $ManE_{Tmar}$ as previously measured in Chapter 6 (at 37°C: 2 nM for $ManE_{Tmar}$ and 536 nM for $ManD_{Tmar}$, see Table 18). The $ManE_{Tmar}$ and $ManD_{Tmar}$ have identical at position corresponding to 536 in $BglE_{Tmar}$. However at position corresponding to 427, $ManD_{Tmar}$ contains an asparagine while the $ManE_{Tmar}$ contains a phenylalanine. Phenylalanine residues are non-polar with similar structures while asparagine is a polar amino acid that does not form hydrophobic contacts, which could explain why $ManD_{Tmar}$ does not bind laminaribiose and

laminaritriose and binds cellobiose with less affinity than $ManE_{Tmar}$ and $BglE_{Tmar}$. At position corresponding to 234, $ManE_{Tmar}$ contains a phenylalanine while $ManD_{Tmar}$ contains a methionine, both hydrophobic residues. The LIGPLOT shows that $BglE_{Tmar}$ contains a phenylalanine at position 234 (Figure 22). However, its side chain does not appear to interact with cellobiose. $ManE_{Tmar}$ has a tryptophan at position corresponding to 511 while $ManD_{Tmar}$ has an alanine. As seen in Figure 22, the tryptophan amino group forms a hydrogen bond with cellobiose and laminaribiose. The change of the residue from tryptophan to alanine might decrease the substrate specificity of the $ManD_{Tmar}$ for cellobiose and prevent the binding for laminaribiose since this residue is not able to form as strong hydrogen bonding as tryptophan, which is important for substrate binding.

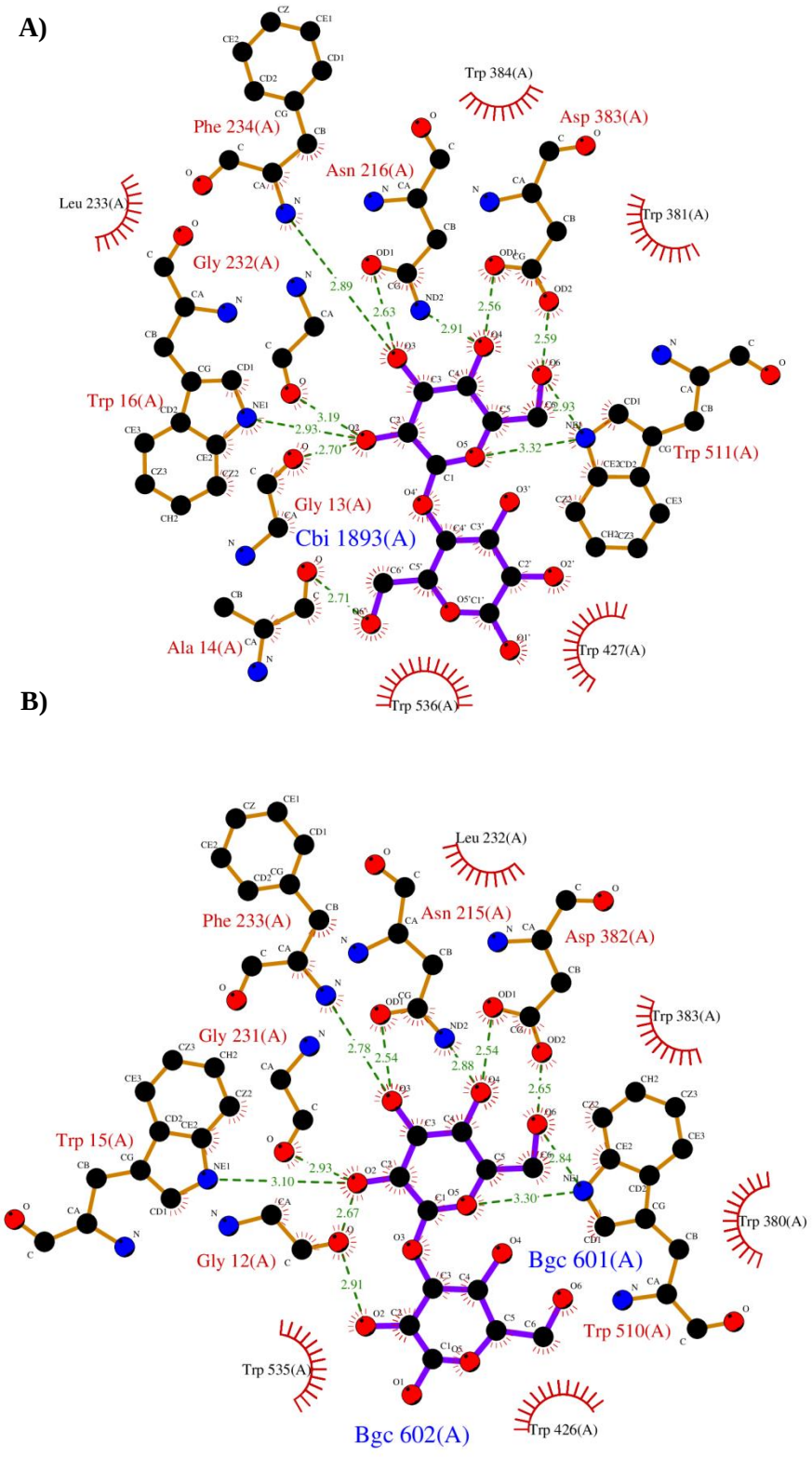


Figure 22. LIGPLOT of $BglE_{Tmar}$ bound to sugars. A) Cellobiose (pdb: 2o7i) and B) laminaribiose (pdb: 4jsd).

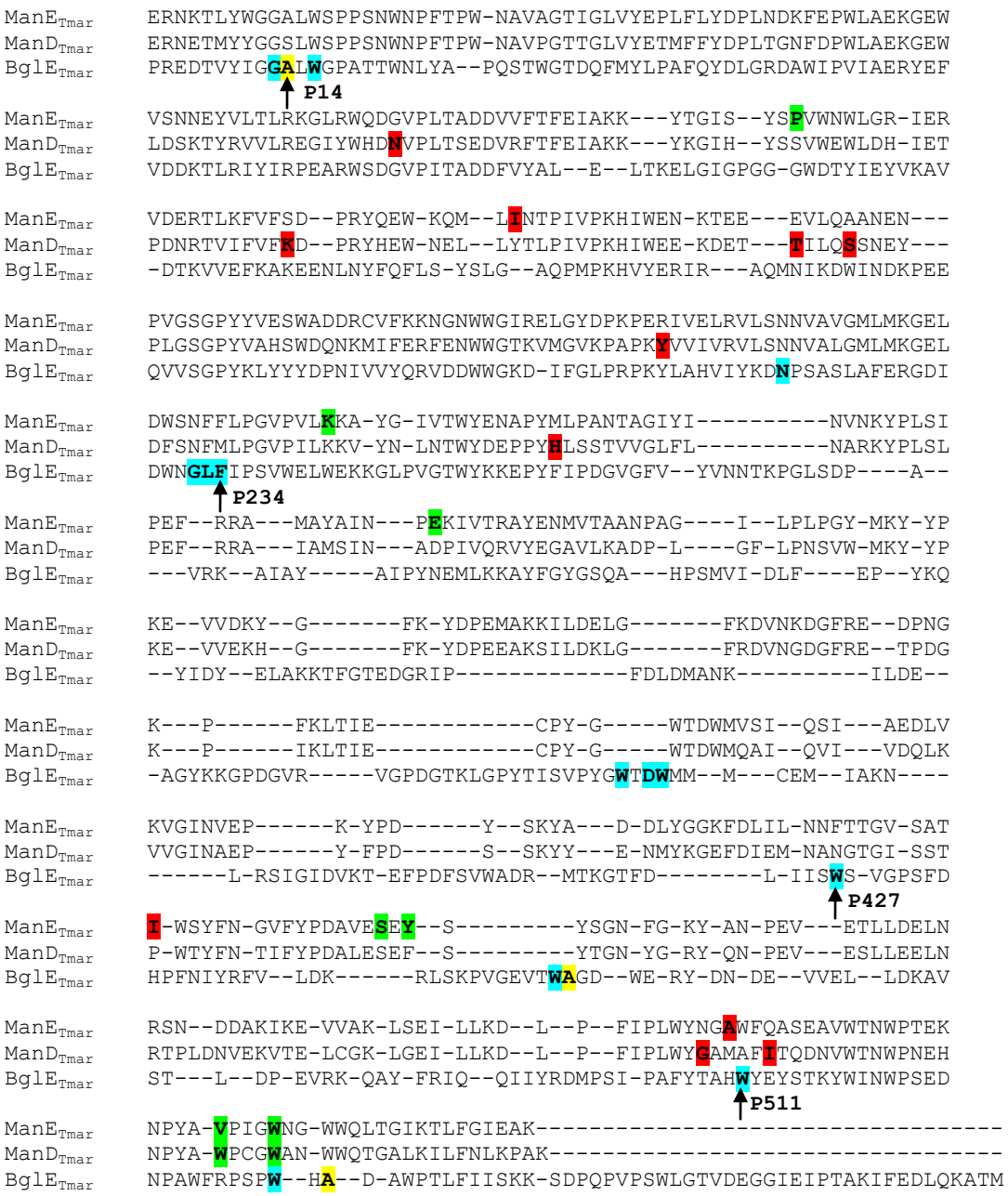


Figure 23. Multi-sequence alignment of ManE$_{Tmar}$, ManD$_{Tmar}$ and BglE$_{Tmar}$. The residues on the BglE$_{Tmar}$ sequence highlighted in aqua indicate that they interact with cellobiose and laminaribiose while those in yellow only interact with cellobiose. The residues highlighted in green and red on the ManE$_{Tmar}$ and ManD$_{Tmar}$ sequences are encoded by codon sites with a $\omega_1 = 1$ and $\omega_2 \geq 1$, respectively. The arrows indicate the residues that are different in ManE$_{Tmar}$ and ManD$_{Tmar}$ sequences that are predicted to interact with the sugar ligand based on the BglE$_{Tmar}$ crystal structures.

Discussion

Horizontal gene transfers in prokaryotes play important roles in evolution, enabling organisms to acquire new phenotypic traits that likely increase survival in changing conditions and perhaps allowing them to colonize new ecological niches. However it is unclear how a horizontally acquired gene adapts to its newfound host and how is the gene is maintained over time. Representatives of the mannoside-binding proteins encoded in the Thermotogales were characterized to study their changes of function leading to substrate adaptability.

An initial study of the binding properties of ManE and ManD in one species, *T. maritima*, indicated that these two proteins have different binding properties, perhaps as a function of their evolutionary histories. However, when the binding properties of several orthologs from different Thermotogales species were examined at temperatures closer to the OGTs of the host organisms, their functions were found to be more similar with the exception that the ManD orthologs do not bind laminaribiose, laminaritriose, sophorose and mannobiose (see Chapter 6). To understand the evolutionary processes leading to substrate selection the residues that interact with cellobiose and laminaribiose were determined by structure superposition and the selective pressures that acted on their codons were examined by branch-site analysis.

T. maritima, *Thermotoga species* RQ2 and *T. naphthophila* have two copies of mannoside-binding protein encoding genes, one each of *manD* and *manE*. A branch-site analysis was performed to determine the codon sites under neutral evolution and positive selection on the *manD* and *manE* branches (Figure 21). The analysis suggests that four codon sites were under neutral evolution (4, 20, 539/535 and 546/542) on both branches

(Table 21). The codon sites at position 4 and 20 are located in a region coding the signal peptide while the codon sites at position 539/535 and 546/542 are located in the region coding the mature protein (Figure 23). Interestingly the codon sites 539/535 and 546/542 encode residues that are located in the binding pocket Figure 24 and Figure 25. The codon site 546/542 encodes a residue that corresponds to the residue at position 536 of $BglE_{Tmar}$ which directly interacts with cellobiose and laminaribiose (Figure 22 and Figure 23). PAML analysis categorized these as codon sites under neutral evolution. These codon sites encode residues located in the binding site, and more so, one (546/542) directly interact with the sugar. It is possible that these codon sites identified as neutral were in transition to become essential for the function of the protein (purifying selection) since these codons encode a tryptophan (except for codon site 535 in ManE), and this residue is able to form strong hydrogen bonding which is important for the substrate binding. Additional codon sites were found to be under neutral evolution (125, 258, 305, 457 and 459) on the branch leading to *manE* (Table 21) while they were not identified on the branch leading to *manD*. These codon sites are likely an artifact from the analysis since they were only present on the branch leading to *manD*.

The PAML analysis suggests that 8 codon sites on the branch leading to *manD* have been under positive selection (102, 148, 177, 181, 516, 224, 274 and 521) while 3 codon sites on the branch leading to *manE* were under positive selection (158, 444, and 514) (Table 21). The residues encoded by these codon sites are mapped on the crystal structures of $ManE_{Tmar}$ (Figure 24) and $ManD_{Tmar}$ (Figure 25). Along both branches, PAML detected codon sites under positive selection that are located in the region coding the binding site (*manE*: 158, 444, and 514, and *manD*: 274, 516 and 521) (Table 21).

The residues encoded by these codon sites could have been involved in the change of substrate specificity between ManD and ManE orthologs. An additional LTR was performed to confirm that the codon sites under positive selection are not the result of a relaxation of the purifying selection.

Interestingly, the branch-site analysis identified a cluster of codon sites (*manE*: 514, and *manD*: 516 and 521) that encode residues found in the binding pocket that were under positive selection in each branch. This suggests that both proteins experienced selective pressures that might result in changes of function. Within this region is a residue that was identified based on the structure superposition to interact with cellobiose and laminaribiose (Figure 23). The residue that corresponds to position 511 of $BglE_{Tmar}$ interacts directly with cellobiose, cellopentaose, laminaribiose and laminaritriose. The LIGPLOT at Figure 22 shows that the residue at position 511 dictates the substrate specificity by forming hydrogen bonds with the first glucosyl unit at the non-reducing end of the sugar. This residue corresponds to residue 519 in $ManD_{Tmar}$ and residue 515 $ManE_{Tmar}$.

In the *manD* branch, PAML detected codon sites under positive selection that were not located in the region encoding the binding pocket (102, 148, 177, 181, 224 and 274). Most of them are located at the protein surface. It is possible that these codon sites encode residues that are important for the protein function by unknown mechanisms. For example, these codon sites might encode residues important for flexibility of the hinge as previously discussed in Chapter 6 (181, 182) and so affect ligand binding indirectly. Alternatively, they may be involved for another function performed by these binding proteins and, perhaps these sites are important for the interaction with the permeases.

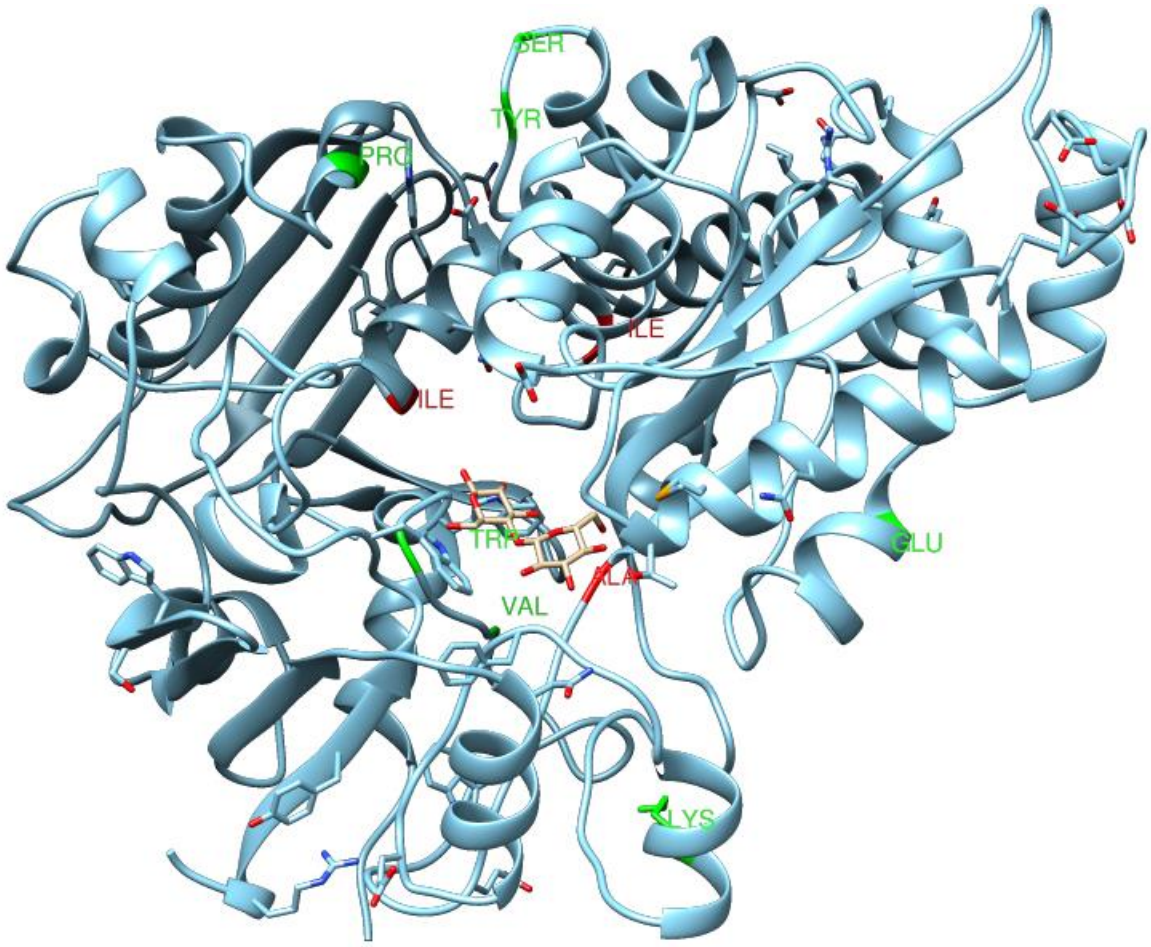


Figure 24. Map of the codon sites under neutral selection and positive selection on $ManE_{Tmar}$ structure superposition. The residues coded by the codons inferred under neutral evolution (class 1) and positive selection (class 2a) are highlighted in green and red, respectively. The residues located in the signal peptide are absent (position 4 and 20). The substrate cellobiose is drawn inside the binding pocket of $ManE_{Tmar}$ (pdb: 1vr5, blue) using the crystal structure of $BglE_{Tmar}$ (pdb: 3i5o).

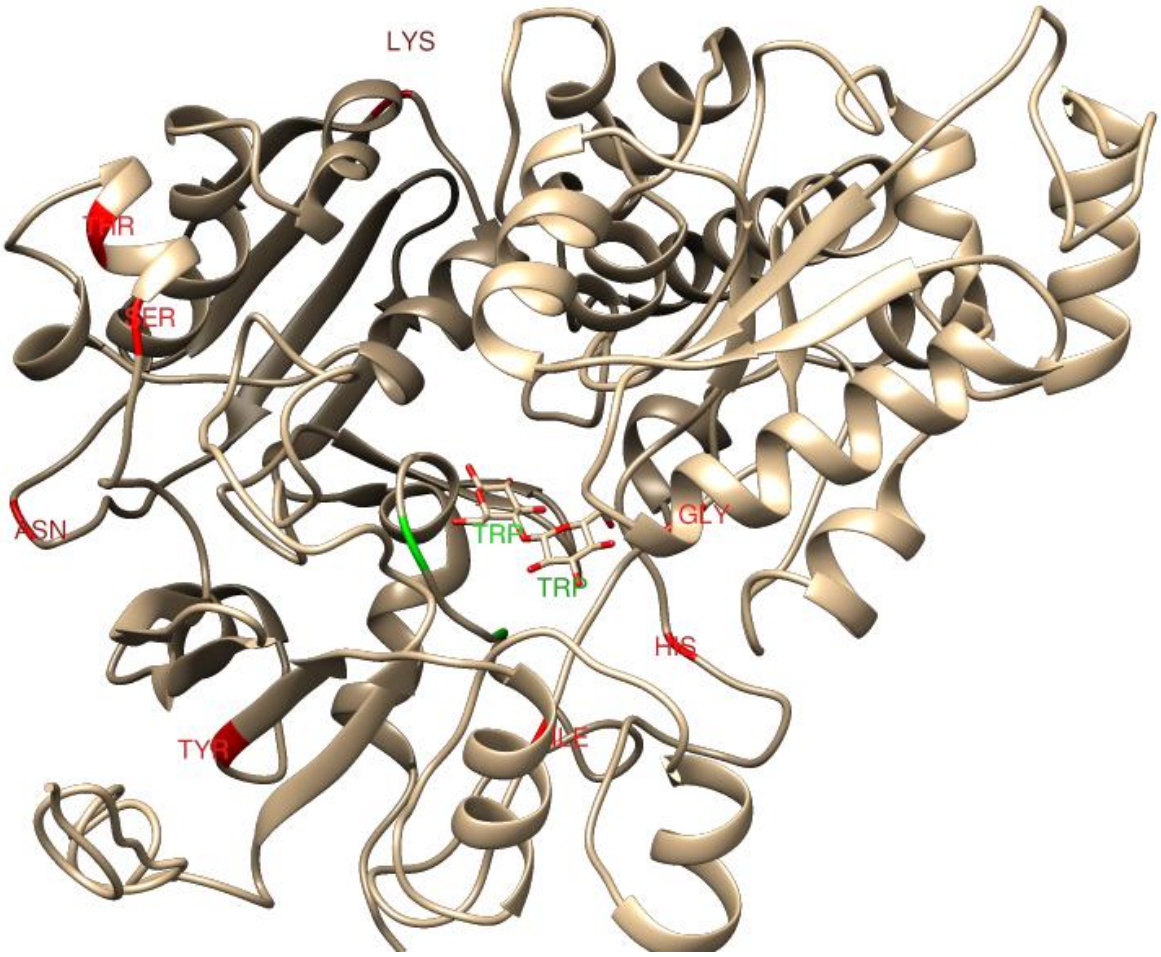


Figure 25. Map of the codon sites under neutral selection and positive selection on $ManD_{Tmar}$ structure superposition. The residues coded by the codons inferred under neutral evolution (class 1) and positive selection (class 2a) are highlighted in green and red, respectively. The residues located in the signal peptide are absent (position 4 and 20). The substrate cellobiose is drawn inside the binding pocket of $ManD_{Tmar}$ (I-TASSER, grey) using the crystal structure of $BglE_{Tmar}$ (pdb: 3i5o)

.

Why are ManD orthologs still maintained?

The PAML analysis suggests that each paralogous copy, *manE* and *manD*, contains codon sites that were under positive selection. An examination of the residues encoded by these codon sites reveals that many are near the substrate binding site. Consequently, this might indicate that each copy followed a different evolutionary path and their encoded proteins have acquired different functions. Although the SBPs still have overlapping binding functions, it is possible that *in vivo* their respective transporters have developed different preferences for specific substrates.

An example of this is the *T. maritima* MalE1 and MalE2. Both SBPs bind maltotriose and maltose, but, in addition, MalE1 binds β-1,4-mannotetraose and MalE2 binds trehalose (93). Transcriptional differences were measured for *malE1* and *malE2* when *T. maritima* was grown on different carbon sources suggesting that their respective transporters are involved in the transport of different sugars. The *malE1* transcripts are the most abundant when cells are grown on lactose and guar gum while the *malE2* transcripts are the most abundant when cells are grown on starch and trehalose (93). This could indicate that Mal1 is involved in the transport of galacto-manno-oligosaccharides from the hydrolysis of guar gum (191) and Mal2 is involved in the transport of maltooligosaccharides from the hydrolysis of starch (93).

A similar situation exists for the mannoside transporters. Microarray data show transcriptional differences between *manE* and *manD* when *T. maritima* is grown on different sugars at 80°C (43). The *manE* transcripts are highly up-regulated in the presence of saccharides containing β-1,4-linked glucose molecules while the *man*D transcripts are up-regulated on saccharides containing a β-D-mannose molecule (43). In

this study, the ManE orthologs bind cellobiose with the highest affinity at 37°C and 60°C (K_d, 1-21 nM) (Table 18) which suggests that the ManE orthologs have a preference for cellobiose. These results are consistent with the microarray data.

Additional evidence for a divergence of transport functions can be found in the functions of the protein-coding genes in proximity to *manD*. The functions of those genes suggest that ManD is involved in the transport of products resulting from the hydrolysis of mannan. In most *Thermotoga* species the operon containing *manD* encodes an endo-1,4-β-mannosidase (ManB) and the transcriptional regulator ManR (Figure 17). These proteins have been demonstrated to be involved in the mannan degradation pathway. ManB has a mannanase activity (174–177, 192), and in *T. maritima*, the enzyme recognizes mannan, galactomannan (carob and guar), azo-galactoman and glucomannan as substrates (192). ManR is known to control transcription of the operon encoding ManD. Mannose is the effector of ManR that prevents ManR from binding upstream of the operon, which allows its transcription (50). ManR also co-regulates the *mtpEFGKL* operon that encodes an ABC transporter of unknown function (TM1746-TM1750) (50).

Unfortunately similar evidence cannot be derived from the functions of genes in the proximity of *manE*. Most hydrolases and the transcription factor encoding genes in the vicinity of *manEFGKL* have not been characterized. Interestingly, the genome of *F. nodosum* encodes a characterized endo-1,4-β-glucanase (Cel5A) downstream of *manEFGKL* (Figure 17) (179, 193, 194). This glucanase has a specific activity for carboxymethyl cellulose, β-D-glucan and galactomannan (179). This is consistent with the fact that I found that $ManE_{Fnod}$ binds cellobiose, -triose and, tetraose which are

products of hydrolysis of β-D-glucans like cellulose The substrate specificities of $ManE_{Fnod}$ and Cel5A suggest that perhaps ManEFGKL is involved in the transport of the products of the hydrolysis of galactomannans and β-D-glucans.

Conclusion

In the previous chapter, the ManD and ManE orthologs were characterized to determine their respective binding functions. The binding properties of ManD and ManE orthologs are similar, but the ManD orthologs do not bind β-mannobiose, laminaribiose, laminaritriose and sophorose while the ManE orthologs do. In this chapter, the residues likely present in the binding sites were identified and the selective pressures that acted on the codons were examined to understand the evolutionary mechanisms that determine substrate adaptability. The protein structure of $ManE_{Tmar}$ and the modeled $ManD_{Tmar}$ superimposed onto $BglE_{Tmar}$ indicate that the residues corresponding to the position 14, 234, 427 and 511 of are important for the binding of cellobiose and laminaribiose. The residue of $ManD_{Tmar}$ that corresponds to the W511 of $BglE_{Tmar}$ is an alanine while the corresponding residue in $ManE_{Tmar}$ is a tryptophan. Tryptophan has more potential to form hydrogen bonding than alanine. All the ManD orthologs have an alanine at this position and having this resdue in $ManD_{Tmar}$ can explain why these proteins do not bind β-mannobiose, laminaribiose, laminaritriose and sophorose. The branch-site analysis suggests that the codon site 514 in the *manE* branch and the codon sites 516 and 521 in *manD* branch were under positive selection. These codon sites are in a region encoding the residues that correspond to W511 of $BglE_{Tmar}$. The *manD* branch hashad more codon sites under positive selection and more codon sites that encode a residue important in the

binding site than the *manE* branch. Therefore, it is possible that the *Thermotoga* species that still encode *manD* have a selective advantage conferred by the ability to utilize other carbon sources in their extreme environment.

Additional codon sites under positive selection were found in the *manD* branch. Many of these codon sites do not appear to encode residues in the binding site, therefore, it is unclear if these codon sites are encoding region important for the proteins' functions. These codon sites can encode residues located in the hinge and be important for the protein flexibility or SBP docking. In some instances, it has been shown that an SBP is able to bind a sugar *in vitro*, but that same substrate is not transported by its ABC transporter system (195, 196) indicating that another component other than the SBP was involved in the transport. Recently, structural analysis of the *E. coli* maltose ABC transporter showed that its permease, MalF, binds 3 glucosyl units at the non-reducing end of the sugar (65). This new finding suggests that the permease is more important for the sugar binding than initially expected. How the SBP interacts with the permease remains unclear. It is possible that the codon sites under positive selection encodefor the residues outside the binding pocket are important for the interaction of the SBP with the permease and for SBP docking.

Despite the fact that the $ManD_{Tmar}$ does not bind any ligands not bound by $ManE_{Tmar}$, the transcriptional data published by Connors et al. suggest that each protein has a distinct function (43). The characterization of representatives of the mannoside-binding proteins of the Thermotogales and the examination of the selective pressures that operated on their genes suggest that the paralogous copies of *manD* have acquired beneficial mutations in their binding-sites that improve the fitness of the organisms.

Additionally, perhaps, the functions of the ManD and ManE homologs are not only dictated by their binding-sites but also by other regions of the protein or other components of the ABC systems might have an impact on the transporters' functions. Future studies to understand the evolutionary aspects of gene maintenance of the SBP-encoding genes might require investigations of all the ABC components rather than only the SBPs.

Future work

Based on their phylogenetic history, the ancestral gene to the *manE* or *manD* could have been acquired from an archaeon. The genome of *P. furiosus* encodes two ManD/ManE orthologs (Figure 21). The *P. furiosus* CbtA ($CbtA_{Pfur}$) shares a common ancestry with ManE/ManD and the SBP has overlapping function with ManE and ManD homologs (7). The ManE, ManD and $CbtA_{Pfur}$ homologs bind cellobiose, cellotriose, cellotetraose, cellopentaose. Only the ManE orthologs and $CbtA_{Pfur}$ bind laminaribiose, laminaritriose, and sophorose. However, the binding properties of $CbtA_{Pfur}$ were not measured in the presence of sugars composed of β-D-mannose such as β-1,4-mannobiose, β-1,4-mannotriose and β-1,4-mannotetraose. Depending whether $CbtA_{Pfur}$ binds the mannan hydrolysis products or not it could provide a better understanding of the common ancestry of the ManD and ManE homologs. To get a more definitive answer, the characterization of the two SBPs paralog present in *P. furiosus* would be necessary (Figure 21).

The residues in the binding pockets of the ManD and ManE orthologs and BglE are mostly conserved. The residues at position 14, 234, 427 and 511 that are different might be associated with the differences of function. The results from the characterizations of the mannoside-binding proteins suggest that proteins related to $ManE_{Tmar}$ have broader sugar specificities than those related to $ManD_{Tmar}$. Future experiments could involve site-directed mutagenesis of these residues. Site-directed mutagenesis was performed to generate $ManD_{Tmar}$ (T373A) and $ManD_{TRQ2}$ (A373T) because preliminary results suggested that these proteins had different functions. However, the analysis of the binding properties of the mutant proteins revealed no changes of their functions. Later, it

was found that the ManD orthologs are prone to photodecomposition, causing a gradual decrease of fluorescence that was confused with ligand binding. However, additional site-directed mutagenesis experiments can confirm or refute the hypothesis that the residue at position 511 is involved in the changes of function noted between the ManD and ManE paralogs.

Transcript analysis using RNA-seq or real-time PCR could be performed to investigate the hypothesis that ManD orthologs are utilized in the transport of sugars composed of β-D-mannose and that ManE orthologs are utilized in the transport of sugars composed of β-D-glucose. This experiment was done in *T. maritima* but not in the other Thermotogales. Based on the function of the glucanase encoded near the *F. nodosum manEFGKL* operon, it is likely that ManEFGKL is utilized in the transport of β-D-glucans and β-D-mannans. These experiments could increase our knowledge of transporter specificity in these organisms.

Appendix 1: List of genomic variations between *T. maritima* MSB8 genomovars TIGR, DSM3109 and ATCC.

The protein functions were determined using the annotation of the genomovar TIGR provided by the Joint Genome Institute (JGI). The difference in the nucleotide sequence at a specific position is indicated in the column "Genotype". The positions are based from the genomovar TIGR (GenBank sequence AE000512.1). The abbreviations of the genotypes are T, genomovar TIGR (40) (AE000512.1); D, genomovar DSM3109 (AGIJ00000000.1); and A, genomovar ATCC (99) (CP004077.1). Residue change is from genomovar TIGR to genomovar DSM3109 or ATCC.

Locus			**Function**	**Position**	**Genotype**			**Residue**
TIGR	**DSM3109**	**ATCC**			**T**	**D**	**A**	**change**
TM0023	Thema_0926	Tmari_0020	methyl-accepting chemotaxis protein	20088	T	C	T	**G→G**
				20241	C	T	C	**K→K**
				20319	C	A	C	**L→L**
				20330	A	C	A	**S→A**
				20412	T	-	T	
				20415	C	-	C	
TM0035	Thema_0914	Tmari_0032	hypothetical	35288	T	-	-	
TM0084	Thema_0866	Tmari_0081	hypothetical	88372	-	T	T	
TMrrnaA16S	N/A	Tmari_R0004	16S rRNA	189134	T	C	C	[1]
				189311	-	C	C	[1]
				189386	-	C	C	[1]
TMrrnaA23S	N/A	Tmari_R0007	23S rRNA	193744	-	G	G	
TM0193	Thema_0746	Tmari_0192	permease	206649	T	C	C	**E→G**
TM0227	Thema_0711	Tmari_0225	ATPase	241704	-	G	G	[2]
TM0254	Thema_0683	Tmari_0252	SsrA-binding protein	265175	G	-	-	
TM0255	Thema_0683	Tmari_0253	ribosomal protein L28	265389	G	A	A	**G→E**
TM0257	Thema_0680	Tmari_0255	hypothetical	266624	-	G	G	[2]
TM0266	Thema_0671	Tmari_0264	DNA-binding protein, HU	277083	C	A	A	**T→N**
TM0277	Thema_0660	Tmari_0275	carbohydrate ABC transporter SBP (CUT1)	291738	C	-	-	[2]
TM0279	Thema_0658	Tmari_0277	carbohydrate ABC transporter permease (CUT1)	294531	-	C	C	
TM0340	Thema_0597	Tmari_0338	hypothetical	359249	T	-	-	

TM0378	Thema_0560	Tmari_0376	glycerol-3-P-dehydrogenase	397783	-	C	C	
TM0380	Thema_0558	Tmari_0278	alkylhydroperoxidase	399871	-	G	G	[2]
TM0424	Thema_0515	Tmari_0421	permeases	442276	T	C	C	**I→V**
TM0429	Thema_0510	Tmari_0426	methyl-accepting chemotaxis sensory transducer	448945	T	C	T	**T→A**
				448946	T	C	T	**R→R**
				448951	G	A	G	**L→L**
				448952	G	A	G	**S→S**
				448960	G	A	G	**L→L**
				449315	A	C	A	**A→A**
				449318	A	G	A	**S→S**
TM0443	Thema_0496	Tmari_0440	gluconate kinase	466024	G	G	A	**R→K**
TM0479	Thema_0454	Tmari_0476	translocase (secG)	506728	-	G	G	
TM0503-TM0504				531878	-	G	G	
				531878	-	-	G	
TM0522	Thema_0408	Tmari_0518	ATP-dependent protease	547852	C	C	T	**T→T**
TM0556-TM0557				585521	-	C	C	
TM0588-TM0589				623237	A	-	-	
TM0637	Thema_0292	Tmari_0638	hypothetical	669776	-	G	G	
TM0641	Thema_0288	N/A	hypothetical	675430	-	C	C	
TM0674	Thema_0256	Tmari_0674	fagellar hook-basal body proteins	703282	-	C	C	[2]
				703301	-	C	C	
TM0680	Thema_0250	Tmari_0680	flagellar motor switch	707405	-	G	G	
TM0785	Thema_0143	Tmari_0786	uncharacterized	808729	-	G	G	
TM0873	Thema_0053	Tmari_0875	ATPases	896001	-	C	C	[2]
TM0878	Thema_0048	Tmari_0880	2-oxoglutarate ferredoxin oxidoreductase	901369	C	C	T	**G→E**
TM0888	Thema_0038	Tmari_0890	thiamine pyrophosphokinase	911310	-	C	C	
TM0984	Thema_1856	Tmari_0887	DNA/RNA helicases, SNF2 family	995364	T	A	A	**N→K**
TM1003	Thema_1837	Tmari_1007	transposase	1020847	A	G	A	**I→T** [3]
TM1006	Thema_1840	Tmari_1010	oxidoreductase	1024239	T	A	A	**L→I**
				1024289	G	C	C	**P→P**
				1024351	G	T	T	**S→I**
				1024390	G	A	A	**R→K**
				1024417	G	T	T	**R→L**
				1024423	C	A	A	**T→K**
TM1007	Thema_1836	Tmari_1011	arabinose efflux permease	1025621	-	G	G	[2]
TM1015-TM1016				1034614	-	T	T	
TM1018	Thema_1826	Tmari_1021	hypothetical	1036173	G	-	-	
TM1020	Thema_1825	Tmari_1022	hypothetical	1038577	A	-	-	
TM1068	Thema_1776	Tmari_1072	alpha-	1083790	C	A	C	**L→L**

			glucosidase	1084018 1084045	A A	G G	A A	**N→N** **H→H**
TM1136	Thema_1706	Tmari_1142	amino acid/amide ABC transporter TM1	1149541	C	C	T	**S→F**
TM1149-TM1150				1163517	G	-	-	
TM1161	Thema_1679	Tmari_1168	MgtE	1175094	A	A	T	**M→K**
TM1168	Thema_1672	Tmari_1175	maltodextrin phosphorylase	1183419	-	G	G	[2]
TM1295	Thema_1545	Tmari_1303	Zn-dependent hydrolase	1322546	G	-	-	
TM1314	Thema_1528	Tmari_1321	hypothetical	1333884	-	-	T	
TM1318	Thema_1525	Tmari_1325	ABC-type multidrug transport system, ATPase and permease	1337345	-	G	G	[2]
TM1322	Thema_1521	Tmari_1328	hypothetical	1340939 1340941 1341026	G C T	A T -	G C T	**V→I** **V→I**
TM1322-TM1323				1342559 1342562	C G	T A	C G	
TM1328	Thema_1514	Tmari_1335	ABC-type multidrug transport system, ATPase and permease	1348605	G	-	G	
TM1341	Thema_1503	Tmari_1348	hypothetical	1361357	-	C	C	
TM1415-TMtRNA-Val				1430817	C	-	-	
TM1421	Thema_1417	Tmari_1428	Fe only hydrogenase large subunit	1436667	-	G	G	
TM1432	Thema_1407	Tmari_1438	glycerol-3-P-dehydrogenase	1449528	A	G	G	**N→D**
TM1438	Thema_1401	Tmari_1445	glycosidase	1453976 1453978 1453989 1453990	A G A G	T T T T	T T T T	***→K** **T→K** **A→E** **A→E**
TM1439-TM1440				1455257	C	-	-	
TM1445	Thema_1394	Tmari_1451	ribosomal protein S1	1460678	-	G	G	
TM1477	Thema_1360	Tmari_1455	bacterial translation initiation factor 1	1486757	T	-	-	
TM1573	Thema_1264	Tmari_1581	quinonprotein alcohol dehydrogenase-like	1560825	A	G	G	**L→L**
TM1574-TM1575				1562661	-	G	G	
TM1691	Thema_1141	Tmari_1699	hypothetical	1670011	G	A	A	**S→S**

TM1697	Thema_1135	Tmari_1705	SAM-dependent methyl-transferase	1675977	G	G	A	**A→A**
TM1707	Thema_1125	Tmari_1715	transcriptional regulator NrdR	1687120	C	G	G	**L→F**
TM1710	Thema_1123	Tmari_1718	hypothetical, cofD-related	1689515	C	-	-	[2]
TM1771-TM1772				1748054	G	-	-	
TM1781	Thema_1048	Tmari_1790	arginine-succinate lyase	1757441	-	T	T	
TM1813	Thema_1015	N/A	CRISPR-associated protein, Csx2 family	1790228	C	-	-	
TM1837	Thema_0992	Tmari_0252	MalF2	1812658	C	T	T	**V→I**[4]
TM1837	Thema_0992			1813792	-	C	C	[1]
TM1838	Thema_0992			1814076	-	C	C	[1]
TM1838-TM1839	Thema_0926			1814255	C	-	C	[4]
				1814255	C	-	C	[4]
				1814255	C	-	C	[4]
				1814255	T	-	T	[4]

[1] Validated sequencing error on genomovar TIGR
[2] Annotated frameshift or point of mutation by NCBI
[3] Annotated as authentic frameshift (99)
[4] Authentic single-nucleotide polymorphism (SNP) on genomovar TIGR and DSM3109 validated by Sanger sequencing

Appendix 2: Details of the differential scanning fluorimetry (DSF) protocol.

Reagent	Final amount, final concentration or recommended volume
protein	2.0-2.5 μg (no more than 5 μg)
5X citric acid/Na_2HPO_4 (100 mM citric acid, pH adjusted with 200 mM dibasic sodium phosphate) (pH 2.0-7.0)	1X
NaCl	150 mM
5000X SYPRO Orange (Life Technologies)	8X
1-10 mM ligand solution or water	2 μl
Total volume	20 μl

Ligand screening protocol

1. Dispense the ligand solutions or the buffer for pH titration in separate wells of a 96 well-plate or into microfuge tubes. Each ligand (and the water control) is measured in triplicate. Place the plate/tubes away from the light at room temperature.

2. Prepare the master mix containing the test protein, citrate buffer, NaCl and SYPRO Orange by adding the reagents in the same order as in the table. Add 18 μl of the master mix to each well or tube.

3. Set the real-time PCR thermocycler to the SYBR green setting with excitation at 490 nm and emission at 530 nm.

4. Typically, the cycling is performed by heating the samples from 25°C to 98°C at a heating rate of 0.5°C per min.

5. Calculate the Tm using the raw data located in the FRET file. The midpoint temperature of the unfolding transition (Tm) is obtained from curve fitting to the following Boltzmann function (109). Where L, U and *a* are the minimum and maximum fluorescence intensities and the slope of the curve, respectively.

$$y = \mathrm{L} + \frac{U - L}{1 + \exp^{\left(\frac{\mathrm{Tm} - x}{a}\right)}}$$

Appendix 3: Thermostabilities of MalE1, MalE2, TreE and XylE2 measured by DSF and represented by ΔTm values.

The assay was performed in triplicate and standard deviations are shown. If the Tm_{ligand} could not be calculated precisely because of incomplete denaturation or inability to observe an unfolding curve, Tm_{ligand} was estimated using the maximum fluorescence intensities (≥).

	ΔTm (°C)			
Ligand	**MalE1**	**MalE2**	**TreE**	**XylE2**
no ligand	0.00 ± 0.04	0.00 ± 0.40	0.00 ± 0.08	0.00 ± 0.40
arabinose	0.14 ± 0.27	-0.07 ± 0.25	-0.06 ± 0.12	3.60 ± 0.25
cellobiose	-0.21 ± 0.03	-0.26 ± 0.24	0.06 ± 0.14	6.03 ± 0.24
fructose	0.04 ± 0.13	-0.27 ± 0.28	0.44 ± 0.10	-0.27 ± 0.28
fucose	-0.16 ± 0.08	-0.18 ± 0.26	-0.15 ± 0.08	8.82 ± 0.26
galactose	-0.77 ± 0.1	-0.05 ± 0.26	1.91 ± 0.09	1.20 ± 0.26
glucose	-0.02 ± 0.07	-0.09 ± 0.24	3.72 ± 0.10	≥14.04
lactose	-0.12 ± 0.09	-0.17 ± 0.25	-0.23 ± 0.07	-0.04 ± 0.25
maltose	4.52 ± 0.09	1.02 ± 0.25	≥15.69	2.24 ± 0.25
maltotetraose	11.75 ± 0.24	8.67 ± 0.42	4.67 ± 0.34	2.07 ± 0.42
maltotriose	10.77 ± 0.48	6.85 ± 0.24	7.57 ± 0.10	0.32 ± 0.24
mannan	0.10 ± 0.15	0.01 ± 0.26	1.04 ± 0.07	0.00 ± 0.26
mannose	-0.50 ± 0.23	0.07 ± 0.31	0.10 ± 0.11	1.18 ± 0.31
mannotetraose	4.22 ± 0.12	0.24 ± 0.30	0.02 ± 0.69	0.69 ± 0.30

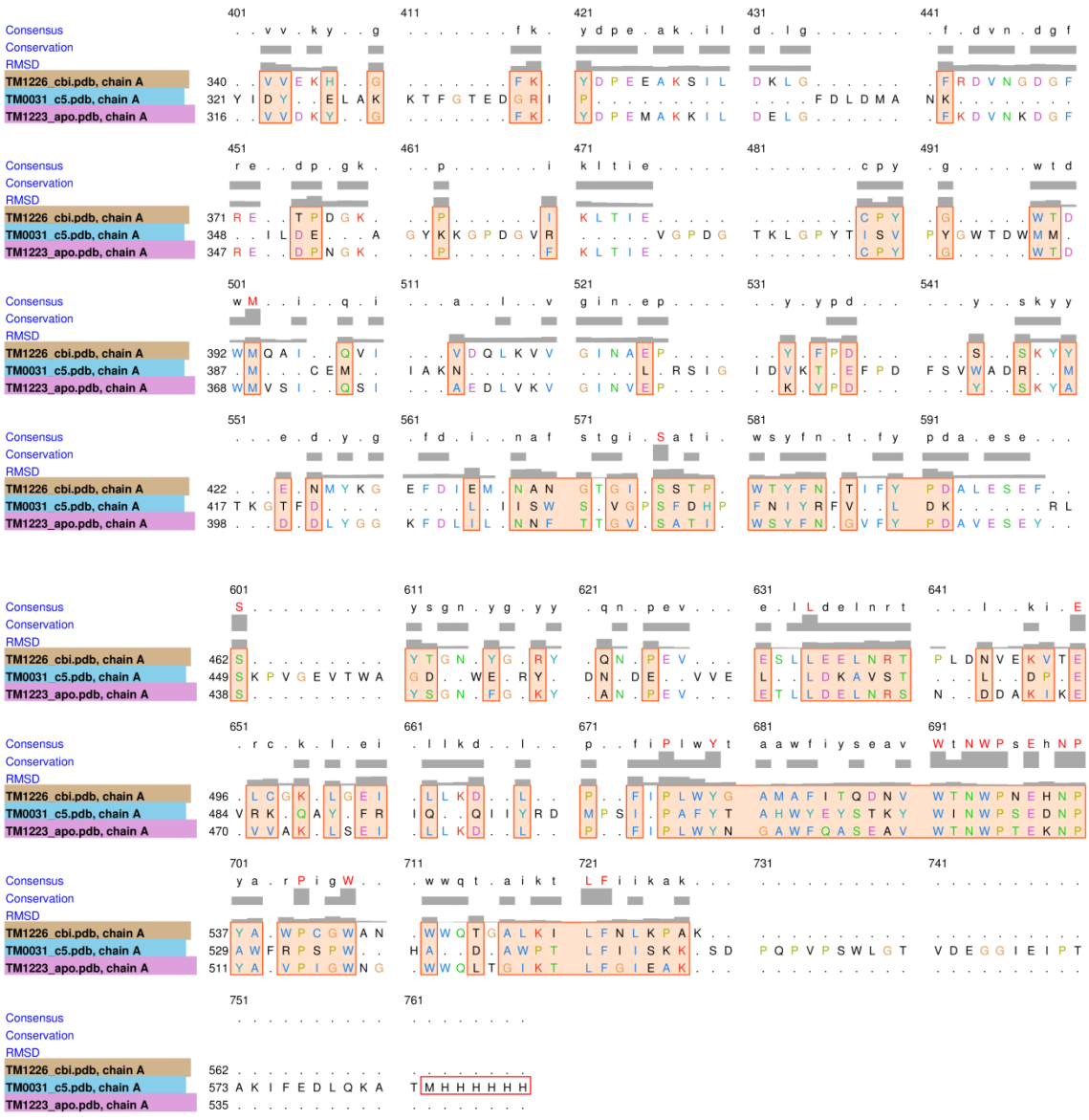
mannotriose	0.11 ± 0.05	-0.08 ± 0.25	0.57 ± 0.79	2.23 ± 0.25
melibiose	-0.18 ± 0.08	-0.20 ± 0.33	-0.61 ± 0.08	0.12 ± 0.33
myo-inositol	-0.16 ± 0.09	-0.08 ± 0.23	-0.18 ± 0.07	5.41 ± 0.23
pullulan	0.23 ± 0.11	0.04 ± 0.25	0.01 ± 0.20	-0.03 ± 0.25
raffinose	-0.23 ± 0.25	-0.03 ± 0.25	1.26 ± 0.11	0.50 ± 0.24
rhamnose	0.06 ± 0.09	-0.17 ± 0.27	-0.05 ± 0.11	-0.14 ± 0.27
ribose	-0.03 ± 0.14	-0.08 ± 0.29	0.28 ± 0.08	-0.01 ± 0.29
sorbitol	-0.16 ± 0.07	-0.20 ± 0.33	0.39 ± 0.10	-0.13 ± 0.33
sucrose	-0.09 ± 0.18	-0.20 ± 0.22	11.41 ± 0.23	10.89 ± 0.22
tagatose	0.04 ± 0.06	-0.23 ± 0.25	-0.15 ± 0.05	-0.15 ± 0.25
trehalose	0.22 ± 0.13	0.26 ± 0.24	≥13.77	0.32 ± 0.24
xyloglucan	0.00 ± 0.13	-0.09 ± 0.31	-0.09 ± 0.11	-0.09 ± 0.31
xylose	0.16 ± 0.32	-0.08 ± 0.24	-0.02 ± 0.25	≥13.14

Appendix 4: Glycosidic bonds of disaccharides and oligosaccharides.

Sugar	Composition
arabinogalactose	β-1,3-arabinose-galactose
cellobiose	β -1,4-glucose disaccharide
cellotriose	β -1,4-glucose trisaccharide
cellotetraose	β -1,4-glucose tetrasaccharide
gentiobiose	β -1,6-glucose disaccharide
lactose	β-1,4-galactose-glucose
laminaribiose	β -1,3-glucose disaccharide
laminaritriose	β -1,3-glucose trisaccharide
maltose	α-1,4-glucose disaccharide
maltotriose	α-1,4-glucose trisaccharide
mannobiose [a]	β -1,4-mannose disaccharide
mannotriose	β-1,4-mannose trisaccharide
mannotetraose	β -1,4-mannose tetrasaccharide
mannobiose	α-1,6-mannose disaccharide
melibiose	α-1,6-galactose-glucose
raffinose	galactose-α-1,6-α-glucopyranosyl-β-fructofuranoside
sucrose	β-fructofuranosyl-α -glucopyranoside
sophorose	β -1,2-glucose disaccharide
trehalose	α-1,1-glucose disaccharide
xyloglucan oligosaccharides	Mixture of heptasaccharide (four β -1,4 glucosyl residues with three α-1,6 xylosyl side-chains), octasaccharide and nonasaccharide (heptasaccharide containing one or two galactosyl residues linked β -1,2 to the xylosyl units) oligomers
glucomannan (Konjac)	β-1,4 linked mannose and glucose units (1:6 ratio)
mannan	α-1,6-mannan with side chains of α-2,1- and α-3,1-linked mannans
pullulan	α-1,4 glucose trisaccharide units linked α-6,1 at sugars 1 and 3

[a] mannobiose in β-linkage was used in this study unless otherwise specified

Appendix 5: Sequence alignment using $BglE_{Tmar}$ (pdb:3i5o), $ManE_{Tmar}$ (pdb: 1vr5) and $ManD_{Tmar}$ (modelled using I-Tasser) from structures superposition using CHIMERA


Consensus
Conservation
RMSD
TM1226_cbi.pdb, chain A
TM0031_c5.pdb, chain A
TM1223_apo.pdb, chain A

Appendix 6: *manD* and *manE* codon alignment for the PAML analysis

```
19 1806 I
Ndas_3658  ---------- -----ATGAG A---ATGCTG CGGCTTCCGG CGGCACTCGC
Xcel_0262  ATGTTCCGCT TACGCACCCG GAGGAGTGCC GCCGTTGCCG CGGTGGCCGC
Tnap_1566  ---------- -----ATGAA AAGATTTTTA CTGGTATTCG TGGTTGTTTT
TM1226     ---------G TGAATATGAA AAGATTTTTA CTGGTATTCG TGGTTGTTTT
TRQ2_1592  ---------- -----ATGAA AAGATTTTTA CTGGTATTCG TGGTTGTTTT
Tlet_1438  ---------- -----ATGAA AAAGTTTCTG GTTCTATCCA TGGTTCTTCT
CTN_1348   ---------- -----ATGAA AAGGTTTTTA GTTGTTCTCG TTCTGGTCCT
TM1223     ---------- -----ATGAA AAGGTTTTTA GTCGTTCTCG TTCTGGTCCT
Tnap_1569  ---------- -----ATGAA AAGGTTTTTA GTCGTTCTCG TTCTGGTCCT
TRQ2_1595  ---------- -----ATGAA AAGGTTTTTA GTCGTTCTCG TTCTGGTCCT
Tpet_1545  ---------- -----ATGAA AAGGTTTTTA GTCGTTCTCG TTCTGGTCCT
Theba_2445 ---------- -----ATGAA AAAGAGGTTC ATTTTACTAC TTCTGGCCAT
Fnod_1553  ---------- -----ATGAA AAAGGTTTTC ATGACATTGT TATTGGTAGC
TeCCSD1    ---------- -----ATGAG AAGGCTTTTA AGTCTATTAT TAGTAGTGAT
Theet_1038 ---------- -----ATGAG AAGGCTGTTA AGTCTATTAT TAGTGTTAAT
Thebr_1980 ---------- -----ATGAG AAGGCTGTTA AGTCTATTAT TAGTGTTAAT
STHERM     ---------- -----ATGAG AAAAGCCTTT ACCGTGGTCC TGAGTCTCTT
Tbis_3512  ---------- -----ATGCG GAGCCGAGTG CGGTTGCTGA TACCGCTCCT
Tfu_0910   ---------- -----ATGCG AAACCGCTAC AGGTTTCTGA TCCCCCTCGC

           GGCCCTGGCC CTCGCGGTCA GCGCCTGCTC GGGTGGCGGC ----------
           CTCCATGCTC ATCCTCGCCG CTGGTTGCTC CAGCGACGGG GGCTCCACCA
           TCTTTTCTCA AGTCTGTTC- ---------- ---------- ----------
           TCTTTTCTCA AGTCTGTTC- ---------- ---------- ----------
           TCTTTTCTCA AGTCTGTTC- ---------- ---------- ----------
           GATAACAATT GCAGCCTTC- ---------- ---------- ----------
           GGCACTGGTT TCTGTTTTC- ---------- ---------- ----------
           GGCACTGGTT TCGGTTTTC- ---------- ---------- ----------
           GGCACTGGTT TCGGTTTTC- ---------- ---------- ----------
           GGCACTGGTT TCGGTTTTC- ---------- ---------- ----------
           GGCACTGGTT TCGGTTTTC- ---------- ---------- ----------
           CATACTCGTT AGTAGCCTCG CAATT----- ---------- ----------
           AGTTTTCTCG ACTTTGGTAC TTTCT----- ---------- ----------
           GCTAATAATG GGCGCTTTAG CAGGTTGTAG CCAGAAGAGT CAAGAGACTG
           GCTAATAATG GGCGCATTAG CAGGCTGTGG ACAAAAGAGT CAAGAAACAA
           GCTAATAATG GGCGCGTTAG CAGGCTGTGG ACAAAAGAGT CAAGAAACAA
           CCTCCTTGCG GGTGTGGTGT TCGCGAGCGG TCAAGCTGCC CAGCAGGCCG
           CGCGGTC--- GGCCTCCTGA CGGCGTGCAC GGGCGGGGCC GAGACGCCCG
           ACTGGTCAGC TTGCTGGCCA CATCCTGCGG GGGTGACACA ----------

           ---------- ---------- ---------- --------GG GAACAGCGAC
           CCGACCCCAC GCCCACCGAG GGCGGCGAGG CCACCGAGGA GGTCCCC---
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
           ---------- ---------- ---------- ---------- ----------
```

```
CTGACAAATC GACACAGCAA GAGAATGCAA ATACCCAACA GGTTGCAGAC
GCGATAAATC GAACCAGCAA AAT---GCAA ATACACAGCA AGTTGCACAG
GCGATAAATC GAACCAGCAA AAT---GCAA ATACACAGCA AGTTGCACAG
GC-------- ---------- ---------- --------GC GGCTCCCTCT
CGCGATCCCC GGCA------ ------CCGG CGGCCGCCGC GGTGCCGGGC
---------- ---------- ---------- --------GA GGGCGACGGT

GGCGGTTCGG GTCAGTACCC CAGGAACGAG ACCCTGTACA CCACGGGTAC
GACGTCGTCG GCGAGTACCC CCGAGCCGAG ACCGTCTTCA CCTCGGGCAC
TCCCAAGTT- -----TTAGA ACGAAACGAA ACTATGTACT ATGGAGGTTC
TCCCAAGTT- -----TTAGA ACGAAACGAA ACTATGTACT ATGGAGGTTC
TCCCAAGTT- -----TTAGA ACGAAACGAA ACTATGTACT ATGGAGGTTC
GGCGAAACA- -----TTTGA AAGAAGCAAA ACAGTCTATG TTGGAGGAGG
GGTCAAACT- -----TTCGA AAGGAACAAG ACACTTTACT GGGGTGGAGC
GGACAGACT- -----TTTGA GAGAAACAAA ACGCTCTACT GGGGTGGAGC
GGACAGACT- -----TTTGA GAGAAACAAA ACGCTCTACT GGGGTGGAGC
GGACAGACT- -----TTTGA GAGAAACAAA ACGCTCTACT GGGGTGGAGC
GGACGGACT- -----TTTGA GAGAAACAAA ACGCTCTACT GGGGTGGAGC
GGTCAAGTT- -----TACGA TCGAAAAGAA ACTCTGTACG CCGGTGGAGG
GACGTAGTT- -----TACAA AAGAGATGAG ACGTTGTACG CTGGAGGAGG
GAGAAAACA- -----TATCA AAGAAACGAA ACAGTATACT TTGGAGGTGG
GAGACAACA- -----TATCA AAGAAACGAA ACAGTATACT TTGGAGGTGG
GAGACAACA- -----TATCA AAGAAACGAA ACAGTATACT TTGGAGGTGG
CAGCAGGTGG TGGAGTTCCC CAGGAACGAG ACCCTGTACT CCAGCGGTAC
GGAGAATCG- -----CTCCC CCGGCACGAG ACCCTCTACA CCAGCGGCAC
GGAGCGACAG GGGACTACCC CCGTGCGGAG ACCCTCTACA CCGGTGGTAC

GGCCTGGGAG GCGCCGACCA GCTGGAACCC GATGATGCGG GGCCAGTTCG
CCAGTGGGGT CCGCCGTCGA GCTGGAACCC GATCCCGGGC TCCGGCGACG
TCTGTGGTCT CCTCCTTCTA ATTGGAATCC GTTCACTCCG TGGAATGCGG
TCTGTGGTCT CCTCCTTCTA ATTGGAATCC GTTCACTCCG TGGAATGCGG
TCTGTGGTCT CCTCCTTCTA ATTGGAATCC GTTCACTCCG TGGAATGCGG
TATGTGGTCT CCACCATCAA ACTGGAATCC GTTCACATCA TGGAACGCTG
TCTCTGGTCT CCTCCATCCA ACTGGAACCC GTTCACACCC TGGAACGCTG
GCTGTGGTCT CCTCCATCCA ACTGGAACCC GTTCACACCA TGGAACGCGG
GCTGTGGTCT CCTCCATCCA ACTGGAACCC GTTCACACCA TGGAACGCGG
GCTGTGGTCT CCTCCATCCA ACTGGAACCC GTTCACACCA TGGAACGCGG
GCTGTGGTCT CCTCCATCCA ACTGGAACCC GTTCACACCA TGGAACGCGG
CCTTTGGAGT CCACCAAACA ACTGGAACCC ATTCACCCCA TGGGCGATTA
TATGTGGGCA CCACCATCAA ACTGGAACCC TATTACTCCT TGGAATGCTG
ATTATGGAGC CCGCCCAGCA ACTGGAATCC GCTAACACCA TGGAACGCTG
ATTATGGAGC CCGCCCAGCA ACTGGAATCC GCTAACACCA TGGAATGCTG
ATTGTGGAGC CCGCCCAGCA ACTGGAATCC GCTAACACCA TGGAACGCTG
CATGTGGGCT CCTCCGTCCA ACTGGAACCC CTATACCCCG TGGGCGGTGG
GCAGTGGGGG CCACCGGCGA ACTGGAACCC GCTCCGGGAA TGGGACTTCG
CCAGTGGGGG CCGCCGAGCA CGTGGAACCC GCTCGACACG GGTAACTACG

CGGTCGGCAC CAACGGCCTG GTCTACGAGT CGCTCTTCCA CTACGACGCG
CGACCGGCAC CCGCGGTCTG CTCTACGAGA CCTTGTTCCA CTTCGACCCG
TACCAGGAAC AACTGGACTT GTCTATGAAA CAATGTTCTT TTACGATCCA
TACCAGGAAC AACTGGACTT GTCTATGAAA CAATGTTCTT TTACGATCCA
TACCAGGAAC AACTGGACTT GTCTATGAAA CAATGTTCTT TTACGATCCA
TGACTGGGAC TATCGGATTG GTTTATGAAA CACTGTTTCT GTATGATCCA
TAGCAGGAAC CATCGGTCTT GTCTATGAAC CCCTGTTTCT CTACGATCCT
TTGCGGGAAC CATCGGTCTT GTCTATGAAC CTCTGTTCCT CTACGATCCT
TTGCGGGAAC CATCGGTCTT GTCTATGAAC CTCTGTTCCT CTACGATCCT
TTGCGGGAAC CATCGGTCTT GTCTATGAAC CTCTGTTCCT CTACGATCCT
```

TTGCGGGAAC CATCGGTCTT GTCTATGAAC CTTTGTTCCT CTACGATCCC
TGACCGGAAC CAACGGTCTA ATCTATGAAT ACCTGTTTAT GTTCGATCCT
TTACTGGTAC TGTTGGGCTA ATCTATGAAA CTCTTTACGG TTACGACCCA
TAACAGGTAC AGTGGGATTA ATTTATGAAA CATTATTCAA CTATGACCCA
TAACAGGTAC AGTTGGATTA ATTTATGAAA CACTATTCAA CTATGACCCG
TAACGGGTAC AGTTGGATTA ATTTATGAAA CACTATTCAA CTATGACCCG
CCACAGGGAC CATCGGTCTC TGCTACGAGC CTCTCTTCCT CTATGATCCG
CCACCGGGAC GAAGGGCCTC GTCTACGAGA CCCTCTTCCT CTACGACCCG
CCACCGGAAC GGTGGGTCTG GTCTACGAGA CCCTGTTCCT CTACGAGCCG

GACGCGGGAG AGTACGTCCA CTGGCTCGCC GAGAGCGACG AGTGGACCTC
AGCACGCTCG AGCTCTCGCC CTGGCTCGCC GAGTCCGGCG AGTGGGTCGA
CTCACTGGAA ATTTTGATCC ATGGCTTGCA GAAAAAGGTG AATGGTTAGA
CTCACTGGAA ATTTTGATCC ATGGCTTGCA GAAAAAGGTG AATGGTTAGA
CTCACTGGAA ATTTTGATCC ATGGCTTGCA GAAAAAGGTG AATGGTTAGA
TTATCAGGTG AATTCGAACC ATGGCTTGCG GAGAGTGGTA GATGGATTAA
CTGAACGACA AATTCGAACC ATGGCTTGCA GAGAGCGGAC AGTGGGTCAG
CTGAACGACA AGTTCGAGCC GTGGCTTGCA GAAAAAGGAG AATGGGTCAG
CTGAACGACA AGTTCGAGCC GTGGCTTGCA GAAAAAGGAG AATGGGTCAG
CTGAACGACA AGTTCGAGCC GTGGCTTGCA GAAAAAGGAG AATGGGTCAG
CTGAACGACA AGTTTGAACC GTGGCTTGCA GAAAAAGGAG AATGGGTCAG
CTGAGCAACG AGATGATACC CTGGCTTGCG GTGGATGGAG GATGGATAGA
CTGAAAGACG AAATGATTCC ATGGCTTGCT GAAAGCGGAA AGTGGACATC
CTTAAAAACG AATTCATACC ATGGTTAGCA GAGAAAGGCG AATGGACCTC
CTTAAAAACG AATTCATACC ATGGTTAGCA GAGAAAGGTG AATGGACCTC
CTTAAAAACG AATTCATACC ATGGTTAGCA GAGAAAGGTG AATGGACCTC
CTCAAGGACG AGTTCATCCC ATGGCTCGCC GAGAGCGGGA AGTGGGTCGA
AGCATCGACC GGCTCATCCC GTGGCTCGCC GAGAGCGGCT CCTGGACCGG
CAGACTGGAG AATACATCCC CTGGCTTGCG GAGAGCGGCG AGTGGGTCGA

GGAGACCGAG CACGTGATCA CCCTGCGCGA GGGCGTCACG TGGAACGACG
CGACCAGACC TACACGGTCA CGCTCCGTGA CAACGCGACC TGGACCGACG
CAGTAAGACT TACAGGGTTG TATTGAGAGA GGGTATATAC TGGCATGATA
CAGTAAGACT TACAGGGTTG TATTGAGAGA GGGTATATAC TGGCATGATA
CAGTAAGACT TACAGGGTTG TATTGAGAGA GGGTATATAC TGGCATGATA
CAGCAATACG TATCAGATCA AACTCAGAGA AGGTATCTCA TGGCAGGATG
CGACAACGAA TATGTGCTCA AACTCAGAAA GGGTCTCAGA TGGCAAGACG
CAACAACGAA TACGTACTCA CGCTCAGAAA GGGTCTCAGA TGGCAGGACG
CAACAACGAA TACGTACTCA CGCTCAGAAA GGGTCTCAGA TGGCAGGACG
CAACAACGAA TACGTACTCA CGCTCAGAAA GGGTCTCAGA TGGCAGGACG
CAACAACGAG TACGTACTCA CGCTCAGAAA GGGTCTCAGA TGGCAGGATG
CGAAAAGACC TACGAGCTGA AGCTTCGAGA CGGAGTATAC TGGACGGATG
AAAGAACACT TACGAGATTA AACTAAGAAA AGGAGTTACA TGGCACGATG
TGATAATACT TATCAAATTG CCTTAAGGGA CGGGCTCACA TGGCAAGATG
TGATAATACT TATCAAATCA CTTTAAGAGA CGGGCTCACA TGGCAAGATG
TGATAATACT TATCAAATCA CTTTAAGAGA TGGGCTCACA TGGCAAGATG
CGACAAGACG TACGAACTGA AGGTGAGAGA GGGGATCACG TGGCAGGACG
GGAGAAGGAG TACACCCTCA AGCTCCGGAA GGGCATCACC TGGGCGGACG
CGACAAGACC TACGAGCTGA AGCTCCGCCA GGGCGTCAAG TGGAGCGACG

GCGAGCCCTT CGTCGCCCAG GACGTGGTCA CCACGCTGGA ACTCGGCCAG
GCGAGGCGCT CGACGCCGAG GACGTCGTGT TCACCACCGA GCTCGGCCAG
ATGTTCCATT GACATCAGAA GACGTTCGAT TTACTTTCGA AATAGCTAAG
ATGTTCCATT GACATCAGAA GACGTTCGAT TTACTTTCGA AATAGCTAAG
ATGTTCCATT GACATCAGAA GACGTTCGAT TTACTTTCGA AATAGCTAAG
GAAAACCACT CACGATCGAT GATGTAATTT TCACTTTCGA GATTGCCAAA
GAGTACCTCT CACGGTGGAC GATGTGATCT TCACCTTCGA GATCGCAAAG

```
GAGTTCCTCT CACGGCAGAC GACGTGGTTT TCACCTTTGA AATCGCCAAG
GAGTTCCTCT CACGGCAGAC GACGTGGTTT TCACCTTTGA AATCGCCAAG
GAGTTCCTCT CACGGCAGAC GACGTGGTTT TCACCTTTGA AATCGCCAAG
GAGTTCCTCT CACGGCAGAC GATGTGGTTT TCACCTTCGA AATCGCCAAG
GAGAGGAGTT CAATGCCGAA GATGTGAAAT TCACATTTGA CATCGCAAAG
GTAAACCATT TACATCAAAA GATGTAAAAT TTACATTTGA GATTGCAAAA
GAAAACCTTT AACATCAGAG GACGTAAAAT TCACCTTTGA AATAGCAAAA
GAAAACCTTT AACATCAGAG GACGTAAAAT TCACCTTTGA AATAGCAAAA
GAAAACCTTT AACATCAGAG GACGTAAAAT TTACATTTGA AATAGCAAAG
GCAAGCCTCT CACCGCCGAG GACGTGAAGT TCACCTTCGA GCTCGCCAAG
GCGAGCCGTT CACCGCCGAG GACGTGGTCT TCACCTTCGA GCTCGGCAAG
GCGAAGACTT CACCGCCGAC GACGTGGTCT TCACCGTCGA GATCGGCAAG

---GTCCCCG GAGTCCCCTA CAGCAACGTC TGGGACTACA TCGAGAGCGT
CAG---CCGG GCGTGCCGTG GCAGAACCTG TGGAACTGGC TCGACTCCGT
AAGTACAAGG GAATACATTA CAGTAGTGTT TGGGAATGGC TTGATCATAT
AAGTACAAGG GAATACATTA CAGTAGTGTT TGGGAATGGC TTGATCATAT
AAGTACAAGG GAATACATTA CAGTAGTGTT TGGGAATGGC TTGATCATAT
AAGTACACCG GAATAAACTA CAGTCCTATC TGGGAGTGGC TCGAGAAAAT
AAGTACACCG GTATCAGTTA CAGTCCCGTC TGGAACTGGC TCGACAGAAT
AAGTACACTG GTATCAGCTA CAGTCCTGTG TGGAACTGGC TCGGCAGGAT
AAGTACACTG GTATCAGCTA CAGTCCTGTG TGGAACTGGC TCGGCAGGAT
AAGTACACTG GTATCAGCTA CAGTCCTGTG TGGAACTGGC TCGACAGGAT
AAGTACACTG GTATCAGCTA CAGTCCTGTG TGGAACTGGC TCGACAGGAT
AAGTATCCCG GAGTCCATTA CAGCTCTATG TGGAACTGGA TGAAGGAAGT
CAAATACCAG AGATTTCCTA CAGCCCAGTT TGGACATGGC TTGCGAAGGT
CAATATTCCG AGATTTATTA TAGCCCAATA TGGCAATGGT TGCAATCTAT
CAATATTCTG AGATTTATTA TAGCCCAATG TGGCAATGGT TGCAATCTAT
CAATATTCAG AGATTTATTA TAGCCCGATG TGGCAATGGT TGCAATCTAT
CAG---CAGA GCGTGTCCCT GAGCGCCATC TGGGACTGGC TCGCCGAGAT
---CTGGAGA CCGTCCCCTA CCACCAGCTC TGGGAGTGGC TGGCGCGGGC
---TACGAGG GCTCCTCCTA CCACTCGCTG TGGGAGTGGC TCGAGAGCGC

CGAGGCCACC GACGAGCGCA CGGTCACCGT CACCTTCTCG GAGAGCCGTC
CGAGGCCGTC GACGCGCACA CCGTCACCTG GCACTTCTCC GAGTCGCGCC
TGAAACACCC GACAACAGAA CCGTCATTTT TGTGTTCAAA GATCCTCGAT
TGAAACACCC GACAACAGAA CCGTCATTTT TGTGTTCAAA GATCCTCGAT
TGAAACACCC GACAACAGAA CCGTCATTTT TGTGTTCAAA GATCCTCGAT
CCAGAAAATT GATAACCTCA CTTTAAATTT CGTATTCTCT GATCCAAGAT
CGAAAAGGTT GATTCCTTCA CATTGAAGTT CGTCTTTTCC GATCCAAGAT
CGAAAGGGTC GATGAACGAA CGCTGAAGTT CGTCTTCTCC GACCCGAGGT
CGAAAGGGTC GATGAACGAA CGCTGAAGTT CGTCTTCTCC GACCCGAGGT
CGAAAGGGTC GATGAACGAA CGCTGAAGTT CGTCTTCTCC GACCCGAGGT
CGAAAGGATC GACGAACGAA CGCTGAAGTT TGTCTTCTCC GACCCGAGGT
GGAGATAGTA GACAGACTGA CTGTACGTGT CCACTTCACC GAACCGCTGT
TGATACTCCA GATGATTACA CTGTTGTTTT CACATTCTCA TCTCCAAGAT
AGAGACACCG GACAACAAAA CAGTAATATT CAAATTTTCA ACTGTAAACT
AGAGACACCG GACAACAAAA CAGTAATATT TAAGTTCTCA ACTGTAAACT
AGAGACACCG GACAACAAAA CAGTAATATT TAAGTTCTCA ACTGTAAACT
CGTGAAGGTC GACGATTACA CCTTGCGGTT CACGTTCAGC GATCCCCGTT
CGAGGCGGTG GACCAGCACA CGGTCAGGTT CACCTTCACT GAGGCCAACC
TGAGGCCGTC GACGACTACA CGGTCAAGTT CACCTTCTCC AAGGCCAACT

CGCAGGAGTG GATGAACTGG GCCTACTCCA ACCCCATCGT CCCGGACCAC
CGCAGGAGTG GGAGAACTGG CTCTACACCC GCACGATCCT GCCGCAGCAC
ATCATGAATG GAATGAACTC CTCTATACAC TTCCAATTGT TCCAAAACAT
ATCATGAATG GAATGAACTC CTCTATACAC TTCCAATTGT TCCAAAACAT
```

ATCATGAATG GAATGAACTC CTCTATACAC TTCCAATTGT TCCAAAACAT
ATCAAGAATG GGGACAACAA CTCATAAGTA TAGCAATCGT ACCCAAACAC
ACCAGGAATG GAAACAAATG CTCATCAACA CACCGATTGT TCCAAAGCAC
ACCAGGAATG GAAACAGATG CTCATCAACA CACCGATCGT ACCAAAACAC
ACCAGGAATG GAAACAGATG CTCATCAACA CACCGATCGT ACCAAAACAC
ACCAGGAATG GAAACAGATG CTCATCAACA CACCGATCGT ACCAAAACAC
ACCAGGAATG GAAACAGATG CTCATCAACA CACCGATCGT ACCAAAACAC
ATCAGCAATG GTCCTTCCAG CTCTACCAGT TGCCCATGGT TCCCGAGCAC
ACCACGAATG GGCATACCAA CTTTACCAGC TTCCAATTAT CCCAGAACAC
ATCACGAATG GACCTATAAC TTGTATCAGA TACCTATTAT TCCAAAGCAC
ATCATGAATG GGCATATAAC TTATATCAAA TACCTATTAT TCCAAAGCAT
ATCATGAATG GGCATATAAC TTGTATCAGA TACCTATTAT TCCAAAGCAT
ACCAGTCGTG GGACAACATC CTCTACACCC AGGGGATCGT GCCCAAGCAC
ACCAGGAGTG GTCGACCCAC CTCTACAGCC GGGCGATCGT GCCCAAGCAC
ACGCGCAGTG GGCCAACTGG CTGTACTTCA ACGCCATTGT GCCCGAACAC

ATCTGGGCCG GCATGGAGGA GAGCCAGGTC GCCGACAGCC CCAACGAGAA
ATCATGGGCG ACTGGACCGG TGAGGACATC ACCACCAACG CCAACCCGGA
ATTTGGGAAG AAAAAGATGA AACCACTATA CTTCAATCAT CAAATGAGTA
ATTTGGGAAG AAAAAGATGA AACCACTATA CTTCAATCAT CAAATGAGTA
ATTTGGGAAG AAAAAGATGA AACCACTATA CTTCAATCAT CAAATGAGTA
ATCTGGGAAA ATAGAACTGA AGAGCAGATA TTACAGGGTG CAAACGAAAA
ATCTGGGAAG GAAAGACAGA GGAAGAAGTG CTTCAGAGCG CCAACGAAAA
ATCTGGGAAA ACAAAACAGA GGAAGAAGTT CTTCAGGCGG CCAATGAAAA
ATCTGGGAAA ACAAAACAGA GGAAGAAGTT CTTCAGGCGG CCAATGAAAA
ATCTGGGAAA ACAAAACAGA GGAAGAAGTT CTTCAGGCGG CCAATGAAAA
ATCTGGGAAA ACAAAACAGA GGAAGAAGTC CTTCAGGCGG CTAACGAAAA
ATATGGAAAA ACAAGACGGA AGCCGAAATT CTGACTGGAG CGAACGAGGG
ATTTGGTCAA AGCGCTCAAA AGATGAAATA TTAAGTGGAG CAAATGAAAA
ATTTGGGCAA ACAAATCCAA AGATGATTTA TTGAATGGTG CAAACGAGAA
ATATGGGCAG ATAAGTCAAA AGATGATATG TTAAATGGTG CAAATGAAAA
ATATGGGCAG ATAAGTCAAA AGATGATATG TTAAATGGTG CAAATGAAAA
ATCTTCGAGG GCAAG----- -GAGGACGTG GCCAACGATG CGAACGAGAA
CTGTGGGAGG TCCGGTCCGA GGAGGAGGTG ATGAACGGCG CCAACGAGAA
ATCTGGTCGG AGCGCTCCGA GGAGGAGATC CTCACCGGCG CCAACGAGGA

CCCGGTCGGC ACCGGCCCCT ACGTCTACGA GTCGCACACC GACGACCGCA
CCCGGTCGGC TCGGGCCCGT TCGTGTACTC GACCCACAGC CAGGACCGCA
TCCATTGGGA TCAGGCCCAT ATGTTGCTCA CTCGTGGGAC CAGAATAAAA
TCCATTGGGA TCAGGCCCAT ATGTTGCTCA CTCGTGGGAC CAGAATAAAA
TCCATTGGGA TCAGGCCCAT ATGTTGCTCA CTCGTGGGAC CAGAATAAAA
ACCTGTTGGC TCTGGGCCAT ATTATGTTGA GAGCTGGGCT GATGACAGAA
TCCAGTCGGA TCCGGTCCTT ACTACATTGA AAGCTGGGCA GACGACAGGT
TCCAGTTGGA TCCGGTCCGT ACTACGTTGA AAGCTGGGCA GACGACAGAT
TCCAGTTGGA TCCGGTCCGT ACTACGTTGA AAGCTGGGCA GACGACAGAT
TCCAGTTGGA TCCGGTCCGT ACTACGTTGA AAGCTGGGCA GACGACAGAT
TCCAGTTGGA TCCGGTCCGT ACTACGTCGA GAGCTGGGCA GACGACAGAT
GCCCATTGGA ACGGGTTGCT ACACAGCCGA AGGTTACGGT CAGGACAGAA
AGCGATAGGA ACAGGGCCTT ATATGTTTGA AACTTACAGT GACGACAGAA
GGCAATAGGT TCGGGACCTT ACTTATTTGA CACTTACAGC AATGATCGTA
GGCTATAGGT TCAGGACCCT ATTTATTTGA CACTTATAGC AATGATCGTA
GGCTATAGGT TCAGGACCCT ATTTATTTGA CACTTATAGC AATGATCGTA
CCCCGTGGGT TCCGGTGCCT ACAAGTATTA CGGGCACAGC GAAGACCGCA
CCCCATCGGC ACCGGGCCGT ACGCCTACCA CTCGCACGAC CAGGACCGCA
CCCCGTCGGC ACCGGCCCGT ACCTCTACGA GACCCACGAC CAGGACCGCC

TGGTCTGGGA GCGCAACGAC GAGTGGTGGG CCATCGAGGC CCTCGACATG

TGGTCTGGGA GCGCAACGAC GACTGGTGGG GCACCGAGGC CATGGGCATC
TGATTTTCGA GCGTTTTGAG AATTGGTGGG GAACAAAAGT TATGGGTGTG
TGATTTTCGA GCGTTTTGAG AATTGGTGGG GAACAAAAGT TATGGGTGTG
TGATTTTCGA GCGTTTTGAG AATTGGTGGG GAACAAAAGT TATGGGTGTG
ATGTGTACAG GAAAAACCCG CAGTGGTGGG GCATCAAA-- -GTTGGATAC
GTGTCTTCAA GAAAAACGAA AACTGGTGGG GCATCAGGGA ACTCGGCTAC
GTGTATTCAA GAAGAACGGG AACTGGTGGG GCATCAGAGA ACTCGGTTAC
GTGTATTCAA GAAGAACGGG AACTGGTGGG GCATCAGAGA ACTCGGTTAC
GTGTATTCAA GAAGAACGGG AACTGGTGGG GCATCAGAGA ACTCGGTTAC
GTGTATTCAA GAAGAACGAG AACTGGTGGG GCATCAGAGA ACTCGGTTAC
TGATCTATTT GAGAAACGAA AACTGGTGGG CTATCGAGCA GCTGGGAATA
TGGTTTACCT CAGAAATGAT AATTGGTGGG GTAACAAAGT TTTTGGG---
TGGTCTACAA GAGGAATGAC AACTGGTGGG GCATAAAAGC AATGAATATG
TGGTTTACAA GAGGAATGAC AACTGGTGGG GCATAAAAGC AATGAATATG
TGGTCTACAA GAGGAATGAC AACTGGTGGG GCATAAAAGC AATGAATATG
TGATCTGGGT GCGGAACGAC GACTGGTGGG CCACGAAGCT CCTCGGCAAG
TGGTCTGGGT GCGCCGGGAC GGCTGGTGGG CGACCAAGGT GATCGGCAAG
AGGTCTGGAA GAAGAACGAG GACTGGTGGG CCACTGAGGC GCTCGGCCAC

ACGATGGACG CCCGCTACAT CGTCGACATC GTCAACGCCT CCAACGAGGT
GAGTTCCCGA TGCGCTACGT CGTCGACATC GTGAACCCGT CGAACGAGGT
AAACCTGCTC CGAAATACGT TGTTATAGTG AGAGTCCTCA GTAACAACGT
AAACCTGCTC CGAAATACGT TGTTATAGTG AGAGTCCTCA GTAACAACGT
AAACCTGCTC CGAAATACGT TGTTATAGTG AGAGTCCTCA GTAACAACGT
GATCCAAAAC CTGAAAGGGT TGTAATTCTC AGGATTCTCA GCAACAATGT
GATCCCAAAC CCGAAAGGAT CGTTGAACTG AGAGTGCTCA GCAACAACGT
GATCCCAAAC CTGAAAGGAT CGTGGAACTG AGAGTGCTCA GCAACAATGT
GATCCCAAAC CTGAAAGGAT CGTGGAACTG AGAGTGCTCA GCAATAATGT
GATCCCAAAC CTGAAAGGAT CGTGGAACTG AGAGTGCTCA GCAACAATGT
GATCCCAAAC CTGAAAGGAT CGTGGAACTG AGAGTGCTCA GCAACAATGT
AAGCCTACGC CCAAGAGAAT AGTCTATCTC ACCGTTTCGG GAAACAACGT
CAACCAAAGC CAAAGAGAGT TGTTTATTTG AGAGTTCTTA GTAACAACGT
ACACCAGCGC CTAAGAGAAT AGTTTATTTA ATTGTGCCCA GTAACAACGT
ACCCCAGCAC CAAAGAGAAT AGTTTATTTA ATTGTGCCTA GTAACAACGT
ACCCCAGCAC CAAAGAGAAT AGTTTATTTA ATTGTGCCTA GTAACAACGT
AAGGTTGCTC CCAAGTACAT CGTCGACATC CGCAACTCGA GCAACAACGT
CGGGTCGCGC CCAAGTACAT CGTGGACATC GTCAACTCGA GCAACGAGGT
GAGGTCAAGC CCACGTACAT CGTCGACGTG GTCAACACCA GCAACGAGGC

CACGATGGGC ATGCTGAACC AGGGCGAGGT CGACCTCTCC AACAACTTCC
CGCGCTCAGC CTGCTCATGC AGGGCACGCT GGACGTCTCG AACAACTTCC
GGCGCTCGGC ATGTTGATGA AAGGAGAACA GGACTTCAGT AATTTCATGC
GGCGCTCGGC ATGTTGATGA AAGGAGAACT GGACTTCAGT AATTTCATGC
GGCGCTCGGC ATGTTGATGA AAGGAGAACT GGACTTCAGT AATTTCATGC
TGCCGTAGGG ATGTTGATGA AAGGTGAGCT TGACTGGAGT AACTTTTTCT
AGCGGTTGGA ATGCTCATGA AAGGAGAACT CGACTGGAGT AACTTCTTCC
CGCAGTAGGA ATGCTCATGA AAGGAGAACT CGACTGGAGC AACTTCTTCC
CGCAGTAGGA ATGCTCATGA AAGGAGAACT CGACTGGAGC AACTTCTTCC
CGCAGTAGGA ATGCTCATGA AAGGAGAACT CGACTGGAGC AACTTCTTCC
CGCAGTAGGA ATGCTCATGA AAGGAGAACT CGACTGGAGC AACTTCTTCC
GGCTCTCGGT ATGATCTTCA AAGGGGAACT GGATATCAGT AACTTCTTCC
TGCACTTGGA ATGATAATGA AAGGTGAACT TGATATATCG AACTTCTTCC
TGCTTTGGGA ATGTTAATGA AAGGTGAATT AGATTTGAGT AACTTTTTCC
TGCATTGGGA ATGCTAATGA AAGGTGAATT AGATTTAAGT AACTTTTTCC
TGCATTGGGA ATGCTAATGA AAGGTGAATT AGATTTAAGT AACTTTTTCC
GGCGCTCGGC ATGGTGGTGA AGGGTGAGCT CGACCTGTCC AACAACTTCC
GGCGATGGAC TGGCTGCTCC AGAAGCACCT CGACCTGAGC AACAACTTCC

```
CGCGCTGAGC CAGGTGCTCC AGGGCAACAT CGACATCAAC AACAACTTCC

TGCCCGGTAT CGACCAGGTC CTCAACAGCA ACGAGACC-- -ATCACCAGC
TGCCCGGCAT CACGCAGCTC GTCGAGTCCG GAGCG----- -GTCGCGACG
TCCCAGGTGT TCCCATTTTG ---AAAAAAG TTTATAAT-- -CTCAATACA
TCCCAGGTGT TCCCATTTTG ---AAAAAAG TTTATAAT-- -CTCAATACA
TCCCAGGTGT TCCCATTTTG ---AAAAAAG TTTATAAT-- -CTCAATACA
TGCCGGGCAT ACCTATACTG ---AAAAAAT CGTATGGA-- -ATCCACACA
TCCCGGGTAT TCCTGTGTTG ---AAGAGAG CCTATGGA-- -ATCGTTACC
TGCCGGGTGT TCCGGTTTTG ---AAGAAAG CATACGGA-- -ATCGTCACC
TGCCGGGTGT TCCGGTTTTG ---AAGAAAG CATACGGA-- -ATCGTCACC
TGCCGGGTGT TCCGGTTTTG ---AAGAAAG CATACGGA-- -ATCGTCACC
TGCCGGGTGT TCCGGTTTTG ---AAGAAAG CATACGGA-- -ATCGTCACC
TTCCAGGAGT CCCAGCCGTC ---AAATCCG CTTACGGA-- -ATCCACACA
TTCCTGGTGT GCCGACACTT ---AAGAAAA CTTACAGTGA TATTCATACA
TTCCAGGTAT AAAGACTTTG ---AAAGCTA ACTATGGT-- -ATAACGACT
TTCCAGGTAT AAAGACTTTG ---AAAGCTA ACTACGGT-- -ATAACGACT
TTCCAGGTAT AAAGACTTTG ---AAAGCTA ACTACGGT-- -ATAACGACT
TGCCCGGGAT CGCCGCCCTG GTGAAGCAGG GGTAC----- -GTGAAGACC
TCCCGGGCGT CGCCAACCTG GTCACCGGTG ACTTCGGC-- -CTCCAGACC
TGCCCGGTAT CGCGCAGCTC GTCCAGGGCG GCTACCAG-- -GTCCAGACC

TTCTACGACG GCCCCCCGTA CATGAAGAGC GCCAACACGG CGTGGCTCAT
TACTACGACG AGGCGCCGTA CATGCTCTCG GCCAACACCG CGATGCTCAT
TGGTACGACG AACCACCGTA TCACCTCTCA TCAACGGTTG TTGGTCTTTT
TGGTACGACG AACCACCGTA TCACCTCTCA TCAACGGTTG TTGGTCTTTT
TGGTACGACG AACCACCGTA TCACCTCTCA TCAACGGTTG TTGGTCTTTT
TGGTACTCGG AAGCACCTTA CATGTTACCA GCTAACACGA CAGGCATATT
TGGTACGAAA ACGCACCTTA CATGCTTCCG GCCAACACCG CAGGGATCTT
TGGTATGAAA ACGCTCCTTA CATGCTCCCG GCCAACACCG CAGGAATCTA
TGGTATGAAA ACGCTCCTTA CATGCTCCCG GCCAACACCG CAGGAATCTA
TGGTATGAAA ACGCTCCTTA CATGCTCCCG GCCAACACCG CAGGAATCTA
TGGTATGAAA ACGCTCCTTA CATGCTCCCG GCCAACACCG CAGGAATCTA
TATTTCGATG GCCCGCCTTA CATGCTTTCC GACAATACGG CAGTTCTCTT
TGGTTTGACA AAGAACCTTA CATGTTGTCA GACAACACAG CTTACTTGTT
TTTTATGATA ATCCACCATA TATGATTCCA GATAATACTG TGTTTATGTT
TTTTATGATA ATCCGCCATA TATGATTCCA GATAATACTG TGTTTATGTT
TTTTATGATA ATCCGCCATA TATGATTCCA GATAATACTG TGTTTATGTT
TACTACGACG GGCCTCCGTA CATGCTCTCG GCTAACACCG CTTTCCTCTG
TACTACAACC GGCCGCCGTA CATGCTCGCC GCGAACACGG CCTGGCTGGT
TACTTCCCTG AAGAGCCCTA CATGCTCGCG GCCAACACCG CGTGGCTGGT

CCCGAACCAC ACCCGTGAGC CGCTCAACGA CACGGCGTTC CGCCAGGCCC
CCCGAACGCC ACCAAGGCGC CGGGCAACGA CGCCGCGTTC CGCCGCGCGC
CCTCAATGCA CGAAAATATC CTCTTAGCCT TCCCGAGTTC AGAAGAGCAA
CCTCAATGCA CGAAAATATC CTCTTAGCCT TCCCGAGTTC AGAAGAGCAA
CCTCAATGCA CGAAAATATC CTCTTAGCCT TCCCGAGTTC AGAAGAGCAA
TCTGAATGTC AAAAAATACC CTCTCAATAT CGCTCAATTT AGGAGGGCTA
TGTAAACGTG AACAAGTATC CTCTCAACAT CGCTGAGTTC AGAAGAGCGA
CATCAACGTG AATAAGTATC CTCTCAGCAT ACCTGAGTTC AGAAGAGCAA
CATCAACGTG AATAAGTATC CTCTCAGCAT ACCTGAGTTC AGAAGAGCAA
CATCAACGTG AGCAAGTATC CTCTCAGCAT ACCTGAGTTC AGAAGAGCAA
CATCAACGTG AACAAGTATC CTCTCAGCAT ACCTGAATTC AGAAGAGCAA
TCTAAACAAC AGCAGAAAGC CAATGGATGA TGTTAACTTC AGGAAGGCCG
CATTAATACA ACAAAGAAAC CTCTGAATGA TCCAAACTTC AGAAGAGCAA
TATCAATACC ACTAAATCTC CATTAAACAA TGTTGAATTA AGGCGAGCAA
TATCAATACC ACTAAATCTC CATTAAACAA TGTTGAATTA AGGAAAGCAA
```

TATCAATACC ACTAAATCTC CATTAAACAA TGTTGAATTA AGGCGAGCAA
GCTCAACCTT ACGAAGAAGC CCCTCAATGA TCCTGCTTTC CGGAGGGCGA
GATGAACACC AAGAAGAAGC CGATGGACGA CCCGGTGTTC CGGAGGGCGC
GCCCAACACC ACCAAGAAGC CGATGGACGA CCCTGAGTTC CGCAAGGCGC

TGGCCCACTC GATCAACATC ACCCAGATCG TCGAGGGCCC GTACGCCAAC
TCGCGCACGC CATCGACATC GACACCATCG TCACGACCGC CTACGGCAAC
TTGCTATGTC GATAAATGCA GATCCAATAG TTCAAAGAGT CTATGAAGGA
TTGCTATGTC GATAAATGCA GATCCAATAG TTCAAAGAGT CTATGAAGGA
TTGCTATGTC GATAAATGCA GATCCAATAG TTCAAAGAGT CTATGAAGGA
TGGCTTTTGC CATTGATCCA AATAAAATTG TTGATAGAGC TTTCGAAAAA
TGGCGTTTGC CATCAACCCG GAAAAGATCG TGACCAGAGC CTACGAGAAC
TGGCTTACGC TATCAATCCC GAGAAGATCG TTACCAGAGC TTACGAGAAC
TGGCTTACGC TATCAATCCC GAGAAGATCG TTACCAGAGC TTACGAGAAC
TGGCTTACGC TATCAATCCC GAGAAGATCG TTACCAGAGC TTACGAGAAC
TGGCTTACGC TATCAATCCC GAAAAGATCG TCACCAGGGC TTACGAGAAC
TTGCCTGGGC AATAAACGCC GACGACATAG TCACAAGGGT TTTCGAGAAC
TAGCATTTGC AATTGATCCA ACAGTTATAG CGAAAACAGT ATTTGAAGGT
TGGCATATGC AATTAATCCT AAGGTAATAG CAGAAAAAGT ATATGAAAAC
TGGCATATGC GATCAATCCT AAGGTAATAG CAGAAAAAGT ATATGAAAAT
TGGCATATGC AATCAATCCT AAGGTAATAG CAGAAAAAGT ATATGAAAAC
TCGCCTTCGC CATTGACACC CAGAAGATCG TGAACGTGGC CTACGCAGGT
TCGCCCACGC CATCGACACC AGGAAGATCG TCGAGGGCGT GTACCAGAAC
TCTCGGCCTC CATCGACATG GACGAGATCG TCAACAACGT CTACGGCGGC

CTGGTCCAGG CGGCCAACCC CACGGGT--- ---------- ----------
ATCGTCCAGG CGGCCAACCC GACCGGTCTG CTGCCCGCG- -----TTCGA
GCCGTCTTAA AAGCAGATCC CCTTGGTTTT CTTCCGAATT CTGTTTGGAT
GCCGTCTTAA AAGCAGATCC CCTTGGTTTT CTTCCGAATT CTGTTTGGAT
GCCGTCTTAA AAGCAGATCC CCTTGGTTTT CTTCCGAATT CTGTTTGGAT
ATGGTAGAAC CATCCAACGC AGTTGGCATC ATGCCAATTC CCGGCTGGAT
ATGGTCACCG CTGCCAATCC TGCAGGAATA CTGCCGCTTC CTGGTTACAT
ATGGTGACGG CTGCCAATCC CGCTGGAATC CTGCCGCTTC CCGGTTACAT
ATGGTGACGG CTGCCAATCC CGCTGGAATC CTGCCGCTTC CCGGTTACAT
ATGGTGACGG CTGCCAATCC CGCTGGAATC CTGCCGCTTC CCGGTTACAT
ATGGTGACGG CTGCCAATCC CGCTGGAATC CTGCCCCTTC CCGGTTACAT
CAGGTAATCA AGTCTAACCC ACTTGGCTTC CTTCCTATCG ACGCCTGGAT
CAGGTCCTTC CATCTAATTC TATTGGTTTC TTACCAATTA AGGGTTGGAT
CAAGTGGAAC CAGCAAATTC ATTAGGATTT GTACCGGCTA AGGCTTGGGA
CAAGTAGAAC CAGCAAATCC ATTAGGATTT GTGCCGGCTA AGGCTTGGGA
CAAGTAGAAC CAGCAAATCC ATTAGGATTT GTACCGGCTA AGGCTTGGGA
CTCGTCCAGG CTGCGGATCC GACCGGCTTG TTGCCCACC- -----TGGAG
CTGGTGCAGG CGGCGAACCC GACCGGGCTC CTCCCGCAG- -----TGGAG
CTGGTCACCG AGGCCGACCC CACCGGTCTG CTCCCTGTC- -----TGGGA

---------- ---------- ---------- ---------- ----------
GACGTACTAC GACCAGGACG TCATCGACGA GCTCGGGTTC ACCTTCGACG
GAAGTACTAT CCAAAAGAAG TTGTAGAAAA GCATGGTTTC AAATACGATC
GAAGTACTAT CCAAAAGAAG TTGTAGAAAA GCATGGTTTC AAATACGATC
GAAGTACTAT CCAAAAGAAG TTGTAGAAAA GCATGGTTTC AAATACGATC
AAAGTATTAT CCATCTGAAG TAGCAGAAAA ATATGGTTTC AAATATGATC
GAAGTACTAT CCCGACGAAG TTGTTGAAAA GTACGGTTTC AGGTATGATC
GAAGTACTAT CCGAAAGAAG TCGTCGATAA GTACGGATTC AAGTACGATC
GAAGTACTAT CCGAAAGAAG TCGTCGATAA GTACGGATTC AAGTACGATC
GAAGTACTAT CCGAAAGAAG TCGTCGATAA GTACGGATTC AAGTACGATC
GAAGTACTAT CCGAAAGAAG TCGTTGATAA GTACGGATTC AAGTACGATC
GAAGTATTAC GATGAAAAGG TTGTCGAGCA GTATGGATTC AAGTACGATC

```
GAAATATTAC CCAGAAAATG CCGTAAAACA ATACGGATTT AGATACGACA
GGAATATTAT GATAAAAATG TAGTTGATAA ATATGGATAT ACCTATGATC
AGAATATTAT GATAAAAATG TAGTTAATAA ATATGGCTAT ACCTATGATC
AGAATATTAT GATAAAAATG TAGTTAATAA ATATGGCTAT ACCTATGATC
TAAGTACGTG GACAAGGACG TGGTGGCGAA GTACGGGTTC AAGTACGATA
CAAGTACATC GACCAGGACG TGGTGAACCG GCTCGGCTTC TTCTACAGCC
CGACTACATC GACACCGAGG TGGTCGAGAA GCACGCCACC CCGCGTGACC

---------- ---------- ---GATGATG CGGGG----- ----------
TCGAGACGGC CAAGCAGCTC CTGGCCGACG CCGGGTACGA GGACAGCGAC
CTGAAGAGGC GAAAAGTATT CTTGATAAGC TTGGATTCAG GGATGTAAAT
CTGAAGAGGC GAAAAGTATT CTTGATAAGC TTGGATTCAG GGATGTAAAT
CTGAAGAGGC GAAAAGTATT CTTGATAAGC TTGGATTCAG GGATGTAAAT
CACAGATGGC TAAAGAGATC CTGGATGAAC TTGGTTTCAA AGATGTGAAC
CAGAGACGGC AAAGAAGATC CTCGATGAAC TTGGATTCAA GGATGTGAAC
CGGAGATGGC AAAGAAGATC CTCGACGAGC TTGGATTCAA AGATGTGAAC
CGGAGATGGC AAAGAAGATC CTCGACGAGC TTGGATTCAA AGATGTGAAC
CGGAGATGGC AAAGAAGATC CTCGACGAGC TTGGATTCAA AGATGTGAAC
CGGAGATGGC AAAGAAGATC CTCGACGAGC TTGGATTCAA AGATGTGAAC
CATCGGTTTC CAAAAAGGTC CTTGCCGATG CGGGCTACAA GGACATTAAC
CAAAGACGGC AAAAGATCTT CTTGATAAGG CAGGGTACAA AGATGTTAAC
CAGAAAAAGC GAAATCAATT TTGGATGCAG CAGGATTTAA A---TTAGGA
CAGAAAAAGC GAAATCAATT TTGGATGCAG CAGGATTTAA A---TTAGGA
CAGAAAAAGC GAAATCAATT TTGGATGCAG CAGGATTTAA A---TTAGGA
CCGCTCGGGC GAAGAAGATC CTCGCCGATG CGGGTTACAA GGACGTGGAC
CGGCCAAGGC GAAGGAGCTG CTCATCGACG CCGGCTACCG GGACCGGGAC
CGAAGGTCGC CAAGAAGATC CTTGAGGAAG CCGGCTACGA GGACACCGAC

---------- ---------- ---------- ---CCAGTTC GC--------
GGCGACGGGT ACGTCGAGAA CCTCGACGGT TCGGAGATGA ACCTCGAGCT
GGAGATGGTT TCAGAGAAAC CCCAGATGGA AAACCCATTA AGCTCACCAT
GGAGATGGTT TCAGAGAAAC CCCAGATGGA AAACCCATTA AGCTCACCAT
GGAGATGGTT TCAGAGAAGC CCCAGATGGA AAACCCATTA AGCTCACCAT
GGAGACGGAC TCAGGGAAGA TCCAAATGGG AAGTCGTTTA AGCTAACGAT
GGGGACGGAT TCAGAGAAGA TCCCAACGGA AAACCGTTCA AACTCACCAT
AAGGATGGAT TCAGAGAAGA TCCGAACGGA AAGCCGTTCA AGCTCACGAT
AAGGATGGAT TCAGAGAAGA TCCGAACGGA AAGCCGTTCA AACTCACCAT
AAGGATGGAT TCAGAGAAGA TCCGAACGGA AAGCCGTTCA AGCTCACGAT
AAGGATGGGT TCAGAGAAGA TCCGAACGGA AAGCCGTTCA AGCTCACGAT
GGAGACGGAT TCGTCGAGGC GCCGGACGGT TCCGAGATAG AGCTCTCTAT
AAAGACGGAT ATAGAGAGGC ACCAGATGGT AGCAAATTCA AAGTTGAAAT
AGTGATGGAG TAAGGACAAC ACCCGATGGG AAAAAATTTA AGTTAGAGAT
AGTGATGGAG TAAGGACAGC ACCCGATGGA AAAAGATTTA AGTTAGAGAT
AGTGATGGAG TAAGGACAGC ACCCGATGGA AAAAGATTTA AGTTAGAGAT
GGTGACGGGT TCGTCGAGGC GCCGGACGGG TCCAAGATCA GGCTCTCGGT
GGGGACGGCT TCATGGAGTC GCCCAGCGGG GCGAAGATCG CGCTCAAGAT
AACGACGGCT TCGTCGAGAC CCCTGACGGT GAGCCGATCG ATCTGACCCT

---------- ---------- ---------- ---------- ----------
CATCGTCCCC GCCGGCTGGA CCGACTGGAT GGACGCGGCG CAGATCATCG
CGAGTGTCCG TATGGATGGA CCGACTGGAT GCAGGCAATT CAGGTGATAG
CGAGTGTCCG TATGGATGGA CCGACTGGAT GCAGGCAATT CAGGTGATAG
CGAGTGTCCG TATGGATGGA CCGACTGGAT GCAGGCAATT CAGGTGATAG
AGAATGCCCG TATGGCTGGA CAGACTGGAT GGTTTCAATA CAATCCATAG
CGAGTGTCCT TACGGCTGGA CTGACTGGAT GGTCTCCATA CAGTCCATCG
TGAGTGTCCG TACGGATGGA CCGACTGGAT GGTTTCTATC CAGTCCATTG
CGAGTGTCCG TACGGATGGA CCGACTGGAT GGTTTCTATC CAGTCCATTG
```

```
TGAGTGTCCG TACGGATGGA CCGACTGGAT GGTTTCTATC CAGTCCATTG
TGAGTGTCCG TACGGATGGA CCGACTGGAT GGTTTCTATC CAGTCTATTG
AATAGTTCCA TTCGGCTGGA CCGACTGGAT GGAGTCGATA AAGATAATTG
AATTGTTCCA TATGGTTGGA CAGATTGGAT GGAATCAATC AAAATAATTG
TAGTGTGCCA TATGGTTGGA CAGATTGGAT GGAAGCAGCT AAAATTGTAG
TAGTGTGCCA TATGGTTGGA CAGATTGGAT GGAAGCAGCT AAAATTGTTG
TAGTGTGCCA TATGGTTGGA CAGATTGGAT GGAAGCAGCT AAGATAGTTG
GATCGTGCCG TTCGGGTGGA CCGACTGGAT GGAGTCCATC AAGATCGTGG
CGCCGTGCCG GCCGGGTGGA CCGACTGGAT GGAGGCCGCC CGGGTGATCA
CATCGTTCCC AGCGGCTGGA CCGACTGGAT GGAGGCCGCC CGGGTCATCA

---------- ---------- ---------- ---------- ----------
CCGAGTCGGC CGGCGAGGCG GGCATCCACA TCACCAACGC CACGCCCGAC
TAGATCAACT CAAGGTGGTT GGAATAAACG CTGAACCATA CTTCCCGGAT
TAGATCAACT CAAGGTGGTT GGAATAAACG CTGAACCATA CTTCCCGGAT
TAGATCAACT CAAGGTGGTT GGAATAAACG CTGAACCATA CTTCCCGGAT
CAGAAGATCT GAGAAAAGTG GGTATAAATG TGGAACCATC TTATCCTGAT
CAGAAGATCT TGTGAAGGTC GGAATCAACG TGGAACCCAA GTACCCTGAT
CAGAAGATCT CGTGAAAGTC GGAATCAACG TCGAACCTAA ATACCCCGAC
CAGAAGATCT CGTGAAAGTC GGAATCAACG TCGAACCTAA ATACCCCGAC
CAGAAGATCT CGTGAAAGTC GGAATCAACG TCGAACCTAA ATACCCCGAC
CAGAAGATCT CGTGAAAGTC GGAATCAACG TCGAACCCAA GTACCCCGAC
CCAACAACCT GAACGCCGTG GGCATTAACG CAAAGGCCGA GTTCCCCGAC
CTTCTCAGCT TAGGATGGTT GGAATTAATG CAGAAGCAAA ATTCCCAGAT
CAGATCAATT AAAGGCTGTA GGGATAGATG CAGAAGCCAA ATTCCCAGAT
CAGATCAATT AAAGGTAGTA GGAATCGATG CAGAAGCTAA ATTTCCAGAT
CAGATCAATT AAAGGCAGTA GGAATCGATG CAGAAGCTAA ATTCCCAGAT
CCGAGGGGTG TAAGGCCGCG GGTATCAACG TGGAACCCGA GTATCCCGAC
GCGAGGGCGC CAAGGGGGCC GGGATCAACC TCGAGCCGGA GTTCCCCGAC
GCGAGAGCGC GAGCGAAGTC GGTATCAAGG TCACCACTGA CTTCCCCGAG

---------- ---------- ---------- ---------- ----------
TCGGGCGCCG TCGACGACGC CCGCACCACG GGCAACTTCG ACCTGGTGAT
TCTTCCAAAT ACTATGAAAA CATGTACAAA GGAGAATTCG ATATAGAAAT
TCTTCCAAAT ACTATGAAAA CATGTACAAA GGAGAATTCG ATATAGAAAT
TCTTCCAAAT ACTATGAAAA CATGTACAAA GGAGAATTCG ATATAGAAAT
GCCTCCAAGT ATAATGATGA TTTGTATGGT GGAAAATTTG ATATCATATT
TACTCCAAGT ACGCAGACGA CCTCTACGGT GGAAAATTCG ACCTGATTCT
TACTCCAAAT ACGCAGACGA CCTCTACGGT GGAAAGTTCG ATCTCATACT
TACTCCAAAT ACGCAGACGA CCTCTACGGT GGAAAGTTCG ATCTCATACT
TACTCCAAAT ACGCAGACGA CCTCTACGGT GGAAAGTTCG ATCTCATACT
TACTCCAAAT ACGCAGACGA CCTCTACGGT GGAAAGTTTG ATCTCATACT
TATTCGAGGT ATCAGGACGA GCTTTACGGC GGCAACTTCG ATATGGCGAT
TACAGCAAAT ACTGGGAAGA TTTGACAACA GGAAAGTTTG ATATGGCAAT
TACAGCAAAT ATTATGAGGA TTTAACAAAG GGCACATTTG ACTTATCGTT
TATAGTAAAT ATTATGAGGA TTTAACGAAA GGAACATTTG ATTTATCATT
TATAGTAAAT ATTATGAGGA TTTAACGAAA GGAACATTTG ATTTATCATT
TTCGGTGGGT ACAGCGATCA GCTCTACGGC GGCACCTTCG ATATGGCGAT
TACAACGCGC TCGTCGACGC CCGCAACTCC GGCAAGTTCG ACATGGTCCT
TTCAACGCCC TGGTGGACCA GCGCAACAGC GGCGAGTTCG ACCTGGTGAT

---------- ---------- ---------- ---------- ----------
GAACAACTGG ---GCGCAGA TGTCCAACAC GCCGTGGACG TACTACAACT
GAATGCCAAT GGAACAGGTA TAAGCAGCAC TCCCTGGACA TATTTCAATA
GAATGCCAAT GGAACAGGTA TAAGCAGCAC TCCCTGGACA TATTTCAATA
GAATGCCAAT GGAACAGGTA TAAGCAGCAC TCCCTGGACA TATTTCAATA
GAACAATTAT GTCACAGGAG TTTCAAGCAC CATATGGTCA TATTTTAATG
```

```
CAACAACTTC GTAACCGGTG TCTCTGCAAC CATATGGTCC TACTTCAACG
CAACAACTTT ACAACCGGTG TTTCCGCTAC CATCTGGTCC TATTTCAACG
CAACAACTTT ACAACCGGTG TTTCCGCTAC CATCTGGTCC TATTTCAACG
CAACAACTTT ACAACCGGTG TTTCCGCTAC CATCTGGTCC TATTTCAACG
CAACAACTTT ACAACCGGTG TTTCCGCTAC CATCTGGTCC TACTTCAACG
AAACAACTTC AACAGCAACC TGTCTAACAC AGTATGGAGT TACTACTACT
CAACAACTTC AACAGTCAGA TGACGGTTTC ACCATGGACA ATGTTTAATT
TAATAACTTT GGAAGCCAAG TAACATCAAC TCCATGGACA CTGTACAACT
TAATAACTTT GGCAGCCAAG TCACGTCAAC TCCATGGACT TTGTATAATT
TAATAACTTT GGCAGCCAAG TCACGTCAAC TCCATGGACT TTGTATAATT
CAACAACTTC GGCTCCGGTC TCAGCAACAC GCCGTGGACC TTCTACAACT
CAACAAC--- GACCGCCAGC TCGCCAGCAC CCCGTGGCGG TACTACGACT
CAACAAC--- GAGCGCCAGA TCAGCAACAC GCCGTGGACC TACTACGACT

---------- ---------- ---------- ---------- ----------
ACCTGTTCAG CATGCCG--- ---ATCCAGG ACTCGATGTG GTCCGGCAAC
CTATTTTCTA TCCTGATGCT TTAGAATCTG AATTCTCTTA CACAGGAAAT
CTATTTTCTA TCCTGATGCT TTAGAATCTG AATTCTCTTA CACAGGAAAT
CTATTTTCTA TCCTGATGCT TTAGAATCTG AATTCTCTTA CACAGGAAAT
CCGTATTTTA CCCTGATGCT GTTGAATCAG AATATTCATA CTCTGGTAAC
GTGTGTTCTA TCCAGATGCA GCAGAATCCG AGTACTCTTA CTCTGGAAAC
GTGTGTTCTA TCCGGATGCA GTAGAATCCG AGTACTCCTA CTCCGGAAAC
GTGTGTTCTA TCCGGATGCA GTAGAATCCG AGTACTCCTA CTCCGGAAAC
GTGTGTTCTA TCCGGATGCA GTAGAATCCG AGTACTCCTA CTCCGGAAAC
GTGTGTTCTA TCCAGATGCA GTAGAATCCG AGTACTCCTA CTCCGGAAAC
GGCTATTCTG GGAT------ ---ATCAGAG AGCAGCAGAC TCAGGGCAAC
GGTTATTCAA TTCTAAC--- ---ATAAGTG ACAATATGTA CAATGGAAAC
GGTTATTTAA CAAAGTA--- ---GAAGGAG ATGCACAATA TAACGGAAAC
GGCTATTTAA CAAAGTTCAA ---GAAAACG GTCCTCAGAA TAATGGAAAC
GGCTATTTAA CAAAGTTCAA ---GAAAACG GTCCTCAGAA TAATGGAAAC
GGGTGTTCTA CCATCCC--- ---ATCTCGG ACAATATGCC CAACGGTAAC
TCATCTTCCG CCTGCCG--- ---GTGCGCA AGCAGCAGAC CACGGCGAAC
ACATCTTCCG CCTCCCG--- ---GTCCAGG ACCAGCAGAC TACGGTCAAC

---GGTCGGC ACCAACGGCC TGGT------ ---------- ----------
TTCGGTCGCT GGGACGCCAC CGAGGCGTTC GCGAAGGTCA ACGATCTCGC
TATGGAAGAT ACCAGAATCC CGAGGTGGAA AGTTTACTTG AAGAACTCAA
TATGGAAGAT ACCAGAATCC CGAGGTGGAA AGTTTACTTG AAGAACTCAA
TATGGAAGAT ACCAGAATCC CGAGGTGGAA AGTTTACTTG AAGAACTCAA
TTTGGAAAAT ATGAAAACCC TGATGTAGAG GTTCTTCTGG ATGAACTCAA
TTCGGAAAGT ACGCCAACCC CGAAGTTGAA ACACTCCTTG ATGAACTCAA
TTTGGAAAGT ACGCCAATCC TGAAGTTGAG ACTCTTCTCG ACGAACTCAA
TTTGGAAAGT ACGCCAATCC TGAAGTTGAG ACTCTTCTCG ACGAACTCAA
TTTGGAAAGT ACGCCAATCC TGAAGTTGAG ACTCTTCTCG ACGAACTCAA
TTTGGAAAGT ACGCCAATCC TGAAGTTGAG ACTCTTCTCG ACGAACTCAA
TACGGAAAAT ACAACAATCC CAAGGCCTTT GAGTTGATGG AAGCCTTCGA
TTCGGTAAAT ATAAAAATCA AAAACTCTTC GATTTGATTA CACAACTTAA
TTTGGACGTT TTGATGTGCC AGGATTACAA GACTTAATAG CAAAATTCAA
TTTGGGCGCT TTGATGTGCC AGGATTACAA GACTTAATAG CAAAATTTAA
TTTGGGCGCT TTGATGTGCC AGGATTACAA GACTTAATAG CAAAATTTAA
TTCGGCCGGT ATAACAACCA GCAGGTCTTC GACCTCGTGG AAGAGCTCAA
TTCGGCCGGT ACGAGAACAA GCAGGCCTGG CGGCTGGTCC GGGAGCTCGA
TTCGGCCGCT ACGAGAACGA AGAGGCCTGG GAGCTGGTGG GTGAACTCGA

---------- ---------- ---------- ---------- ----------
TCGCGCCCAG TCCGGTTCGC CGGAGTTC-- -CAGGCGGCG ATCAGCGACC
CAGGACACCA CTTGACAATG TTGAGAAAGT CACCGAACTC TGTGGAAAAC
```

```
CAGGACACCA CTTGACAATG TTGAGAAAGT CACCGAACTC TGTGGAAAAC
CAGGACACCA CTTGACAATG TTGAGAAAGT CACCGAACTC TGTGGAAAAC
CCGAACACCC TTCACAAATG AAGAGAAAAT AAGGCAAACC GTTGCACAGC
CAGAAGCAGA ---GACGATG CTAAAATT-- -AAAGAGATA GTAGCCAAAC
CAGAAGC--- AATGATGATG CTAAAATT-- -AAAGAAGTA GTAGCCAAGC
CAGAAGC--- AATGATGATG CTAAAATT-- -AAAGAAGTA GTAGCCAAGC
CAGAAGC--- AATGATGATG CTAAAATT-- -AAAGAAGTA GTAGCCAAGC
CAGAAGC--- AATGATGATG CTAAAATT-- -AAAGAAGTA GTAGCCAAGC
CAGAACGCCT GTCGATGACT ACGAGACTGG TCAAAAGATC ATGTCGGAAC
CTCAACACCA ATGGAAGATA TAGCTGCAAA CAAGAAAGTA TTGGAGCAAA
CCAAACAAAG TTAGGAAGCC CAGAAGCT-- -AAACAAGCA GCTGCACAAC
CCAAACAAAG CTAGGAAGTC CAGAAGCC-- -AAGCAAGCA GCTGCACAAC
CCAAACAAAG CTAGGAAGTC CAGAAGCT-- -AAGCAAGCA GCTGCACAAC
CCGTGTGCCC ACCGACGATA CCGAAGGCAT GAAAGCCGTG ATCTCCAAGA
CGGCGTCCGG ACCGACGACG TCGAGGGGAT GAAGCGGATC ATCTCCCGGC
CGGCATCCCG GTTGAGGACA CCGAGAAGAT CCGCGAGGTG GCGGGCAAGA

---------- ---------- ---------- ---------- ----------
TCCAGCGCAT CTCGCTGGAA GAGATGCCGA TGATCCCCAT GTGGTACAAC
TTGGAGAGAT CCTTTTAAAA GATCTACCTT TCATCCCTCT CTGGTATGGA
TTGGAGAGAT CCTTTTAAAA GATCTACCTT TCATCCCTCT CTGGTATGGA
TTGGAGAGAT CCTTTTAAAA GATCTACCTT TCATCCCTCT CTGGTATGGA
TCTCAGAAAT ATTGCTCAGA GAATTGCCGT TTATTCCTCT CTGGTATAAC
TGTCGGAGAT TCTGCTCAAA GATCTGCCAT TCATACCGCT CTGGTACAAC
TGTCAGAGAT ACTGCTCAAG GATCTGCCGT TCATTCCTCT GTGGTACAAC
TGTCAGAGAT ACTGCTCAAG GATCTGCCGT TCATTCCTCT GTGGTACAAC
TGTCAGAGAT ACTGCTCAAG GATCTGCCGT TCATTCCTCT GTGGTACAAC
TTTCAGAGAT ACTGCTCAAG GATCTGCCGT TCATTCCTCT GTGGTACAAC
TGGAAGAGCT TTTCTTGAAA GAGATTCCCT ACGTTCCCCT CTGGTTCAAC
TCGCTGAAAT TTTCTTAAAA GAAATGCCTG CTATTCCTCT TTGGTACAAC
TTGAAGAGAT TTTCTTAAAG AATATGCCAG CAATACCACT TTGGTACAAT
TTGAAGAGAT TTTCTTAAAG AATATGCCAG CTATACCACT TTGGTACAAT
TTGAAGAGAT TTTCTTAAAG AATATGCCAG CTATACCACT TTGGTACAAT
TCCAGGAGAT CCAGCTCAAG GAGATGGTGG CCATCCCGCT CTGGTTCAAC
TCCAGGAGAT CCACCTCCGG GAGATGCCGA TCATCCCGCT CTGGTACAAC
TCCAGAAGAT CCAGCTCGAG GAGATGCCGG TCATCCCGCT CTGGTACAAC

---------- ---------- ---------- ---------- ----------
GGCCTGTGGT CGCAGGTCAC GGACGGTACG TGGACCAACT GGCCGTCGTC
GCGATGGCGT TTATAACACA AGATAACGTT TGGACTAACT GGCCCAACGA
GCGATGGCGT TTATAACACA AGATAACGTG TGGACTAACT GGCCCAACGA
GCGATGGCGT TTATAACACA AGATAACGTT TGGACTAACT GGCCCAACGA
GGAGCATGGT TCCAGGCCGT AGAAAACGTT TGGACCAACT GGCCAGATGA
GGTGCGTGGT TCCAGGCTTC TGAAGCCGTC TGGACCAACT GGCCAACAGA
GGTGCATGGT TCCAGGCTTC TGAAGCTGTG TGGACCAACT GGCCAACGGA
GGTGCATGGT TCCAGGCTTC TGAAGCTGTG TGGACCAACT GGCCAACGGA
GGTGCATGGT TCCAGGCTTC TGAAGCTGTG TGGACCAACT GGCCAACGGA
GGTGCATGGT TCCAGGCCTC TGAAGCTGTG TGGACCAACT GGCCAACGGA
GGAATGTGGT TCCAGGCAAG TACCAATGTT TGGACCAACT GGCCGAGCGA
GGTATGTGGT TCCAAGCAAG TACGCAAGTA TGGAAGAACT GGCCAAGTGA
GGGCTTTGGT TCCAAGCATC AAATGCAGCA TGGGAGAATT GGCCAACAGA
GGTCTCTGGT TCCAAGCATC AAATGCAGCA TGGGAGAATT GGCCAACAGA
GGTCTCTGGT TCCAAGCATC AAATGCAGCA TGGGAGAATT GGCCAACAGA
GGCCTCTGGG CCCAGTTCAA CACCTCGGTG TGGACCAACT GGCCCACCTC
GGGCTGTGGG CGCAGATGAC CAGCGCGGTC TGGACGAACT GGCCGTCCGA
GGCCTGTGGG CGCAGTACAA CAACTCGGTG TGGACCAACT GGCCCTCCAG
```

```
-------AAC GAGTCA---T ATATATCTAC GACGAGGTCG CGGAAGAGAA
T------GAC GGCCCGGTCA CCGCCTTCCC GTCGACGTGG CACGGCTACT
A------CAT AATCCA---T ATGCCTGGCC ATGTGGTTGG GCCAACTGGT
A------CAT AATCCA---T ATGCCTGGCC ATGTGGTTGG GCCAACTGGT
A------CAT AATCCA---T ATGCCTGGCC ATGTGGTTGG GCCAACTGGT
G------AAC AATCCT---T ATGCATGGCC TATCTCCTGG GGAGGTCACT
G------AAG AATCCG---T ACGCTGTGCC GATAGGTTGG AACGGCTGGT
G------AAG AATCCG---T ACGCTGTCCC GATAGGCTGG AACGGCTGGT
G------AAG AATCCG---T ACGCTGTCCC GATAGGCTGG AACGGCTGGT
G------AAG AATCCG---T ACGCTGTCCC GATAGGCTGG AACGGCTGGT
G------AAG AATCCG---T ACGCTGTCCC GATAGGCTGG AACGGCTGGT
A------CAC GGCCCT---C ACTACTATCC ATGTACGTGG AACGGAAAAT
G------AAA AAACCA---T ATGCATATCC AGTTACATGG GGTGGAAGAT
G------CAT GATCCA---T ATGCGTATCC AGTTACTTGG GGTGGCAGAT
A------CAT GACCCA---T ATGCTTATCC AGTTACTTGG GGTGGCAGAT
A------CAT GACCCA---T ATGCTTATCC AGTTACTTGG GGTGGCAGAT
AGCCAAGGAT ACGCCC---G ACTACATGCC GTGTCTGTGG GGCGGGTACG
GGCGATGGGA GCCCCC---A AGCACGCTCC GAGCATGTGG CGGGACTGGA
CGAG---TCC GACAAC---A ACTACCTGCC CTCGACCTGG CGCGGCTACT

CGTCCACTGG C--------- ---------- ---------- ----------
GGGAGATGGG CGGCCTCGAG ATGCTCGCCA GCCTGACGCC GGCCGGC---
GGCAAACTGG TGCCTTGAAA ATTCTATTTA ATCTCAAACC GGCAAAA---
GGCAAACTGG TGCCTTGAAA ATTCTATTTA ATCTCAAACC GGCAAAA---
GGCAAACTGG TGCCTTGAAA ATTCTATTTA ATCTCAAACC GGCAAAA---
GGCAACTCTG CGGTGTAAAA GTGCTTTTCA ACATAAAAGC AAAA------
GGCAGCTCAC AGGAATCAAG ACACTGTTTG GTATCGAAGC AAAA------
GGCAGCTCAC AGGAATCAAG ACGCTCTTTG GTATTGAAGC AAAG------
GGCAGCTCAC AGGAATCAAG ACGCTCTTTG GTATTGAAGC AAAG------
GGCAGCTCAC AGGAATCAAG ACGCTCTTTG GTATTGAAGC AAAG------
GGCAGTTCAC AGGAATCAAG ACACTCTTCG GTATTGAAGC AAAG------
GGCAGCTCGG TGGTATATTC ATGCTGACAG CCCTGGAGGC TAAG------
GGCAGACTGG TGGAGTGTTA ATGCTTATAG GAATTACCAA TAAA------
GGCAGACAGG TGGAACAATG ATGTTGATGA AAATAAAACC GAAGACTTCG
GGCAAACAGG TGGGACAATG ATGTTGATGA AGATAAAACC GAAAGCTTCG
GGCAAACAGG TGGGACAATG ATGTTGATGA AGATAAAACC GAAAGCTTCG
TCCAGGTGGG CGGTCTCCTC ACCCTCACCG AACTCAAGCC CGCACAGCAG
TGGAGATGGG CGGCCTCCTC ATGCTGACCG AGCTCCGGCC GGCGGCGGGC
TCCAGCTCGG CAGCATCCTG ATGCTGACCG AACTGGAGCC CGCCTCGGAG

------
------
------
------
------
------
------
------
------
------
------
------
------
CAG---
CAA---
CAA---
------
```

```
------
GAAGAG
```

Appendix 7: Newick trees with each *manD* and *manE* branch flagged for PAML dN/dS ratio analysis

manD branch

(Tnap_1566:0.0017757516,TM1226:0.0000000001,(TRQ2_1592:0.00000002948,(((Fnod_1553:0.2533291520,(TeCCSD1:0.0105505983,(Theet_1038:0.0032265210,Thebr_1980:0.0000001132):0.0249640354):0.3364401484):0.1322806005,(Theba_2445:0.4011430862,(STHERM:0.2140141937,(Tbis_3512:0.3249905892,(Tfu_0910:0.3244443743,(Ndas_3658:0.5843141874,Xcel_0262:0.4914507159):0.1319804891):0.1113220760):0.2551016948):0.2552789352):0.1085570425):0.2219440451,(Tlet_1438:0.1782150369,(CTN_1348:0.0188083256,(Tpet_1545:0.0053935123,(TRQ2_1595:0.0018074054,(TM1223:0.0000000001,Tnap_1569:0.0000000001):0.0018117306):0.0018635322):0.0333640404):0.1469916775):0.1872377288)**#1**:0.3849462796):0.0017791687);

manE branch

(Tnap_1566:0.0017757516,TM1226:0.0000000001,(TRQ2_1592:0.00000002948,(((Fnod_1553:0.2533291520,(TeCCSD1:0.0105505983,(Theet_1038:0.0032265210,Thebr_1980:0.0000001132):0.0249640354):0.3364401484):0.1322806005,(Theba_2445:0.4011430862,(STHERM:0.2140141937,(Tbis_3512:0.3249905892,(Tfu_0910:0.3244443743,(Ndas_3658:0.5843141874,Xcel_0262:0.4914507159):0.1319804891):0.1113220760):0.2551016948):0.2552789352):0.1085570425):0.2219440451,(Tlet_1438:0.1782150369,(CTN_1348:0.0188083256,(Tpet_1545:0.0053935123,(TRQ2_1595:0.0018074054,(TM1223:0.0000000001,gb|CP00183:0.0000000001):0.0018117306):0.0018635322):0.0333640404):0.1469916775)**#1**:0.1872377288):0.3849462796):0.0017791687);

Appendix 8: PAML control file for *manD* under hypothesis testing branch-model H_0

```
      seqfile = nt2.phy * sequence data filename
     treefile = manD_aa2.phy_phyml_tree.txt    * tree structure file name
      outfile = manD_H0.out        * main result file name

        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

        ndata = 1
        clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
       aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
   aaRatefile = /opt/paml44/dat/wag.dat  * only used for aa seqs with model=empirical(_F)
                   * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

        model = 2
                   * models for codons:
                       * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                   * models for AAs or codon-translated AAs:
                       * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                       * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 2  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0
```

```
        icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
        Mgene = 0
                   * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                   * AA: 0:rates, 1:separate


    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 1.06  * initial or fixed kappa
    fix_omega = 1  * 1: omega or omega_1 fixed, 0: estimate
        omega = 1 * initial or fixed omega, for codons or codon-based AAs


    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0 * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 1  * # of categories in dG of NSsites models


        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)


   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time


* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Appendix 9: PAML control file for *manD* under hypothesis testing branch-model H_1

```
      seqfile = nt2.phy * sequence data filename
     treefile = manD_aa2.phy_phyml_tree.txt    * tree structure file name
      outfile = manD_H1.out        * main result file name

        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

        ndata = 1
        clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
       aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
   aaRatefile = /opt/paml44/dat/wag.dat  * only used for aa seqs with model=empirical(_F)
                   * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

        model = 2
                   * models for codons:
                       * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                   * models for AAs or codon-translated AAs:
                       * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                       * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 2  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0

        icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
```

```
        Mgene = 0
                    * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                    * AA: 0:rates, 1:separate

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 1.06  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = 1 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0 * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 1  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Appendix 10: PAML control file for *manE* under hypothesis testing branch-model H_0

```
      seqfile = nt2.phy * sequence data filename
     treefile = manE_aa2.phy_phyml_tree.txt    * tree structure file name
      outfile = manE_H0.out        * main result file name

        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

        ndata = 1
        clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
       aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
   aaRatefile = /opt/paml44/dat/wag.dat  * only used for aa seqs with model=empirical(_F)
                   * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

        model = 2
                   * models for codons:
                       * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                   * models for AAs or codon-translated AAs:
                       * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                       * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 2  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0

        icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
```

```
        Mgene = 0
                    * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                    * AA: 0:rates, 1:separate

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 1.06  * initial or fixed kappa
    fix_omega = 1  * 1: omega or omega_1 fixed, 0: estimate
        omega = 1 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0 * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 1  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Appendix 11: PAML control file for *manE* under hypothesis testing branch-model H_1

```
      seqfile = nt2.phy * sequence data filename
     treefile = manE_aa2.phy_phyml_tree.txt    * tree structure file name
      outfile = manE_H1.out        * main result file name

        noisy = 9  * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1  * 0: concise; 1: detailed, 2: too much
      runmode = 0  * 0: user tree;  1: semi-automatic;  2: automatic
                   * 3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise

      seqtype = 1  * 1:codons; 2:AAs; 3:codons-->AAs
    CodonFreq = 2  * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table

        ndata = 1
        clock = 0  * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
      aaDist = 0  * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
  aaRatefile = /opt/paml44/dat/wag.dat  * only used for aa seqs with model=empirical(_F)
                   * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

        model = 2
                   * models for codons:
                       * 0:one, 1:b, 2:2 or more dN/dS ratios for branches
                   * models for AAs or codon-translated AAs:
                       * 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
                       * 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr=189)

      NSsites = 2  * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
                   * 5:gamma;6:2gamma;7:beta;8:beta&w;9:beta&gamma;
                   * 10:beta&gamma+1; 11:beta&normal>1; 12:0&2normal>1;
                   * 13:3normal>0

        icode = 0  * 0:universal code; 1:mammalian mt; 2-10:see below
```

```
        Mgene = 0
                    * codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
                    * AA: 0:rates, 1:separate

    fix_kappa = 0  * 1: kappa fixed, 0: kappa to be estimated
        kappa = 1.06  * initial or fixed kappa
    fix_omega = 0  * 1: omega or omega_1 fixed, 0: estimate
        omega = 1 * initial or fixed omega, for codons or codon-based AAs

    fix_alpha = 1  * 0: estimate gamma shape parameter; 1: fix it at alpha
        alpha = 0 * initial or fixed alpha, 0:infinity (constant rate)
       Malpha = 0  * different alphas for genes
        ncatG = 1  * # of categories in dG of NSsites models

        getSE = 0  * 0: don't want them, 1: want S.E.s of estimates
 RateAncestor = 0  * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

   Small_Diff = .5e-6
    cleandata = 0  * remove sites with ambiguity data (1:yes, 0:no)?
*  fix_blength = -1  * 0: ignore, -1: random, 1: initial, 2: fixed
       method = 0  * Optimization method 0: simultaneous; 1: one branch a time

* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GEN
```

Appendix 12: PAML branch-site model output: the *manD* branch set as foreground

Codon site	Residue	Site class			
		0	1	2a	2b
1	-	0.71883	0.14208	0.11585	0.02323
2	-	0.71883	0.14208	0.11585	0.02323
3	-	0.71883	0.14208	0.11585	0.02323
4	-	0.74874	0.12525	0.10616	0.01985
5	-	0.49355	0.26039	0.20347	0.04259
6	M	0.99376	0	0.00624	0
7	R	0.98998	0.00001	0.01001	0
8	-	0.91864	0.01439	0.0662	0.00078
9	M	0.01666	0.95892	0.00021	0.02421
10	L	0.70637	0.2767	0.01058	0.00635
11	R	0.07098	0.75763	0.04151	0.12988
12	L	0.06451	0.84686	0.01414	0.0745
13	P	0.79469	0.16392	0.03408	0.0073
14	A	0.71921	0.22063	0.03971	0.02046
15	A	0.02527	0.85761	0.00781	0.10931
16	L	0.96042	0.01881	0.01943	0.00134
17	A	0.14159	0.69506	0.06111	0.10224
18	A	0.00183	0.88898	0.00034	0.10885
19	L	0.35835	0.53281	0.05688	0.05197
20	A	0.0001	0.77962	0.00083	0.21945
21	L	0.00478	0.88182	0.00054	0.11286
22	A	0.01242	0.85736	0.00572	0.1245
23	V	0.97274	0.01596	0.01074	0.00056
24	S	0.01409	0.84683	0.00225	0.13684
25	A	0.13288	0.72803	0.02124	0.11785
26	C	0.86068	0.00024	0.13905	0.00004
27	S	0.03321	0.8277	0.0053	0.13378
28	G	0.74654	0.11437	0.12037	0.01871
29	G	0.55191	0.309	0.08871	0.05037
30	G	0.6317	0.22922	0.10167	0.03742
31	-	0.79532	0.0656	0.12834	0.01074
32	-	0.19467	0.66625	0.03113	0.10795
33	-	0.74719	0.11372	0.12048	0.01861
34	-	0.08676	0.77415	0.01386	0.12523
35	-	0.2373	0.62361	0.03797	0.10111

36	-	0.42486	0.43605	0.06816	0.07092
37	-	0.7419	0.11902	0.11961	0.01947
38	-	0.2411	0.61981	0.03858	0.1005
39	-	0.55883	0.30209	0.08984	0.04925
40	-	0.79044	0.07048	0.12755	0.01154
41	-	0.11449	0.74642	0.01829	0.12079
42	-	0.6279	0.23302	0.10105	0.03804
43	-	0.48401	0.3769	0.07772	0.06137
44	-	0.78481	0.0761	0.12663	0.01246
45	-	0.81493	0.04599	0.13155	0.00753
46	-	0.49939	0.36152	0.08021	0.05888
47	G	0.71439	0.14652	0.11513	0.02396
48	N	0.44015	0.42077	0.07063	0.06845
49	S	0.25779	0.60312	0.04126	0.09782
50	D	0.00826	0.85265	0.00132	0.13777
51	G	0.05682	0.01237	0.92535	0.00547
52	G	0.02113	0.95544	0.00018	0.02324
53	S	0.00536	0.84774	0.00113	0.14577
54	G	0.84722	0.0137	0.13684	0.00225
55	Q	0.81136	0.04955	0.13097	0.00812
56	Y	0.77551	0.00032	0.22414	0.00003
57	P	0.98615	0.00023	0.0136	0.00001
58	R	0.96293	0	0.03707	0
59	N	0.95587	0.02936	0.01248	0.00229
60	E	0.97836	0	0.02164	0
61	T	0.9816	0	0.0184	0
62	L	0.70654	0.00101	0.29233	0.00011
63	Y	0.99131	0	0.00869	0
64	T	0.07842	0.6676	0.15344	0.10054
65	T	0.98714	0.0012	0.01163	0.00003
66	G	0.98163	0	0.01837	0
67	T	0.08448	0.00032	0.91509	0.00011
68	A	0.8526	0.13064	0.0126	0.00416
69	W	0.99202	0	0.00798	0
70	E	0.52229	0.45391	0.00969	0.0141
71	A	0.98606	0.00001	0.01393	0
72	P	0.98178	0	0.01822	0
73	T	0.29423	0.67884	0.00439	0.02254
74	S	0.98332	0	0.01668	0
75	W	0.99202	0	0.00798	0
76	N	0.98732	0	0.01268	0

77	P	0.99069	0	0.00931	0
78	M	0.99014	0.00268	0.00709	0.00009
79	M	0.94479	0.03702	0.0167	0.00149
80	R	0.71999	0.26184	0.01015	0.00802
81	G	0.99127	0.00074	0.00798	0.00001
82	Q	0.96403	0.01913	0.01543	0.00142
83	F	0.95458	0.02404	0.01994	0.00144
84	A	0.98067	0	0.01933	0
85	V	0.60966	0.00014	0.39019	0.00002
86	G	0.98806	0	0.01194	0
87	T	0.98185	0	0.01815	0
88	N	0.11423	0.73647	0.04874	0.10056
89	G	0.98677	0	0.01323	0
90	L	0.98416	0	0.01583	0
91	V	0.96994	0.00141	0.02854	0.00011
92	Y	0.98982	0	0.01018	0
93	E	0.98937	0	0.01063	0
94	S	0.976	0.00703	0.01657	0.0004
95	L	0.7535	0	0.24649	0
96	F	0.99097	0	0.00903	0
97	H	0.63089	0.08739	0.27101	0.01072
98	Y	0.99076	0	0.00924	0
99	D	0.98821	0	0.01179	0
100	A	0.98819	0	0.01181	0
101	D	0.87203	0.10598	0.01658	0.00541
102	A	0.4752	0.43457	0.0437	0.04653
103	G	0.61872	0.06467	0.30947	0.00713
104	E	0.37129	0.00006	0.62863	0.00001
105	Y	0.94429	0.0399	0.01408	0.00173
106	V	0.75584	0.00368	0.23999	0.00048
107	H	0.98857	0	0.01143	0
108	W	0.99202	0	0.00798	0
109	L	0.98876	0	0.01124	0
110	A	0.98473	0	0.01527	0
111	E	0.98515	0	0.01485	0
112	S	0.22908	0.24458	0.49975	0.02659
113	D	0.98543	0	0.01457	0
114	E	0.4938	0.4602	0.00862	0.03738
115	W	0.99202	0	0.00798	0
116	T	0.35227	0.35249	0.23969	0.05555
117	S	0.40139	0.49203	0.02922	0.07736

118	E	0.68325	0.07764	0.22961	0.0095
119	T	0.9128	0.00001	0.08719	0
120	E	0.95141	0.03103	0.01601	0.00155
121	H	0.9897	0	0.0103	0
122	V	0.00868	0.80125	0.04041	0.14965
123	I	0.68473	0.00329	0.31158	0.00039
124	T	0.10389	0.00168	0.89397	0.00047
125	L	0.88409	0	0.1159	0
126	R	0.99349	0	0.00651	0
127	E	0.98024	0.00384	0.01562	0.0003
128	G	0.98868	0	0.01132	0
129	V	0.92708	0.01025	0.06141	0.00126
130	T	0.04505	0.02637	0.91293	0.01564
131	W	0.99202	0	0.00798	0
132	N	0.10452	0.83306	0.01682	0.0456
133	D	0.98796	0	0.01204	0
134	G	0.03348	0	0.96652	0
135	E	0.30457	0.00034	0.69502	0.00007
136	P	0.96619	0.02137	0.01194	0.0005
137	F	0.88313	0.00007	0.11679	0.00001
138	V	0.98467	0.00024	0.01508	0.00001
139	A	0.87124	0.00446	0.12407	0.00023
140	Q	0.97913	0.00001	0.02086	0
141	D	0.98244	0	0.01756	0
142	V	0.97744	0	0.02256	0
143	V	0.25205	0.00245	0.74444	0.00106
144	T	0.98749	0	0.01251	0
145	T	0.98527	0	0.01473	0
146	L	0.99184	0.00001	0.00814	0
147	E	0.98506	0	0.01494	0
148	L	0.9813	0	0.0187	0
149	G	0.98188	0	0.01812	0
150	Q	0.98759	0	0.01241	0
151	-	0.9819	0.00002	0.01809	0
152	V	0.80472	0.1818	0.00723	0.00625
153	P	0.27632	0.23291	0.42308	0.0677
154	G	0.98776	0.00023	0.01201	0.00001
155	V	0.9844	0.00002	0.01558	0
156	P	0.76335	0.14728	0.07177	0.0176
157	Y	0.99145	0.00008	0.00847	0
158	S	0.99292	0.00004	0.00704	0

159	N	0.0046	0.49756	0.12575	0.37209
160	V	0.87725	0.04616	0.07291	0.00368
161	W	0.99202	0	0.00798	0
162	D	0.4455	0.45389	0.059	0.04161
163	Y	0.99198	0.00003	0.00798	0
164	I	0.98613	0.00001	0.01386	0
165	E	0.53206	0.38268	0.05162	0.03363
166	S	0.0003	0.80547	0.00169	0.19254
167	V	0.97978	0.00096	0.01921	0.00006
168	E	0.98543	0	0.01457	0
169	A	0.96546	0.00495	0.02928	0.00031
170	T	0.23485	0.41548	0.28711	0.06257
171	D	0.98303	0	0.01697	0
172	E	0.00917	0.90791	0.00022	0.08269
173	R	0.00041	0.68082	0.00696	0.31181
174	T	0.98396	0	0.01604	0
175	V	0.95189	0.00002	0.04809	0
176	T	0.48775	0.4224	0.04221	0.04765
177	V	0.98564	0.00001	0.01435	0
178	T	0.73789	0.18877	0.05313	0.02021
179	F	0.99313	0	0.00687	0
180	S	0.02931	0.00107	0.96869	0.00093
181	E	0.98527	0.00143	0.0132	0.0001
182	S	0.92304	0.05821	0.01677	0.00198
183	R	0.68634	0.2743	0.02333	0.01603
184	P	0.98917	0.00004	0.01079	0
185	Q	0.90949	0.00003	0.09048	0
186	E	0.9893	0.00005	0.01064	0
187	W	0.99202	0	0.00798	0
188	M	0.00014	0.82264	0.00048	0.17674
189	N	0.70142	0.00237	0.29593	0.00028
190	W	0.00097	0.82842	0.00114	0.16948
191	A	0.99154	0	0.00846	0
192	Y	0.95836	0	0.04164	0
193	S	0.01249	0.79625	0.01169	0.17956
194	N	0.02315	0.88456	0.00097	0.09131
195	P	0.97863	0.00913	0.01202	0.00022
196	I	0.98775	0	0.01225	0
197	V	0.98023	0.00001	0.01976	0
198	P	0.98834	0	0.01166	0
199	D	0.98807	0.00015	0.01177	0.00001

200	H	0.98548	0	0.01452	0
201	I	0.98595	0	0.01405	0
202	W	0.99046	0.00154	0.00797	0.00002
203	A	0.96237	0.02535	0.01072	0.00155
204	G	0.00267	0.89855	0.0047	0.09408
205	M	0.79462	0.17656	0.01765	0.01117
206	E	0.03862	0.10791	0.78267	0.0708
207	E	0.98755	0	0.01244	0
208	S	0.06273	0.00104	0.93587	0.00036
209	Q	0.02367	0.00049	0.97561	0.00023
210	V	0.97762	0.01008	0.01179	0.00051
211	A	0.37453	0.5765	0.01215	0.03682
212	D	0.92272	0.03357	0.04134	0.00237
213	S	0.02371	0.00727	0.96341	0.00562
214	P	0.61079	0	0.38921	0
215	N	0.98656	0	0.01344	0
216	E	0.98381	0	0.01619	0
217	N	0.38197	0.00464	0.61265	0.00074
218	P	0.98738	0	0.01262	0
219	V	0.48518	0.00021	0.51458	0.00003
220	G	0.98895	0	0.01105	0
221	T	0.97892	0.00174	0.01929	0.00005
222	G	0.98439	0	0.01561	0
223	P	0.98618	0.00027	0.01354	0.00001
224	Y	0.98924	0	0.01076	0
225	V	0.00267	0.82447	0.00178	0.17109
226	Y	0.58302	0.00647	0.40938	0.00113
227	E	0.1982	0.16672	0.60777	0.0273
228	S	0.0611	0.2635	0.55481	0.12059
229	H	0.98638	0	0.01361	0
230	T	0.57315	0.21055	0.18334	0.03296
231	D	0.45832	0.02076	0.51891	0.00202
232	D	0.75975	0	0.24025	0
233	R	0.79515	0	0.20484	0
234	M	0.98127	0.00003	0.01869	0
235	V	0.73885	0.00001	0.26113	0
236	W	0.89933	0.00019	0.10047	0.00001
237	E	0.14769	0.68382	0.07292	0.09557
238	R	0.49465	0.00009	0.50523	0.00003
239	N	0.06103	0	0.93897	0
240	D	0.98159	0.00045	0.01793	0.00003

241	E	0.98467	0.00004	0.01529	0
242	W	0.99202	0	0.00798	0
243	W	0.99202	0	0.00798	0
244	A	0.97992	0	0.02008	0
245	I	0.66987	0.00006	0.33006	0.00001
246	E	0.98929	0	0.0107	0
247	A	0.71246	0.21236	0.04954	0.02564
248	L	0.63309	0.03774	0.32408	0.00509
249	D	0.98705	0.00001	0.01294	0
250	M	0.10795	0.59699	0.12953	0.16553
251	T	0.49466	0.44195	0.02069	0.0427
252	M	0.98375	0.00164	0.01457	0.00004
253	D	0.02056	0.72145	0.04064	0.21735
254	A	0.98449	0.00008	0.01543	0
255	R	0.97874	0.00004	0.02122	0
256	Y	0.00581	0.00001	0.99417	0.00001
257	I	0.84965	0.00003	0.15032	0
258	V	0.98166	0	0.01834	0
259	D	0.41909	0.00096	0.57962	0.00033
260	I	0.68998	0.00013	0.30987	0.00001
261	V	0.9819	0.01088	0.00701	0.00022
262	N	0.9804	0.00001	0.01959	0
263	A	0.95832	0.02969	0.01091	0.00108
264	S	0.99131	0.00001	0.00868	0
265	N	0.98814	0	0.01186	0
266	E	0.98623	0	0.01377	0
267	V	0.979	0	0.021	0
268	T	0.98422	0	0.01578	0
269	M	0.9768	0.00015	0.02303	0.00001
270	G	0.98176	0.00001	0.01823	0
271	M	0.99376	0	0.00624	0
272	L	0.95755	0.01617	0.02559	0.00069
273	N	0.99332	0.00033	0.00634	0.00001
274	Q	0.99064	0	0.00936	0
275	G	0.98508	0.00001	0.01492	0
276	E	0.98819	0.00034	0.01145	0.00002
277	V	0.98111	0.00011	0.01877	0
278	D	0.98543	0	0.01457	0
279	L	0.59568	0.21069	0.16996	0.02368
280	S	0.77116	0.22004	0.00478	0.00402
281	N	0.98333	0	0.01667	0

282	N	0.99259	0	0.00741	0
283	F	0.16681	0	0.83319	0
284	L	0.98535	0	0.01465	0
285	P	0.98746	0	0.01254	0
286	G	0.98951	0	0.01049	0
287	I	0.86534	0.00048	0.13414	0.00004
288	D	0.98334	0.00413	0.01241	0.00012
289	Q	0.73374	0.21328	0.03482	0.01816
290	V	0.984	0.00005	0.01596	0
291	L	0.8565	0.00441	0.13836	0.00072
292	N	0.98807	0.00003	0.01191	0
293	S	0.00034	0.94833	0	0.05132
294	N	0.28065	0.48347	0.15018	0.0857
295	E	0.98906	0.00009	0.01085	0
296	T	0.03233	0.29354	0.55657	0.11756
297	-	0.71883	0.14208	0.11585	0.02323
298	I	0.74703	0.00006	0.2529	0.00001
299	T	0.00442	0.8853	0.00145	0.10884
300	S	0.98646	0	0.01354	0
301	F	0.9892	0.00207	0.00869	0.00003
302	Y	0.99055	0	0.00945	0
303	D	0.97197	0.0028	0.02494	0.00029
304	G	0.69735	0.25187	0.03029	0.02049
305	P	0.89871	0.07833	0.01991	0.00305
306	P	0.98266	0	0.01734	0
307	Y	0.98666	0	0.01334	0
308	M	0.00912	0	0.99088	0
309	K	0.98578	0.00069	0.0135	0.00003
310	S	0.70023	0.26315	0.025	0.01162
311	A	0.43483	0.00001	0.56516	0
312	N	0.39108	0	0.60892	0
313	T	0.06555	0	0.93445	0
314	A	0.64362	0.00003	0.35635	0
315	W	0.14559	0.79807	0.00523	0.05111
316	L	0.95784	0.00001	0.04215	0
317	I	0.98396	0.00324	0.01267	0.00013
318	P	0.97897	0.00715	0.01346	0.00041
319	N	0.98868	0	0.01132	0
320	H	0.22295	0.67079	0.02415	0.0821
321	T	0.20792	0.0055	0.78467	0.00191
322	R	0.98822	0	0.01178	0

323	E	0.93097	0.0491	0.01842	0.00151
324	P	0.98831	0	0.01169	0
325	L	0.97406	0.00942	0.01617	0.00035
326	N	0.81306	0.00005	0.18689	0
327	D	0.51298	0.00001	0.487	0
328	T	0.09592	0.86032	0.00622	0.03754
329	A	0.84372	0.12976	0.01617	0.01034
330	F	0.99288	0	0.00712	0
331	R	0.99107	0	0.00893	0
332	Q	0.99216	0.00072	0.0071	0.00001
333	A	0.98474	0	0.01526	0
334	L	0.93875	0.00002	0.06123	0
335	A	0.98479	0	0.01521	0
336	H	0.02914	0.60568	0.17693	0.18824
337	S	0.59741	0.00002	0.40256	0
338	I	0.98091	0	0.01909	0
339	N	0.9808	0.00001	0.01919	0
340	I	0.6411	0.00168	0.35699	0.00023
341	T	0.0018	0.90363	0.00048	0.0941
342	Q	0.03956	0.27423	0.57312	0.11309
343	I	0.987	0	0.013	0
344	V	0.9804	0	0.01959	0
345	E	0.01664	0.84544	0.00952	0.1284
346	G	0.5506	0.43807	0.0036	0.00773
347	P	0.96225	0.00101	0.03665	0.00009
348	Y	0.97723	0.00001	0.02276	0
349	A	0.98768	0.0002	0.01211	0.00001
350	N	0.04186	0.36812	0.5242	0.06582
351	L	0.06224	0.00147	0.93566	0.00063
352	V	0.98046	0	0.01954	0
353	Q	0.00076	0.79131	0.00071	0.20721
354	A	0.07602	0.03598	0.87389	0.01411
355	A	0.92813	0.00006	0.0718	0.00001
356	N	0.8012	0	0.19879	0
357	P	0.98784	0.00016	0.01199	0.00001
358	T	0.87649	0.02065	0.10114	0.00172
359	G	0.98638	0	0.01362	0
360	-	0.9793	0.00001	0.02069	0
361	-	0.97489	0.00025	0.02485	0.00001
362	-	0.98984	0	0.01016	0
363	-	0.00617	0.85206	0.0022	0.13958

364	-	0.30439	0.41476	0.2181	0.06275
365	-	0.64952	0.04171	0.30403	0.00474
366	-	0.98907	0.00288	0.008	0.00004
367	-	0.92053	0.0703	0.00598	0.0032
368	-	0.98945	0.00002	0.01052	0
369	-	0.9886	0	0.0114	0
370	-	0.98328	0.00619	0.0104	0.00014
371	-	0.97975	0.00423	0.01591	0.00011
372	-	0.01847	0.89828	0.00061	0.08264
373	-	0.97691	0.00019	0.02289	0.00001
374	-	0.98082	0	0.01917	0
375	-	0.98073	0.00002	0.01925	0
376	-	0.87593	0.10802	0.01027	0.00578
377	-	0.98574	0.00058	0.01364	0.00004
378	-	0.76835	0.00034	0.23128	0.00003
379	-	0.98786	0	0.01214	0
380	-	0.99282	0	0.00718	0
381	-	0.34218	0.61783	0.00424	0.03575
382	-	0.98962	0.00003	0.01034	0
383	-	0.98857	0	0.01143	0
384	-	0.97474	0.00667	0.01836	0.00023
385	-	0.87351	0.10065	0.01647	0.00937
386	-	0.00131	0.83999	0.00437	0.15433
387	-	0.98363	0	0.01637	0
388	-	0.98946	0	0.01054	0
389	-	0.09972	0.33787	0.48634	0.07606
390	-	0.98538	0.00051	0.01407	0.00003
391	-	0.98743	0	0.01257	0
392	D	0.98718	0.00079	0.01198	0.00004
393	D	0.78865	0.00016	0.21118	0.00002
394	A	0.96835	0.00001	0.03163	0
395	G	0.98806	0	0.01194	0
396	-	0.99056	0	0.00944	0
397	-	0.76144	0.00007	0.23849	0.00001
398	-	0.98845	0	0.01155	0
399	-	0.71773	0.25134	0.01472	0.01621
400	-	0.98233	0.00001	0.01766	0
401	-	0.6894	0.29339	0.00806	0.00915
402	-	0.9851	0	0.0149	0
403	-	0.97808	0	0.02192	0
404	-	0.99293	0.00004	0.00702	0

405	-	0.99208	0.0004	0.00751	0.00001
406	-	0.98835	0	0.01165	0
407	-	0.9045	0.06423	0.02522	0.00605
408	-	0.98809	0.00001	0.01191	0
409	-	0.9759	0.00001	0.02409	0
410	-	0.98763	0	0.01237	0
411	-	0.48559	0.47133	0.0099	0.03319
412	P	0.01491	0.88172	0.00179	0.10158
413	V	0.74693	0.00002	0.25304	0
414	R	0.93683	0.05063	0.01009	0.00244
415	-	0.9903	0	0.0097	0
416	-	0.39745	0.5453	0.01129	0.04597
417	-	0.98282	0.00002	0.01716	0
418	-	0.91988	0.01247	0.06635	0.0013
419	-	0.98057	0.00013	0.0193	0
420	-	0.99013	0	0.00987	0
421	-	0.96768	0.02123	0.0106	0.00048
422	-	0.98294	0	0.01706	0
423	-	0.99202	0	0.00798	0
424	-	0.9846	0	0.0154	0
425	-	0.98584	0	0.01416	0
426	-	0.99202	0	0.00798	0
427	-	0.99376	0	0.00624	0
428	-	0.55399	0.00001	0.44599	0
429	-	0.61636	0.00067	0.38288	0.00009
430	-	0.9821	0.00047	0.0174	0.00003
431	-	0.98508	0.0004	0.0145	0.00001
432	-	0.49482	0.00332	0.50105	0.0008
433	-	0.98453	0.00001	0.01546	0
434	-	0.70037	0.0042	0.29497	0.00046
435	-	0.96811	0.00061	0.03124	0.00004
436	-	0.11946	0.78269	0.03861	0.05924
437	-	0.97999	0.00674	0.01302	0.00025
438	-	0.49534	0.45546	0.01296	0.03624
439	-	0.1057	0.71173	0.05151	0.13106
440	-	0.98004	0.00001	0.01995	0
441	-	0.98878	0	0.01122	0
442	-	0.98524	0	0.01476	0
443	-	0.98565	0	0.01434	0
444	-	0.95551	0.00137	0.043	0.00012
445	-	0.98907	0.00012	0.0108	0.00001

446	-	0.9793	0.00528	0.01528	0.00015
447	-	0.20278	0.07544	0.70931	0.01246
448	-	0.98685	0.00008	0.01306	0
449	-	0.98498	0	0.01502	0
450	-	0.98835	0	0.01165	0
451	-	0.68804	0.01652	0.29366	0.00178
452	-	0.49351	0.48393	0.00577	0.01679
453	-	0.98609	0.00004	0.01387	0
454	-	0.99011	0.00007	0.00982	0
455	-	0.00006	0.91044	0	0.08949
456	-	0.93445	0	0.06555	0
457	-	0.77141	0.02603	0.19923	0.00332
458	-	0.76447	0.00012	0.2354	0.00001
459	-	0.98986	0.00053	0.00959	0.00002
460	-	0.0047	0.76636	0.03567	0.19328
461	-	0.98944	0	0.01056	0
462	-	0.79804	0.01059	0.19023	0.00114
463	-	0.99056	0	0.00944	0
464	-	0.98857	0	0.01143	0
465	-	0.95113	0.00045	0.04839	0.00004
466	-	0.63653	0.01464	0.34598	0.00285
467	-	0.90398	0.00425	0.09133	0.00044
468	-	0.98452	0	0.01548	0
469	-	0.14834	0	0.85166	0
470	-	0.10005	0.00143	0.89806	0.00046
471	-	0.01653	0.87888	0.00367	0.10092
472	-	0.91602	0.04396	0.0372	0.00283
473	-	0.7795	0.18341	0.02682	0.01027
474	-	0.70278	0.01114	0.28476	0.00133
475	-	0.15931	0.39572	0.26195	0.18302
476	-	0.39096	0.50728	0.03475	0.06701
477	-	0.98392	0	0.01608	0
478	-	0.93908	0.0043	0.05612	0.0005
479	-	0.99202	0	0.00798	0
480	-	0.97028	0.00169	0.02792	0.00011
481	-	0.98642	0.00105	0.01249	0.00004
482	-	0.98958	0	0.01042	0
483	-	0.98819	0	0.01181	0
484	-	0.05019	0.09668	0.80182	0.05131
485	-	0.59015	0.04293	0.35975	0.00716
486	-	0.99258	0	0.00742	0

487	-	0.9328	0.05471	0.01094	0.00156
488	-	0.06242	0.91359	0.00078	0.0232
489	-	0.98388	0.00226	0.01371	0.00016
490	-	0.84805	0.12043	0.02372	0.0078
491	-	0.41973	0.38528	0.14433	0.05065
492	-	0.96178	0.01544	0.02177	0.00101
493	-	0.00131	0.94684	0.00005	0.0518
494	-	0.98206	0.00038	0.01754	0.00002
495	-	0.00319	0.84433	0.002	0.15048
496	-	0.03351	0.91567	0.00159	0.04923
497	-	0.87083	0.11748	0.00779	0.00389
498	-	0.75608	0.15762	0.06174	0.02456
499	-	0.98889	0.00004	0.01106	0
500	-	0.98333	0	0.01667	0
501	-	0.71862	0	0.28138	0
502	G	0.98929	0	0.01071	0
503	R	0.95353	0.00057	0.04586	0.00004
504	H	0.98916	0.00006	0.01077	0
505	Q	0.16735	0.59646	0.13641	0.09978
506	R	0.98602	0.00155	0.01234	0.00009
507	P	0.74722	0.23787	0.00829	0.00663
508	G	0.959	0.02139	0.01807	0.00154
509	-	0.97767	0.00051	0.02178	0.00004
510	-	0.02	0.8862	0.00131	0.09249
511	-	0.0141	0.82055	0.03872	0.12663
512	-	0.96329	0.00036	0.03633	0.00002
513	-	0.97849	0.00011	0.0214	0
514	-	0.58855	0.37133	0.01444	0.02568
515	-	0.98887	0.0001	0.01102	0.00001
516	-	0.99142	0	0.00857	0
517	-	0.98801	0	0.01199	0
518	-	0.84801	0.08966	0.05154	0.0108
519	-	0.97489	0.01094	0.01374	0.00043
520	-	0.83337	0.15224	0.01003	0.00435
521	-	0.00184	0.89448	0.00069	0.10299
522	-	0.95812	0.01777	0.02215	0.00195
523	-	0.91326	0.00734	0.07889	0.00051
524	-	0.00023	0.88625	0.00013	0.11339
525	-	0.9857	0	0.01429	0
526	-	0.00157	0.89111	0.00026	0.10705
527	-	0.39967	0.42682	0.10819	0.06532

528	-	0.36164	0.01308	0.62253	0.00276
529	-	0.22729	0.71503	0.00722	0.05047
530	-	0.15727	0.66717	0.06245	0.11311
531	-	0.01889	0.26596	0.58911	0.12603
532	-	0.52339	0.13637	0.32148	0.01876
533	-	0.89572	0.04082	0.06019	0.00327
534	-	0.98616	0.00022	0.01361	0.00001
535	-	0.57242	0.01192	0.41296	0.00271
536	-	0.98732	0.00001	0.01266	0
537	-	0.98297	0	0.01703	0
538	-	0.82421	0.10787	0.05881	0.00911
539	-	0.97073	0	0.02926	0
540	-	0.98966	0.00009	0.01024	0
541	-	0.94476	0.00001	0.05522	0
542	-	0.92993	0	0.07007	0
543	-	0.98533	0.00009	0.01458	0
544	-	0.65628	0.31699	0.0099	0.01684
545	-	0.98367	0	0.01633	0
546	-	0.98831	0	0.01168	0
547	-	0.99133	0	0.00867	0
548	-	0.99202	0	0.00798	0
549	-	0.9876	0	0.01239	0
550	-	0.02849	0	0.97151	0
551	-	0.55763	0	0.44237	0
552	-	0.96019	0.00224	0.03728	0.00029
553	-	0.05485	0.00001	0.94514	0
554	-	0.98504	0.00001	0.01495	0
555	-	0.04193	0	0.95807	0
556	-	0.51865	0.21117	0.24407	0.0261
557	-	0.00014	0.66282	0.00441	0.33262
558	-	0.76516	0.10273	0.12097	0.01114
559	-	0.00023	0.88284	0.00005	0.11688
560	-	0.98086	0.00012	0.01901	0.00001
561	-	0.99202	0	0.00798	0
562	-	0.98257	0.00001	0.01742	0
563	-	0.98788	0	0.01212	0
564	-	0.99202	0	0.00798	0
565	-	0.98329	0	0.01671	0
566	-	0.7481	0.11824	0.11933	0.01433
567	-	0.93172	0.05051	0.01465	0.00312
568	-	0.77934	0.08158	0.12573	0.01336

569	-	0.73151	0.1294	0.11792	0.02117
570	N	0.1801	0.74198	0.01131	0.06662
571	E	0.68524	0.29044	0.0089	0.01542
572	S	0.98726	0.00016	0.01258	0
573	-	0.71883	0.14208	0.11585	0.02323
574	Y	0.98695	0.00232	0.01068	0.00005
575	I	0.89225	0.08894	0.01364	0.00517
576	S	0.00014	0.96066	0.00001	0.03919
577	T	0.98804	0	0.01195	0
578	T	0.08139	0.75435	0.07197	0.09229
579	R	0.14085	0.59548	0.15933	0.10433
580	S	0.99201	0.00001	0.00798	0
581	R	0.63668	0.10105	0.2476	0.01468
582	K	0.13108	0.00009	0.8688	0.00003
583	R	0.00047	0.95341	0.00004	0.04608
584	T	0.51579	0.47327	0.00404	0.0069
585	S	0.98766	0.00012	0.01222	0
586	T	0.09194	0.7084	0.07147	0.1282
587	G	0.98465	0.00002	0.01534	0
588	-	0.59162	0.00001	0.40837	0
589	-	0.37127	0.12165	0.49284	0.01424
590	-	0.86265	0.09858	0.02983	0.00894
591	-	0.87555	0.00002	0.12443	0
592	-	0.98043	0	0.01957	0
593	-	0.98194	0.00035	0.01769	0.00001
594	-	0.02006	0.84347	0.00904	0.12743
595	-	0.79956	0.00002	0.20042	0
596	-	0.89457	0.09064	0.00979	0.005
597	-	0.93404	0.00637	0.05906	0.00053
598	-	0.10397	0.00004	0.89599	0.00001
599	-	0.00207	0.85189	0.00159	0.14445
600	-	0.06525	0.79566	0.01042	0.12867
601	-	0.78899	0.07193	0.12731	0.01178
602	-	0.71883	0.14208	0.11585	0.02323

Appendix 13: PAML branch-site model output: the *manE* branch set as foreground

Codon site	Residue	Class site			
		0	1	2a	2b
1	-	0.76336	0.16514	0.05882	0.013
2	-	0.76336	0.16514	0.05882	0.013
3	-	0.76336	0.16514	0.05882	0.013
4	-	0.7893	0.1392	0.06081	0.011
5	-	0.61021	0.31829	0.04706	0.024
6	M	0.99322	0	0.00678	0.000
7	R	0.98848	0.00001	0.01151	0.000
8	-	0.97126	0.01378	0.01458	0.000
9	M	0.01661	0.95657	0.00037	0.026
10	L	0.6947	0.27894	0.01852	0.008
11	R	0.06827	0.84732	0.02008	0.064
12	L	0.06681	0.88988	0.00243	0.041
13	P	0.82507	0.14292	0.0252	0.007
14	A	0.68624	0.26335	0.03598	0.014
15	A	0.02951	0.93057	0.00078	0.039
16	L	0.95678	0.01972	0.02261	0.001
17	A	0.146	0.7782	0.02816	0.048
18	A	0.0015	0.91255	0.00066	0.085
19	L	0.35254	0.549	0.05988	0.039
20	A	0.00013	0.87868	0.00018	0.121
21	L	0.0016	0.81862	0.02917	0.151
22	A	0.01685	0.90325	0.00194	0.078
23	V	0.9691	0.01556	0.01488	0.000
24	S	0.01307	0.91543	0.00101	0.070
25	A	0.12544	0.80307	0.0097	0.062
26	C	0.92821	0.0003	0.07147	0.000
27	S	0.0314	0.89711	0.00243	0.069
28	G	0.78833	0.14017	0.06074	0.011
29	G	0.56151	0.367	0.04331	0.028
30	G	0.6532	0.2753	0.05036	0.021
31	-	0.84966	0.07885	0.06545	0.006
32	-	0.19027	0.73823	0.01471	0.057
33	-	0.79415	0.13436	0.06119	0.010
34	-	0.08548	0.84303	0.00661	0.065
35	-	0.23578	0.69272	0.01822	0.053

36	-	0.428	0.5005	0.03304	0.038
37	-	0.78717	0.14133	0.06065	0.011
38	-	0.24144	0.68706	0.01866	0.053
39	-	0.57862	0.34988	0.04463	0.027
40	-	0.84602	0.08248	0.06517	0.006
41	-	0.11265	0.81585	0.00871	0.063
42	-	0.65555	0.27296	0.05054	0.021
43	-	0.50084	0.42766	0.03865	0.033
44	-	0.83696	0.09154	0.06447	0.007
45	-	0.87454	0.05396	0.06736	0.004
46	-	0.51389	0.41462	0.03965	0.032
47	G	0.75285	0.17566	0.05801	0.013
48	N	0.44552	0.48298	0.03439	0.037
49	S	0.25153	0.67698	0.01944	0.052
50	D	0.00789	0.92062	0.00061	0.071
51	G	0.84084	0.1195	0.03244	0.007
52	G	0.02175	0.95357	0.00034	0.024
53	S	0.00131	0.87532	0.02225	0.101
54	G	0.9119	0.0166	0.07022	0.001
55	Q	0.86945	0.05905	0.06696	0.005
56	Y	0.98038	0.00036	0.01924	0.000
57	P	0.98521	0.00025	0.01453	0.000
58	R	0.98393	0	0.01607	0.000
59	N	0.95176	0.0317	0.01491	0.002
60	E	0.74333	0	0.25667	0.000
61	T	0.9839	0	0.0161	0.000
62	L	0.93999	0.00117	0.05877	0.000
63	Y	0.989	0	0.011	0.000
64	T	0.1046	0.78443	0.04504	0.066
65	T	0.98236	0.00123	0.01638	0.000
66	G	0.9831	0	0.0169	0.000
67	T	0.97063	0.00311	0.02611	0.000
68	A	0.84536	0.13088	0.01975	0.004
69	W	0.98247	0	0.01753	0.000
70	E	0.51175	0.45684	0.01538	0.016
71	A	0.98058	0.00001	0.01942	0.000
72	P	0.98109	0	0.0189	0.000
73	T	0.30798	0.66848	0.00493	0.019
74	S	0.99044	0	0.00956	0.000
75	W	0.98247	0	0.01753	0.000
76	N	0.98976	0	0.01024	0.000

77	P	0.98426	0	0.01574	0.000
78	M	0.98843	0.00257	0.00894	0.000
79	M	0.94648	0.03661	0.01591	0.001
80	R	0.71028	0.26465	0.01702	0.008
81	G	0.98174	0.00072	0.01752	0.000
82	Q	0.9678	0.02039	0.01084	0.001
83	F	0.95659	0.0264	0.01619	0.001
84	A	0.98413	0	0.01586	0.000
85	V	0.96984	0.00018	0.02998	0.000
86	G	0.98411	0	0.01589	0.000
87	T	0.98529	0	0.01471	0.000
88	N	0.19208	0.76519	0.00332	0.039
89	G	0.98375	0	0.01625	0.000
90	L	0.97996	0	0.02004	0.000
91	V	0.96823	0.00155	0.03014	0.000
92	Y	0.98343	0	0.01657	0.000
93	E	0.98965	0	0.01035	0.000
94	S	0.97106	0.00678	0.02189	0.000
95	L	0.97941	0	0.02059	0.000
96	F	0.9868	0	0.0132	0.000
97	H	0.8636	0.07489	0.0572	0.004
98	Y	0.98825	0	0.01175	0.000
99	D	0.9895	0	0.0105	0.000
100	A	0.98104	0	0.01896	0.000
101	D	0.85586	0.12001	0.02032	0.004
102	A	0.51282	0.44515	0.01615	0.026
103	G	0.91999	0.05922	0.01799	0.003
104	E	0.97709	0.00009	0.02281	0.000
105	Y	0.95772	0.03163	0.00969	0.001
106	V	0.77246	0.00364	0.22363	0.000
107	H	0.98113	0	0.01887	0.000
108	W	0.98247	0	0.01753	0.000
109	L	0.98152	0	0.01848	0.000
110	A	0.98465	0	0.01535	0.000
111	E	0.98877	0	0.01123	0.000
112	S	0.593	0.35793	0.03356	0.016
113	D	0.98384	0	0.01616	0.000
114	E	0.4534	0.49946	0.0213	0.026
115	W	0.98247	0	0.01753	0.000
116	T	0.53978	0.41672	0.01747	0.026
117	S	0.2722	0.55935	0.11552	0.053

118	E	0.89008	0.08246	0.02298	0.004
119	T	0.94781	0.00001	0.05218	0.000
120	E	0.92114	0.02593	0.05112	0.002
121	H	0.98571	0	0.01429	0.000
122	V	0.02044	0.92104	0.00077	0.058
123	I	0.97031	0.00312	0.02644	0.000
124	T	0.97358	0.00917	0.01677	0.000
125	L	0.97774	0	0.02226	0.000
126	R	0.98743	0	0.01257	0.000
127	E	0.97634	0.00434	0.01914	0.000
128	G	0.98379	0	0.01621	0.000
129	V	0.95824	0.01099	0.03024	0.001
130	T	0.65299	0.30187	0.02702	0.018
131	W	0.98247	0	0.01753	0.000
132	N	0.14096	0.83379	0.00314	0.022
133	D	0.98927	0	0.01073	0.000
134	G	0.97628	0	0.02371	0.000
135	E	0.96841	0.00071	0.03083	0.000
136	P	0.95878	0.02208	0.0186	0.001
137	F	0.82997	0.00007	0.16996	0.000
138	V	0.98469	0.00023	0.01507	0.000
139	A	0.19591	0.00108	0.80282	0.000
140	Q	0.83498	0.00001	0.16501	0.000
141	D	0.98895	0	0.01105	0.000
142	V	0.98424	0	0.01576	0.000
143	V	0.45297	0.00608	0.54005	0.001
144	T	0.98966	0	0.01034	0.000
145	T	0.98509	0	0.01491	0.000
146	L	0.99009	0.00001	0.0099	0.000
147	E	0.98872	0	0.01128	0.000
148	L	0.98713	0	0.01287	0.000
149	G	0.98594	0	0.01406	0.000
150	Q	0.98935	0	0.01065	0.000
151	-	0.97903	0.00002	0.02095	0.000
152	V	0.80692	0.17974	0.00871	0.005
153	P	0.52082	0.43224	0.0215	0.025
154	G	0.98307	0.00022	0.0167	0.000
155	V	0.9874	0.00002	0.01258	0.000
156	P	0.49105	0.14682	0.34668	0.015
157	Y	0.98968	0.00008	0.01025	0.000
158	S	0.98794	0.00004	0.01203	0.000

159	N	0.0096	0.95548	0.00027	0.035
160	V	0.92891	0.04723	0.02176	0.002
161	W	0.98247	0	0.01753	0.000
162	D	0.47902	0.47725	0.01765	0.026
163	Y	0.98244	0.00003	0.01753	0.000
164	I	0.98664	0.00001	0.01335	0.000
165	E	0.61488	0.35749	0.01136	0.016
166	S	0.00075	0.91628	0.00005	0.083
167	V	0.98335	0.00093	0.01569	0.000
168	E	0.98437	0	0.01563	0.000
169	A	0.61397	0.00637	0.37924	0.000
170	T	0.14298	0.32447	0.48599	0.047
171	D	0.9885	0	0.0115	0.000
172	E	0.00815	0.93804	0.00016	0.054
173	R	0.001	0.89954	0.00031	0.099
174	T	0.98421	0	0.01579	0.000
175	V	0.51516	0.00003	0.48481	0.000
176	T	0.26411	0.49989	0.19124	0.045
177	V	0.99064	0	0.00935	0.000
178	T	0.7365	0.19693	0.05371	0.013
179	F	0.99116	0	0.00884	0.000
180	S	0.84	0.03608	0.12238	0.002
181	E	0.98597	0.00144	0.01252	0.000
182	S	0.91979	0.06111	0.01761	0.001
183	R	0.68312	0.29372	0.01367	0.009
184	P	0.98471	0.00004	0.01524	0.000
185	Q	0.93665	0.00003	0.06332	0.000
186	E	0.98957	0.00005	0.01038	0.000
187	W	0.98247	0	0.01753	0.000
188	M	0.0002	0.93146	0.00004	0.068
189	N	0.65441	0.00229	0.34314	0.000
190	W	0.00147	0.94335	0.00006	0.055
191	A	0.98876	0	0.01124	0.000
192	Y	0.03002	0	0.96998	0.000
193	S	0.01202	0.90813	0.00461	0.075
194	N	0.01057	0.86033	0.01656	0.113
195	P	0.9695	0.00858	0.02167	0.000
196	I	0.98737	0	0.01263	0.000
197	V	0.98246	0.00001	0.01753	0.000
198	P	0.98157	0	0.01843	0.000
199	D	0.98804	0.00016	0.0118	0.000

200	H	0.99142	0	0.00858	0.000
201	I	0.98782	0	0.01218	0.000
202	W	0.98095	0.00151	0.0175	0.000
203	A	0.96374	0.0246	0.01058	0.001
204	G	0.00569	0.9447	0.00023	0.049
205	M	0.79004	0.18858	0.01378	0.008
206	E	0.26725	0.64503	0.0459	0.042
207	E	0.98789	0	0.0121	0.000
208	S	0.84572	0.01022	0.14338	0.001
209	Q	0.95768	0.00772	0.03408	0.001
210	V	0.9668	0.01078	0.02197	0.000
211	A	0.38734	0.5851	0.00853	0.019
212	D	0.91579	0.03076	0.05177	0.002
213	S	0.79538	0.17849	0.01759	0.009
214	P	0.98032	0	0.01968	0.000
215	N	0.98926	0	0.01074	0.000
216	E	0.98809	0	0.01191	0.000
217	N	0.97077	0.00864	0.0202	0.000
218	P	0.98126	0	0.01874	0.000
219	V	0.87538	0.00027	0.12433	0.000
220	G	0.98391	0	0.01609	0.000
221	T	0.97706	0.00162	0.02128	0.000
222	G	0.98524	0	0.01476	0.000
223	P	0.98108	0.00027	0.01864	0.000
224	Y	0.98477	0	0.01523	0.000
225	V	0.00095	0.81961	0.02558	0.154
226	Y	0.93114	0.00903	0.05928	0.001
227	E	0.64827	0.30619	0.03001	0.016
228	S	0.1325	0.67402	0.11383	0.080
229	H	0.97111	0	0.02889	0.000
230	T	0.43562	0.2265	0.31403	0.024
231	D	0.50852	0.02112	0.46879	0.002
232	D	0.98597	0	0.01403	0.000
233	R	0.98584	0	0.01416	0.000
234	M	0.06357	0.00001	0.93642	0.000
235	V	0.92932	0.00001	0.07067	0.000
236	W	0.98494	0.00018	0.01487	0.000
237	E	0.17299	0.74562	0.02712	0.054
238	R	0.54061	0.00014	0.45923	0.000
239	N	0.98149	0.00002	0.01849	0.000
240	D	0.96614	0.00043	0.0334	0.000

241	E	0.97367	0.00004	0.02629	0.000
242	W	0.98247	0	0.01753	0.000
243	W	0.98247	0	0.01753	0.000
244	A	0.98866	0	0.01134	0.000
245	I	0.97346	0.00005	0.02648	0.000
246	E	0.98002	0	0.01998	0.000
247	A	0.58894	0.23616	0.15836	0.017
248	L	0.88126	0.04959	0.06657	0.003
249	D	0.98483	0.00001	0.01516	0.000
250	M	0.01211	0.23674	0.70465	0.047
251	T	0.07478	0.28991	0.61084	0.024
252	M	0.98068	0.00163	0.01765	0.000
253	D	0.01117	0.83082	0.0684	0.090
254	A	0.98142	0.00009	0.01849	0.000
255	R	0.72815	0.00006	0.27179	0.000
256	Y	0.96313	0.00111	0.03572	0.000
257	I	0.97887	0.00003	0.0211	0.000
258	V	0.98344	0	0.01656	0.000
259	D	0.92004	0.00251	0.07728	0.000
260	I	0.97513	0.00014	0.02472	0.000
261	V	0.97579	0.01069	0.01328	0.000
262	N	0.9796	0.00001	0.02039	0.000
263	A	0.953	0.03132	0.01472	0.001
264	S	0.98842	0.00001	0.01157	0.000
265	N	0.9913	0	0.0087	0.000
266	E	0.98894	0	0.01106	0.000
267	V	0.98391	0	0.01609	0.000
268	T	0.98568	0	0.01432	0.000
269	M	0.54362	0.00019	0.45617	0.000
270	G	0.98437	0.00001	0.01562	0.000
271	M	0.99321	0	0.00678	0.000
272	L	0.96317	0.01478	0.02151	0.001
273	N	0.99269	0.00035	0.00695	0.000
274	Q	0.98987	0	0.01013	0.000
275	G	0.98382	0.00001	0.01618	0.000
276	E	0.98862	0.00036	0.01101	0.000
277	V	0.98324	0.00012	0.01664	0.000
278	D	0.98969	0	0.01031	0.000
279	L	0.2766	0.13232	0.57209	0.019
280	S	0.77454	0.21213	0.00911	0.004
281	N	0.99044	0	0.00956	0.000

282	N	0.9903	0	0.0097	0.000
283	F	0.9879	0	0.0121	0.000
284	L	0.97689	0	0.02311	0.000
285	P	0.98302	0	0.01698	0.000
286	G	0.98442	0	0.01558	0.000
287	I	0.90259	0.00053	0.09686	0.000
288	D	0.97707	0.0044	0.01842	0.000
289	Q	0.72595	0.22076	0.03992	0.013
290	V	0.98215	0.00004	0.01781	0.000
291	L	0.9232	0.00531	0.07109	0.000
292	N	0.98902	0.00003	0.01096	0.000
293	S	0.00031	0.95841	0	0.041
294	N	0.4339	0.5177	0.01331	0.035
295	E	0.9841	0.00009	0.01581	0.000
296	T	0.11594	0.83435	0.0046	0.045
297	-	0.76336	0.16514	0.05882	0.013
298	I	0.98655	0.00006	0.01338	0.000
299	T	0.00672	0.92566	0.00034	0.067
300	S	0.98472	0	0.01528	0.000
301	F	0.97864	0.0021	0.01921	0.000
302	Y	0.98848	0	0.01152	0.000
303	D	0.82895	0.00405	0.16673	0.000
304	G	0.68709	0.2782	0.02118	0.014
305	P	0.47514	0.06889	0.4493	0.007
306	P	0.97997	0	0.02003	0.000
307	Y	0.98907	0	0.01093	0.000
308	M	0.98743	0	0.01257	0.000
309	K	0.98083	0.00081	0.01834	0.000
310	S	0.36094	0.24391	0.37172	0.023
311	A	0.9789	0.00001	0.02108	0.000
312	N	0.9849	0	0.0151	0.000
313	T	0.97922	0	0.02078	0.000
314	A	0.96961	0.00003	0.03036	0.000
315	W	0.16287	0.79029	0.00617	0.041
316	L	0.54406	0.00001	0.45593	0.000
317	I	0.98354	0.0035	0.01288	0.000
318	P	0.97108	0.00804	0.02057	0.000
319	N	0.98957	0	0.01043	0.000
320	H	0.12637	0.68786	0.12709	0.059
321	T	0.72995	0.02157	0.24688	0.002
322	R	0.989	0	0.01099	0.000

323	E	0.92532	0.04738	0.02586	0.001
324	P	0.97992	0	0.02008	0.000
325	L	0.97629	0.00932	0.01415	0.000
326	N	0.98695	0.00004	0.01301	0.000
327	D	0.7894	0.00002	0.21058	0.000
328	T	0.08844	0.86391	0.01068	0.037
329	A	0.82994	0.14292	0.0202	0.007
330	F	0.99049	0	0.00951	0.000
331	R	0.98782	0	0.01218	0.000
332	Q	0.98662	0.00071	0.01267	0.000
333	A	0.98439	0	0.01561	0.000
334	L	0.91531	0.00002	0.08467	0.000
335	A	0.9837	0	0.0163	0.000
336	H	0.06187	0.90729	0.00128	0.030
337	S	0.98316	0.00003	0.01682	0.000
338	I	0.98677	0	0.01323	0.000
339	N	0.98469	0.00001	0.01531	0.000
340	I	0.96256	0.00246	0.0349	0.000
341	T	0.00215	0.95199	0.00006	0.046
342	Q	0.14886	0.71392	0.06372	0.074
343	I	0.98487	0	0.01513	0.000
344	V	0.98273	0	0.01727	0.000
345	E	0.01524	0.89637	0.00732	0.081
346	G	0.55391	0.43019	0.00713	0.009
347	P	0.62956	0.00165	0.3687	0.000
348	Y	0.98427	0	0.01573	0.000
349	A	0.98856	0.0002	0.01123	0.000
350	N	0.18293	0.68361	0.06869	0.065
351	L	0.32266	0.00677	0.66977	0.001
352	V	0.98503	0	0.01497	0.000
353	Q	0.00047	0.88829	0.00226	0.109
354	A	0.68031	0.25602	0.05097	0.013
355	A	0.98078	0.00007	0.01915	0.000
356	N	0.98663	0	0.01337	0.000
357	P	0.97943	0.00017	0.02039	0.000
358	T	0.47887	0.01757	0.50157	0.002
359	G	0.98517	0	0.01483	0.000
360	-	0.62353	0.00001	0.37646	0.000
361	-	0.9724	0.00022	0.02737	0.000
362	-	0.98354	0	0.01646	0.000
363	-	0.00858	0.93187	0.00028	0.059

364	-	0.36165	0.45033	0.15212	0.036
365	-	0.87819	0.05002	0.06985	0.002
366	-	0.97398	0.00266	0.02328	0.000
367	-	0.91695	0.072	0.00869	0.002
368	-	0.98994	0.00002	0.01004	0.000
369	-	0.98598	0	0.01402	0.000
370	-	0.97747	0.00606	0.01633	0.000
371	-	0.97521	0.00383	0.02085	0.000
372	-	0.01559	0.90508	0.00274	0.077
373	-	0.97155	0.00019	0.02825	0.000
374	-	0.98399	0	0.01601	0.000
375	-	0.97673	0.00002	0.02325	0.000
376	-	0.87783	0.10811	0.01014	0.004
377	-	0.98709	0.0006	0.01228	0.000
378	-	0.98203	0.00035	0.01761	0.000
379	-	0.98377	0	0.01622	0.000
380	-	0.99106	0	0.00894	0.000
381	-	0.33149	0.64123	0.00397	0.023
382	-	0.98563	0.00004	0.01433	0.000
383	-	0.98969	0	0.01031	0.000
384	-	0.97426	0.00684	0.01873	0.000
385	-	0.87458	0.10449	0.01577	0.005
386	-	0.00235	0.93784	0.00018	0.060
387	-	0.9845	0	0.0155	0.000
388	-	0.98889	0	0.01111	0.000
389	-	0.32858	0.60946	0.02211	0.040
390	-	0.98739	0.00048	0.01211	0.000
391	-	0.98322	0	0.01678	0.000
392	D	0.98849	0.00085	0.01063	0.000
393	D	0.97943	0.00015	0.02042	0.000
394	A	0.95296	0.00001	0.04703	0.000
395	G	0.98379	0	0.01621	0.000
396	-	0.98738	0	0.01262	0.000
397	-	0.9855	0.00006	0.01443	0.000
398	-	0.98959	0	0.01041	0.000
399	-	0.71208	0.26573	0.01147	0.011
400	-	0.98973	0.00001	0.01026	0.000
401	-	0.69016	0.2898	0.01162	0.008
402	-	0.98943	0	0.01057	0.000
403	-	0.98355	0	0.01645	0.000
404	-	0.98784	0.00004	0.01212	0.000

405	-	0.98684	0.00039	0.01276	0.000
406	-	0.98895	0	0.01105	0.000
407	-	0.46739	0.08036	0.44505	0.007
408	-	0.9818	0.00001	0.01819	0.000
409	-	0.75003	0.00001	0.24995	0.000
410	-	0.98423	0	0.01577	0.000
411	-	0.46326	0.50274	0.00979	0.024
412	P	0.0204	0.91739	0.00121	0.061
413	V	0.94688	0.00003	0.05309	0.000
414	R	0.93449	0.05419	0.00961	0.002
415	-	0.98681	0	0.01319	0.000
416	-	0.44267	0.52439	0.00929	0.024
417	-	0.98764	0.00002	0.01234	0.000
418	-	0.90997	0.01173	0.07749	0.001
419	-	0.96865	0.00012	0.03123	0.000
420	-	0.98397	0	0.01603	0.000
421	-	0.96358	0.02098	0.01495	0.000
422	-	0.98705	0	0.01295	0.000
423	-	0.98247	0	0.01753	0.000
424	-	0.98557	0	0.01443	0.000
425	-	0.99004	0	0.00996	0.000
426	-	0.98247	0	0.01753	0.000
427	-	0.99322	0	0.00678	0.000
428	-	0.33957	0.00001	0.66041	0.000
429	-	0.97147	0.00103	0.02747	0.000
430	-	0.98553	0.00048	0.01397	0.000
431	-	0.97955	0.00039	0.02005	0.000
432	-	0.05617	0.00071	0.94296	0.000
433	-	0.98715	0.00001	0.01284	0.000
434	-	0.976	0.00485	0.01897	0.000
435	-	0.84934	0.00088	0.14972	0.000
436	-	0.05909	0.63198	0.24436	0.065
437	-	0.97611	0.00744	0.01626	0.000
438	-	0.39062	0.50756	0.0728	0.029
439	-	0.02887	0.59516	0.30604	0.070
440	-	0.98505	0.00001	0.01494	0.000
441	-	0.98365	0	0.01635	0.000
442	-	0.98766	0	0.01234	0.000
443	-	0.98994	0	0.01005	0.000
444	-	0.64	0.0018	0.35808	0.000
445	-	0.98938	0.00012	0.01049	0.000

446	-	0.97562	0.00465	0.01961	0.000
447	-	0.77791	0.16522	0.04597	0.011
448	-	0.67166	0.00011	0.32822	0.000
449	-	0.98126	0	0.01874	0.000
450	-	0.98961	0	0.01039	0.000
451	-	0.96218	0.01514	0.02203	0.001
452	-	0.51335	0.46248	0.00922	0.015
453	-	0.98841	0.00004	0.01155	0.000
454	-	0.98744	0.00008	0.01248	0.000
455	-	0.00004	0.89358	0.00003	0.106
456	-	0.93478	0	0.06522	0.000
457	-	0.96095	0.0241	0.01382	0.001
458	-	0.97703	0.00012	0.02284	0.000
459	-	0.98732	0.00048	0.01218	0.000
460	-	0.00655	0.87135	0.00986	0.112
461	-	0.98366	0	0.01634	0.000
462	-	0.97621	0.00987	0.01355	0.000
463	-	0.98911	0	0.01089	0.000
464	-	0.98965	0	0.01035	0.000
465	-	0.95458	0.00044	0.04495	0.000
466	-	0.08342	0.00489	0.91103	0.001
467	-	0.63416	0.00593	0.35961	0.000
468	-	0.98995	0	0.01005	0.000
469	-	0.98404	0	0.01596	0.000
470	-	0.96971	0.01034	0.01959	0.000
471	-	0.0131	0.90891	0.00622	0.072
472	-	0.94171	0.03587	0.02096	0.001
473	-	0.80435	0.16291	0.0253	0.007
474	-	0.89614	0.01123	0.09191	0.001
475	-	0.12421	0.71519	0.10424	0.056
476	-	0.38605	0.56256	0.01545	0.036
477	-	0.98698	0	0.01302	0.000
478	-	0.02352	0.00037	0.97602	0.000
479	-	0.98247	0	0.01753	0.000
480	-	0.50371	0.00165	0.49449	0.000
481	-	0.98332	0.00089	0.01576	0.000
482	-	0.98653	0	0.01347	0.000
483	-	0.9896	0	0.0104	0.000
484	-	0.31436	0.52418	0.10443	0.057
485	-	0.76473	0.05305	0.17873	0.003
486	-	0.9903	0	0.0097	0.000

487	-	0.93506	0.04932	0.01427	0.001
488	-	0.0609	0.91035	0.00137	0.027
489	-	0.98473	0.00252	0.01263	0.000
490	-	0.8689	0.10516	0.02112	0.005
491	-	0.46455	0.41263	0.09267	0.030
492	-	0.95448	0.01681	0.0279	0.001
493	-	0.00136	0.95353	0.00006	0.045
494	-	0.98335	0.00037	0.01627	0.000
495	-	0.00398	0.95143	0.00012	0.044
496	-	0.03696	0.91625	0.00187	0.045
497	-	0.86735	0.12002	0.00953	0.003
498	-	0.33357	0.12687	0.52179	0.018
499	-	0.98359	0.00004	0.01637	0.000
500	-	0.99043	0	0.00957	0.000
501	-	0.98372	0	0.01628	0.000
502	G	0.98364	0	0.01636	0.000
503	R	0.75951	0.0008	0.23966	0.000
504	H	0.9865	0.00007	0.01342	0.000
505	Q	0.2298	0.66738	0.04753	0.055
506	R	0.98747	0.00165	0.01082	0.000
507	P	0.73178	0.24872	0.01303	0.006
508	G	0.95387	0.02347	0.02163	0.001
509	-	0.98264	0.00053	0.0168	0.000
510	-	0.0188	0.91536	0.00153	0.064
511	-	0.01817	0.89862	0.00706	0.076
512	-	0.84204	0.00037	0.15755	0.000
513	-	0.97321	0.00011	0.02668	0.000
514	-	0.41971	0.46044	0.09187	0.028
515	-	0.98907	0.00011	0.01082	0.000
516	-	0.98845	0	0.01154	0.000
517	-	0.99111	0	0.00889	0.000
518	-	0.88505	0.09164	0.02026	0.003
519	-	0.96995	0.01084	0.0188	0.000
520	-	0.83171	0.13663	0.02553	0.006
521	-	0.00249	0.93387	0.00025	0.063
522	-	0.94252	0.02057	0.03564	0.001
523	-	0.97643	0.00694	0.01635	0.000
524	-	0.00033	0.93416	0.00004	0.065
525	-	0.98003	0	0.01997	0.000
526	-	0.00139	0.93509	0.00016	0.063
527	-	0.48067	0.46164	0.02386	0.034

528	-	0.95637	0.02007	0.02258	0.001
529	-	0.20715	0.76119	0.00474	0.027
530	-	0.17738	0.74705	0.0257	0.050
531	-	0.10263	0.81342	0.00907	0.075
532	-	0.81004	0.16686	0.01588	0.007
533	-	0.93154	0.04883	0.01795	0.002
534	-	0.98465	0.0002	0.01514	0.000
535	-	0.10121	0.00336	0.89463	0.001
536	-	0.98913	0.00001	0.01086	0.000
537	-	0.98551	0	0.01449	0.000
538	-	0.84741	0.11293	0.03492	0.005
539	-	0.83534	0	0.16465	0.000
540	-	0.98329	0.00009	0.01662	0.000
541	-	0.98284	0.00001	0.01715	0.000
542	-	0.97058	0	0.02941	0.000
543	-	0.98272	0.00008	0.01719	0.000
544	-	0.66052	0.31184	0.01406	0.014
545	-	0.98665	0	0.01335	0.000
546	-	0.98	0	0.02	0.000
547	-	0.98846	0	0.01154	0.000
548	-	0.98247	0	0.01753	0.000
549	-	0.98693	0	0.01306	0.000
550	-	0.98001	0	0.01999	0.000
551	-	0.98174	0	0.01826	0.000
552	-	0.0357	0.00041	0.96383	0.000
553	-	0.98043	0.00012	0.01944	0.000
554	-	0.99087	0.00001	0.00912	0.000
555	-	0.98165	0	0.01835	0.000
556	-	0.74625	0.22986	0.01437	0.010
557	-	0.00027	0.88664	0.00029	0.113
558	-	0.62214	0.1187	0.25112	0.008
559	-	0.00025	0.93263	0.00001	0.067
560	-	0.98386	0.00012	0.01601	0.000
561	-	0.98247	0	0.01753	0.000
562	-	0.98707	0.00001	0.01293	0.000
563	-	0.99111	0	0.00889	0.000
564	-	0.98247	0	0.01753	0.000
565	-	0.9812	0	0.0188	0.000
566	-	0.81893	0.12749	0.04529	0.008
567	-	0.93641	0.05054	0.01117	0.002
568	-	0.83213	0.09638	0.0641	0.007

569	-	0.77723	0.15128	0.05989	0.012
570	N	0.10124	0.76742	0.07523	0.056
571	E	0.66981	0.30934	0.00904	0.012
572	S	0.98138	0.00016	0.01845	0.000
573	-	0.76336	0.16514	0.05882	0.013
574	Y	0.98224	0.00228	0.01543	0.000
575	I	0.86972	0.11249	0.01368	0.004
576	S	0.00015	0.95681	0.00001	0.043
577	T	0.98161	0	0.01839	0.000
578	T	0.0296	0.38382	0.45594	0.131
579	R	0.21044	0.72735	0.01167	0.051
580	S	0.98246	0.00001	0.01753	0.000
581	R	0.86341	0.10716	0.02311	0.006
582	K	0.98374	0.00052	0.01571	0.000
583	R	0.00045	0.96136	0.00002	0.038
584	T	0.51532	0.46546	0.00921	0.010
585	S	0.98742	0.00011	0.01247	0.000
586	T	0.0424	0.60812	0.27155	0.078
587	G	0.70823	0.00001	0.29176	0.000
588	-	0.98194	0.00001	0.01805	0.000
589	-	0.79524	0.14661	0.04962	0.009
590	-	0.86179	0.10056	0.03095	0.007
591	-	0.97196	0.00001	0.02802	0.000
592	-	0.98168	0	0.01831	0.000
593	-	0.97722	0.00034	0.02243	0.000
594	-	0.02394	0.92028	0.00117	0.055
595	-	0.89427	0.00002	0.10571	0.000
596	-	0.88714	0.09319	0.01583	0.004
597	-	0.59044	0.00787	0.40111	0.001
598	-	0.97796	0.00021	0.02182	0.000
599	-	0.00245	0.92605	0.00019	0.071
600	-	0.06216	0.86635	0.00481	0.067
601	-	0.84422	0.08428	0.06503	0.006
602	-	0.76336	0.16514	0.05882	0.013

References

1. **Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C**. 2009. The Transporter Classification Database: recent advances. Nucleic Acids Res. **37**:D274–278.
2. **Higgins CF**. 1992. ABC transporters: from microorganisms to man. Annu. Rev. Cell Biol. **8**:67–113.
3. **Saurin W, Dassa E**. 1994. Sequence relationships between integral inner membrane proteins of binding protein-dependent transport systems: evolution by recurrent gene duplications. Protein Sci. Publ. Protein Soc. **3**:325–344.
4. **Tomii K, Kanehisa M**. 1998. A comparative analysis of ABC transporters in complete microbial genomes. Genome Res. **8**:1048–1059.
5. **Schneider E**. 2001. ABC transporters catalyzing carbohydrate uptake. Res. Microbiol. **152**:303–310.
6. **Gage DJ, Long SR**. 1998. Alpha-galactoside uptake in *Rhizobium meliloti*: isolation and characterization of agpA, a gene encoding a periplasmic binding protein required for melibiose and raffinose utilization. J. Bacteriol. **180**:5739–5748.
7. **Koning SM, Elferink MGL, Konings WN, Driessen AJM**. 2001. Cellobiose uptake in the hyperthermophilic archaeon *Pyrococcus furiosus* is mediated by an inducible, high-affinity ABC transporter. J. Bacteriol. **183**:4979–4984.
8. **Nanavati DM, Thirangoon K, Noll KM**. 2006. Several archaeal homologs of putative oligopeptide-binding proteins encoded by *Thermotoga maritima* bind sugars. Appl. Environ. Microbiol. **72**:1336–1345.
9. **Davidson AL, Dassa E, Orelle C, Chen J**. 2008. Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol. Mol. Biol. Rev. MMBR **72**:317–364.
10. **Isenbarger TA, Carr CE, Johnson SS, Finney M, Church GM, Gilbert W, Zuber MT, Ruvkun G**. 2008. The most conserved genome segments for life detection on Earth and other planets. Orig. Life Evol. Biosphere J. Int. Soc. Study Orig. Life **38**:517–533.
11. **Kuan G, Dassa E, Saurin W, Hofnung M, Saier MH**. 1995. Phylogenetic analyses of the ATP-binding constituents of bacterial extracytoplasmic receptor-dependent ABC-type nutrient uptake permeases. Res. Microbiol. **146**:271–278.
12. **Tam R, Saier MH**. 1993. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol. Rev. **57**:320–346.
13. **Imamura H, Jeon B-S, Wakagi T**. 2004. Molecular evolution of the ATPase subunit of three archaeal sugar ABC transporters. Biochem. Biophys. Res. Commun. **319**:230–234.
14. **Lapierre P, Butzin NC, Noll KM**. 2013. Application of a new mapping algorithm to reevaluate evidence of interdomain lateral gene transfer in the genome of *Thermotoga maritima* lateral gene transfer in evolution. Springer Science.
15. **Noll KM, Thirangoon K**. 2009. Interdomain transfers of sugar transporters overcome barriers to gene expression. Methods Mol. Biol. Clifton NJ **532**:309–322.

16. **Noll KM, Lapierre P, Gogarten JP, Nanavati DM**. 2008. Evolution of mal ABC transporter operons in the Thermococcales and Thermotogales. BMC Evol. Biol. **8**:7.
17. **Quiocho FA**. 1990. Atomic structures of periplasmic binding proteins and the high-affinity active transport systems in bacteria. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **326**:341–351; discussion 351–352.
18. **Vyas NK, Vyas MN, Quiocho FA**. 1991. Comparison of the periplasmic receptors for L-arabinose, D-glucose/D-galactose, and D-ribose. J. Biol. Chem. **266**:5226–5237.
19. **Sharff AJ, Rodseth LE, Spurlino JC, Quiocho FA**. 1992. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. Biochemistry (Mosc.) **31**:10657–10663.
20. **Fukami-Kobayashi K, Tateno Y, Nishikawa K**. 1999. Domain dislocation: A change of core structure in periplasmic binding proteins in their evolutionary history. J. Mol. Biol. **286**:279–290.
21. **Quiocho FA, Ledvina PS**. 1996. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. Mol. Microbiol. **20**:17–25.
22. **Conte LL, Brenner SE, Hubbard TJP, Chothia C, Murzin AG**. 2002. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. **30**:264–267.
23. **Murzin AG, Brenner SE, Hubbard T, Chothia C**. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**:536–540.
24. **Nesbo CL, L'Haridon S, Stetter KO, Doolittle WF**. 2001. Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. **18**:362–375.
25. **Huber R, Langworthy TA, König H, Thomm M, Woese CR, Sleytr UB, Stetter KO**. 1986. *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. Arch. Microbiol. **144**:324–333.
26. **Patel BKC, Morgan HW, Daniel RM**. 1985. *Fervidobacterium nodosum* gen. nov. and spec. nov., a new chemoorganotrophic, caldoactive, anaerobic bacterium. Arch. Microbiol. **141**:63–69.
27. **Jannasch HW, Huber R, Belkin S, Stetter KO**. 1988. *Thermotoga neapolitana* sp. nov. of the extremely thermophilic, eubacterial genus *Thermotoga*. Arch. Microbiol. **150**:103–104.
28. **Nesbø CL, Bradnan DM, Adebusuyi A, Dlutek M, Petrus AK, Foght J, Doolittle WF, Noll KM**. 2012. *Mesotoga prima* gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. Extremophiles **16**:387–393.
29. **Dipippo JL, Nesbø CL, Dahle H, Doolittle WF, Birkland N-K, Noll KM**. 2009. *Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. Int. J. Syst. Evol. Microbiol. **59**:2991–3000.
30. **Takahata Y, Nishijima M, Hoaki T, Maruyama T**. 2001. *Thermotoga petrophila* sp. nov. and *Thermotoga naphthophila* sp. nov., two hyperthermophilic

bacteria from the Kubiki oil reservoir in Niigata, Japan. Int. J. Syst. Evol. Microbiol. **51**:1901–1909.

31. **Balk M, Weijma J, Stams AJM**. 2002. *Thermotoga lettingae* sp. nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. Int. J. Syst. Evol. Microbiol. **52**:1361–1368.
32. **Nesbø CL, Dlutek M, Zhaxybayeva O, Doolittle WF**. 2006. Evidence for existence of "*mesotogas*," members of the order Thermotogales adapted to low-temperature environments. Appl. Environ. Microbiol. **72**:5061–5068.
33. **Ben Hania W, Postec A, Aüllo T, Ranchou-Peyruse A, Erauso G, Brochier-Armanet C, Hamdi M, Ollivier B, Saint-Laurent S, Magot M, Fardeau M-L**. 2013. *Mesotoga infera* sp. nov., a mesophilic member of the order Thermotogales, isolated from an underground gas storage aquifer. Int. J. Syst. Evol. Microbiol. **63**:3003–3008.
34. **Liebl W, Winterhalter C, Baumeister W, Armbrecht M, Valdez M**. 2008. Xylanase attachment to the cell wall of the hyperthermophilic bacterium *Thermotoga maritima*. J. Bacteriol. **190**:1350–1358.
35. **Schumann J, Wrba A, Jaenicke R, Stetter KO**. 1991. Topographical and enzymatic characterization of amylases from the extremely thermophilic eubacterium *Thermotoga maritima*. FEBS Lett. **282**:122–126.
36. **Petrus AK, Swithers KS, Ranjit C, Wu S, Brewer HM, Gogarten JP, Pasa-Tolic L, Noll KM**. 2012. Genes for the major structural components of Thermotogales species' togas revealed by proteomic and evolutionary analyses of OmpA and OmpB homologs. PLoS ONE **7**.
37. **Engel AM, Cejka Z, Lupas A, Lottspeich F, Baumeister W**. 1992. Isolation and cloning of Omp alpha, a coiled-coil protein spanning the periplasmic space of the ancestral eubacterium *Thermotoga maritima*. EMBO J. **11**:4369–4378.
38. **Lupas A, Müller S, Goldie K, Engel AM, Engel A, Baumeister W**. 1995. Model structure of the Omp alpha rod, a parallel four-stranded coiled coil from the hyperthermophilic eubacterium *Thermotoga maritima*. J. Mol. Biol. **248**:180–189.
39. **Rachel R, Engel AM, Huber R, Stetter K-O, Baumeister W**. 1990. A porin-type protein is the main constituent of the cell envelope of the ancestral eubacterium *Thermotoga maritima*. FEBS Lett. **262**:64–68.
40. **Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM**. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature **399**:323–329.
41. **Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbø CL, Doolittle WF, Gogarten JP, Noll KM**. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc. Natl. Acad. Sci. **106**:5865–5870.
42. **Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr**. 2000. Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J. Mol. Biol. **301**:75–100.

43. **Conners SB, Montero CI, Comfort DA, Shockley KR, Johnson MR, Chhabra SR, Kelly RM**. 2005. An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium *Thermotoga maritima*. J. Bacteriol. **187**:7267–7282.
44. **Ledl F, Beck J, Sengl M, Osiander H, Estendorfer S, Severin T, Huber B**. 1989. Chemical pathways of the Maillard reaction. Prog. Clin. Biol. Res. **304**:23–42.
45. **Driskill LE, Kusy K, Bauer MW, Kelly RM**. 1999. Relationship between glycosyl hydrolase inventory and growth physiology of the hyperthermophile *Pyrococcus furiosus* on carbohydrate-based media. Appl. Environ. Microbiol. **65**:893–897.
46. **Kim KW, Lee SB**. 2003. Inhibitory effect of Maillard reaction products on growth of the aerobic marine hyperthermophilic archaeon *Aeropyrum pernix*. Appl. Environ. Microbiol. **69**:4325–4328.
47. **Schröder C, Selig M, Schönheit P**. 1994. Glucose fermentation to acetate, CO2 and H2 in the anaerobic hyperthermophilic eubacterium *Thermotoga maritima*: involvement of the Embden-Meyerhof pathway. Arch. Microbiol. **161**:460–470.
48. **Selig M, Xavier KB, Santos H, Schönheit P**. 1997. Comparative analysis of Embden-Meyerhof and Entner-Doudoroff glycolytic pathways in hyperthermophilic archaea and the bacterium *Thermotoga*. Arch. Microbiol. **167**:217–232.
49. **Swithers KS, DiPippo JL, Bruce DC, Detter C, Tapia R, Han S, Saunders E, Goodwin LA, Han J, Woyke T, Pitluck S, Pennacchio L, Nolan M, Mikhailova N, Lykidis A, Land ML, Brettin T, Stetter KO, Nelson KE, Gogarten JP, Noll KM**. 2011. Genome sequence of *Thermotoga* sp. strain RQ2, a hyperthermophilic bacterium isolated from a geothermally heated region of the seafloor near Ribeira Quente, the Azores. J. Bacteriol. **193**:5869–5870.
50. **Kazanov MD, Li X, Gelfand MS, Osterman AL, Rodionov DA**. 2013. Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the *Thermotogae* phylum. Nucleic Acids Res. **41**:790–803.
51. **Rodionova IA, Yang C, Li X, Kurnasov OV, Best AA, Osterman AL, Rodionov DA**. 2012. Diversity and versatility of the *Thermotoga maritima* sugar kinome. J. Bacteriol. **194**:5552–5563.
52. **Lee S-J, Engelmann A, Horlacher R, Qu Q, Vierke G, Hebbeln C, Thomm M, Boos W**. 2003. TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. J. Biol. Chem. **278**:983–990.
53. **Lee S-J, Moulakakis C, Koning SM, Hausner W, Thomm M, Boos W**. 2005. TrmB, a sugar sensing regulator of ABC transporter genes in *Pyrococcus furiosus* exhibits dual promoter specificity and is controlled by different inducers. Mol. Microbiol. **57**:1797–1807.
54. **Lee S-J, Surma M, Hausner W, Thomm M, Boos W**. 2008. The role of TrmB and TrmB-like transcriptional regulators for sugar transport and metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*. Arch. Microbiol. **190**:247–256.

55. **Sutcliffe IC, Harrington DJ**. 2002. Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. Microbiology **148**:2065–2077.
56. **Albers S-V, Koning SM, Konings WN, Driessen AJ**. 2004. Insights into ABC transport in Archaea. J. Bioenerg. Biomembr. **36**:5–15.
57. **Lee S-J, Böhm A, Krug M, Boos W**. 2007. The ABC of binding-protein-dependent transport in Archaea. Trends Microbiol. **15**:389–397.
58. **Albers S-V, Konings WN, Driessen AJM**. 1999. A unique short signal sequence in membrane-anchored proteins of Archaea. Mol. Microbiol. **31**:1595–1596.
59. **Zolghadr B, Klingl A, Rachel R, Driessen AJM, Albers S-V**. 2011. The bindosome is a structural component of the *Sulfolobus solfataricus* cell envelope. Extremophiles **15**:235–244.
60. **Li Y-D, Xie Z-Y, Du Y-L, Zhou Z, Mao X-M, Lv L-X, Li Y-Q**. 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. Gene **436**:8–11.
61. **Chen J, Lu G, Lin J, Davidson AL, Quiocho FA**. 2003. A Tweezers-like motion of the ATP-Binding Cassette dimer in an ABC transport cycle. Mol. Cell **12**:651–661.
62. **Khare D, Oldham ML, Orelle C, Davidson AL, Chen J**. 2009. Alternating access in maltose transporter mediated by rigid-body rotations. Mol. Cell **33**:528–536.
63. **Lu G, Westbrooks JM, Davidson AL, Chen J**. 2005. ATP hydrolysis is required to reset the ATP-binding cassette dimer into the resting-state conformation. Proc. Natl. Acad. Sci. U. S. A. **102**:17969–17974.
64. **Oldham ML, Khare D, Quiocho FA, Davidson AL, Chen J**. 2007. Crystal structure of a catalytic intermediate of the maltose transporter. Nature **450**:515–521.
65. **Oldham ML, Chen S, Chen J**. 2013. Structural basis for substrate specificity in the *Escherichia coli* maltose transport system. Proc. Natl. Acad. Sci. **110**:18132–18137.
66. **Abrahams JP, Leslie AGW, Lutter R, Walker JE**. 1994. Structure at 2.8 Â resolution of F1-ATPase from bovine heart mitochondria. Nature **370**:621–628.
67. **Walker JE, Saraste M, Runswick MJ, Gay NJ**. 1982. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1**:945–951.
68. **Fetsch EE, Davidson AL**. 2002. Vanadate-catalyzed photocleavage of the signature motif of an ATP-binding cassette (ABC) transporter. Proc. Natl. Acad. Sci. U. S. A. **99**:9685–9690.
69. **Manavalan P, Dearborn DG, McPherson JM, Smith AE**. 1995. Sequence homologies between nucleotide binding regions of CFTR and G-proteins suggest structural and functional similarities. FEBS Lett. **366**:87–91.
70. **Buchaklian AH, Klug CS**. 2006. Characterization of the LSGGQ and H motifs from the *Escherichia coli* lipid A transporter MsbA. Biochemistry (Mosc.) **45**:12539–12546.

71. **Schmees G, Stein A, Hunke S, Landmesser H, Schneider E**. 1999. Functional consequences of mutations in the conserved "signature sequence" of the ATP-binding-cassette protein MalK. Eur. J. Biochem. FEBS **266**:420–430.
72. **Szakács G, Ozvegy C, Bakos E, Sarkadi B, Váradi A**. 2001. Role of glycine-534 and glycine-1179 of human multidrug resistance protein (MDR1) in drug-mediated control of ATP hydrolysis. Biochem. J. **356**:71–75.
73. **George AM, Jones PM**. 2012. Perspectives on the structure-function of ABC transporters: the switch and constant contact models. Prog. Biophys. Mol. Biol. **109**:95–107.
74. **Mao B, Pear MR, McCammon JA, Quiocho FA**. 1982. Hinge-bending in L-arabinose-binding protein. The "Venus's-flytrap" model. J. Biol. Chem. **257**:1131–1133.
75. **Jardetzky O**. 1966. Simple allosteric model for membrane pumps. Nature **211**:969–970.
76. **Hollenstein K, Frei DC, Locher KP**. 2007. Structure of an ABC transporter in complex with its binding protein. Nature **446**:213–216.
77. **Orelle C, Ayvaz T, Everly RM, Klug CS, Davidson AL**. 2008. Both maltose-binding protein and ATP are required for nucleotide-binding domain closure in the intact maltose ABC transporter. Proc. Natl. Acad. Sci. U. S. A. **105**:12837–12842.
78. **Orelle C, Alvarez FJD, Oldham ML, Orelle A, Wiley TE, Chen J, Davidson AL**. 2010. Dynamics of α-helical subdomain rotation in the intact maltose ATP-binding cassette transporter. Proc. Natl. Acad. Sci. **107**:20293–20298.
79. **Rees DC, Johnson E, Lewinson O**. 2009. ABC transporters: the power to change. Nat. Rev. Mol. Cell Biol. **10**:218–227.
80. **Shilton BH**. 2008. The dynamics of the MBP–MalFGK2 interaction: A prototype for binding protein dependent ABC-transporter systems. Biochim. Biophys. Acta BBA - Biomembr. **1778**:1772–1780.
81. **Boucher N, Noll KM**. 2011. Ligands of thermophilic ABC transporters encoded in a newly sequenced genomic region of *Thermotoga maritima* MSB8 screened by differential scanning fluorimetry. Appl. Environ. Microbiol. **77**:6395–6399.
82. **Rodionova IA, Leyn SA, Burkart MD, Boucher N, Noll KM, Osterman AL, Rodionov DA**. 2013. Novel inositol catabolic pathway in *Thermotoga maritima*. Environ. Microbiol. **15**:2254–2266.
83. **Koning SM, Konings WN, Driessen AJM**. 2002. Biochemical evidence for the presence of two alpha-glucoside ABC-transport systems in the hyperthermophilic archaeon *Pyrococcus furiosus*. Archaea Vanc. BC **1**:19–25.
84. **Xavier KB, Martins LO, Peist R, Kossmann M, Boos W, Santos H**. 1996. High-affinity maltose/trehalose transport system in the hyperthermophilic archaeon *Thermococcus litoralis*. J. Bacteriol. **178**:4773–4777.
85. **Liebl W, Gabelsberger J, Schleifer KH**. 1994. Comparative amino acid sequence analysis of *Thermotoga maritima* beta-glucosidase (BglA) deduced from the nucleotide sequence of the gene indicates distant relationship between beta-glucosidases of the BGA family and other families of beta-1,4-glycosyl hydrolases. Mol. Gen. Genet. MGG **242**:111–115.
86. **Rozen S, Skaletsky H**. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. Clifton NJ **132**:365–386.

87. **Lukashin AV, Borodovsky M**. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26**:1107–1115.
88. **Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L**. 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res. **31**:28–33.
89. **Abreu-Goodger C, Merino E**. 2005. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res. **33**:W690–692.
90. **Krogh A, Larsson B, von Heijne G, Sonnhammer EL**. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**:567–580.
91. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.
92. **Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG**. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.
93. **Nanavati DM, Nguyen TN, Noll KM**. 2005. Substrate specificities and expression patterns reflect the evolutionary divergence of maltose ABC transporters in *Thermotoga maritima*. J. Bacteriol. **187**:2002–2009.
94. **Wassenberg D, Liebl W, Jaenicke R**. 2000. Maltose-binding protein from the hyperthermophilic bacterium *Thermotoga maritima*: stability and binding properties. J. Mol. Biol. **295**:279–288.
95. **Cuneo MJ, Changela A, Warren JJ, Beese LS, Hellinga HW**. 2006. The crystal structure of a thermophilic glucose binding protein reveals adaptations that interconvert mono and di-saccharide binding sites. J. Mol. Biol. **362**:259–270.
96. **Rossello R, Garcia-Valdes E, Lalucat J, Ursing J**. 1991. Genotypic and phenotypic diversity of *Pseudomonas stutzeri*. Syst. Appl. Microbiol. **14**:150–157.
97. **Ursing J, Aleksić S**. 1995. *Yersinia frederiksenii*, a genotypically heterogeneous species with few differential characteristics. Contrib. Microbiol. Immunol. **13**:112–116.
98. **Vandamme P, Mahenthiralingam E, Holmes B, Coenye T, Hoste B, De Vos P, Henry D, Speert DP**. 2000. Identification and population structure of *Burkholderia stabilis* sp. nov. (formerly *Burkholderia cepacia* genomovar IV). J. Clin. Microbiol. **38**:1042–1047.
99. **Latif H, Lerman JA, Portnoy VA, Tarasova Y, Nagarajan H, Schrimpe-Rutledge AC, Smith RD, Adkins JN, Lee D-H, Qiu Y, Zengler K**. 2013. The genome organization of *Thermotoga maritima* reflects its lifestyle. PLoS Genet **9**:e1003485.
100. **Van Dijk EL, Jaszczyszyn Y, Thermes C**. 2014. Library preparation methods for next-generation sequencing: tone down the bias. Exp. Cell Res. **322**:12–20.
101. **Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, Polozov RV, Nechipurenko YD, Grokhovsky SL**. 2014. Non-random DNA fragmentation in next-generation sequencing. Sci. Rep. **4**.
102. **Taub MA, Bravo HC, Irizarry RA**. 2010. Overcoming bias and systematic errors in next generation sequencing data. Genome Med. **2**:87.

103. **Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT**. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE **7**:e30087.
104. **Minoche AE, Dohm JC, Himmelbauer H**. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. **12**:R112.
105. **Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB**. 2013. Characterizing and measuring bias in sequence data. Genome Biol. **14**:R51.
106. **Koshland DE**. 1958. Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. U. S. A. **44**:98–104.
107. **Linderstrøm–Lang K, Schellman J**. 1959. Protein structure and enzymatic activity, p. 443–510. *In* The Enzymes, 2nd ed. New York: Academic Press.
108. **Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P, Salemme FR**. 2001. High-density miniaturized thermal shift assays as a general strategy for drug discovery. J. Biomol. Screen. **6**:429–440.
109. **Niesen FH, Berglund H, Vedadi M**. 2007. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. **2**:2212–2221.
110. **Simpson RJ**. 2010. SYPRO orange fluorescent staining of protein gels. Cold Spring Harb. Protoc. **2010**:pdb.prot5414.
111. **Lo M-C, Aulabaugh A, Jin G, Cowling R, Bard J, Malamas M, Ellestad G**. 2004. Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. Anal. Biochem. **332**:153–159.
112. **Matulis D, Kranz JK, Salemme FR, Todd MJ**. 2005. Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using ThermoFluor. Biochemistry (Mosc.) **44**:5258–5266.
113. **Goto Y, Calciano LJ, Fink AL**. 1990. Acid-induced folding of proteins. Proc. Natl. Acad. Sci. U. S. A. **87**:573–577.
114. **Ames GFL**. 1986. Bacterial periplasmic transport systems: structure, mechanism, and evolution. Annu. Rev. Biochem. **55**:397–425.
115. **Giuliani SE, Frank AM, Corgliano DM, Seifert C, Hauser L, Collart FR**. 2011. Environment sensing and response mediated by ABC transporters. BMC Genomics **12**:S8.
116. **Giuliani SE, Frank AM, Collart FR**. 2008. Functional assignment of solute-binding proteins of ABC transporters using a fluorescence-based thermal shift assay. Biochemistry (Mosc.) **47**:13974–13984.
117. **Cuneo MJ, Beese LS, Hellinga HW**. 2008. Ligand-induced conformational changes in a thermophilic ribose-binding protein. BMC Struct. Biol. **8**:50.
118. **Cuneo MJ, Beese LS, Hellinga HW**. 2009. Structural analysis of semi-specific oligosaccharide recognition by a cellulose-binding protein of *Thermotoga maritima* reveals adaptations for functional diversification of the oligopeptide periplasmic binding protein fold. J. Biol. Chem. **284**:33217–33223.

119. **Dolgikh DA, Gilmanshin RI, Brazhnikov EV, Bychkova VE, Semisotnov GV, Venyaminov SYu, Ptitsyn OB**. 1981. Alpha-lactalbumin: compact state with fluctuating tertiary structure? FEBS Lett. **136**:311–315.
120. **Dolgikh DA, Kolomiets AP, Bolotina IA, Ptitsyn OB**. 1984. "Molten-globule" state accumulates in carbonic anhydrase folding. FEBS Lett. **165**:88–92.
121. **Ohgushi M, Wada A**. 1983. "Molten-globule state": a compact form of globular proteins with mobile side-chains. FEBS Lett. **164**:21–24.
122. **Ganesh C, Shah AN, Swaminathan CP, Surolia A, Varadarajan R**. 1997. Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein. Biochemistry (Mosc.) **36**:5020–5028.
123. **Novokhatny V, Ingham K**. 1997. Thermodynamics of maltose binding protein unfolding. Protein Sci. Publ. Protein Soc. **6**:141–146.
124. **Lee JC, Timasheff SN**. 1981. The stabilization of proteins by sucrose. J. Biol. Chem. **256**:7193–7201.
125. **Simpson RJ**. 2010. Stabilization of proteins for storage. Cold Spring Harb. Protoc. **2010**:pdb.top79.
126. **Higashiyama T**. 2002. Novel functions and applications of trehalose. Pure Appl. Chem. **74**:1263–1269.
127. **Butzin NC, Secinaro MA, Swithers KS, Gogarten JP, Noll KM**. 2013. *Thermotoga lettingae* can salvage cobinamide to synthesize vitamin B12. Appl. Environ. Microbiol. **79**:7006–7012.
128. **Butzin NC, Lapierre P, Green AG, Swithers KS, Gogarten JP, Noll KM**. 2013. Reconstructed ancestral myo-inositol-3-phosphate synthases Indicate that ancestors of the *Thermococcales* and *Thermotoga* species were more thermophilic than their descendants. PLoS ONE **8**:e84300.
129. **Fiala G, Stetter KO**. 1986. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. Arch. Microbiol. **145**:56–61.
130. **Montelione GT**. 2012. The Protein Structure Initiative: achievements and visions for the future. F1000 Biol. Rep. **4**.
131. **Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C**. 2009. PSI-2: structural genomics to cover protein domain family space. Struct. Lond. Engl. 1993 **17**:869–881.
132. **Tian Y, Cuneo MJ, Changela A, Hocker B, Beese LS, Hellinga HW**. 2007. Structure-based design of robust glucose biosensors using a *Thermotoga maritima* periplasmic glucose-binding protein. Protein Sci. Publ. Protein Soc. **16**:2240–2250.
133. **Bendtsen JD, Nielsen H, von Heijne G, Brunak S**. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340**:783–795.
134. **Horlacher R, Xavier KB, Santos H, DiRuggiero J, Kossmann M, Boos W**. 1998. Archaeal binding protein-dependent ABC transporter: molecular and biochemical analysis of the trehalose/maltose transport system of the hyperthermophilic archaeon *Thermococcus litoralis*. J. Bacteriol. **180**:680–689.
135. **Goldstein A, Barrett RW**. 1987. Ligand dissociation constants from competition binding assays: errors associated with ligand depletion. Mol. Pharmacol. **31**:603–609.

136. **Swillens S**. 1995. Interpretation of binding curves obtained with high receptor concentrations: practical aid for computer analysis. Mol. Pharmacol. **47**:1197–1203.
137. **Edgar RC**. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.
138. **Edgar RC**. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.
139. **Stamatakis A**. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinforma. Oxf. Engl. **22**:2688–2690.
140. **Saier MH Jr**. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. MMBR **64**:354–411.
141. **Diruggiero J, Dunn D, Maeder DL, Holley-Shanks R, Chatard J, Horlacher R, Robb FT, Boos W, Weiss RB**. 2000. Evidence of recent lateral gene transfer among hyperthermophilic archaea. Mol. Microbiol. **38**:684–693.
142. **Silva Z, Sampaio M-M, Henne A, Böhm A, Gutzat R, Boos W, da Costa MS, Santos H**. 2005. The high-affinity maltose/trehalose ABC transporter in the extremely thermophilic bacterium *Thermus thermophilus* HB27 also recognizes sucrose and palatinose. J. Bacteriol. **187**:1210–1218.
143. **Herman P, Staiano M, Marabotti A, Varriale A, Scirè A, Tanfani F, Vecer J, Rossi M, D'Auria S**. 2006. D-trehalose/D-maltose-binding protein from the hyperthermophilic archaeon *Thermococcus litoralis*: the binding of trehalose and maltose results in different protein conformational states. Proteins **63**:754–767.
144. **Atkins WM, Stayton PS, Villafranca JJ**. 1991. Time-resolved fluorescence studies of genetically engineered *Escherichia coli* glutamine synthetase. Biochemistry (Mosc.) **30**:3406–3416.
145. **Loewenthal R, Sancho J, Fersht AR**. 1991. Fluorescence spectrum of barnase: contributions of three tryptophan residues and a histidine-related pH dependence. Biochemistry (Mosc.) **30**:6775–6779.
146. **Ryu S-I, Kim J-E, Huong NT, Woo E-J, Moon S-K, Lee S-B**. 2010. Molecular cloning and characterization of trehalose synthase from *Thermotoga maritima* DSM3109: syntheses of trehalose disaccharide analogues and NDP-glucoses. Enzyme Microb. Technol. **47**:249–256.
147. **Rodionov DA, Rodionova IA, Ravcheev DA, Tarasova Y, Zengler K**. 2013. Transcriptional regulation of the carbohydrate utilization network in *Thermotoga maritima*. Front. Microb. Physiol. Metab. **4**:244.
148. **Liebl W, Wagner B, Schellhase J**. 1998. Properties of an alpha-galactosidase, and structure of its gene galA, within an alpha-and beta-galactoside utilization gene cluster of the hyperthermophilic bacterium *Thermotoga maritima*. Syst. Appl. Microbiol. **21**:1–11.
149. **Comfort DA, Bobrov KS, Ivanen DR, Shabalin KA, Harris JM, Kulminskaya AA, Brumer H, Kelly RM**. 2007. Biochemical analysis of *Thermotoga maritima* GH36 alpha-galactosidase (TmGalA) confirms the mechanistic commonality of clan GH-D glycoside hydrolases. Biochemistry (Mosc.) **46**:3319–3330.

150. **Kim C s., Ji E-S, Oh D-K**. 2004. Characterization of a thermostable recombinant β-galactosidase from *Thermotoga maritima*. J. Appl. Microbiol. **97**:1006–1014.
151. **Yang H, Ichinose H, Yoshida M, Nakajima M, Kobayashi H, Kaneko S**. 2006. Characterization of a thermostable endo-beta-1,4-D-galactanase from the hyperthermophile *Thermotoga maritima*. Biosci. Biotechnol. Biochem. **70**:538–541.
152. **Hansen T, Schlichting B, Schönheit P**. 2002. Glucose-6-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: expression of the g6pd gene and characterization of an extremely thermophilic enzyme. FEMS Microbiol. Lett. **216**:249–253.
153. **McCarthy JK, O'Brien CE, Eveleigh DE**. 2003. Thermostable continuous coupled assay for measuring glucose using glucokinase and glucose-6-phosphate dehydrogenase from the marine hyperthermophile *Thermotoga maritima*. Anal. Biochem. **318**:196–203.
154. **Childers SE, Noll KM**. 1994. Characterization and regulation of sulfur reductase activity in *Thermotoga neapolitana*. Appl. Environ. Microbiol. **60**:2622–2626.
155. **Nanavati D, Noll KM, Romano AH**. 2002. Periplasmic maltose- and glucose-binding protein activities in cell-free extracts of *Thermotoga maritima*. Microbiology **148**:3531–3537.
156. **Rodionova IA, Scott DA, Grishin NV, Osterman AL, Rodionov DA**. 2012. Tagaturonate–fructuronate epimerase UxaE, a novel enzyme in the hexuronate catabolic network in *Thermotoga maritima*. Environ. Microbiol. **14**:2920–2934.
157. **Rodionov DA, Rodionova IA, Li X, Ravcheev DA, Tarasova Y, Portnoy VA, Zengler K, Osterman AL**. 2013. Transcriptional regulation of the carbohydrate utilization network in *Thermotoga maritima*. Front. Microbiol. **4**.
158. **Turner BL, Papházy MJ, Haygarth PM, McKelvie ID**. 2002. Inositol phosphates in the environment. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **357**:449–469.
159. **Martins LO, Carreto LS, Da Costa MS, Santos H**. 1996. New compatible solutes related to di-myo-inositol-phosphate in members of the order Thermotogales. J. Bacteriol. **178**:5644–5651.
160. **Ramakrishnan V, Verhagen M, Adams M**. 1997. Characterization of di-myo-inositol-1,1(prm1)-phosphate in the hyperthermophilic bacterium *Thermotoga maritima*. Appl. Environ. Microbiol. **63**:347–350.
161. **Santos H, da Costa MS**. 2001. Organic solutes from thermophiles and hyperthermophiles, p. 302–315. *In* Michael W. W. Adams, RMK (ed.), Methods in Enzymology. Academic Press.
162. **Rodionov DA, Kurnasov OV, Stec B, Wang Y, Roberts MF, Osterman AL**. 2007. Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. Proc. Natl. Acad. Sci. U. S. A. **104**:4279–4284.
163. **Ji E-S, Park N-H, Oh D-K**. 2005. Galacto-oligosaccharide production by a thermostable recombinant β-galactosidase from *Thermotoga maritima*. World J. Microbiol. Biotechnol. **21**:759–764.
164. **De Geus D, Hartley AP, Sedelnikova SE, Glynn SE, Baker PJ, Verhees CH, van der Oost J, Rice DW**. 2003. Cloning, purification, crystallization and

preliminary crystallographic analysis of galactokinase from *Pyrococcus furiosus*. Acta Crystallogr. D Biol. Crystallogr. **59**:1819–1821.

165. **Hartley A, Glynn SE, Barynin V, Baker PJ, Sedelnikova SE, Verhees C, de Geus D, van der Oost J, Timson DJ, Reece RJ, Rice DW**. 2004. Substrate specificity and mechanism from the structure of *Pyrococcus furiosus* galactokinase. J. Mol. Biol. **337**:387–398.
166. **Inagaki E, Sakamoto K, Obayashi N, Terada T, Shirouzu M, Bessho Y, Kuroishi C, Kuramitsu S, Shinkai A, Yokoyama S**. 2006. Expression, purification, crystallization and preliminary X-ray diffraction analysis of galactokinase from *Pyrococcus horikoshii*. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. **62**:169–171.
167. **Nguyen TN, Ejaz AD, Brancieri MA, Mikula AM, Nelson KE, Gill SR, Noll KM**. 2004. Whole-genome expression profiling of *Thermotoga maritima* in response to growth on sugars in a chemostat. J. Bacteriol. **186**:4824–4828.
168. **Chhabra SR, Shockley KR, Conners SB, Scott KL, Wolfinger RD, Kelly RM**. 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278**:7540–7552.
169. **Dunten P, Mowbray SL**. 1995. Crystal structure of the dipeptide binding protein from *Escherichia coli* involved in active transport and chemotaxis. Protein Sci. Publ. Protein Soc. **4**:2327–2334.
170. **Klepsch MM, Kovermann M, Löw C, Balbach J, Permentier HP, Fusetti F, de Gier JW, Slotboom DJ, Berntsson RP-A**. 2011. *Escherichia coli* peptide binding protein OppA has a preference for positively charged peptides. J. Mol. Biol. **414**:75–85.
171. **Bauer MW, Bylina EJ, Swanson RV, Kelly RM**. 1996. Comparison of a β-glucosidase and a β-mannosidase from the hyperthermophilic archaeon *Pyrococcus furiosus* purification, characterization, gene cloning, and sequence analysis. J. Biol. Chem. **271**:23749–23755.
172. **Park SH, Park KH, Oh BC, Alli I, Lee BH**. 2011. Expression and characterization of an extremely thermostable β-glycosidase (mannosidase) from the hyperthermophilic archaeon *Pyrococcus furiosus* DSM3638. New Biotechnol. **28**:639–648.
173. **Petersen TN, Brunak S, von Heijne G, Nielsen H**. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods **8**:785–786.
174. **Chhabra S, Parker KN, Lam D, Callen W, Snead MA, Mathur EJ, Short JM, Kelly RM**. 2001. β-Mannanases from *Thermotoga* species, p. 224–238. *In* Michael W.W. Adams, RMK (ed.), Methods in Enzymology. Academic Press.
175. **Parker KN, Chhabra SR, Lam D, Callen W, Duffaud GD, Snead MA, Short JM, Mathur EJ, Kelly RM**. 2001. Galactomannanases Man2 and Man5 from *Thermotoga* species: growth physiology on galactomannans, gene sequence analysis, and biochemical properties of recombinant enzymes. Biotechnol. Bioeng. **75**:322–333.
176. **Dos Santos CR, Paiva JH, Meza AN, Cota J, Alvarez TM, Ruller R, Prade RA, Squina FM, Murakami MT**. 2012. Molecular insights into substrate

specificity and thermal stability of a bacterial GH5-CBM27 endo-1,4-β-D-mannanase. J. Struct. Biol. **177**:469–476.

177. **Santos CR, Squina FM, Navarro AM, Ruller R, Prade R, Murakami MT**. 2010. Cloning, expression, purification, crystallization and preliminary X-ray diffraction studies of the catalytic domain of a hyperthermostable endo-1,4-beta-D-mannanase from *Thermotoga petrophila* RKU-1. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. **66**:1078–1081.
178. **Thomas GH**. 2010. Homes for the orphans: utilization of multiple substrate-binding proteins by ABC transporters. Mol. Microbiol. **75**:6–9.
179. **Wang Y, Wang X, Tang R, Yu S, Zheng B, Feng Y**. 2010. A novel thermostable cellulase from *Fervidobacterium nodosum*. J. Mol. Catal. B Enzym. **66**:294–301.
180. **Katsuraya K, Okuyama K, Hatanaka K, Oshima R, Sato T, Matsuzaki K**. 2003. Constitution of konjac glucomannan: chemical analysis and 13C NMR spectroscopy. Carbohydr. Polym. **53**:183–189.
181. **Herman P, Staiano M, Marabotti A, Varriale A, Scirè A, Tanfani F, Vecer J, Rossi M, D'Auria S**. 2006. D-trehalose/D-maltose-binding protein from the hyperthermophilic archaeon *Thermococcus litoralis*: the binding of trehalose and maltose results in different protein conformational states. Proteins **63**:754–767.
182. **Herman P, Barvik I Jr, Staiano M, Vitale A, Vecer J, Rossi M, D'Auria S**. 2007. Temperature modulates binding specificity and affinity of the d-trehalose/d-maltose-binding protein from the hyperthermophilic archaeon *Thermococcus litoralis*. Biochim. Biophys. Acta **1774**:540–544.
183. **Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE**. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. **25**:1605–1612.
184. **Zhang Y**. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics **9**:40.
185. **Wallace AC, Laskowski RA, Thornton JM**. 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. **8**:127–134.
186. **Yang Z**. 2007. PAML 4: phylogenetic analysis by Maximum Likelihood. Mol. Biol. Evol. **24**:1586–1591.
187. **Yang Z**. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. CABIOS **13**:555–556.
188. **Abascal F, Zardoya R, Telford MJ**. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. gkq291.
189. **Yang Z, Wong WSW, Nielsen R**. 2005. Bayes Empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22**:1107–1118.
190. **Zhang J, Nielsen R, Yang Z**. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. **22**:2472–2479.
191. **Kurakake M, Sumida T, Masuda D, Oonishi S, Komaki T**. 2006. Production of galacto-manno-oligosaccharides from guar gum by beta-mannanase from *Penicillium oxalicum* SO. J. Agric. Food Chem. **54**:7885–7889.

192. **Chhabra SR, Shockley KR, Ward DE, Kelly RM**. 2002. Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium *Thermotoga maritima* grown on glucan- and mannan-based polysaccharides. Appl. Environ. Microbiol. **68**:545–554.
193. **Zheng B, Yang W, Wang Y, Feng Y, Lou Z**. 2009. Crystallization and preliminary crystallographic analysis of thermophilic cellulase from *Fervidobacterium nodosum* Rt17-B1. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. **65**:219–222.
194. **Zheng B, Yang W, Wang Y, Lou Z, Feng Y**. 2011. Influence of the N-terminal peptide on the cocrystallization of a thermophilic endo-β-1,4-glucanase with polysaccharide substrates. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. **67**:1218–1220.
195. **Ferenci T**. 1980. The recognition of maltodextrins by *Escherichia coli*. Eur. J. Biochem. FEBS **108**:631–636.
196. **Ferenci T, Muir M, Lee KS, Maris D**. 1986. Substrate specificity of the *Escherichia coli* maltodextrin transport system and its component proteins. Biochim. Biophys. Acta **860**:44–50.