Characterisation of the TATA-box binding protein (TBP) from the Antarctic archaeon, Methanococcoides burtonii

Author:
Chong, Wai Yin Kevin

Publication Date:
2011

DOI:
https://doi.org/10.26190/unsworks/23753

Characterisation of the TATA-box binding protein (TBP) from the Antarctic archaeon, *Methanococcoides burtonii*

Wai Yin CHONG

Student number: 3040337

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Biotechnology and Biomolecular Sciences

Faculty of Science

University of New South Wales

Sydney, Australia

2011

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

Copyright Statement

Signed ……..

Date ……..

Authenticity Statement

‘I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.’

Signed ……..

Date ……..

Acknowledgements

No man is an island, but rather it is through a collective effort that science progresses. As such, I would like to extend my thanks to those who have helped me with my PhD project. First and foremost, I would like to thank the School of Biotechnology and Biomolecular Sciences (BABS), as well as the Graduate Research School for awarding me a scholarship to pursue my PhD studies. Without the monetary support from UNSW, I would not have been able to contribute to the scientific community through my research, meet many wonderful people, or experience firsthand the sights and sounds of Australia.

I would also like to express my deepest gratitude to my supervisor, Professor Ricardo Cavicchioli, for taking me under his tutelage and for his patience, support and invaluable guidance throughout my PhD candidature. In addition, I would also like to extend my gratitude to my co-supervisor, Professor Staffan Kjelleberg for his feedback and advice.

Further, I would like to thank the many people that have contributed to my PhD project with their important input. They include: Professor Paul Curmi, Dr. Sohail Siddiqui, Dr. Oliver Pilak and Dr. Davide De Francisci for their expertise in recombinant protein production, purification and biophysical characterisation; Dr. Tim Williams for his help with cloning and taking time off to read my thesis drafts; Dr. Dominic Burg for imparting me the methodology of culturing *Methanococcoides burtonii*; Geoff Kornfeld for his guidance in the gel shift assays; and Bill O'Sullivan for his invaluable comments on my thesis.

In addition, I would also like to thank Dr. Charmaine Ng, Dr. Nico Wanandy, Dr. Sabine Matallana Surget, Catherine Seah, Emily Chew, Arjun Verma, and Suhaila Mohd Omar, all of whom have been amazing and helpful lab colleagues, for their warm friendship.

No words can describe the appreciation and gratitude I have towards my wife, Ms. Li Min Chua, for being there for me when I was feeling down and depressed during the initial trying years when I could not obtain any good results for my project. Subsequently, she had to juggle between completing her own PhD thesis and editing my PhD thesis. Her sacrifices have not gone unnoticed, and I am forever indebted to her.

Lastly, I want to reiterate my thanks to all who had helped me in one way or another to make my PhD candidature a success, and an enjoyable experience.

Abstract

The TATA-box binding protein (TBP) plays an important role in transcription initiation. TBPs from many eucaryotes and archaea have been studied. However the structure and function of TBP from psychrophilic (cold-adapted) organisms is less understood. *Methanococcoides burtonii* is a psychrophilic archaeon that has served well for studies of cold adaption in archaea. The complete genome sequence for *M. burtonii* is published, and a single *tbp* gene was identified.

Translation of the *M. burtonii tbp* open reading frame was found to occur from two separate AUG codons, producing TBP isoforms containing 206 (MbTBP206) and 183 (MbTBP183) amino acids. Additionally, MbTBP206 contains a sequence repeat (RepA) in the N-terminal region that is absent in MbTBP183. Both isoforms possess identical putative DNA-interacting residues and are predicted to form a saddle-like conformation, typical of TBPs. Both isoforms were detected *in vivo* by Western blotting, and found to be over-abundant at 4°C than 23°C. Recombinant MbTBP isoforms produced were shown by gel shift assays to preferentially interact with oligonucleotides containing TATA-box sequences; a functional trait typical of TBPs.

Biophysical characterisation of both recombinant isoforms included evaluation of secondary structure and protein hydrophobicity by circular dichroism (CD) and ANS-fluorescence, respectively. Both isoforms were found to contain secondary structures typical of TBPs, with an unstructured RepA region in MbTBP206. Hydrophobicity studies also implied that MbTBP206 was more thermostable than MbTBP183. Further, heat-induced loss of protein conformation in both recombinant MbTBP from 4°C to 28°C was reduced by the addition of glutamate and aspartate, both of which are previously found intracellular solutes of *M. burtonii*. As such, functional characterisations of both isoforms were conducted using buffers that mimic the intracellular milieu of *M. burtonii*.

In order to elucidate functional differences between the isoforms and begin to determine putative binding locations throughout the genome, *in vitro* pull-down of fragmented genomic DNA was performed. The consensus sequences bound by MbTBP183 and MbTBP206 were TTnAAAww and TkymArAA, respectively. The consensus sequence for MbTBP183 had more conserved residues than MbTBP206 indicating MbTBP183 might have more specific promoter recognition. A diverse array of genes potentially regulated by MbTBP183 and/or MbTBP206 was identified.

Table of Contents

Chapter I

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

Chapter IX

Appendix

List of Figures

List of Tables

List of Abbreviations

^{32}P	Radioisotopic phosphate
%	Percent
°C	Degrees Celsius
ε	Extinction coefficient
Δε	Differential extinction coefficient
µg	Microgram
µL	Microliter
µM	Micromolar
×	Times
aIF	Archaeal translation initiation factor
ANS	1-anilino-8-naphthalene sulfonate
APS	Ammonium persulfate
At	*Arabidopsis thaliana*
ATP	Adenosine triphosphate
BGI	Beijing Genomics Institute
bp	Base pairs
BSA	Bovine serum albumin
CD	Circular dichroism
ChIP	Chromatin immunprecipitation
COG	Clusters of orthologous groups
CoM	Coenzyme M
Csp	Cold-shock protein
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophorectic mobility shift assay
ER	Evidence rating

ESI-MS	Electrospray ionisation-mass spectrometry
g	Gram / Gravity
h	Hour
His	Hexa-histidine tag
Hsp70	Heat shock protein 70
HTV	High tension voltage
IMAC	Immobilised metal-affinity chromatography
IMG	Integrated Microbial Genome database
Inr	Initiator element
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kav	Partition coefficient
kb	Kilo-base pairs
KDa	Kilo-dalton
L	Liter
LB	Luria-Bertani
m	Meter
M	Molar
MbTBP	*M. burtonii* TBP
MfTBP	*M. frigidum* TBP
mg	Milligram
min	Minute
mL	Milliliter
mM	Millimolar
mRNA	Messenger RNA
MRW	Mean residue weight
Mw	Molecular weight
ncRNA	Non-coding RNA
ng	Nanogram
nm	Nanometer
NPC	No protein control

NRMSD	Normalised root mean square deviation
NTSR	N-terminal sequence repeats
OD	Optical density
ORF	Open reading frame
p-value	Probability value
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PDB	Protein database
pmol	Picomole
PIC	Preinitiation complex
PVDF	Polyvinylidene fluoride
Pw	*Pyrococcus woesei*
Rep	Sequence repeat
RNA	Ribonucleic acid
RNAP	RNA polymerase
rpm	Revolutions per minute
rRNA	Ribosomal RNA
s	Second
Sc	*Saccharomyces cerevisiae*
S-D	Shine-Dalgarno
SDS	Sodium dodecyl sulfate
SEC	Size-exclusion chromatography
SELEX	Systematic evolution of ligands by exponential enrichment
SOC	Super optimal broth with catabolite repression
TAF	TBP-associated factor
TBP	TATA-box binding protein
TEMED	Tetramethylethylenediamine
TF	Transcription factor
TFB	Transcription factor B
TFE	Transcription factor E

TIE	Translation initiation elements
U	Units
UV	Ultraviolet
V	Volts
V_0	Void volume
v/v	Volume per volume
w/v	Weight per volume

CHAPTER I

Introduction

1.1 The Archaea

1.1.1 The birth of Archaea

The prokaryote-eucaryote dichotomy was the prevailing concept behind the classification of all organisms in the pre-1977 era [1]. However, this bipartite paradigm was revisited with the development of molecular sequencing techniques [2-3], and better understanding of the role of ribosomal RNA (rRNA) in molecular phylogeny [3-5]. By analysing and cataloging the RNase-digested products of rRNA from various organisms, Woese and Fox (1977) had found that a group of methanogenic 'bacteria' to be phylogenetically distinct from eucaryotes and typical bacteria [6-7]. As these anaerobic methanogenic 'bacteria' thrive in environments reminiscent of the Archeon Eon (3 – 4 billion years ago) on Earth, they were grouped and designated as 'archaebacteria' [7]. Consequently, the tree of life was modified from a dichotomy, to tripartite kingdoms of the urkaryote, eubacteria, and archaebacteria.

In addition, due to the more profound differences exhibited between the three kingdoms, than the differences between organisms within each kingdom, the tripartite status was elevated from kingdoms to domains of life. Further, there was also a need to segregate the apparent relatedness between the 'eubacteria' and 'archaebacteria', which was implied by the presence of 'bacteria' in both names. As such, the three domains of life, namely Bacteria, Eucarya, and Archaea were proposed [8]. The domain Archaea was further subdivided into two kingdoms/phyla by Woese *et al*. (1990), the Euryarchaeota and Crenarchaeota [8]. Subsequently, the kingdoms of 'Nanoarchaeota' [9], 'Korarchaeota' [10-11] and 'Thaumarchaeota' [12] were recently proposed with the discovery of the respective type specimens. However, statistical analysis of the phylogenetic tree of the Archaea revealed that the 'Nanoarchaeota', 'Korarchaeota' and 'Thaumarchaeota' phyla are subsumed under the Euryarchaeota and Crenarchaeota clade, and thus questions the possibility of 'Nanoarchaeota', 'Korarchaeota' and 'Thaumarchaeota' as being distinct phyla in their own right [13].

1.1.2 Unique archaeal features

The Archaea domain is a monophyletic lineage, and its uniqueness within the tree of life is clearly exemplified not only in their 16S rRNA sequences but also in their cellular and biochemical features. For example, the components of the archaeal phospholipid membrane are drastically different from those of the Bacteria and Eucarya [14-15]. The archaeal phospholipid is composed of a glycerol and highly branched hydrocarbon chains containing isoprene side chains. These components are bound together by ether linkages. Phospholipid from the Bacteria and Eucarya on the other hand, are made up of a glycerol with straight hydrocarbon chains, and are bound together by ester linkages. In addition, thermophilic and acidophilic archaeon often possess membrane-spanning monolayer phospholipid made up of tetraether lipids, as compared to the phospholipid bilayer membrane structure in bacteria and eucaryotes.

Another unique aspect of the archaeal cellular and biochemical features, involves the cell wall. All archaeal species lack murein, which is an essential component in the peptidoglycan cell wall of bacteria. Instead, most of the archaeal species have an envelope of proteineaceous surface layer (S-layer) [16]. Some bacteria, such as the *Clostridium*, *Bacillus* and *Lactobacillus* species do contain an S-layer in addition to their cell wall as well [17]. However, the S-layers in most archaeal species have hexagonal lattice symmetries, in contrast to those from the Bacteria that contain either oblique or tetragonal lattice symmetries [17-18].

The uniqueness of archaea, as compared to the other two domains of life, also extends to genomic and proteomic features [19]. As an example, the heat shock protein 70 (hsp70) is found in most bacteria and eukaryote, but not in archaea. The presence of hsp70 only in the archaeal *Methanobacterium thermoautotrophicum*, *Thermoplasma acidophilum*, and *Halobacterium* has been attributed by the authors to lateral gene transfer [19].

Finally, other uniquely archaeal features also include methanogenesis (Chapter 1.2), and the absence of pathogenic archaeal species thus far [13,20].

1.1.3 Common features between Archaea, Bacteria and Eucarya

Archaea also share some features with bacteria and eucaryotes. Archaea have a similar cell size to bacteria. Together with bacteria, Archaea lack organelles and a nucleus, and contain a single circular chromosome copy [19,21]. Plasmids are also present in archaea, such as *Haloferax volcanii* [22] and *Halobacterium sp.* strain NRC-1 [23]. In addition, the arrangements of the genes within some operons, such as those for rRNA and ribosomal protein, are also similar between archaea and bacteria as well [21,24].

The domains of Archaea and Eucarya are similar in several aspects. The effects of antibiotics targeting the ribosomes of both domains of life are good illustrations of this similarity. Archaea and eucaryotes are resistant to the 70S ribosome-targeting streptomycin, while bacteria are sensitive to it [21]. The DNA-dependent RNA polymerase from the archaeon *Sulfolobus acidocaldarius* was sensitive to eukaryotic, but not bacterial, transcriptional inhibitors [25]. Another example would be the multi-subunit proteosome, where targeted cellular proteins are degraded. The proteosomes have been reported in both archaea and eukaryote, but not in bacteria except for the actinomycetes [21]. However, the acquiring of the proteosome by the actinomycetes is likely to be a consequence of lateral gene transfer from either the Archaea or Eucarya [26].

The archaeal transcription and translation systems are essentially a mosaic of both bacterial and eukaryotic features [27]. The archaeal gene promoter and transcriptional machineries such as the TATA-box binding protein (TBP), transcription factor B (TFB) and E (TFE), and DNA-dependent RNA polymerase are all eukaryotic-like [28-32]. However, archaeal gene regulations are generally bacteria-like [33-34]. The archaeal mRNA products are also reminiscent to that of the Bacteria, in that it lacks introns, 5’ methyl-caps and 3’ poly-A tail processing.

In the archaeal translation system, the translation machineries, such as the aIF-1 and aIF-2 are homologous to their eukaryotic counterparts [27,35]. Nevertheless, the archaeal ribosome-binding element is similar to that of the Shine-Dalgarno sequences in bacteria. In addition, analogous to the Bacteria, the start codons used by the Archaea is not limited to only AUG, but uses the GUG and UUG codons in some situations [35].

1.1.4 Diversity of archaeal habitats

Archaea have been discovered in anoxic sediments [36-37], hot springs [38-40], hydrothermal vents [41-42], hypersaline lakes [43-45], acid mines [46-47], marine [48] and cold environments [49-51], where these harsh environments would be dire to most organisms. As such, archaea have long been viewed as somewhat on parity with the term 'extremophile', due to the extreme environmental conditions that they thrive in. However, non-extremophilic archaea are also abundant in various non-extreme biospheres [52].

There have also been reports of the existence of archaeal species within other organisms. For example, archaeal species from the *Methanobrevibacter* genus has been reported to be prevalent in the guts of termites [53], cows [54], and humans [55]. Such interaction with other organsims was even extended to the endosymbiosis of the archaea, *Nanoarchaeum equitans*, into another archaeal host, the *Ignicoccus hospitalis* [9].

1.1.5 Significance of archaea

The significance of archaea is not trivial. Archaea accounts for at least 20% of the biomass in the ocean [56], or approximately 1.3×10^{28} archaeal cells within the world's oceans [48]. Such sheer abundance of archaea in the global ocean ecology therefore illustrates its importance, possibly as part of the food chain or in oceanic biogeochemical cycles [57]. The discovery of an ammonia-oxidizing, inorganic carbon-fixing marine archaeon emphasises the role of archaea in regulating oceanic nitrogen and carbon cycles [58].

Arguably, one of the more significant species of Archaea in recent decades is the methanogenic archaea and their contribution to the global warming of the Earth. Methane, a greenhouse gas, is a consequence of anaerobic respirations by the methanogens. For example, methane emission from rice paddy fields accounts for 10 – 25% of the total global methane emission, and archaea such as *Methanosaeta thermophila* are key methane producers in these fields [59-61].

1.2 Methanogen and methanogenesis

All methanogens belong exclusively to the domain of Archaea, and specifically, to the phylum of Euryarchaeota. The methanogens are also restricted to the orders of Methanococcales, Methanopyrales, Methanobacteriales, Methanosarcinales, and Methanomicrobiales. Recently, the order Methanocellales had been included in the methanogen-specific orders as well [62]. These methanogens are broadly classified into two classes based on their phylogenies [63-64]. The archaea-of-interest discussed in this thesis, *Methanococcoides burtonii*, belongs to the order of Methanosarcinales which is categorised under the class II methanogens.

Methane production is the catabolic consequence when methanogens extract energy for anaerobic growth. The majority of the methanogens metabolise acetate by the aceticlastic pathway, while the rest produce methane through either the CO_2-reduction or methylotrophic pathway [65]. The aceticlastic pathway involves the cleavage of acetate into its carbonyl and methyl components, where the former provides electrons for the reduction of the latter into methane. Methanogens that utilise the CO_2-reduction pathway obtain electrons from hydrogen, carbon monoxide or formate, for the reduction of CO_2 to methane [65]. *Methanosaeta* [66] and *Methanocella paludicola* [62] are classic examples of methanogens that utilise the aceticlastic and CO_2-reduction pathways, respectively.

The methylotrophic pathway, employed by *Methanococcoides burtonii*, extracts electrons from the methyl group of methanol and methylamines for the reduction of methyl to methane [65,67]. As depicted in Figure 1, methyl groups from the methylated substrates are first captured by coenzyme M (HS-CoM) to form the molecule CH_3-SCoM. The molecule CH_3-SCoM then proceeds to either generate electrons through the oxidation of CH_3-H_4SPT to CO_2, or to be reduced to methane with the generation of CoMS-SCoB. The heterodisulfide reductase subsequently cleaves CoMS-SCoB into coenzyme M and coenzyme B, with the generation of a proton gradient for ATP synthesis. Finally, the CO_2 generated in the methylotrophic process is used in its biomass production [67].

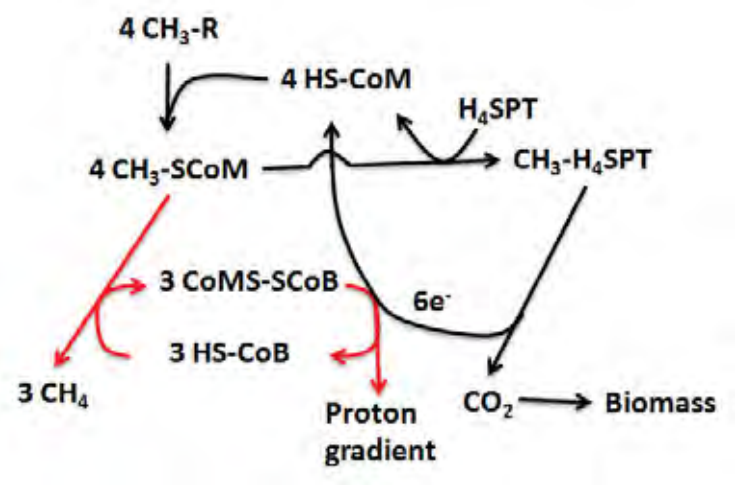


Figure 1: Overview of the methylotropic pathway.

The pathyway is adapted and modified from Ferry (2010) [65] and Williams *et al*., (2009) [67]. Pathways in black are unique to methylotrophic archaea, where R refers to '-OH' and '-$N(CH_3)_3$' in the case of *Methanococcoides burtonii*. Pathways in red are common to all three methanogenesis methods.

1.3 Cold adaptation in archaea

Most of the Earth's biospheres are cold, and are permanently below 5°C [68-69]. The majority of the cold environments are found on the deep ocean beds since the ocean makes up about 70% of the Earth's surface. The other cold habitats include the polar, sub-polar, glacial and alpine regions. Despite the freezing temperatures, microbial communities are still found thriving in these environments [37,49-51,68,70-71].

Organisms that lack homeothermic regulation, and survive only in cold environments are generally designated as psychrophiles. Specifically, organisms that grow over a narrow and wide temperature range in a cold environment are termed as stenopsychrophiles and eurypsychrophiles, respectively [69]. Stenopsychrophiles do not survive at temperatures beyond their limited low growth temperatures. Eurypsychrophiles, on the other hand, are tolerant of temperatures beyond its preferred low growth temperature, and therefore has a wider growth temperature range [68-69].

Microorganisms thriving in cold environments face several challenges. The reduction in membrane fluidity would lead to a loss of membrane permeability and function, and consequentially, altered nutrient and waste transport between the intra and extracellular milieu, as well as electron transport. In addition, according to the Q_{10} rule, the reaction rate decreases by two to four-fold with every decrease of 10°C. This is likely due to the increased structural rigidity and thermostability of enzymes and proteins with decreasing temperature. The slowdown of reaction kinetics, at extreme cold, undermines the ability of the microorganism to respond and survive.

Psychrophilic archaea have employed several adaptations to counter such disastrous effects of cold. For example, both *Methanococcoides burtonii* and *Halorubrum lacusprofundi* contain abundant levels of unsaturated lipids on their cellular membrane, which prevents the onset of membrane rigidity under cold temperatures [72-73]. The psychrophilic archaea also express cold-induced proteins, whose roles are to aid in the crucial transcriptional process, which would otherwise not be functional at cold temperature. The cold-shock protein (Csp) and DEAD-box RNA helicase, for example, are used by *Methanogenium frigidum* and *Methanococcoides burtonii*, respectively to reduce secondary structure formation in the mRNA transcripts [74-75], and hence allow accessibility of the transcripts to the ribosomes.

Another strategy engaged by psychrophilic archaea, and even bacteria, is to modify the cellular proteins, in particular, enzymes to be cold-active or tolerant. Specifically, psychrophilic enzymes have increased localised conformational flexibility in their active sites as compared to their non-psychrophilic counterparts. However, between of both the psychrophilic and the non-psychrophilic enzymes, there are no significant changes to the conformational flexibility in non-active sites [76]. The consequences of an increased conformational flexibility and decreased thermostability are a decrease in the activation energy necessary for reactions, and thus greater enzymatic activities in cold temperature [77-79].

The above correlation between conformational flexibility and higher protein activity therefore begs the question of what contributes to the increase in conformation flexibility in pyschrophilic proteins and enzymes. Studies from an analysis of 1,111 protein homology models, from nine methanogenic archaea with different optimal growth temperatures, revealed that proteins from cold-adapted archaea contained more glutamine, threonine, and hydrophobic residues, but fewer charged residues, on the protein surface [80]. The replacement of charged residues with glutamine, threonine, and hydrophobic residues on the surface would not only destabilise protein conformation (hence increasing conformational flexibility) but also prevent protein aggregation at cold temperatures [80].

An insight into the archaeal cold adaptation strategy is also made possible by deciphering how hyperthermophilic archaeal proteins differ from their 'colder' mesophilic counterparts. For example, the mesophilic TATA-box binding protein (TBP) differs from those of highly thermostable hyperthermophilic counterparts, in that the former lack disulphide bridges, and hydrophobic residues (valine, leucine, and isoleucine) in the interior of TBP [81]. In addition, there is also a positive correlation between the optimal growth temperature and surface hydrophilicity of TBP [81]. Therefore, low surface hydrophilicity and a lack of interior hydrophobic residues, lower the thermostability of a TBP, thus making it more adaptable to low temperatures.

1.4 Methanococcoides burtonii

Methanococcoides burtonii is a motile eurypsychrophilic archaea, from the Euryarchaeota phylum. It was first isolated and described by Franzman *et al.* (1992), and has a broad growth temperature range of -2°C – 28°C [49,68]. Despite having an optimal growth temperature of 23°C, it is widely viewed that *M. burtonii* is subjected to thermal stress under this temperature [82-84]. This is evident by the differential abundance of the molecular chaperones DnaK and Cpn60 at 23°C compared to 4°C [82-84]. Molecular chaperones are essential aids to the refolding of stress-denatured proteins, and are indicators of heat stress. Therefore, the optimal growth of *M. burtonii* at 23°C is likely to be a physiological consequence of the increase in reaction kinetics within the archaea, and not a genuine marker of the preferred growth temperature of the archaea [68]. Further, since the ecological niche of *M. burtonii* is at a constant temperature of 1°C – 2°C [68] and *M. burtonii* is also at a low state of stress at 4°C, [82-84] these findings confirm its psychrophilicity.

Methanococcoides burtonii is isolated in the bottom of the meromitic Ace Lake, in Vestfold Hills, in Antarctica, although the genus itself appears to be cosmopolitan in nature with other discoveries in Tanzania [85], Japan [70], and United States [86-87]. Water from the bottom layer (25m) of Ace Lake is anoxic, with a high salinity of 4.3% [68]. It is also saturated with methane [68], which is possibly due, in part, from the methanogenesis process by the obligate methylotrophic *M. burtonii*.

A total of 2494 genes were identified in *M. burtonii*, of which approximately a thousand were functionally annotated with an evidence rating of ER1 – ER3 [88]. An evidence rating of ER1 – ER3 generally signifies a good agreement between the functional assignments and the experimentally characterised function of the gene, with ER1 being the best agreement. At the other end of the spectrum however, with an ER5 status, approximately 600 genes were of unknown function in *M. burtonii* [88]. Nevertheless, this library of manually-curated genetic information from *M. burtonii*

would allow for an in depth study of individual gene and protein with greater confidence, and eliminate any mis-annotation by computational analysis [89].

Finally, comparative genomics analysis of the draft, and subsequently completed, genome of the 2.6 million base-pair *M. burtonii* with that of other microorganisms, yielded valuable insights into the thermal adaption, horizontal gene transfer from ε- and δ-proteobacteria, metabolic pathways, and evolution of *M. burtonii* [80,88,90].

1.5 Transcription initiation and regulation in the three domains of life

Transcription is at the heart of information transfer from its static encrypted state within the genome to a dynamically functional state within a cell. Genes are static, containing the cellular blueprint of its host but are generally not capable of direct interference with cellular processes. Proteins on the other hand, are the workhorses within the cell and their expression is delicately balanced.

The transcription system, which results in the production of mRNA transcripts, is generally divided into three stages, transcription initiation, elongation, and termination. As the transcription initiation involves a complex series of recruitment and interactions of protein factors with the gene-of-interest, it is often the regulatory checkpoint for the expression of most genes [91-92]. In addition, transcription initiation and regulation between the domains of Bacteria and Eucarya are distinctly different, while that of the Archaea possess a mosaic of features from the other two domains [27].

The essential transcriptional components from each domain are outlined below to illustrate the chimeric nature of the Archaea transcription system.

1.5.1 Eucaryal transcription initiation and regulation

The transcription initiation process in all three domains of life begins with the interaction of a promoter-binding protein with the promoter elements upstream of a gene. In Eucarya, the promoter element contains a highly conserved 'TATA' sequence

(termed as the TATA box) that is approximately 30 nucleotides upstream of the transcription initiation site. Together with an initiator element (Inr), which holds the initiation site, the TATA box constitutes the minimal promoter region of a general eukaryotic gene [27,91].

Eucaryal transcription initiation (Figure 2A) begins with the binding of a TATA-box binding protein (TBP), which is part of the larger TFIID complex, with the TATA box. Subsequent binding of the transcription factor, TFIIA to the TFIID/TATA complex, stabilizes the whole preinitiation complex (PIC). This is followed by recruitment of TFIIB into the PIC, which occurs through interactions between the TFIIB with both the TBP and the DNA. The TFIIB mediated by TFIIF, proceeds to recruit the DNA-dependent RNA polymerase II (RNAP II) holoenzyme into the PIC, and onto the Inr element in preparation for transcription initiation. Finally, the recruitment of both TFIIE and TFIIH into the PIC serves to denature the double-stranded DNA that allows for transcription of the gene by RNAP II to begin [27,91,93].

Transcription regulation in eukaryotes is mostly enhancer-directed. Transcriptional enhancers are cis-acting DNA regulatory sequences that when bound by specific transcription factors, provide a stimulatory role to the transcription level of the genes [94]. The effects of the enhancers can be either transcription activation or repression, and are dependent on the types of regulatory transcription factors involved [95-97]. In addition, the effects of the transcriptional enhancers are not limited by distance, but by insulator sequences.

1.5.2 Bacterial transcription initiation and regulation

Compared to the eukaryotic transcription initiation system, that of the bacterial system is a much simpler affair. Bacterial gene promoters include four principle promoter elements, which are the Pribnow box, -35 element, extended -10 element, and UP element. The Pribnow box contains sequences that are rich in thymine and adenine, and are located at approximately 10 nucleotides upstream of the initiator

nucleotide. The immediate upstream region (3 – 4 nucleotides) of the pribnow box is the extended -10 element, while the -35 element is situated at 35 nucleotides upstream of the initiator site. The UP element is located upstream of the -35 element. Collectively, these four promoter elements dictate the transcription initiation site of a bacterial gene [98].

Initiation of bacterial transcription (Figure 2B) begins with the interaction of the sigma factor with the promoter elements. Specifically, the domains 2, 3 and 4 of the σ factor binds to the pribnow box, extended -10 element, and -35 element of a gene, respectively. Thereafter, the bacterial DNA-dependent RNA polymerase (RNAP) holoenzyme is recruited into a σ/DNA complex to initiate transcription [98].

Regulation of the bacterial transcription is achieved by several means [98]. Upon receiving the appropriate signaling, transcription is activated through binding of the activators with the RNAP or domain 4 of the σ factor, or altering the conformation of the promoter to enable the RNAP to interact with the promoters. Transcriptional repression on the other hand, is accomplished through steric hindrance of the RNAP to the preinitiation complex by repressors, sequestering of the promoter elements by the repressors, or inactivation of the transcriptional activators.

1.5.3 Archaea transcription initiation and regulation: a chimera system

The transcription initiation system of the Archaea is similar to that of the eucaryal system, with respect to its promoter element and basal transcription machineries. The archaeal promoter element has a consensus 'TATA' sequence, and resides at approximately 26 nucleotides upstream of the transcription initiation site [29,99]. In addition, the initiation of archaeal transcription involves the archaeal homolog of the eucaryal TBP, TFIIB, TFIIE, and RNAP II as their basal transcriptional machineries [30-31,100-104]. However, archaeal homologs of other eukaryotic transcription proteins such as TFIIA, TFIIF and TFIIH are absent [105].

Transcription initiation of an archaeal gene (Figure 2C) proceeds with the interaction of an archaeal TBP with the TATA-box. The transcription factor B (TFB), an archaeal homolog of TFIIB, is subsequently recruited into the TBP/TATA complex by making contact with both the TBP and BRE (TFB responsive element), and thus directing the orientation of the preinitiation complex [106]. Binding of the TFB to the TBP/TATA-box complex also stabilises the complex [107]. Finally, the archaeal RNAP is recruited to initiate transcription. In addition, although the Archaea possesses transcription factor E (TFE), an archaeal homolog of TFIIE, it is not essential to the initiation of archaeal transcription [108].

Given the resemblance of the archaeal basal transcription apparatus to that of the eucaryal, it is intriguing to find an array of bacterial-like transcriptional regulators within Archaea. In fact, a substantial number, amounting to approximately 60% of the total archaeal transcription regulators, are homologous only to the Bacteria [109]. For example, the Mdr1 transcriptional regulator from the archaeon *Archaeoglobus fulgidus* is a homolog of the bacterial metal-dependent transcriptional repressor DtxR [34]. In addition, the mode of archaeal transcriptional regulation is reminiscent to that of the bacterial *lac* operon, by targeting the promoter or operator sites for transcription repression [34,110].

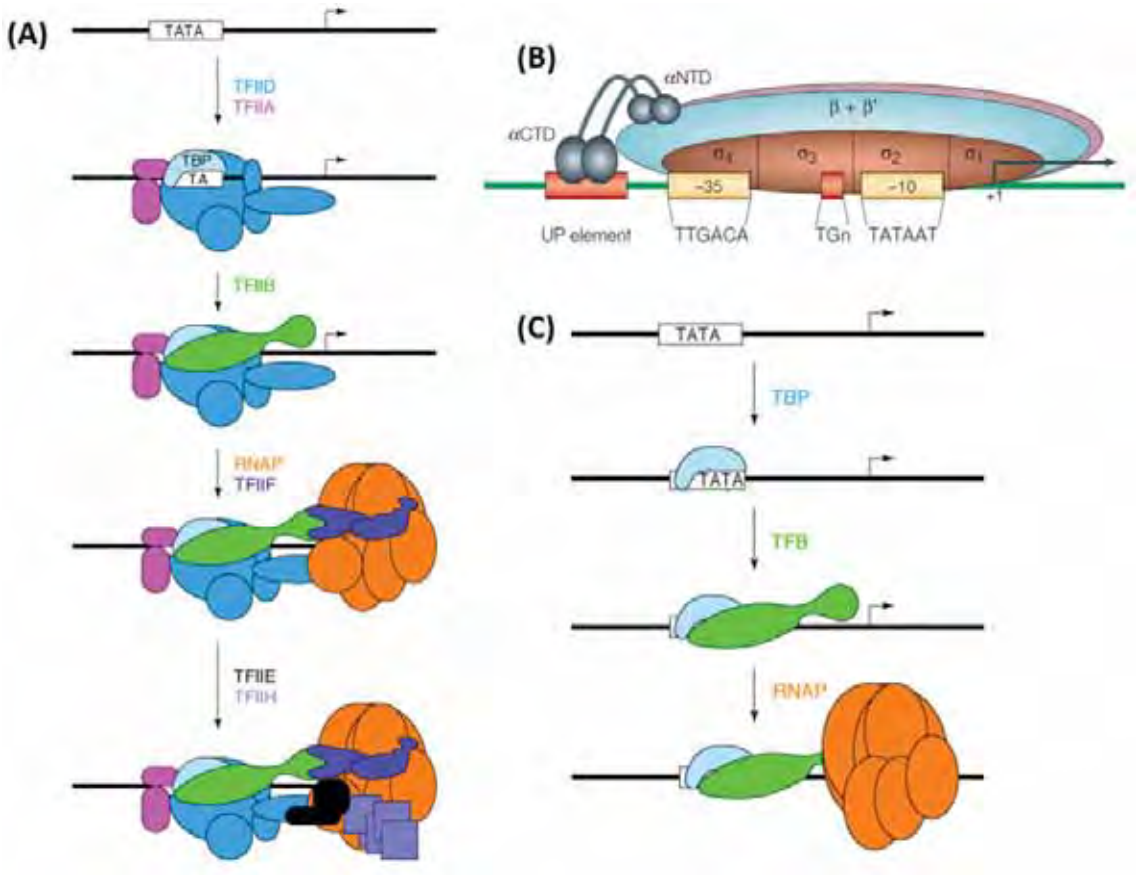


Figure 2: Schematics of the transcription initiation process in the three domains of life.

Schematics for the eucaryal (A) and archaeal (C) transcription initiation are adapted from Bell and Jackson (1998) [27], while that of the bacterial (B) is adapted from Browning and Busby (2004) [98].

1.6 TATA-box binding protein (TBP)

The TATA-box binding protein (TBP) is a minor groove DNA-binding protein [111-114], that specifically binds to an eight base-paired promoter signature, the TATA-box, on the DNA. Upon interacting with the TATA-box, the TBP unravels and bends the DNA helix towards its major groove, and thus exposes its minor groove for binding to the concave surface of the protein. Subsequently, the TBP recruits TFB to continue the transcription initiation process.

The TBP plays such a critical role in the recognition of gene promoters that it is highly conserved among the Eucarya and Archaea domains. Between the two domains, there is a sequence similarity of 20% to 40% [105]. Sequence similarity within the archaeal TBPs

are higher at 40% to 55%, while that of the eucaryal TBPs is only about 27% between the lower and higher eucaryotes [105]. Nevertheless, there is a much higher degree of TBP conservation among the higher eucaryotes [105]. In addition, the amino acid residues that are involved in the TBP/TATA-box interactions are highly conserved, especially those that reside on the concave surface, and the C-terminal stirrup region of the protein. [105,115].

An obvious difference between the archaeal and eucaryal TBPs lies in the N-terminal domain of the proteins. Archaeal TBPs are generally 180 amino acids in length, with the whole protein being constituted as a TATA-binding domain. Archaeal TBPs from the *Pyrococcus*, *Thermococcus* and *Sulfolobus* species however, contain a short acidic tail of unknown significance [116-117]. In addition, archaeal TBPs generally contain an excess of acidic residues, while that of eucaryal TBPs have an excess of basic residues [105,115].

Eucaryal TBPs are longer than that of the archaeal, and contain a species-specific N-terminal domain in addition to the 180 amino acid long C-terminal TATA-binding domain [105]. Functions of the N-terminal domain in the eucaryal TBP varies, from being associated with human neurodegenerative disease [118-119], maternal immune-tolerance of fetus in mice [120], to an autoinhibitory effect on the TBP/TATA-box binding in yeast [121-123].

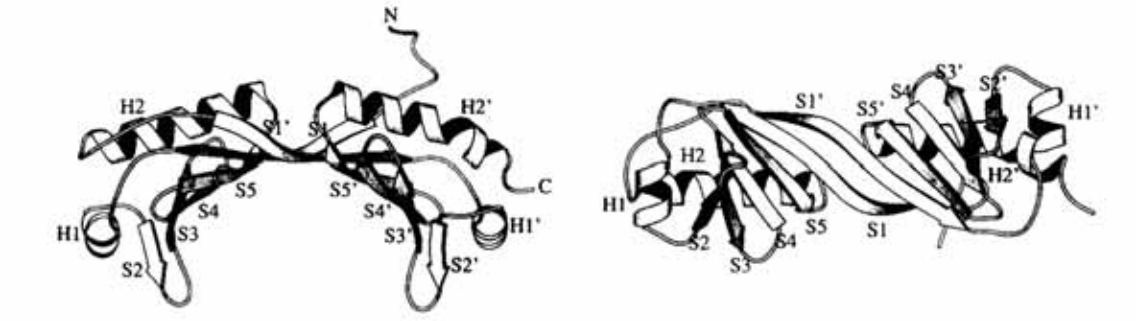


Figure 3: Crystal structure of a TATA-box binding protein from *Arabidopsis thaliana*.

The crystal structure is adapted from Nikolov *et al.*, (1992) [124]. The protein is viewed from the front (left) and underside of the molecular saddle (right).

All archaeal TBPs and eucaryal TBPc (C-terminal domain of the TBP) adopt a saddle-like conformation with a twofold intramolecular symmetry (Figure 6), where each cyclin A-like structural subdomain is approximately 90 amino acids in length [81,112,115,124-126]. In addition, each of the two structural subdomains is comprised of two α-helices and five β-strands, with a molecular stirrup structure occurring in the loop connecting β-strands S2 with S3 [124]. Further, there is also a high level of sequence and topological similarities between the two structural subdomains [105,114,124]. Based on these high homology levels, a duplication event of a precursor gene had potentially occurred to generate an ancient *tbp* gene, before the ancient gene itself was delineated into an archaeal and eucaryal form [105].

1.7 Interactions between TBP and TATA-box

The consensus sequence of the archaeal TATA-box appears to be dependent on whether the archaea is halophilic, or methanogenic or from the Crenarchaeota phylum [99]. A single nucleotide mutation within the TATA-box is sufficient to cause a drastic variation in the angle and flexibility of the TBP-bent DNA, and consequentially affects the transcriptional efficiency of the gene controlled by the mutant promoter [127-129]. Therefore, the differential transcription efficiency of genes would, in part, be dependent on the DNA sequences of the TATA-box. In addition, differences in the TBP species also contributes to the variation in the bent angle of the DNA helix, and hence the transcriptional efficiency [129].

Previous studies had reported that eucaryal TBP binds to the TATA-box, in particular the 'TATAAAAG' and 'TATATATA' sequences, via a three-step binding mechanism, with two intermediates [123,130-131]. Both TBP/TATA-box complex intermediates were postulated to be similar to that of the final binding complex, except that the TBP is not fully locked into position with the bent DNA in the intermediates [131]. In addition, the binding of the TBP to the TATA-box occurs simultaneously with the bending of the DNA helix [123,131].

Crystal structures from the TBP/TATA-box complex yielded further insights into the binding between the protein and the TATA-box. The effect of DNA binding on the saddle-like conformation of TBP is modest, if any, with only a 10° twist of one structural subdomain with respect to the other [113-114]. The binding interface between the TBP and DNA is primarily hydrophobic [113-114], with only approximately 35% of the interface registering a hydrophilic profile [114]. In addition, the binding interaction is locked in place by two pairs of highly conserved phenylalanine residues, where one pair is inserted in between the first two base-pairs, and the next in between the last two base-pairs of the TATA-box sequence [112-114]. The ends of the TATA-box containing minor groove are therefore pried apart, thus resulting in the severe bending of the DNA helix.

Several studies had also reported that the proximal sequences to the TATA-box influence the outcome of a TBP/TATA-box interaction, although the proximal sequences themselves do not come in contact with the TBP [107,132-133]. By using 'TATATATA' as the TATA-box sequence, Faiger *et al.* (2006) had reported that the TATA-flanking sequences partake in a more important role than the TATA-box itself, in determining the TBP/TATA-box stability [133]. Further, it has been reported that the TATA-flanking sequences affects the disassociation, but not the association, kinetics of the TBP/TATA-box complex [107].

Consequentially, the presence of the flanking sequences inevitably provides more variability to the core promoter region without the need to enlarge the TBP-binding sites [133]. This extended (core promoter and flanking) region would therefore allow Nature a greater degree of freedom in tailoring the appropriate expression levels of different genes, than it otherwise would if the transcription efficiency is only dependent on the TATA-box sequence.

1.8 Research motivations and aims

There is a lack of knowledge on the transcriptional process and regulation in cold-adapted archaea, despite non-psychrophilic archaeal transcription and regulation having been the focus of many studies. Of particular interest is the TATA-box binding protein (TBP), an essential member in the transcription process, as it would have a potent and direct effect on the transcriptome and hence proteome of a cold-adapted archaea.

Since the RNA polymerase subunit E/F, an essential components of the archaeal transcription initiation machineries, of the psychrophilic *Methanococcoides burtonii* has been characterised [134], it was appropriate to follow up with the characterisation of the TBP from *M. burtonii*, to build on the knowledge already gained on transcription initiation in the cold-adapted archaea.

In addition, although the gene locus Mbur_1496 has been functionally annotated to be a *tbp* gene of *M. burtonii* [88], there is still no functional evidence which indicates that it is indeed a TBP. Functional characterisation of the Mbur_1496 gene product as a TBP would thus irrevocably confirm its annotation.

Cell-free archaea transcription assay have been widely used to study promoters [135-136], demonstrate transcriptional activity [137-139], and transcription regulation [140], and can potentially be applied to study transcriptional regulation in *M. burtonii*. However, current transcription assays are unsuitable for studying transcription regulation involved in cold-adaptation, as they were designed based on a thermophilic, instead of a psychrophilic, archaeal system. In particular, the cold-adapted heat-labile transcriptional regulator would be denatured at the high temperatures used for the current cell-free transcription assay. Therefore, in depth knowledge of the cold-adapted TBP from *M. burtonii* would be an essential step towards the future development of a low temperature-based cell-free transcription assay, for studies of the cold-adapted transcriptional apparatus.

As such, the aims of the research described in this thesis were to verify that the Mbur_1496 gene product is a TBP, to ascertain the presence of the TBP *in vivo* in *M. burtonii*, and to characterise the biophysical and functional aspects of the protein. Finally, a genome-wide analysis of the promoter regions in *M. burtonii* was conducted to elucidate the consensus of the natural TATA-box sequence for binding by the TBP.

CHAPTER II

Materials and Methods

2.1 Materials

The components of commonly used buffers and solutions are described below.

1. Mineral salt stock solution (100×) – One litre of solution was prepared with 1.5g of nitrilotriacetic acid, 3g of $MgSO_4 \cdot 7H_2O$, 0.5g of $MnSO_4 \cdot 7H_2O$, 1g of NaCl, 0.1g of $FeSO_4 \cdot 7H_2O$, 0.1g of $CoSO_4$, 0.1g of $CaCl_2 \cdot 2H_2O$, 1g of $ZnSO_4$, 10mg of $CuSO_4 \cdot 5H_2O$, 10mg of $AlK(SO_4)_2$, 10mg of H_3BO_3, and 10mg of $Na_2MoO_4 \cdot 2H_2O$. The mineral stock solution was filtered with a 0.22µm filter and stored at -20°C.
2. Vitamin stock solution (100×) – One liter of solution was prepared with 2mg of biotin, 2mg of folic acid, 2mg of pyridoxine HCl, 10mg of thiamine HCl, 5mg of riboflavin, 5mg of nicotinic acid, 5mg of DL-Ca pantothenate, 0.1mg of vitamin B_{12}, 5mg of p-aminobenzoic acid, and 5mg of lipoic acid. The vitamin stock solution was filtered with a 0.22µm filter and stored at -20°C.
3. MFM medium - One liter of media contained 0.335g of KCl, 6g of $MgCl_2 \cdot 6H_2O$, 1g of $MgSO_4 \cdot 7H_2O$, 0.25g of NH_4Cl, 0.14g of $CaCl_2 \cdot 2H_2O$, 23.37g of NaCl, 2mg of $Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$, 1mg of resazurin, 5g of trimethylamine HCl, 2g of yeast extract, 10mL of 100× vitamin stock solution, 10mL of 100× mineral stock solution, 0.1g of sodium acetate, 0.14g of K_2HPO_4, 0.5g of cysteine·HCl, and finally 2.52g of Na_2CO_3. During the entire course of the preparation of MFM media, 80% N_2:20% CO_2 gas was used to purge the oxygen content within the culture to create an anaerobic culture condition. The pH of MFM was adjusted to 6.8 before dispensing into culture bottles at 100mL of culture per bottle. The bottle was seal with a butyl rubber stopper, autoclaved, and swirled overnight to dissolve the precipitated salts.
4. XS buffer – The buffer was prepared with 0.1g of potassium ethylxanthogenate, 2mL of ammonium acetate (4M), 1mL of Tris (1M) at pH7.4, 0.4mL of EDTA (0.5M), and 1mL of 10% (v/v) SDS, and topped up with distilled water to make up a final solution of 10mL. Precipitates that form during the process were dissolved with heating the solution at 65°C. The XS buffer was filtered sterilised, stored in 4°C, and used within a month of preparation.

5. Non-denaturing extraction buffer (2×) – The buffer was prepared with 20mM of Tris at pH 7.8, 2mM of EDTA, 0.2M of NaCl, and 20% (v/v) of glycerol.
6. SOC medium at pH7.5 – The media was prepared with 20g/L of tryptone, 5g/L of yeast extract, 10mM of NaCl, 2.5mM of KCl, 10mM of $MgCl_2$, 10mM of $MgSO_4$, and 20mM of glucose.
7. TAE buffer (50×) – The buffer was prepared with 242g/L of Tris, 5.71% (v/v) of glacial acetic acid, and 18.6g/L of EDTA. 1× TAE buffer was subsequently prepared by diluting 20mL of 50× TAE buffer with 980mL of distilled water.
8. *E. coli* lysis buffer – The buffer was prepared with 50mM of phosphate (KH_2PO_4/K_2HPO_4) buffer system at pH 7.8, 400mM of NaCl, 100mM of KCl, 10% (v/v) of glycerol, and 0.5% (v/v) of triton X-100.
9. Urea buffer – The buffer was prepared with 100mM of NaH_2PO_4 at pH 8.0, 10mM of Tris, and 8M of urea.
10. Wash buffers for His-tag purification – The buffer contains the same components as *E. coli* lysis buffer, but without triton X-100. However it contains a final concentration of 5mM of β-mercaptoethanol.
11. Elution buffer for His-tag purification – The buffer contains the same components as wash buffer, but added with various concentrations of imidazole.
12. GF buffer – The buffer was prepared with 50mM of phosphate (KH_2PO_4/K_2HPO_4) system at pH 7.8, 400mM of NaCl, 10% (v/v) of glycerol, and 1mM of DTT.
13. Stacking gel (4%) – The gel was prepared with 1mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 2.52mL of 0.5M of Tris (pH 6.8), 6.36mL of water, 100µL of 10% (w/v) SDS, 10µL of TEMED, and 50µL of 10% (w/v) APS.
14. Resolving gel (12%) – The gel was prepared with 3mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 2.5mL of 1.5M of Tris (pH 8.8), 4.35mL of water, 100µL of 10% (w/v) SDS, 5µL of TEMED, and 50µL of 10% (w/v) APS.
15. SDS loading buffer (1×) – The buffer was prepared with 6.25% (v/v) 0.5M Tris at pH 6.8, 10% (v/v) glycerol, 2% (v/v) β-mercaptoethanol, 0.1% (w/v)

bromophenol blue, and 2% (w/v) SDS. The amount required for the preparation of 2× and 5× SDS loading buffer were proportionately increased.

16. SDS running buffer (5×) – The buffer was prepared with 15.1g/L of Tris, 72g/L of glycine, and 5g/L of SDS. No pH adjustment was necessary. 1× SDS running buffer was obtained by diluting the 5× SDS running buffer to 1×.
17. Fixing solution – The solution was prepared with 25% (v/v) isopropanol and 10% (v/v) acetic acid.
18. Coomassie staining solution – The solution was prepared with 60mg/L brilliant blue R in 10% (v/v) acetic acid. The solution was filtered to remove any undissolved precipitates.
19. Transfer buffer – The buffer was prepared with 3g/L of Tris, 14.4g/L of glycine, and 20% (v/v) of methanol.
20. TBST – The buffer was prepared with 1× TBS (Bio-Rad) and 0.1% (v/v) of Tween 20.
21. Light Blue cathode buffer (1×) – The buffer was prepared with 10mL of 20× NativePAGE running buffer (Invitrogen), and 1mL of 20× NativePAGE cathode additive (Invitrogen), in a total volume of 200mL.
22. ANS stock solution (10mM) – The solution was prepared with 3.2mg/mL of 1,8-ANS NH_4 (Sigma) in methanol.
23. MB buffer – The buffer was prepared with 50mM of phosphate (KH_2PO_4/K_2HPO_4) system at pH 7.6, 200mM of KCl, 10% (v/v) of glycerol, 1mM of DTT, 100mM of potassium aspartate, and 100mM of potassium glutamate.
24. MBMg buffer (1.25×) – The buffer was prepared with 62.5mM of phosphate (KH_2PO_4/K_2HPO_4) system at pH 7.6, 250mM of KCl, 3.125mM of $MgCl_2$, 12.5% (v/v) of glycerol, 1.25mM of DTT, 125mM of potassium aspartate, and 125mM of potassium glutamate.
25. TBE buffer (10×) – The buffer was prepared with 108g/L of Tris, 55g/L of boric acid, and 9.3g/L of EDTA. 1× TBE buffer was subsequently prepared by diluting 500mL of 10× TBE buffer with 4500mL of distilled water.

26. Native DNA gel (10%) – The gel was prepared with 12.5mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 5mL of 10× TBE, 5mL of glycerol, 27.5mL of water, 40µL of TEMED, and 800µL of 10% (w/v) APS.
27. TBEM buffer – The buffer was prepared with 0.02% (v/v) β-mercaptoethanol in 1× TBE buffer.

2.2 Methods

2.2.1 Culturing, harvesting, and DNA/protein extraction of *M. burtonii*

2.2.1.1 Culturing of *M. burtonii* cultures

A volume of 10mL of fresh Na_2S was then injected into each MFM culture bottle and swirled for another four hours before injecting 1mL of *M. burtonii* culture into each bottle. The cultures were then incubated at 4°C and 23°C without shaking until they were ready for harvesting.

2.2.1.2 Genomic DNA extraction from *M. burtonii*

The cell pellet (from 50mL of culture) of *M. burtonii* was first centrifuged at 3,200×g at 4°C for 35min, before being resuspended in 1mL of XS buffer. In order to aid in the complete resuspension of the cells, the cell pellet and buffer were pass through a 19G syringe for ten times. Subsequently, the cell suspension was heated at 70°C for an hour before placing on ice for 30min. Cell debris was pelleted by centrifugation for 10min, at 4°C, at 14,000×g. The supernatant was collected and mixed with 1 volume of TE equilibrated phenol: chloroform: isoamyl alcohol (Sigma). The DNA-containing aqueous phase was collected after centrifugation at 12,000×g for 5min at 4°C. The DNA was subsequently precipitated with 3 volumes of absolute ethanol and 0.1 volume of sodium acetate (pH 5.2) at -20°C for 30min. The precipitated DNA was pelleted at 12,000×g, for 15min at 4°C, and washed with 0.5mL of 70% (v/v) ethanol. The DNA was again pelleted down at 12,000×g, for 10min, at 4°C before being resuspended in 300µL of EB buffer (Qiagen).

The extracted genomic DNA was subsequently treated with 100units of RNase One ribonuclease (Promega) for 30min at 37°C to completely eliminate the presence of RNA. After which, the treated genomic DNA was extracted with 1 volume of TE equilibrated phenol: chloroform: isoamyl alcohol (Sigma) and precipitated using ethanol precipitation methods as described previously. Finally, the purified, RNase-treated genomic DNA was resuspended in 100µL – 300µL of EB buffer (Qiagen), depending on the concentration that was required. The DNA suspension was stored in -20°C.

2.2.1.2.1 PicoGreen quantification of genomic DNA from M. burtonii

Quantification of the genomic DNA extracted from *M. burtonii* was performed with the PicoGreen dsDNA reagent and kit (Invitrogen), and the readings were measured with a fluorescence microplate reader (Perkin Elmer) with excitation and emission at 485nm and 535nm wavelength, respectively. The DNA quantifications were performed as per the manufacturer's recommendations. Briefly, the lambda DNA that was provided in the kit was used as the DNA standard. The lambda DNA was diluted down to 100µL of 0ng/mL – 1600ng/mL, and subsequently mixed with an equal volume of the picogreen dye. Genomic DNA samples were diluted by 64× before mixing with an equal volume of the picogreen dye. All quantifications were performed in triplicates. Estimation of the genomic DNA concentration was based on the standard curve generated by the lambda DNA, and multiplied by the dilution factor.

2.2.1.3 Non-denaturing lysis of *M. burtonii* for soluble native proteins

Cultures of *M. burtonii* cells were harvested by centrifugation at 3200×g for 35 min at 4°C, after the culture has reached an optical density of 0.3 – 0.4 at 620nm wavelength. The cell pellets were stored in -80°C until they were ready to be lysed by sonication. Each cell pellet was resuspended with 0.5mL of 2× non-denaturing extraction buffer, and topped up to 1mL volume with sterile distilled water. The cell suspension was subsequently sonicated on ice with a digital ultrasonic cell disruptor (Branson), with an instrument setting of 0.5s pulse on and off, at 30% amplitude, for 3 cycles of 30s each.

The cell lysate was then pelleted at 10,000×g at 4°C for 25min, with the supernatant which is the soluble protein fraction stored in -80°C until further uses.

2.2.1.3.1 Quantification of soluble M. burtonii protein fraction

Soluble protein fractions were quantified with the Quick Start Bradford 1× dye reagent (Bio-Rad), as per the manufacturer's instructions. BSA (Sigma) standards from 0 - 50µg/mL, at an interval of 10µg/mL, were prepared in 25µL of 1× non-denaturing extraction buffer, and subsequently made up to a final volume of 100µL with distilled water. All protein assays were performed in duplicates. Finally a calibration graph was obtained by plotting the absorbance of the BSA standards at 595nm against their corresponding concentrations. The protein concentration of a *M. burtonii* protein sample was subsequently determined based on the calibration graph.

2.2.2 Gene cloning

2.2.2.1 Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR) was performed in 50µL volume, with a final concentration of 1× Pwo mastermix (Roche Diagnostics), 0.4µM of forward and reverse primers (Geneworks) each, and 0.5µg of *M. burtonii* genomic DNA. The sequences of the primers used are illustrated in Table 1. The PCR cycling parameters were as follows: an initial denaturation phase at 95°C for 2min, and followed by 25 – 30 cycles of denaturation, annealing, and extension at 95°C for 30s, 55°C for 30s, and 72°C for 45s – 65s, respectively. The resultant PCR products were treated with a final extension phase at 72°C for 2min, before being kept in 4°C till further uses.

Primer name (Forward/Reverse)	Oligonucleotide sequences	Gene locus target	Remarks
TBPMbur1496F (Forward)	5'-CAC CAT GTC GGA ATC TAA TAT CAA AAT-3'	Mbur_1496	
TBPMbur1496R-His (Reverse)	5'-TCA TCA ATG GTG ATG GTG ATG ATG CAG AAG TCC CAT GTT GTC AA-3'	Mbur_1496	Resultant PCR product contains a 3' hexa-histidine sequence on the coding strand.
His-Mbur1496F long (Forward)	5'-CAC CAT GCA TCA TCA CCA TCA CCA TAT GTC GGA ATC TAA TAT CAA AAT-3'	Mbur_1496	Resultant PCR product contains a 5' hexa-histidine sequence on the coding strand.
His-Mbur1496F short (Forward)	5'-CAC CAT GCA TCA TCA CCA TCA CCA TAT GTC GGA ATA TAA TAT CAA AAT-3'	Mbur_1496	Resultant PCR product contains a 5' hexa-histidine sequence on the coding strand.
TBPMbur1496R (Reverse)	5'-TTA CAG AAG TCC CAT GTT GTC AA-3'	Mbur_1496	

Table 1: Oligonucleotides used in the cloning of *M. burtonii* genes.

Oligonucleotides were resuspended in sterile distilled water to 100µM stock solution. The stock solutions were subsequently diluted to 10µM working solution with sterile distilled water.

2.2.2.2 Purification of PCR products from agarose gel matrix

The desired PCR products were purified from agarose gel with the QIAquick Gel Extraction Kit (Qiagen), as per manufacturer's recommendations. Briefly, the gel slices containing the PCR products were dissolved, and the PCR products were subsequently captured and purified with silica-membrane-based spin columns. The purified PCR products were then eluted and stored in 4°C till further uses.

2.2.2.3 Cloning of PCR product into pET101/D-TOPO vector

Purified PCR products were cloned into the vector pET101/D-TOPO (Invitrogen) as per manufacturer's recommendation. Briefly, the purified PCR products were ligated into

the pET101/D-TOPO vector at 23°C for 5min, and the resultant ligation products were chemically transformed into TOP10 competent cells. The transformed cells were cultured on agar plates with a final concentration of 100µg/mL ampicillin as the selective agent.

2.2.2.4 Colony PCR screening

Cells from a single overnight colony were obtained, and resuspended in 50µL of sterile distilled water. The cell suspension was subsequently heated at 95°C for 5min. PCR was carried out in 25µL volume containing a final concentration of 0.4µM of the universal T7 forward and reverse primers, 2mM $MgCl_2$, 0.2mM deoxynucleotide mix, 1× GoTaq flexi green buffer, 0.625U of GoTaq DNA polymerase (Promega), as well as 10µL of the heat-treated cell suspension. The PCR cycling parameters were as follows: an initial denaturation phase at 95°C for 2min, and followed by 25 cycles of denaturation, annealing, and extension at 95°C for 30s, 50°C for 30s, and 72°C for 1min, respectively. The resultant PCR product were treated with a final extension phase at 72°C for 2min, before being kept in 4°C until further use.

2.2.2.5 Plasmid extraction

A volume of 100mL of overnight culture was used to harvest plasmid for subsequent downstream experimentation. Plasmid extraction was performed with HiSpeed Plasmid Midi Kit (Qiagen) as per manufacturer's recommendation. The extracted plasmids were aliquoted into smaller volumes, and stored in -20°C till further uses.

2.2.2.6 DNA gel electrophoresis

All PCR products were assessed for their amplicon size with DNA gel electrophoresis, with 1.5 – 2% (w/v) agarose gel. The electrophoresis was then performed under a TAE buffer system, for an hour at 100V. The gels were subsequently treated with ethidium bromide, and visualized under the Gel Doc system (Bio-Rad) to detect the PCR products.

2.2.2.7 DNA sequencing

All DNA sequencing was carried out by the Micromon DNA sequencing facility, Monash University (Australia). At least 1µg of plasmids was send to Micromon per sequencing reaction, with the appropriate sequencing primers. The resultant chromatogram was loaded into the sequence scanner software version 1.0 (Applied Biosystems), and assessed visually to eliminate any miscalled nucleotide.

2.2.2.8 Nanodrop quantification of plasmid DNA and protein

Both DNA and protein quantification were performed with a Nanodrop ND-1000 (Thermo Scientific) spectrophotometer. A volume of 3µL of DNA or protein sample was quantified, after zeroing the instrument with the appropriate buffer blank. In the case of double stranded DNA samples, one absorbance unit at 260nm wavelength correspond to a concentration of 50µg/mL. Protein concentration in mg/mL, were calculated based on equation 1.

Equation 1:

$$\text{Protein concentration} = [(\text{absorbance at } 280\text{nm}) \times (\text{Mw of protein})] / (\text{theoretical } \varepsilon \text{ of protein})$$

Mw refers to the molecular weight of the protein, while the theoretical ε of protein was determined based on the number of tryptophan, tyrosine and cysteine residues on the protein [141].

2.2.3 Custom gene synthesis

Custom synthetic gene synthesis and cloning of a HisTBP206 mutant gene into a pJexpress404 vector was outsourced to DNA2.0 in Menlo Park, California, USA. The HisTBP206 mutant gene was codon optimized [142] for expression in *Escherichia coli*, and contained a methionine to leucine substitution on the 31st amino acid. In addition, a hexa-histidine amino acid sequence, 'MHHHHHH' was added to the amino-terminal of

TBP to create the HisTBP206 mutant gene (Figure 4). Genes cloned into the ampicillin-resistance gene-containing pJexpress404 vector were controlled by the T5 promoter.

Figure 4: Amino acid sequences of the amino-terminal of TBP from *M. burtonii*, and its recombinant counterparts.

All recombinant TBPs contained the hexa-histidine sequences. MbHisTBP206 mutant differed from MbHisTBP206 where the former contained a Met → Leu substitution (in red).

Upon receiving the synthetic clone, the clone was transformed into a TOP10 competent cell strain and upscaled to a 100mL culture. The HisTBP206 mutant gene-containing plasmid was subsequently extracted, and the synthetic gene sequence verified to ascertain that the translated gene product is that of the expected HisTBP206 mutant. The sequencing primers used were pTR and pTF, which corresponds to 5'-TGGTAGTGTGGGGACTC-3' and 5'-CTCGAAAATAATAAAGGGAAAATCAG-3', respectively. Procedures for the transformation, culturing, plasmid extraction and sequencing were as described previously.

2.2.4 Competent cells preparation and transformation

2.2.4.1 Competent cells preparation

A volume of 0.5mL of overnight competent cell culture was inoculated into 20mL of LB broth. The inoculated broth was cultured 37°C at 250rpm, until the culture reached an optical density reading of 0.5 – 0.6 at 600nm wavelength. The culture was subsequently centrifuged at 4,000×g, at 4°C for 5min. The cell pellet was resuspended firstly with 0.8mL, and a further 9.2mL of 0.1M of cold sterile $CaCl_2$. The cell suspension was next incubated on ice for 20 min before being centrifuged at 4,000×g, at 4°C for 5 min. The cell pellet was then resuspended with 2mL of cold sterile 0.1M $CaCl_2$ containing 20% (v/v) glycerol, aliquoted into 100µL volume, flash frozen in liquid

nitrogen, and stored in -80°C until further use. The competent cell strains, TOP10, Rosetta (DE3), and BL21 were successfully prepared with this method, with the preparation of the Rosetta (DE3) strain requiring the addition of chloramphenicol during culturing.

2.2.4.2 Competent cells transformation

A volume of 1µL of plasmid was added to 100µL of competent cells that were already thawed on ice. The cells were subsequently mixed by tapping the tube gently, and incubated on ice for 20min. The cells were then heat shocked at 42°C for 45s, and placed immediately on ice for 2min. The cells were recovered with the addition of 250µL of SOC medium, and incubated at 37°C for 45min at 160rpm. Finally the recovered cells were cultured overnight at 37°C, on a LB agar plate containing the appropriate antibiotics. The final concentration of the antibiotics used for ampicillin and chloramphenicol were at 100µg/mL and 30µg/mL, respectively.

2.2.5 Production of recombinant proteins

2.2.5.1 Culturing of recombinant cells

Competent cells were freshly transformed for each batch of recombinant protein production, as described above. A single colony of transformed cells was cultured onto a LB agar plate containing the appropriate antibiotics, to ensure that the cells did indeed contain the recombinant plasmid. A single colony was subsequently selected and inoculated into 25mL of LB broth containing the appropriate antibiotics, and cultured at 160rpm overnight. Finally a volume of 2.5mL of the resultant suspension culture was then inoculated into 500mL of LB broth containing the appropriate antibiotics. All cultures were incubated at 37°C.

2.2.5.2 Induction and expression of recombinant proteins

Incubation of 500mL of cultures was performed in a 2L-capacity flask, with swivelling at 350rpm in 37°C. The culture conditions were subsequently adjusted to 20°C as the induction temperature and 250rpm once the optical density of the culture reaches 0.5

– 0.6 absorbance units at 600nm wavelength. The culture was allowed to grow for another 30 min before IPTG were added at a final concentration of 1mM, to initiate the induction of recombinant proteins. The culture was subsequently incubated overnight prior to harvesting.

Pilot studies to determine the suitable induction temperature and competent cell strain were performed on a smaller scale, with 10mL of culture in a 50mL-capacity falcon tube instead. Cultures induced with IPTG at 15°C, 20°C, and 25°C were incubated overnight, while those at 30°C and 37°C were incubated for 5 hours before harvesting.

Both the pilot and actual induction of recombinant proteins were monitored by collecting 0.75 absorbance unit of the cells, with and without IPTG induction, at the point of induction as well as prior to harvesting. The collected samples were centrifuged at 12,000×g for 10min at 4°C. The cell pellets were subsequently resuspended in 100µL of 2× SDS loading buffer, and stored in -20°C until they are ready for SDS-PAGE to evaluate the success of recombinant protein expression.

2.2.5.3 Purification of recombinant proteins

2.2.5.3.1 Non-denaturing recombinant protein extraction

Four liters of bacteria culture expressing recombinant proteins were harvested by centrifugation at 4470×g at 4°C for 8min. The cell pellet was then resuspended in 80mL of ice-cold *E. coli* lysis buffer. The cell suspension was subsequently applied through a French pressure cell press (Thermo Scientific) for four times to ensure thorough lysis of the bacteria cell. The resultant crude cell lysate was subsequently centrifuged at 20,000×g for 20 min at 4°C. The soluble protein-containing supernatant was collected, and the pellet was further washed twice with 70mL of ice-cold lysis buffer and centrifugation at 20,000×g for 20 min at 4°C per wash. The supernatants were collected in both these washes. A total volume of 200mL of soluble protein fraction was eventually collected, filtered with 0.45µm filters, and immediately proceeded to

purify the recombinant protein with His-tag purification. The final insoluble cell pellet was subsequently resuspended to homogenization with 70mL of urea buffer.

2.2.5.3.2 His-tag purification

A volume of 15mL and 12mL of Talon metal affinity resin (Clontech) at 50% (v/v) slurry, per 200mL of soluble protein fraction, were used in the purification of recombinant HisTBP206 mutant and HisTBP183, respectively. The resin was washed twice with ice-cold *E. coli* lysis buffer with centrifugation at 700×g at 4°C for 5min, before being resuspended in 5mL of the ice-cold lysis buffer. The soluble protein fraction was mixed with the resuspended resin at 4°C for 45min, before pelleting the protein-laden resin at 700×g at 4°C for 5min. The resin pellet was subsequently loaded into an empty Econo-Pac column (Bio-Rad). The resin was washed four times, with 50mL of ice-cold wash buffer per wash. Each individual wash was carried out by thoroughly mixing the resin with the wash buffer for 5min at 4°C, before eluting out the wash buffer. The recombinant protein was then eluted from the resin with a total of 15mL and 12mL of ice-cold elution buffer for HisTBP206 mutant and HisTBP183, respectively. The resin and elution buffer were mixed thoroughly for 5min at 4°C prior to eluting the recombinant proteins. The eluted recombinant proteins were stored in 4°C.

The eluted protein fractions were evaluated for the enrichment of the recombinant proteins with SDS-PAGE and Coomassie staining, by comparing with the soluble and insoluble protein fractions, and the flowthrough. A volume of 20µL of soluble and insoluble protein fractions, flowthrough, as well as eluted protein fractions was collected for evaluation. The latter was mixed with 5µL of 5× SDS loading buffer and 10µL of the mixture was loaded per lane in a SDS polyacrylamide gel. The rest (soluble and insoluble protein fractions, and flowthrough) were mixed with 20µL of 2× SDS loading buffer with 2.5µL of the mixture loaded into each lane.

2.2.5.3.3 Preparative size-exclusion chromatography

His-tag purified recombinant proteins that were eluted with 500mM of imidazole were pooled, buffer exchanged thrice with cold GF buffer, and subsequently concentrated with a 3KDa-cutoff Amicon centrifugal filter unit (Millipore) at 4,000×g at 4°C.

The concentrated recombinant proteins were loaded into a GF buffer-equilibrated HiLoad 26/60 superdex 75 column (GE Healthcare), using a Biologic system (Bio-Rad), in 4°C. The column was programmed for isocratic flow with 10mL of GF buffer at 1mL/min, followed by sample injection with 15mL of GF buffer at 0.5mL/min, sample run with 320mL of GF buffer at 1mL/min, and finally a wash at 2mL/min with 500mL of distilled water. All solutions were degassed and filtered before use. Fractions were collected during the run and analysed by SDS-PAGE for the presence of the recombinant protein, with 4µL of each protein fraction.

2.2.5.3.4 Storage of concentrated purified recombinant proteins

Selected fractions of recombinant proteins were first pooled and concentrated with a 10KDa-cutoff Amicon centrifugal filter unit (Millipore) at 4,000×g at 4°C. The proteins were subsequently washed with 10mL of GF buffer, and concentrated again with the centrifugal filter unit until the protein reaches a volume of approximately 1mL. The proteins were transferred into a new tube and centrifuged at 12,000×g at 4°C for 5min to remove any protein precipitate that might have resulted from the Amicon concentration process. The purified proteins were then aliquoted into 10µL per tube, and put on ice for 30min before flash freezing it in liquid nitrogen. The frozen purified proteins were finally stored in -80°C.

The purified protein were also quantified with the Nanodrop ND-1000 (Thermo Scientific) spectrophotometer, as per described in the previous section. Thereafter, 5µg of the purified protein were subjected to SDS-PAGE analysis to verify the purity of the purified recombinant protein.

2.2.6 Biophysical characterisations of recombinant proteins

2.2.6.1 Electrospray ionisation-mass spectrometric (ESI-MS) analysis

At least 10µg of proteins were analyzed per ESI-MS run, to accurately determine the molecular weight of the proteins. All the ESI-MS runs were outsourced to the Bioanalytical Mass Spectrometry Facility (BMSF), located within the University of New South Wales, Australia.

2.2.6.2 SDS polyacrylamide gel electrophoresis (SDS-PAGE)

Proteins were mixed at a ratio of 4:1 (volume: volume) in 5× SDS loading buffer, or 1:1 (volume: volume) in 2× SDS loading buffer, heated at 95°C for 5min and subsequently cooled to room temperature. Thereafter, the protein-SDS buffer mixture was centrifuged at 12,000×g for 5s, and the supernatants were loaded into a 4% stacking and 12% resolving SDS polyacrylamide gel. The Precision Plus Kaleidoscope protein standard (Bio-Rad) and SDS-PAGE Broad range protein standard (Bio-Rad) were used as protein molecular weight standards. The gel was allowed to run in 1× SDS running buffer system at 70V until it reached the resolving gel, when the voltage was increased to 150V. The electrophoresis was stopped when the dye front reached the bottom of the gel.

2.2.6.3 Native-blue PAGE

Native polyacrylamide gel electrophoresis was performed on a pre-casted 4% – 16% gradient Bis-Tris gel (Invitrogen). Protein samples were resuspended in NativePAGE sample buffers (Invitrogen) and loaded directly into the gel. The NativeMark unstained protein standard (Invitrogen) was used as the protein standards for molecular weight estimation. The native gel was run on a non-denaturing buffer system that uses 1 × NativePAGE anode buffer (Invitrogen) and 1 × Light Blue cathode buffer. Gel electrophoresis was performed with chilled buffers, and in a 4°C cold room, at 150V.

2.2.6.4 Coomassie staining of proteins in polyacrylamide gels

Proteins within the gel were fixed by soaking the gel in the fixing solution for 20min. Subsequently the gel was placed in the Coomassie staining solution for at least 30min. The gel was then washed with 10% (v/v) acetic acid until the background was clear. Finally the gel was rinsed twice with distilled water, and ready for image capturing with a scanner.

2.2.6.5 Western blot

Resolved proteins were transferred from a denaturing SDS polyacrylamide gel onto a polyvinylidene difluoride (PVDF) membrane with transfer buffer at 100V for an hour. The PVDF membrane (Bio-Rad) was subsequently blocked with 4% (w/v) non-fat milk in TBST for 40min, washed thrice, incubated for an hour with rabbit anti-TBP antiserum (1 in 10,000 dilutions in TBST), washed thrice and incubated for 45min with anti-rabbit (Bio-Rad) secondary antibody (1 in 40,000 dilution in blocking solution). The protein-of-interest was then detected using Immun-Stars HRP chemiluminescent substrates (Bio-Rad) and captured with the Intelligent Dark Box (Fujifilm LAS-3000).

The western blotting procedure for proteins resolved in a native-blue polyacrylamide gel was identical to that of denaturing SDS polyacrylamide gel, except the following as listed herein. Transfer of proteins from native-blue polyacrylamide gel to PVDF membrane was carried out with 1× NuPAGE transfer buffer (Invitrogen). In addition, a dilution of 1 in 2,000 and 1 in 20,000 of anti-TBP antiserum and anti-rabbit secondary antibody were used, respectively. The incubation period of the PVDF membrane with anti-TBP antiserum was also extended to overnight.

2.2.6.5.1 Anti-TBP antiserum generation

Generation of the polyclonal antiserum against *M. burtonii* TBP was outsourced to the Institute of Medical and Veterinary Science (IMVS), in Adelaide, Australia. Recombinant *M. burtonii* full length TBP with an N-terminal hexa-histidine tag was used as an antigen. Briefly a rabbit was immunized with the antigen emulsified in

Freunds Complete Adjuvant, and was then given three booster shots over the course of a few weeks. Exsanguination was subsequently performed, and the blood was allowed to clot to obtain the antiserum. The antiserum was aliquoted into smaller volumes, and stored in -20°C.

2.2.6.6 Analytical size exclusion chromatography

The analytical size exclusion chromatography was performed on a MB buffer-equilibrated superdex 75 10/300 GL column (GE Healthcare), using an AKTA FPLC system (GE Healthcare), with a 200µL injection loop. The column was first pre-equilibrated with 12mL of MB buffer. The injection loop was then flashed with 0.5mL of MB buffer, after the protein sample was injected into the column. Subsequently, the column was run with 36mL of MB buffer, at a rate of 0.45mL/min – 0.5mL/min. All buffer solutions were degassed and filtered before use. The low molecular weight gel filtration calibration kit (GE Healthcare) was used to calibrate the column.

Briefly, at least 1mg of purified recombinant proteins was resuspended in 500µL of the MB buffer. Protein molecular weight standards were resuspended in MB buffer as well, to a final concentration of 3mg/mL – 4mg/mL. All protein solutions were centrifuged at 12,000×g for 5min at 4°C to remove any precipitates prior to loading into the column.

The calibration graph for the column was obtained by plotting the logarithm of the molecular weight of the protein, against the partition coefficient Kav of the corresponding protein. The Kav of the proteins were calculated as per equation 2. Finally, the molecular weights of the recombinant proteins were determined based on the calibration graph.

Equation 2:

Kav = (elution volume – void volume) / (column volume – void volume)

2.2.6.7 Circular dichroism (CD) analysis

Circular dichroism were performed on a temperature-controlled CD spectropolarimeter (Jasco model J-810), using a 1mm pathlength quartz cuvette. Spectra of the purified proteins, at a final protein concentration of 0.3mg/mL in 20mM of sodium phosphate buffer system (Na_2HPO_4/NaH_2PO_4; pH 7.9), were measured under the program parameters as follows: sensitivity of 100mdeg, scan range of 260nm – 190nm, data pitch of 0.1nm, a continuous scan mode with a scanning speed of 50nm/min, response time of 1s, bandwidth of 1nm, and in triplicates. An equivalent volume of GF buffer (same volume as the volume of protein used to obtain a concentration of 0.3mg/mL) in 20mM of sodium phosphate buffer was used as the blank for CD measurement.

Protein melting analysis based on CD methods was performed using the program parameters as follows: wavelength of 208nm, temperature range of 4°C - 80°C with a slope of 2°C/min, data pitch of 0.1nm, no delay time, sensitivity of 100mdeg, response time of 1s, and a bandwidth of 1nm. The raw data obtained in a CD spectra measurement was subsequently converted to the molar differential extinction coefficient Δε, in $M^{-1}cm^{-1}$ based on equation 3 [143]. This conversion allows for the comparison of CD spectra of different proteins to be better appreciated.

Equation 3:

$$\Delta\varepsilon = (\theta \text{ of protein} \times \text{MRW of protein}) / (L \times \text{Concentration of protein in mgmL}^{-1} \times 3298)$$

The abbreviations θ, and L refers to the measured CD value in millidegrees, and pathlength of cuvette in millimetres, respectively. MRW refers to the mean residue weight of protein, which was determined by dividing the molecular weight of proteins, in Daltons, with the resultant value of number of amino acids minus one [143].

2.2.6.8 ANS fluorescence spectrometry

ANS fluorescence spectrometry was performed with a luminescence spectrometer LS50B (Perkin Elmer) that was connected to a peripheral peltier temperature controller unit. A volume of 2mL of protein at a final concentration of 125µg/mL was loaded into a clean quartz cuvette, placed into the LS50B, and stirred for 5min to allow the temperature within the cuvette to equilibrate to the required temperature as set in the instrument. 1,8-ANS NH_4 (Sigma) solution was subsequently added into the cuvette, to a final concentration of 25µM. The solution was further stirred for another 30min within LS50B to ensure thorough mixing.

Florescence spectra readings were performed with 5nm slit for both excitation and emission slit size, with a scan rate of 100nm/min, at an excitation wavelength of 380nm and an emission wavelength range of 400nm – 500nm. All florescence spectra readings were performed thrice, and averaged automatically by the instrument software. ANS florescence melting was performed from 3°C – 50°C, with an excitation and emission wavelength of 380nm and 465nm, respectively, with a 0.5°C temperature step-up with zero second equilibration, and a 5nm slit size for both excitation and emission wavelength. All florescence intensities were blanked, and subsequently normalized to per micromolar concentration of protein.

2.2.7 Functional characterisation of recombinant proteins

2.2.7.1 Electrophoretic mobility shift assay

2.2.7.1.1 Annealing of complementary oligonucleotides

Single stranded oligonucleotides (Geneworks) were resuspended to a concentration of 100µM with 50mM NaCl-containing TE buffer. Complementary single stranded oligonucleotides were mixed with an equal molar concentration of both oligonucleotide sheets. Specifically, volumes of 100µL of 100µM of single stranded oligonucleotide were mixed to produce 200µL of 50µM double stranded oligonucleotides upon annealing. The annealing reaction was performed in a thermal cycler (Hybaid) by heating the complementary oligonucleotide mixture to 95°C for

5min, cooling it down to 94°C for a minute, and further cooling to 28°C at a rate of 2°C per step with a holding duration of 1min per step. The resultant annealed double stranded oligonucleotides were stored in -20°C.

2.2.7.1.2 Radioactive 5'-end labelling and precipitation of double stranded oligonucleotides

The radioactive labelling reaction was performed in 50µL volume, with 50pmol of double stranded oligonucleotides, 40pmol of ATP [γ-^{32}P] (111KBq/pmol; PerkinElmer), T4 polynucleotide kinase (Promega) and its reaction buffer. The labelling reaction was carried out for two hours in room temperature, before 2µL of 0.5M of EDTA solution was used to stop the reaction. The labelled oligonucleotide probes were subsequently ethanol precipitated. Briefly, 5µL of 3M of sodium acetate (pH 5.2) and 100µL of cold absolute ethanol were added to the labelling reaction, and was kept in -80°C overnight. The precipitated labelled oligonucleotides were then pelleted down, washed with 1mL of cold 70% (v/v) ethanol, and pelleted down again before air-drying it. The labelled oligonucleotides were finally resuspended in 50µL of 10mM of Tris buffer at pH 8.5, to give an approximated concentration of 1µM solution. It was subsequently stored in -20°C till further uses.

2.2.7.1.3 DNA-protein interaction and gel shift assay

Recombinant protein isoforms were diluted to a working concentration of 320pmol/µL with MB buffer. Subsequently, the DNA-protein interaction reaction was performed in 20µL volume on ice for an hour, with 2pmol of the ^{32}P-labeled double stranded oligonucleotides, 1µL of the recombinant protein, and 16µL of 1.25× MBMg buffer. MB buffers, instead of proteins, were added to the 'no protein control' reactions. In the event that denatured recombinant proteins were used, the proteins were denatured by heating at 95°C for 10min before adding it to the interaction reaction mix.

Gel shift assay was performed with 20µL of the interaction reaction mix per well, on a 4°C pre-cooled 10% native DNA gel, and run on a TBEM buffer system at 4°C, at 80V.

Thereafter, the gels were blot-dried on a filter paper, and wrapped in a clear transparent film. Subsequently, the gels were exposed overnight at -80°C on a Hyperfilm ECL (GE Healthcare), with the aid of a BioMax MS intensifying screen (PerkinElmer). Finally the film was developed with the film developer (Kodak) and fixer (Kodak), by rinsing the film in the developer, water and fixer. The film was then scanned to capture a softcopy of the autoradiograph.

2.2.7.2 Recombinant protein-DNA pull-down assay

2.2.7.2.1 Small genomic DNA fragments – sonication and purification

Genomic DNA from *M. burtonii* was first topped up with sterile distilled water, to a final volume of 1mL before sonication. The sonication procedure was performed on ice with a digital ultrasonic cell disruptor (Branson), with an instrument setting of 0.5s pulse on and off, at 50% amplitude, for 20 cycles of 30s each. After sonication, the genomic DNA was vacuum dried in a DNA speedvac (Savant) until approximately 200µL of volume was left. Subsequently, all the sonicated DNA were run on a 2% (w/v) agarose gel electrophoresis and DNA regions pertaining to 100 – 400 base-pairs were excised and collected. The genomic DNA fragments were extracted from within the excised gel matrix with the QIAquick Gel Extraction Kit (Qiagen), as per manufacturer's recommendations. A total of four silica spin columns were used to maximise DNA recovery, and 50µL of TE buffer was used to elute the DNA in each column. The purified 100bp – 400bp genomic DNA fragments were stored in -20°C.

2.2.7.2.2 Promoter DNA pull-down

The recombinant HisMbTBP was mixed with genomic DNA fragments (100bp – 400bp) in a 200µL volume, with 6.4nmol of protein, 750ng of DNA fragments, in 1× MBMg buffer. The mixture was incubated on ice for 45min. Concurrently, 200µL of the Talon resin slurry (50% slurry; Clontech) was pre-equilibrated with 500µL of 1× MBMg buffer. The resin was then pelleted with 700×g for 5min at 4°C.

All the protein-DNA mixture were added to the resin pellet, and allowed to bind to the resin for 15min on ice, before centrifugation at 700×g for 5min at 4°C. Thereafter, the protein-DNA complex-laden resin was washed twice with 0.5mL of 1× MBMg buffer per wash, to remove unbound DNA. Elution of the protein-DNA complex was achieved with 400µL of 1× MBMg buffer that was supplemented with 0.5M of imidazole.

2.2.7.2.3 Pull-down DNA purification, quantification and sequencing

The protein-DNA complex was added with 20µL of sodium acetate (pH 5.2) before being subjected to DNA extraction and purification with the QIAquick PCR purification kit (Qiagen). The purification was conducted as per manufacturer's instructions. The DNA was eluted with 40µL of TE buffer, and quantified with the PicoGreen dsDNA reagent and kit (Invitrogen) as per Chapter 2.2.1.2.1. The purified DNA, corresponding to a minimum amount of 75ng, was vacuum-dried and despatched for sequencing as dried DNA pellet. Sequencing was carried out with Illumina sequencing technology, by Beijing Genomics Institute (BGI) in Hong Kong.

2.2.7.2.4 Statistical analysis of sequenced data

Sequencing reads were mapped onto the genome of *M. burtonii*, in sectors of 50bp in length by BGI. Sectors containing less than 50 mapped reads were disregarded. Subsequently a ratiometric analysis [144] was performed to calculate the read ratio for each sector. Briefly, sequencing reads from each isoform were divided by that of the mock control for the same sector. Statistical Z-test analysis was then performed (equation 4; Appendix C) to obtain a Z-score for the read ratios of sectors for each isoform. Based on the Z-score, a sector was considered significant only if the p-value was less than 0.005 or 0.001.

Equation 4:

$$\text{Z-score}_{(\text{sector n})} = [\text{read ratio}_{(\text{sector n})} - \text{mean of read ratio}_{(\text{for all sectors})}] \,/\, [\text{standard deviation of read ratio}_{(\text{for all sectors})}]$$

CHAPTER III

Bioinformatic analysis of Mbur_1496 and its gene product

3.1 Introduction

The gene locus tag Mbur_1496 had been annotated as having an ER3 status [88], which means that the gene product is homologous to a functionally verified protein, albeit the sequence identity between the two proteins are less than 35%. Mbur_1496 had also been annotated to be a TATA-box binding protein (TBP).

Bioinformatics analysis of Mbur_1496 and its gene product is an important initial step in any experimental project, aimed at gaining insight into the putative structure, functions and possible regulations. Such data would also complement the subsequent experimental studies of its gene product that are described in following chapters.

This chapter describes the dissection of the gene locus Mbur_1496 and its gene product bioinformatically, by analysing its genetic and amino acid sequences, comparing it with other TBPs, assessing its phylogenetic relationship with other TBPs, and finally carrying out predictive modelling of the protein tertiary structure.

3.2 Results

3.2.1 Gene locus tag Mbur_1496 and its gene product

The gene locus tag Mbur_1496 (621 nucleotides in length) has been annotated as encoding a 206 amino acid residue-long TBP from *Methanococcoides burtonii DSM 6242* [88]. The encoded protein product of Mbur_1496, abbreviated as MbTBP, has a theoretical molecular weight and isoelectric point of 22.3KDa and 4.5, respectively.

The gene Mbur_1496 is flanked by an upstream condensing subunit ScpB gene (Mbur_1497) and a downstream glycosyl transferase gene (Mbur_1495). As determined by Campanaro *et al*. (2010) [145] using microarrays, the transcript levels of Mbur_1495 and Mbur_1496 were similar to each other. Genes are assigned to the same operon when they are adjacent in the genome, in the same orientation, and have similar RNA expression levels [145]. Therefore both Mbur_1495 and Mbur_1496

were deemed to be part of the same operon [145]. This conclusion is further substantiated by the lack of a TATA-box between the two genes.

3.2.2 Mbur_1496 and MbTBP contain a pair of terminal repeat and core sequence identity

Nucleotide and amino acid sequences of Mbur_1496 were aligned using ClustalW version 2.0 [146], to establish any sequence or domain repeats within the gene and its encoded gene product.

The nucleotide sequence of Mbur_1496 contains a pair of sequence repeats. They are located directly at the 5'-end of the gene – spanning a total of 129 nucleotides separated by a nine nucleotide spacer. The repeats have only four mismatch sites, which is equivalent to approximately 93% nucleotide sequence identity between the pair of repeats. Its amino acid counterpart registers 90% residue conservation with only two mismatched sites (Figure 5). The repeats are here denoted as RepA and RepB (Figure 5B).

The Mbur_1496 also displays a high level of sequence identity between its two halves – characteristics typical of a TBP [105,147]. There is an approximately 50% nucleotide sequence identity and 37% amino acid sequence identity between the two halves (Figure 5).

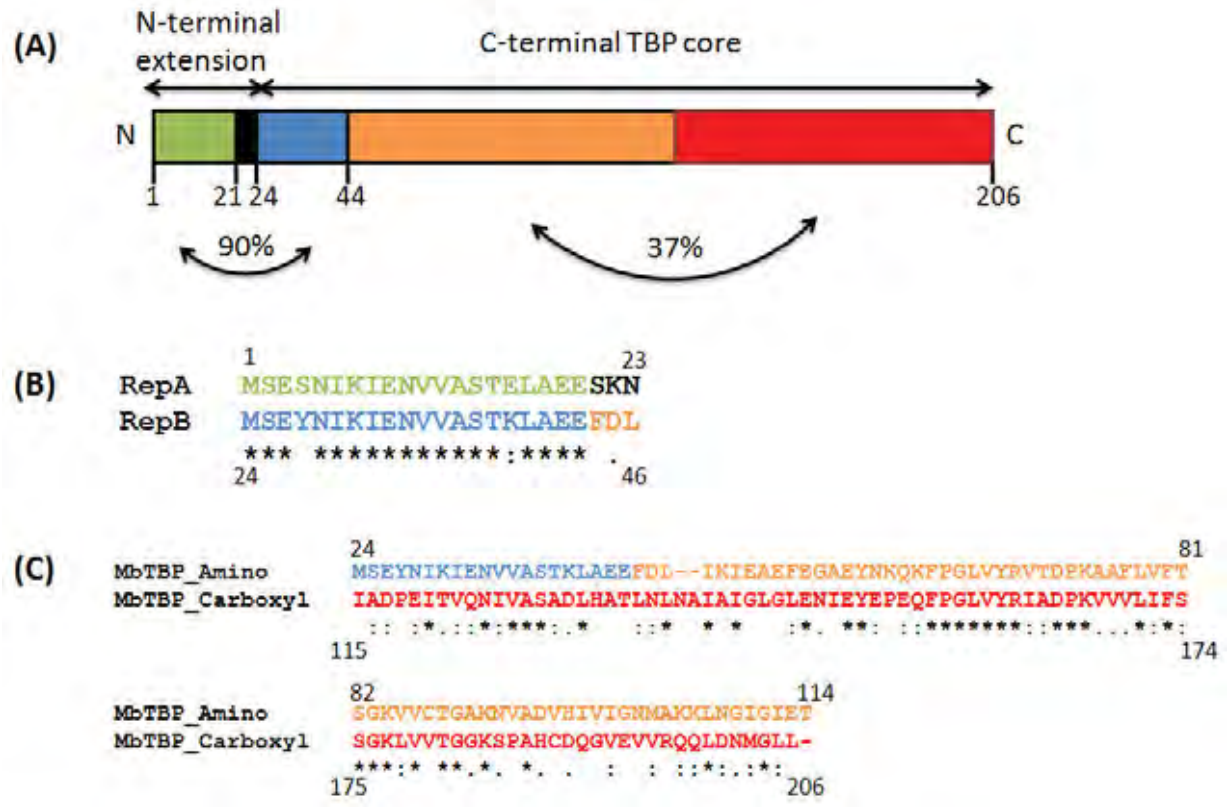


Figure 5: Structural features of MbTBP (TBP from *M. burtonii*).

(A) A schematic representation detailing the structural features of MbTBP. Numbering at the bottom refers to the position of the amino acid residues. (B) and (C) Pairwise alignment of the amino acid sequences of the N-terminal repeats, and the amino against carboxyl halves of MbTBP, respectively. (*) denotes identical alignment of residue; (:) denotes strong conservation of physico-chemical properties of mismatch residues; (.) denotes weak conservation of physico-chemical properties of mismatch residues.

3.2.3 Possibility of a short version of MbTBP

A striking feature of the MbTBP is the presence of a second methionine that is just 23 residues away from the N-terminal methionine (Figure 5B). This second methionine residue, if translation was initiated at this AUG codon, would generate a protein product that is 183 amino acid residues in length, which is similar in size to most archaeal TBPs.

A closer inspection of the nucleotide sequences of Mbur_1496 revealed a couple of interesting details. Firstly, there is no TATA-box signature between the first and second

AUG codon (Figure 6), which means that a shorter version of an MbTBP would likely be translated from the same mRNA transcript as the full length MbTBP. Secondly, and more importantly, the presence of a stretch of purine-rich nucleotides, just nine nucleotides upstream of the second AUG codon (Figure 6) signifies the existence of a putative ribosome-binding site.

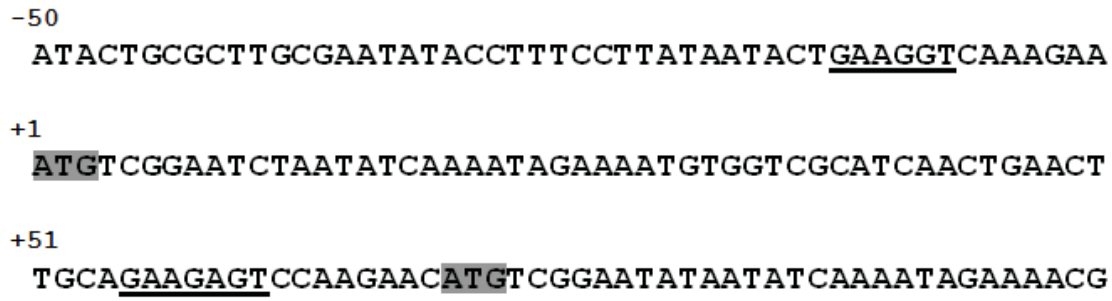


Figure 6: DNA nucleotide sequences of Mbur_1496.

The start codons are highlighted, while the Shine-Dalgarno (S-D) sequences are underlined. The full length MbTBP is translated from a near perfect S-D motif, while that of a shorter version of MbTBP is from an imperfect S-D motif.

M. burtonii has three copies of genes encoding for the 16S rRNA: Mbur_R0023, Mbur_R0039, and Mbur_R0055. All three 16S rRNA genes harbour the same 3' terminal nucleotide sequences of UCACCUCCU (3'-OH), which would complement a Shine-Dalgarno (S-D) sequence of 5'-AGGAGGUGA on an mRNA. The S-D motifs that control the translation of the full length MbTBP is a near perfect GAAGGU sequence, whereas that of the shorter MbTBP is an imperfect GAAGAGU sequence (Figure 6). In *Halobacterium salinarum*, translation was still possible even when the S-D sequence was highly altered [148]. However, the efficiency of translation was significantly reduced.

3.2.4 MbTBP contains a typical TBP core, and an atypical N-terminal repeat

Amino acid residues from 40 randomly chosen TBPs, including MbTBP were aligned and compared, using ClustalW version 2.0 [146]. This array of TBPs was phylogenetically diverse, with four from the Eucarya domain of life, 26 from the Euryarchaeota phylum, and the rest from the Crenarchaeota, 'Thaumarchaeota', and 'Nanoarchaeota' phyla.

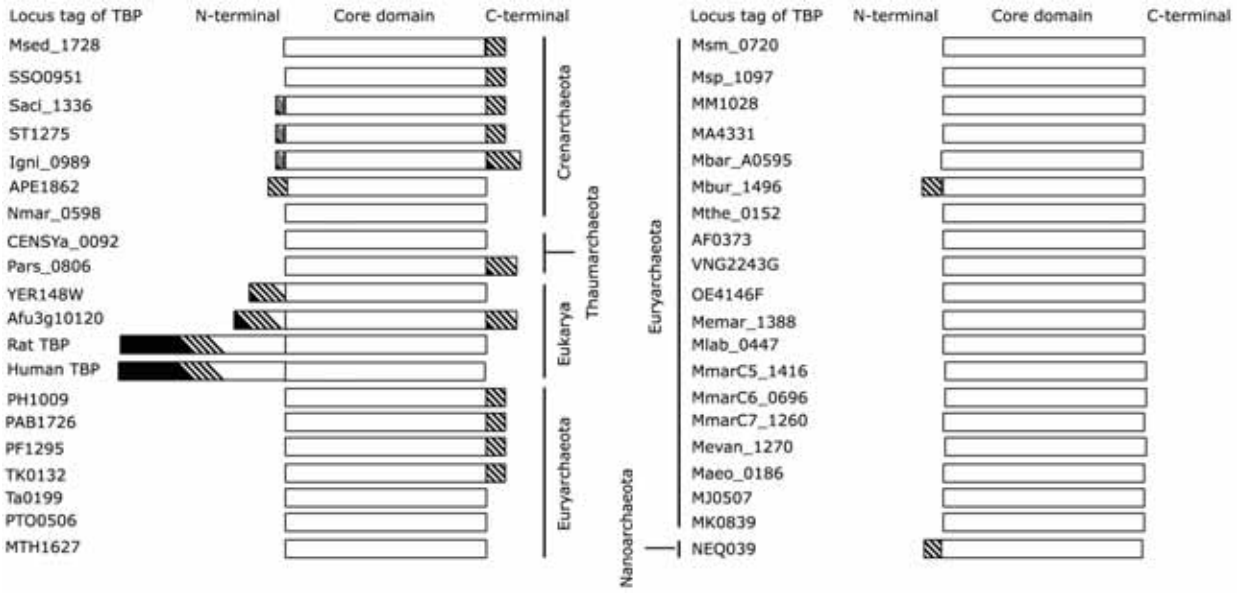


Figure 7: Graphical representation of the alignment of 40 TBPs based on amino acid sequence.

All TBPs are aligned according to the core domains of approximately 180 amino acids in length. MbTBP displays an unusually long N-terminal segment among the analysed species in the Euryarchaeota phyla. All the C-terminal tails are acidic i.e. contain at least a continuous stretch of four acidic residues, except for Afu3g10120 where no acidic C-terminal is present. Each locus tag and its corresponding organism identity are given in Appendix A and B for archaeal and eucaryal TBPs, respectively.

Alignment results indicate that all TBPs contain a central core domain approximately 180 amino acids in length (Figure 7). This core domain is conserved among all the 40 analysed TBPs, with approximately 11% amino acid sequence identity match and an additional 17% of the core having a conserved mismatch. Specifically, all the 38 known DNA-interacting sites, based on crystal structures of the TBP-promoter interacting complex from yeast [114], *Arabidopsis thaliana* [113] and *Pyrococcus woesei* [112], falls within the central core domain. MbTBP contains 34 of the 38 known DNA-interacting residues (Figure 8). Interestingly, none of the DNA-interacting sites occurs on the α-helix structures of the protein, and sequence identity in the α-helix structures between ScTBP, AtTBP, PwTBP and MbTBP is minimal (Figure 8).

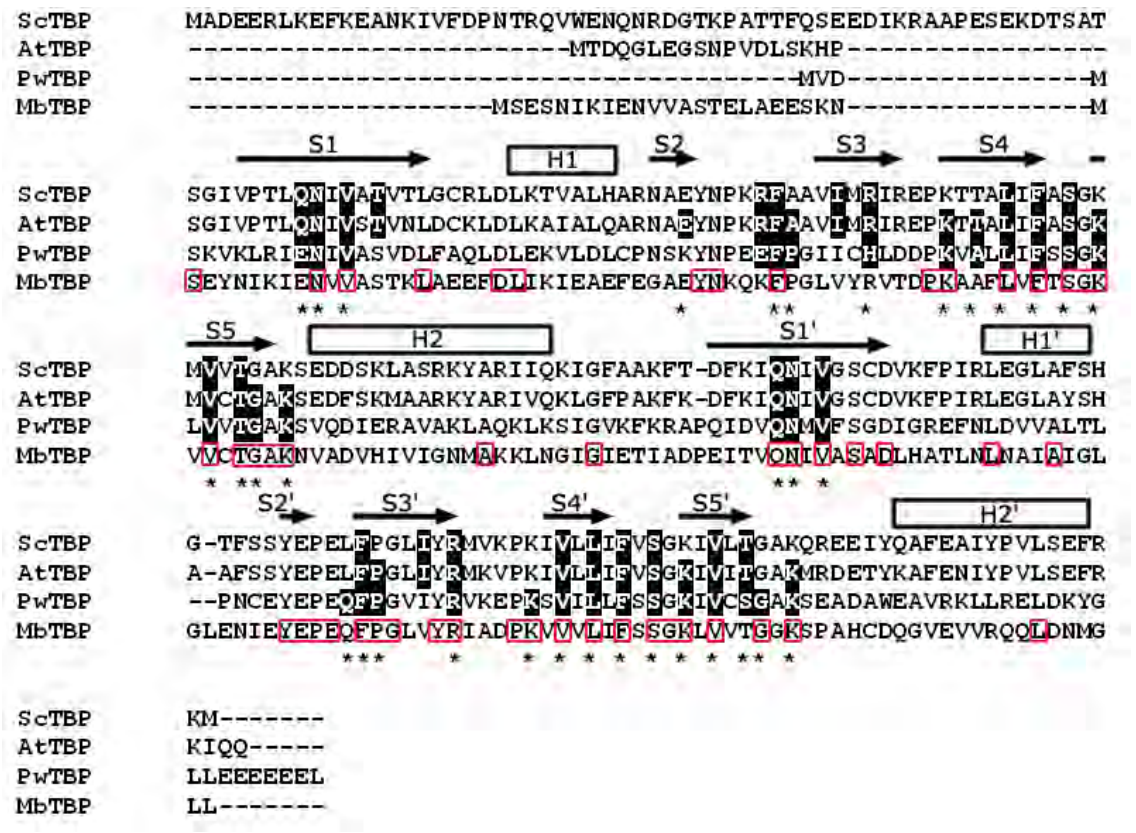


Figure 8: Multiple sequence alignment of TBPs against MbTBP.

Amino acid residues highlighted in black are known DNA-interacting sites, elucidated from the crystal structures of the TBP-promoter complex from its respective organism. (*) denotes residues on MbTBP that are identical to at least one of the known DNA-interacting residues from ScTBP [114], AtTBP [113] or PwTBP [112]. Boxes in red indicate the residues that are conserved among all the four different TBPs. Arrows and boxes in black refers to β-sheet and α-helix secondary structure, respectively. Labeling of the sheets and helices, as well as the positioning of these secondary structures were based on the crystal structure from PwTBP [126].

Among all the 84 archaeal TBPs in the transcription COG 2101 under the Integrated Microbial Genome (IMG) web database [149], about 14% of the archaeal TBP contains a long N-terminal segment. This includes TBPs from various archaeal species and phyla, such as *Aeropyrum pernix* (Crenarchaeota), *Methanococcoides burtonii* (Euryarchaeota), *Nanoarchaeum equitans* ('Nanoarchaeota'), and *Candidatus Korarchaeum cryptofilum* ('Korarchaeota') (Figure 9). Unexpectedly, MbTBP contains the longest N-terminal segment within the Euryarchaeota. In addition, among the

archaeal TBPs with long N-terminal segment, none of the TBPs, except MbTBP harbours N-terminal repeats.

Figure 9: Alignment of archaeal TBPs, showcasing the different length of the N-terminal segments.

(#) denotes the position of the average N-terminal start residue, based on 84 archaeal TBPs within the transcription COG2101. TBP from the locus tag VNG6438G is a typical example where the protein starts at position #. 81% of the archaeal TBPs contain an N-terminal start site in the region ±3 residues from #. Locus tags of the TBPs with the top 10 longest N-terminal segment are listed above VNG6438G. The shortest TBP is encoded by the locus tag Tpen_0293. The locus tag and its corresponding organism identity are given in Appendix A.

3.2.5 MbTBP core exhibits a typical saddle-like tertiary structure

Predictive homology modeling of MbTBP was carried out using the Swiss-Model workspace [150], to determine the tertiary structures of the full length protein. The homology model of the core region of MbTBP, ranging from the 23rd to the 206th amino acid, was generated by the Swiss-Model server (Figure 10). As expected, the MbTBP contains all the secondary structures and a saddle-like tertiary structure *in silico*, typical of TBPs [113-115,126]. TBP-essential structural features such as the protruding phenylalanine side chains [114] were also observed in the predicted model. Nevertheless, the lack of modelled structural information of the RepA sequences (1st to 22nd amino acid) limits any useful comparison and interpretation to the core of MbTBP, rather than the full protein.

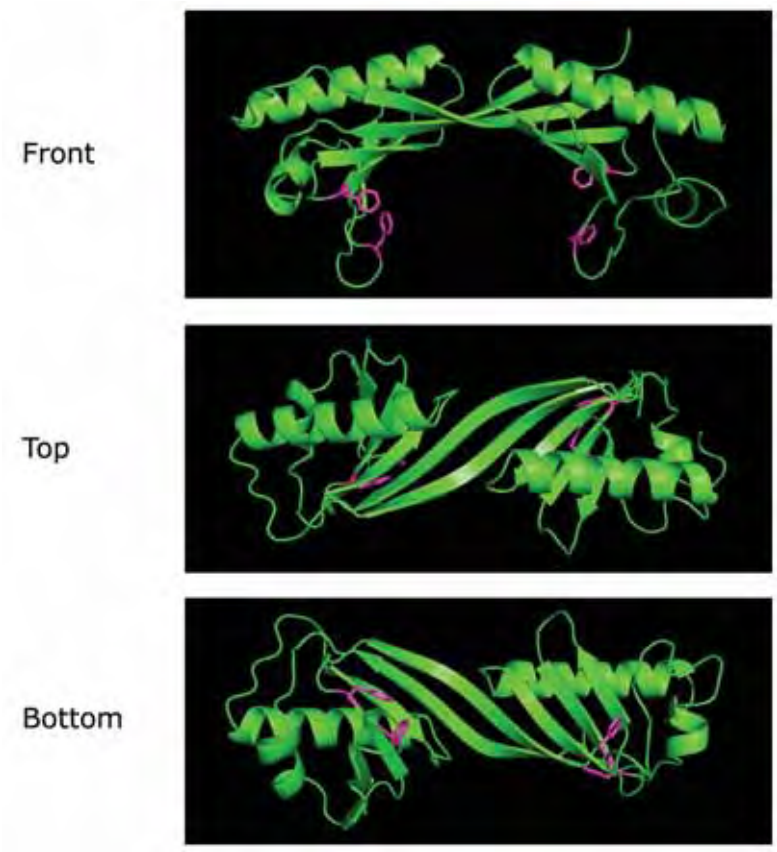


Figure 10: Different views of the homology model of MbTBP, spanning from the 23rd – 206th amino acid residues.

MbTBP has a typical saddle-like tertiary structure. The essential phenylalanine side chains, which penetrate between nucleotide bases and form kinks along the promoter [114], are highlighted in purple.

3.2.6 Eight phylogenetically distinct groups of TBP

A total of 99 full length TBPs from 56 archaea and 15 eucaryotes, where some of these organisms contain multiple TBP variants, were used to examine the evolutionary relationship of TBPs. The TBPs were aligned and bootstrapped with ClustalW version 2.0 [146], and the resultant phylogenetic tree (Figure 11 and 12) was displayed with Dendroscope software [151]. Eight phylogenetically distinct groups of TBPs were immediately evident and are highlighted in Figure 11. The group Methanosarcinales, to which *M. burtonii* belongs, was selected for further analysis.

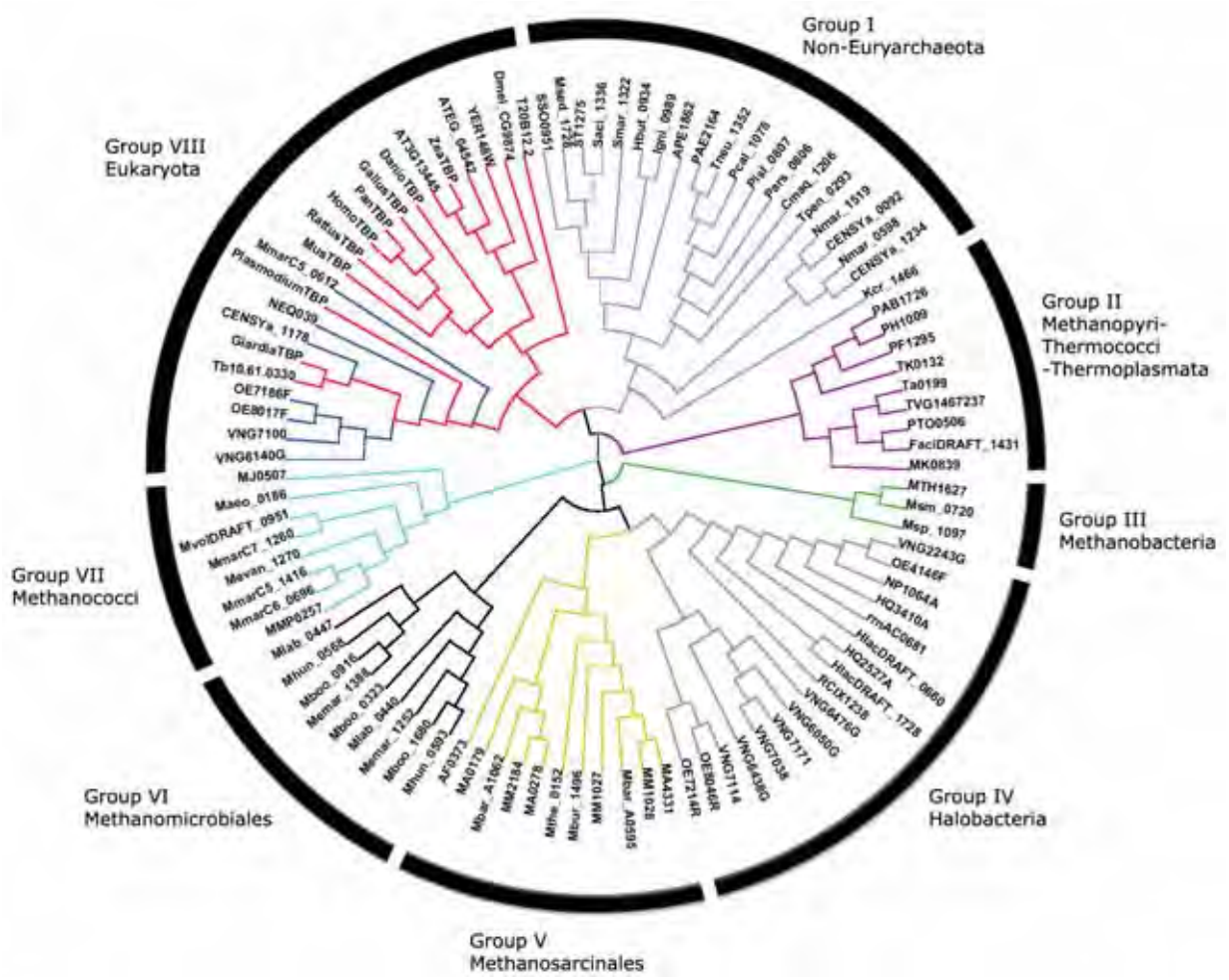


Figure 11: A circular cladogram of the TBP phylogenetic tree, based on 99 full length TBPs.

The TBP cladogram can be divided into eight distinct groups, where MbTBP belongs to Group V along with TBPs from *Methanosarcina* species. Horizontal gene transfer of TBP from eukaryotic to archaeal system is evident as illustrated in Group VIII. Within Group VIII, red lines denote TBP of eukaryotic origin, while blue lines denote TBP of archael origin. Each locus tag and its corresponding organism identity are given in Appendix A and B for archaeal and eucaryal TBPs, respectively

3.2.6.1 Group V: Methanosarcinales

The TBPs from the Methanosarcinales group form a phylogenetically robust group, as evident from the high bootstrap values (Figure 12A).

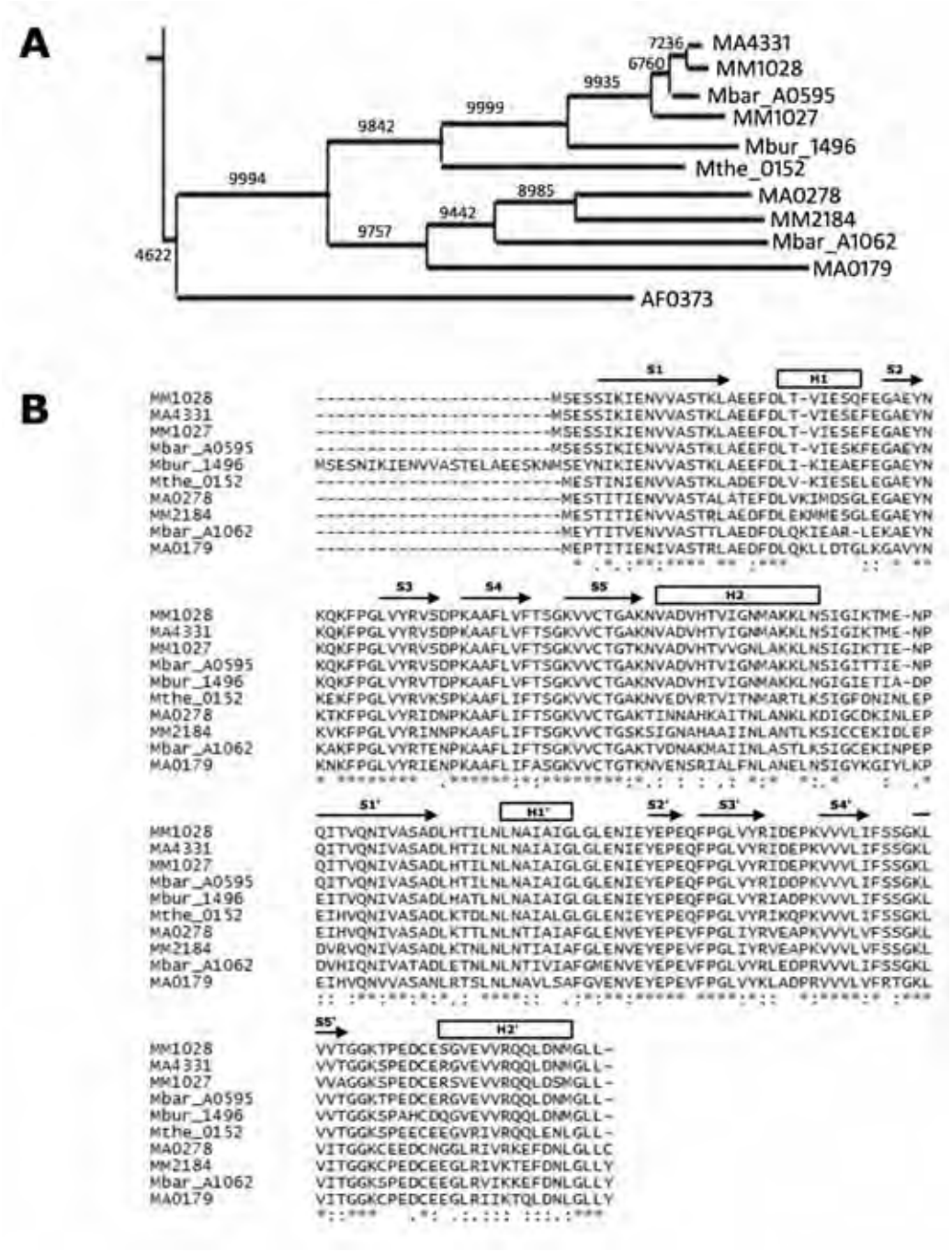


Figure 12: Phylogenetic and multiple sequence analysis of Group V TBPs.

The various TBPs are denoted by their gene locus tag. (A) Close up of the phylogenetic tree of Group V TBPs, with bootstrap values. (B) Multiple sequence alignment of Group V TBPs. Arrows and boxes in black refer to β-sheet and α-helix secondary structure, respectively. Labelling of the β-sheets and α-helices, as well as the positioning of these secondary structures were based on crystal structures from PwTBP [126]. (*) denotes identical alignment of residue; (:) denotes strong conservation of physico-chemical properties of mismatch residues; (.) denotes weak conservation of physico-chemical properties of mismatch residues.

Interestingly, MbTBP which has a psychrophilic origin forms a cluster with TBP (gene locus tag Mthe_0152) from the thermophilic *Methanosaeta thermophila*, as well as TBPs (gene locus tag MA4331, MM1028, Mbar_A0595, MM1027, MA0278, MM2184, Mbar_A1062, and MA0179) from the three mesophilic *Methanosarcina* species. In fact, these 10 TBPs share approximately 48% sequence identity in their TBP core domains (Figure 12B). The majority of the mismatches fall in the non β-sheet secondary structures of the TBP core domains.

3.3 Discussion

The hallmarks of a TBP include the following: the presence of an approximately 180 amino acid residues long TBP core containing a somewhat symmetrical topology between the amino and carboxyl halves [105,147], saddle-like tertiary protein structure [112-114], presence of two pairs of highly conserved phenylalanine residues on the two stirrups of TBP [112-115,126], as well as a high degree of conservation of DNA-interacting amino acid residues [112-114]. MbTBP contains all these hallmarks. However, MbTBP also has unique features as well. MbTBP is the only TBP among the analysed 84 full length archaeal TBPs that contains a pair of sequence repeats (Figure 5B). As a consequence, MbTBP is approximately 23 amino acid residues longer than most archaeal TBPs.

The presence of the RepA and RepB repeat sequence are intriguing. It might involve a partial duplication of a shorter ancestral *tbp* gene, thereby elongating the *tbp* gene to its current full length of 206 amino acid residues, as well as containing a pair of sequence repeats (RepA and RepB).

A high degree of identity between the RepA and RepB sequences could indicate that the partial gene duplication is a recent event. Concurrently, it could also be that the function of RepA is essential for the survival of *M. burtonii* so that any changes to its sequences are selected against. The fact that a pair of N-terminal repeats were also recently found in a TBP from the draft genome of *Methanogenium frigidum*, a

psychrophilic archaea in Ace Lake, could lent creditability to the functional importance of the N-terminal repeats in cold adaptation.

Similar to all other TBP cores, the MbTBP core contains all the necessary α-helices and β-sheets to form a saddle-like protein tertiary structure. However, the predicted model lacks the structural prediction for the long N-terminal segment as defined by the RepA sequence, thus rendering the structure somewhat incomplete. Nevertheless this region of the protein could have taken up a β-sheet secondary structure, since the RepB sequence forms a predicted β-sheet and the RepA sequence is about 90% identical to it. The RepA could also be structurally different from RepB. Similar to the N-terminal domain of yeast TBP, the structure of RepA could also be dependent on the interaction status of the TBP core with another TBP [152] or DNA [153-154]. Although a small number of archaeal TBPs also contain a long N-terminal segment, no studies have been reported thus far to elucidate its structure or role.

Eukaryotic TBPs are generally characterised by a relatively long N-terminal segment, albeit of various sequence and length. The long N-terminal domains are not essential to the promoter binding functionality of the protein [155-156] and in the case of yeast, viability of the organism [157]. Nonetheless, recent studies have shown that the long N-terminal domain of yeast TBP does actively play a role in modulating the TATA-box binding to TBP, through an autoinhibitory mechanism [121-122]. Specifically, competition between the N-terminal segment and DNA for the DNA-binding saddle of TBP reduces the binding affinity of the TBP for DNA [121-122]. In the mouse, the long N-terminal of the TBP goes even as far as to confer protection to the mouse placenta against maternal rejection, by regulating the transcription of a placental beta2m protein [120]. Overall, the functions of the long N-terminal segment appear to be diverse, and organism-specific.

The structure-function relationship of the secondary structures within the TBP core are apparent after comparison with the solved crystal structures of TBPs of yeast [114], *Arabidopsis* [113], and *Pyrococcus* [112], with that of TBP sequence of *M. burtonii* (Figure 8).

The DNA-interacting amino acid residues are predominantly located on the β-sheets of TBP. In addition, conservation of amino acid residues among TBPs of phylogenetically distant organisms is also predominantly along the β-sheets. A large proportion of this conservation relates to the DNA-interacting residues. These conclusions hint at the possibility that the β-sheet structures within the TBP core have a crucial role in TBP-TATA-box binding, which therefore dictates a high degree of amino acid residue conservation across eukaryotic and archaeal systems, since TATA-boxes are highly conserved.

The comparison of the TBPs of phylogenetically distant origins also hints at a structural role of the α-helices in maintaining the overall tertiary structure of MbTBP. Since there is minimal sequence identity between these phylogenetically distant TBPs (Figure 8), it would indicate that the ability of the H1, H1', H2, and H2' regions to form a helical structure is much more important than the actual sequence itself. In addition, TBPs of closely related archaea (*M. burtonii*, *Methanosarcina acetivorans*, *Methanosarcina bakeri*, *Methanosarcina mazei* and *Methanosaeta thermophila*) that have diverse natural growth temperatures, have demonstrated that the α-helices, especially H1 and H2 helices (Figure 12B) are much more varied than the β-sheets. This malleability of the α-helices across species of vast phylogenetic distance and wide temperature range is in contrast to that of the β-sheets. Therefore, modifications to the structural framework of TBPs, through α-helices, have conferred TBPs some degree of adaptability, while conserving its DNA-binding function.

Phylogenetic analysis of TBP sequences, including that of MbTBP also yielded insights into the TBP family. The evolution of TBP adheres closely to that of the archaeal tree of life [158], where TBP from the Euryarchaeota species forms a distinct lineage from others, such as Crenarchaeota species. Unexpectedly, several archaeal TBPs were found within the eukaryotic TBP group. For example, the genome of the marine sponge symbiont *Cenarchaeum symbiosum* [159] contains three *tbp* genes, where one of the gene products is much more closely related to eucaryal TBP than archaeal TBP (Figure 11).

The endosymbiotic relationship between the marine sponge host and the archaeal symbiont therefore provides a plausible avenue for horizontal gene transfer to occur. Indeed, numerous gene transfers between the symbiont and host have been documented [160-162]. Eucaryote to bacteria and archaea gene transfer [163], though rare, is not impossible. At least in the case of *tbp* gene, it appears that such horizontal gene transfer is not so rare after all.

In conclusion, the dissection of the sequences of MbTBP and its comparison with other TBPs has not only demonstrated that MbTBP conforms to the hallmarks of TBP, but also contains unique features. In addition, it has also allowed greater insights into the TBP family. But perhaps the most significant revelation from the sequence of Mbur_1496 is the plausible presence of two isoforms of MbTBP *in vivo* – the full length version that is 206 amino acids and a shorter version that is 183 amino acid residues long. Experimental validations of the presence of two MbTBP isoforms are addressed in the following chapter.

CHAPTER IV

In vivo presence of dual TBP isoforms in *Methanococcoides burtonii*

4.1 Introduction

Bioinformatics analysis established that the gene locus Mbur_1496 encodes for a TATA-box binding protein (TBP) from the archaeon, *Methanococcoides burtonii*. In addition, there was also the suggestion of the presence of two isoforms of MbTBP, which differ in the length of the protein. The full length and its shorter counterpart are referred to as MbTBP206 and MbTBP183, respectively. In addition, 'His' was prefixed to the recombinant MbTBP that contains an N-terminal hexa-histidine tag, for example HisMbTBP206.

The appropriate method to embark upon the validations is through the use of the Western blotting technique [164], and in particular a indirect Western blotting procedure. An indirect Western procedure would require the use of a secondary antibody against the primary antibody that is raised, based on the antigen in question (Figure 13). This additional step serves as a signal amplification for the subsequent signal detection step, as multiple secondary antibodies can be bound to a single primary antibody. Although the Western technique has been widely used in the past three decades, there is no single universal immunoblotting procedure that works on all proteins of interest [165]. The method has to be established empirically for each and every protein-of-interest, depending on the antibodies used, and the nature of the protein of interest, among other factors [165]. Therefore the optimisation process of the Western blotting protocol of MbTBP is also described herein.

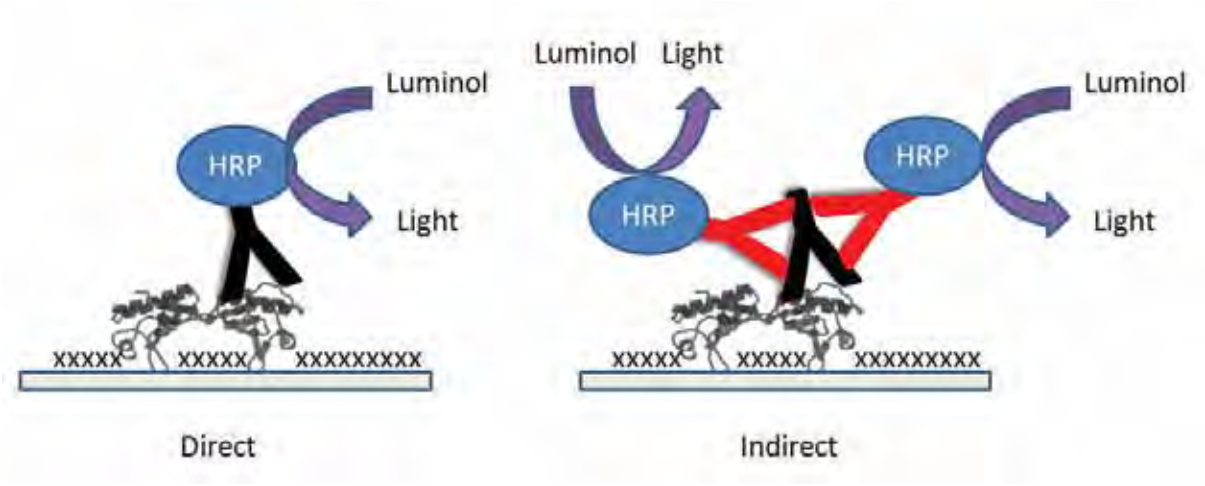


Figure 13: Schematics of a direct and indirect Western blotting procedure.

In the direct procedure, the primary antibody is labelled with an enzyme, such as a horseradish peroxidise (HRP) in this example. In the indirect procedure, a secondary antibody labelled with HRP is used instead. More than one secondary antibody can be bound to a primary antibody. Chemilluminescence detection of the membrane-immobilised antigen would be much more sensitive in the indirect procedure than the direct counterpart due to signal amplification in the indirect method.

Subsequently, the optimised Western blotting protocol was utilised to shed light on the presence of MbTBP isoforms, as well as their abundances in *M. burtonii* cultured under different growth conditions.

4.2 Results

4.2.1 Optimisations of Western blotting protocol for detection of MbTBP

The initial optimisation of the Western blotting protocol was focused on the antisera titer of anti-MbTBP and the detection range of MbTBP. Specifically, two different antisera titers at 5,000 and 10,000 were used to detect both HisMbTBP206 and HisMbTBP183, in approximately 1.5µg – 15ng of His-tag purified recombinant protein samples. In addition, a preliminary detection of MbTBP with 360µg of a soluble protein fraction from *M. burtonii* was also carried out. All the other variables such as membrane types and secondary antibody titer were kept constant.

The antisera was able to specifically recognise and detect the two recombinant isoforms of MbTBP. Differentiation of the two isoforms however, was determined by the molecular weight differences between the two isoforms, with HisMbTBP206 and HisMbTBP183 registering 23.2KDa and 20.7KDa, respectively. The Western procedure, with the anti-MbTBP antisera was able to detect nanogram levels of recombinant MbTBP. As illustrated in Figure 14, intense recombinant MbTBP bands were detected in 15ng, and beyond, of His-tag purified recombinant protein samples. However ghost bands, which contained no meaningful data, were prominent when 150ng or more of His-tag purified protein samples were used, in both antisera titers of 5,000 and 10,000.

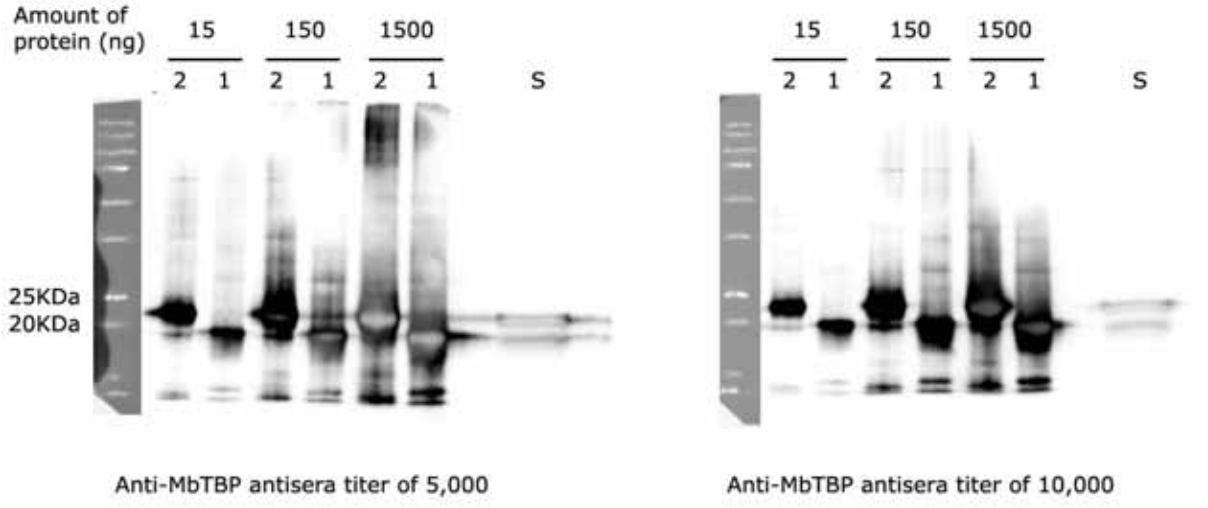


Figure 14: Chemiluminescence detection of HisMbTBP and MbTBP isoforms, with different antisera titers, and different amount of recombinant protein samples.

'2' and '1' refer to his-tag purified samples of HisMbTBP206 and HisMbTBP183, respectively. 'S' refers to the soluble protein fraction of *M. burtonii*.

Ghost bands were also present in the preliminary detection of MbTBP in the soluble protein fraction of *M. burtonii* (Figure 14), therefore necessitating further optimisation of the Western blotting procedure with reference to the amount of soluble protein fraction to be used. As such, subsequent optimisations were performed with different amounts of the soluble protein fraction from *M. burtonii* cultured in both MFM and M-media, with an anti-MbTBP antisera titer of 10,000. These would allow for the

determination of the maximum amount of soluble protein fraction required, without generating any ghost bands.

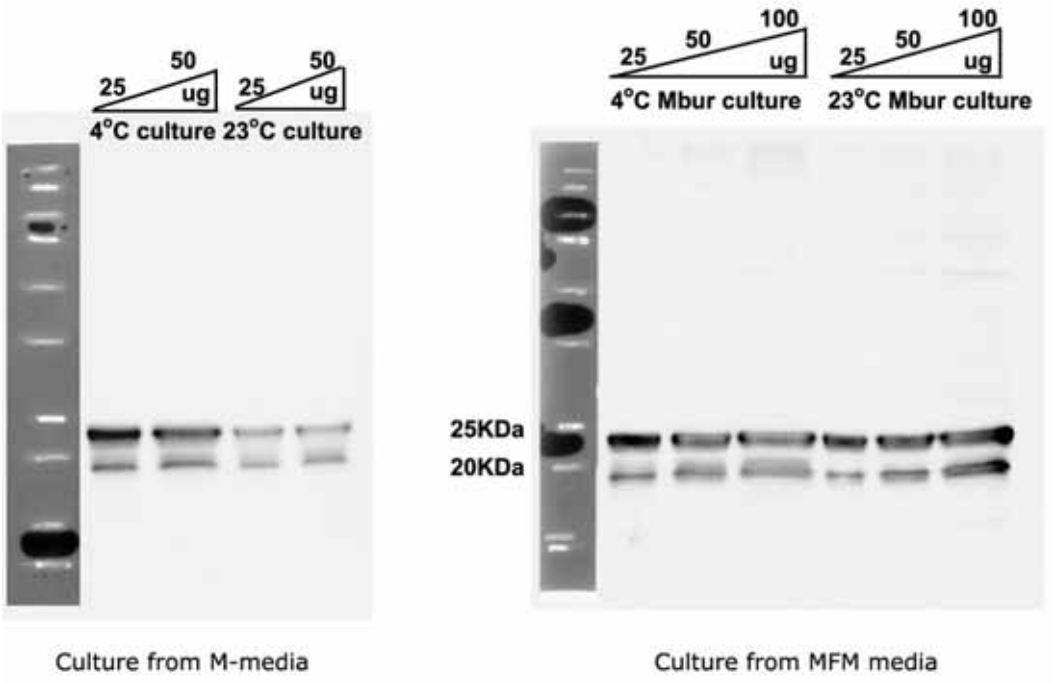


Figure 15: Chemilluminescence detection of MbTBP from various amount of soluble protein fraction of *M. burtonii* grown in different culture conditions.

Ghost bands were clearly present when more than 25µg of soluble protein fractions were used in Western blotting, regardless of the growth conditions of *M. burtonii* culture (Figure 15). Together with the information in Figure 14, the optimised Western blotting protocol with the chemilluminescence detection of MbTBP would involve the use of an anti-MbTBP antisera titer of 10,000, with at most 15ng and 25µg of recombinant protein samples and soluble protein fraction of *M. burtonii*, respectively.

4.2.2 MbTBP is translated *in vivo* and in two isoforms

The detection of MbTBP and its isoforms *in vivo* was achieved with the optimised indirect Western blotting procedure as described above, and in Chapter 2 (Materials and Methods Chapter 2.2.6.5). Soluble protein fractions of *M. burtonii* from different growth conditions were immunoblotted and chemilluminescencely detected on the

same PVDF blot so as to allow meaningful comparison of the MbTBP intensity between growth conditions. The growth conditions investigated pertain to the use of either methanol (M-media) or trimethylamine (MFM media) as the methylated substrates for growth, as well as in high (23°C) and low (4°C) growth temperatures. Such comparison would reveal any differential abundance of MbTBP under the different conditions. Soluble protein fractions of *M. burtonii* cultured in M-media were kindly provided by Williams *et al* (2010) [82].

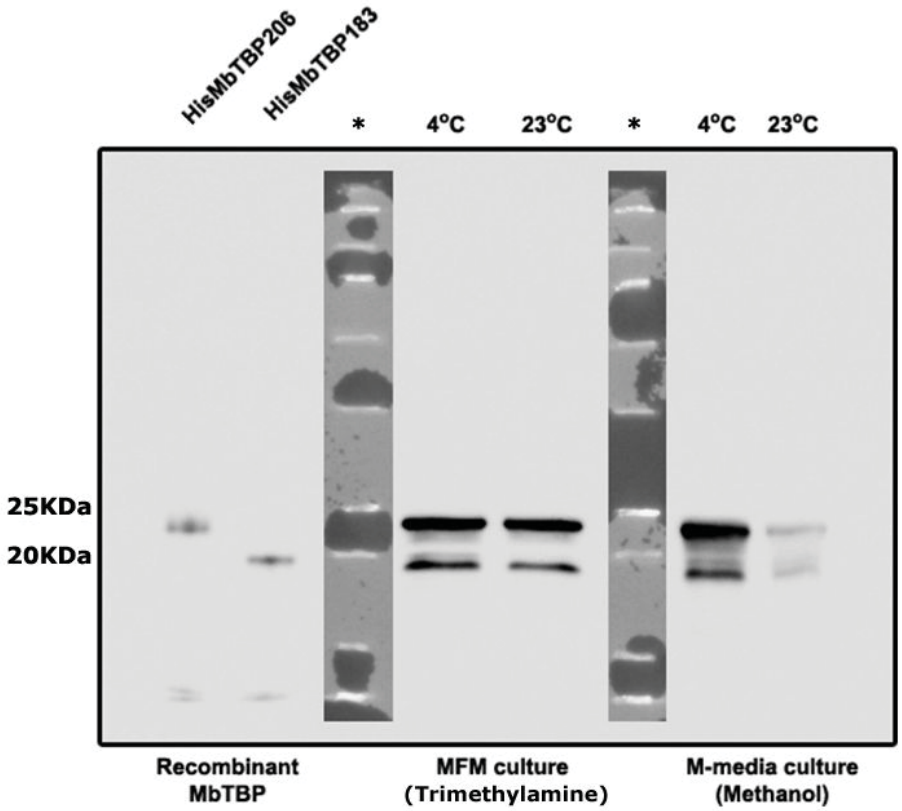


Figure 16: Chemilluminescence detection of MbTBP from soluble protein fractions of *M. burtonii* from different growth conditions on the same blot.

Recombinant HisMbTBPs, at 15ng were included as references to the expected position of the two MbTBP isoforms. '*' refers to protein molecular weight ladder.

As illustrated in the Western blot (Figure 16), the MbTBP, or rather gene product of Mbur_1496 is expressed and translated *in vivo* in *M. burtonii*. In addition, it also showed that there are two isoforms of MbTBP *in vivo*, as predicted based on the

bioinformatics analysis. The two isoforms of MbTBP were present in all growth conditions investigated.

	MFM media	M-media	4° C	23° C
	4° C/23° C	4° C/23° C	MFM/M	MFM/M
MbTBP206	1.2	9.8	1.0	8.3
MbTBP183	1.5	7.5	1.1	5.5
Total MbTBP	1.3	9.0	1.0	7.3

Table 2: Differential abundance of MbTBP, as a function of growth temperature, culture medium, and isoform type.

	MFM media		M-media	
Temperature	4° C	23° C	4° C	23° C
MbTBP206/183 Ratio	2.1	2.7	2.4	1.8

Table 3: Ratio of MbTBP isoforms under different growth conditions.

Based on densitometry analysis of the bands for the MbTBP isoforms (Table 2 and 3), a few observations could be made. Foremost, there was an over-abundance of MbTBPs, regardless of isoforms, in cultures grown at 4°C rather than 23°C, albeit that the differences for the MFM culture was rather small compared to M-media culture. Secondly, there were approximately equal abundances of MbTBP at 4°C, between the MFM and M-media culture. At 23°C however, there was approximately 7× more MbTBP detected in MFM than M-media. Thirdly, there was a higher abundance of the full length isoform, MbTBP206 across all growth conditions.

4.3 Discussion

Bioinformatics analysis of the gene locus Mbur_1496 had previously postulated the presence of a putative shorter isoform of MbTBP, in addition to the full length MbTBP.

As such, the Western blotting technique [164] was utilised to provide evidence of the *in vivo* presence of two MbTBP isoforms.

However many factors affect the success of a Western blotting experiment [165], resulting in a rather narrow range of optimised conditions for successful detection of a protein of interest. As illustrated in Figure 14 and 15, one of the most common outcomes associated with a sub-optimal Western blotting protocol is the presence of ghost bands. Ghost bands are regions of low signal intensity [165], and are frequently identified as white protein bands with an occasional occurrence of a dark halo. Nevertheless, proteins are actually present within the ghost bands and the resultant densitometry analysis for the presence of the protein would always be an underestimated value.

A few factors are known to produce ghost bands in Western blotting. With specific references to the detection of MbTBPs using anti-MbTBP antisera (Figure 14 and 15) herein, the amounts of antisera as well as the antigen (MbTBP) load were the crucial factors. Saturation of the antisera and MbTBP load resulted in extremely rapid and localised depletion of the chemilluminescence substrate before the luminescence signal could be captured [165]. A high anti-MbTBP titer of 10,000, coupled with low MbTBP load, was subsequently successful in eliminating ghost bands. Using these optimised Western blotting conditions, two MbTBP isoforms were detected within *M. burtonii*.

MbTBP has often being detected using proteomics [82-83,166] and microarray [167] approaches. However, what was lacking in both these global analytical approaches as compared to Western blotting was their inability to distinguish between two nearly identical proteins and mRNAs respectively. For example, the detection of a peptide fragment 'FPGLVYR' from the MbTBP core was not sufficient to distinguish between MbTBP206 and MbTBP183 since both isoforms contain the exact MbTBP core domain.

Similarly, the detection of a peptide fragment from the extra 23 N-termini amino acid residues of MbTBP206 can only signify the presence of the full length MbTBP206.

Therefore it is apparent that the presence or absence of a second, shorter MbTBP isoform would never be resolved by proteomics and microarray. There could be a substantial number of protein isoforms within *M. burtonii* that were underestimated when its proteome and genome were analysed with these global approaches. MbTBP is one of such proteins where two isoforms co-exist *in vivo*, and Western blotting was useful in identifying these isoforms.

Besides *M. burtonii*, several other archaea species also harbour multiple copies of TBP. For example, *Methanosarcina acetivorans*, a close relative of *M. burtonii*, contains three TBPs [168]. In fact, archaea such as the *Halobacterium* species can contain up to nine TBPs (Figure 11). Using knockout studies coupled with microarray analysis, TbpD from the *Halobacterium* species was found to uniquely regulate over 4,000 genes [169]. Similarly, TBP2 and TBP3 from *M. acetivorans* were also shown to have unique transcriptional targets [168]. Therefore, functional diversification of the two MbTBP isoforms is also plausible. The extra RepA sequences in MbTBP206 could function to interact with additional transcriptional regulators or MbTBP-interacting proteins, and thus possibly regulate a different set of genes as compared to MbTBP183.

The Western blotting results (Table 2) are also in agreement with the over-expression [167] and abundance [82-83] of MbTBP at low temperature. Together with the over-abundance of other transcription-related proteins in *M. burtonii* grown in 4°C than 23°C [82], it appears that the increase in these proteins compensates the cold-induced reduction in transcriptional capability by an increased abundance of transcriptional proteins.

Interestingly, the extent of differential abundance of MbTBP in low temperature was much more prominent in culture grown in M-media (defined, methanol-containing)

than MFM (complex, trimethylamine-containing) media. This occurrence was due to a drastic decrease in the MbTBP protein level at 23°C, when the *M. burtonii* was cultured in M-media as compared to MFM media (Table 2). In fact, the protein level of another key basal transcriptional protein of *M. burtonii*, the transcription factor B (TFB) was also significantly reduced in M-media over MFM media at 23°C [67]. It would therefore appear that methanol, at 23°C, has a potent dampening effect on the abundance of transcription initiation proteins such as MbTBP and TFB. It is possible that both proteins were more susceptible to denaturation by methanol, and thus degraded to a larger extent at 23°C than 4°C.

The above study also provided the first evidence for the differential abundance of MbTBP isoforms (Table 3). The higher abundance of MbTBP206, compared to MbTBP183, is consistent with the fact that the former is regulated by a ribosome-binding site that bears a greater resemblance to a perfect Shine-Dalgarno sequence than the latter [148].

In summary, the work described in this chapter set out to detect the *in vivo* presence of MbTBP, the gene product of Mbur_1496. Using Western blotting technique, two isoforms of TBP were discovered in *M. burtonii*. In addition, the MbTBPs were differentially abundant in cultures grown at 4°C compared to 23°C. At 23°C, cultures grown in M-media contained significantly less MbTBP than those grown in MFM media. Finally, MbTBP206 was more abundant than its shorter counterpart, in all the growth conditions that were analysed.

CHAPTER V

Protocol development for the large scale expression and purification of recombinant MbTBP isoforms

5.1 Introduction

The optimal conditions for the production of heterologous proteins in *E. coli* is dependent on many factors, which include expression strain of *E. coli*, growth conditions and plasmid copy number, among many others [170-173]. As such, small scale pilot expression studies are often performed to determine the optimal conditions for the expression of heterologous proteins in *E. coli* [172]. These optimised conditions can then be applied to the large scale production and purification of proteins. Therefore, small scale expression studies were conducted to determine the optimal conditions for the expression of recombinant MbTBP isoforms in *E. coli*, before the production system was up-scaled to produce large amounts of recombinant proteins.

5.1.1 Chapter objectives

This chapter reports on the development of an expression and purification method for the large scale production of recombinant MbTBP183 and MbTBP206. Briefly, recombinant hexa-histidine tagged MbTBP isoforms were created, over-expressed in *E. coli* and purified to homogeneity using both immobilised metal-affinity chromatography (IMAC) [174-175] and size-exclusion chromatography (SEC) [176]. The method developed in this chapter allowed for rapid and large scale production of both recombinant MbTBP isoforms that were used for biophysical and functional studies in Chapter 6 and 7, respectively.

5.2 Results

5.2.1 Cloning of recombinant MbTBP isoforms into pET101/D-TOPO vector

Amino-terminal hexa-histidine tags were introduced to the MbTBP isoforms by PCR, using primers containing hexa-histidine-coding nucleotide sequences (Materials and Methods, Chapter 2.2.2). Histidine tags were used as fusion tags, since histidine tags are small and thus allow for minimal functional interference with MbTBP [172]. It also served as a platform for using IMAC techniques to specifically capture and purify the recombinant MbTBP isoforms.

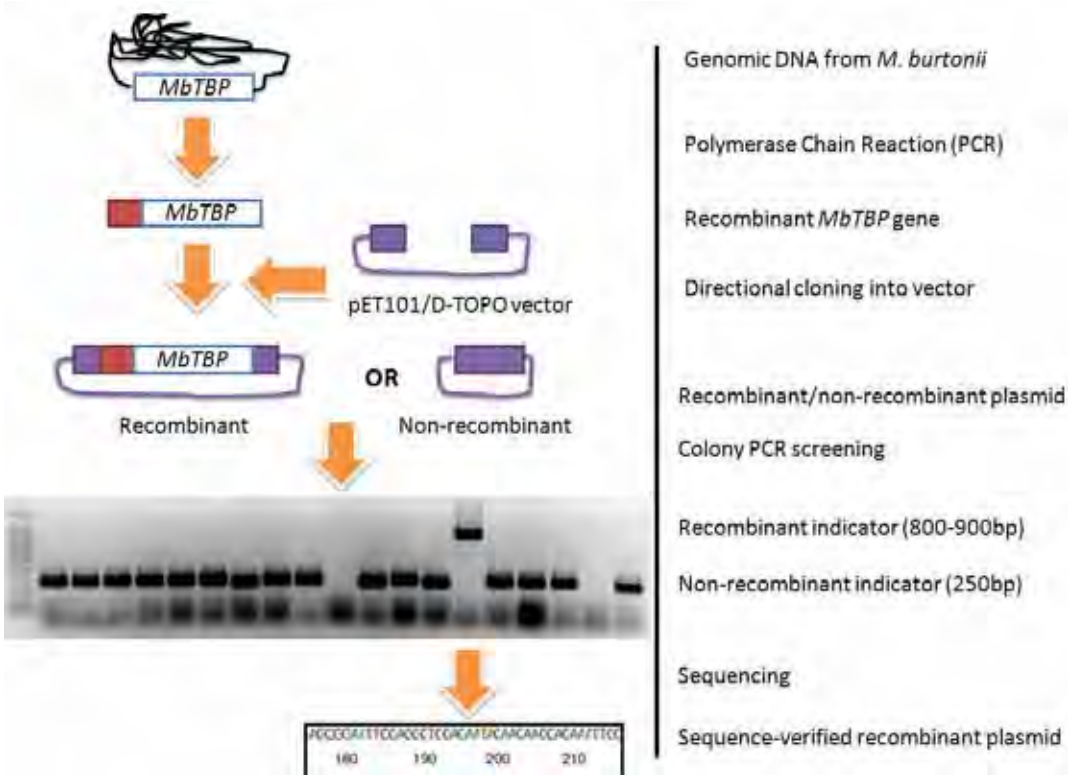


Figure 17: Procedures for cloning a recombinant hexa-histidine tagged MbTBP gene into a pET vector. A general description of the cloning process is provided at the side of the figure detailing the schematics of the cloning procedure.

The resultant recombinant MbTBP gene was subsequently cloned into a pET101 vector through directional TOPO cloning. Colony PCR was used to screen the recombinant plasmids. Recombinant and non-recombinant clones were differentiated based on the size of the PCR amplicons. Specifically, the former would produce a PCR amplicon that was 800bp – 900bp long, while the amplicon size of the latter was approximately 250 base-pair long (Figure 17). A total of about 150 – 200 colonies were PCR-screened until a single recombinant colony was identified, for each recombinant MbTBP isoform. The isolated recombinant plasmids were subsequently sequence verified to contain the recombinant MbTBP isoforms. MbTBP isoforms with hexa-histidines tagged onto the N- and C-terminal of the proteins were assigned a 'His' prefix (such as HisMbTBP206) and suffix (such as MbTBP206His), respectively.

5.2.2 Expression of recombinant HisMbTBP183 in *E. coli*

Pilot expression of the recombinant HisMbTBP183 (MbTBP183 with an N-terminal 'His' tag) was performed using the Rosetta (DE3) *E. coli* strain that was cultured and induced with 1mM IPTG at 20°C, 25°C, 30°C and 37°C respectively. The induction periods at 20°C and 25°C were 24h each while that at 30°C and 37°C were 5h each. Expression of the HisMbTBP183 was subsequently analysed using SDS-PAGE.

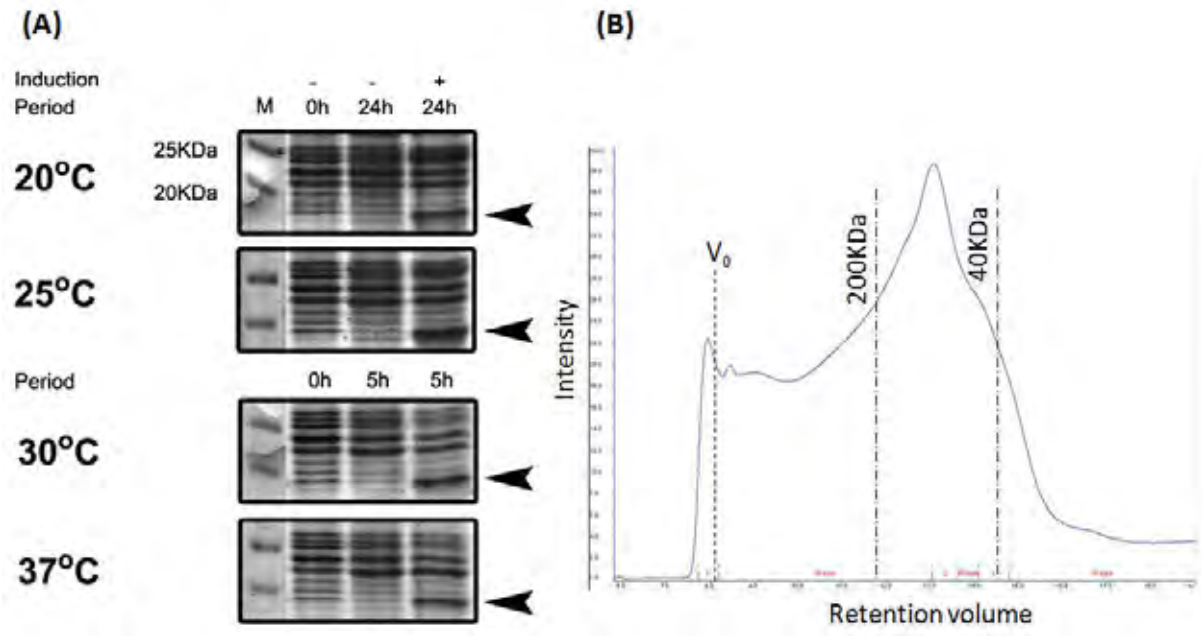


Figure 18: Expression and purification of HisMbTBP183 from Rosetta (DE3) strain of *E. coli*.

(A) Expression levels of the recombinant HisMbTBP183 at different induction temperatures. The arrowheads denote the expected positions of HisMbTBP183. The '-' and '+' refers to the absence and presence of IPTG induction, respectively. Lane 'M' refers to the lane containing the protein molecular weight marker. (B) A typical chromatogram from the size-exclusion chromatography of His-tag purified HisMbTBP183 that was induced and expressed at 37°C. V_0 refers to the void volume of the Superdex 200 10/300GL column. The retention volumes for the 200KDa and 40KDa protein standards are marked. The column was attached to an AKTA system (GE Healthcare) with a flow rate of 0.1mL/min.

The Rosetta (DE3) strain was able to express high levels of HisMbTBP183 at all the induction temperatures tested (Figure 18A). However, the His-tag purified HisMbTBP183 could not be resolved into a single discrete peak when it was expressed at 37°C (Figure 18B). Further, a considerable amount of the protein was eluted in the

void volume, indicating that these proteins exceeded 600KDa in molecular weight. These observations suggest that the recombinant psychrophilic proteins (of approximately 20KDa) might have aggregated when being expressed at 37°C. As such, subsequent large scale expression of HisMbTBP183 was performed at 20°C to prevent aggregation and denaturation of the psychrophilic protein.

5.2.3 MbTBP183 is co-expressed with HisMbTBP206 in *E. coli* transformed with HisMbTBP206 gene

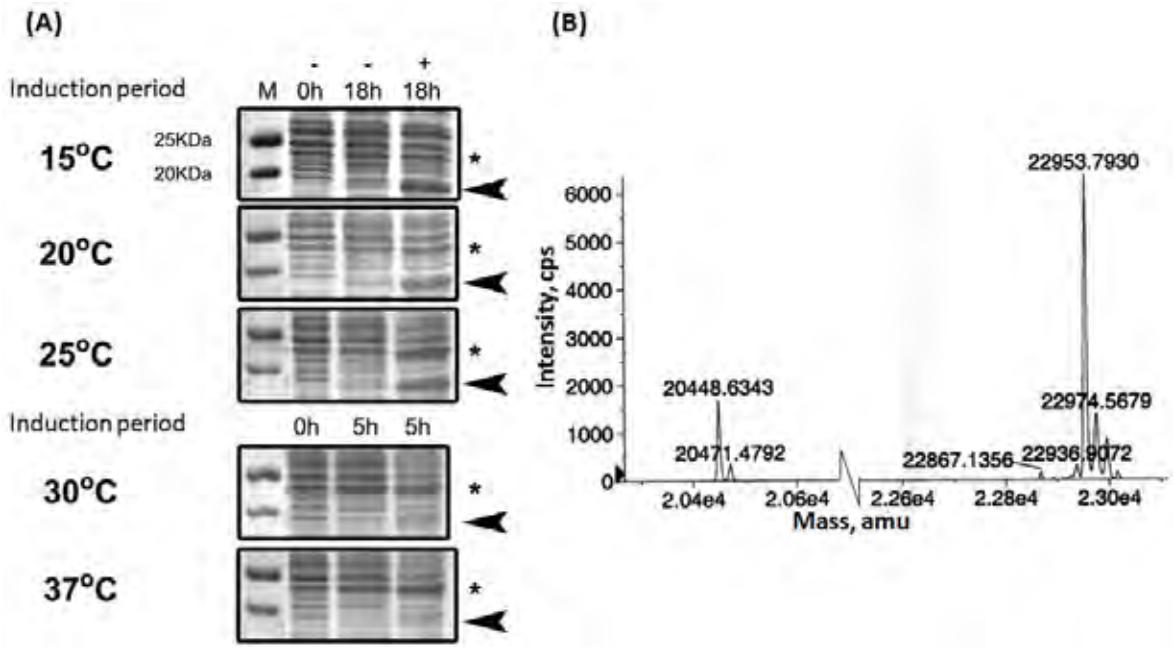


Figure 19: Co-expression of HisMbTBP206 and MbTBP183 from a single recombinant HisMbTBP206 gene.

(A) Expression levels of the recombinant HisMbTBP206 at different induction temperatures. The asterisks and arrowheads denote the expected positions of HisMbTBP206 and MbTBP183, respectively. The '-' and '+' refers to the absence and presence of IPTG induction, respectively. The lane 'M' refers to the lane containing a protein molecular weight marker. (B) ESI-MS chromatogram of a histidine-tag purified protein sample originating from cells transformed with MbTBP206His gene-containing plasmid and expressed at 37°C. Two peaks at 20.4KDa and 23.0KDa corresponded to MbTBP183His and MbTBP206His, respectively.

A similar pilot expression study was conducted in parallel for the recombinant HisMbTBP206 (full length MbTBP containing 'His' tag), also using the Rosetta (DE3) *E. coli* strain culture, which was cultured and induced with 1mM of IPTG, at temperatures of 15°C, 20°C, 25°C, 30°C and 37°C. The induction period for cultures grown at 15°C – 25°C was approximately 18h (overnight), while that at 30°C – 37°C was 5h each. Expression of the recombinant protein was subsequently assessed with SDS-PAGE.

At induction temperatures of less than 30°C, HisMbTBP206 was co-expressed as a minor protein species, together with an unidentified 18KDa protein that was produced in abundance. However at higher induction temperatures of 30°C and 37°C, the HisMbTBP206 was predominantly expressed instead (Figure 19A). Although there was the risk of protein aggregation, the option of overexpressing HisMbTBP206 at 37°C was still explored since the unidentified 18KDa protein was not produced at that temperature. Moreover functional psychrophilic proteins have also been successfully expressed previously at 37°C [177-179]. As such, a trial was subsequently conducted to assess if the HisMbTBP206 that was expressed at 37°C was present in aggregates.

Briefly, His-tag purification of recombinant HisMbTBP206 was performed with the purified protein being subjected to size-exclusion chromatography. The resultant chromatogram revealed a wide peak with the calculated molecular weight of the recombinant protein ranging from approximately 40KDa – 200KDa (Figure 20). The values obtained differed from the expected molecular weight of HisMbTBP206, which was 23KDa. This observation suggested that the recombinant proteins might have aggregated. As such, an attempt was made to resolve the chromatogram into a single discrete peak by manipulating buffer conditions that might prevent aggregation, such as increasing the amount of KCl in the buffer. Despite that, the chromatogram remained unresolved (data not shown). These findings indicated that the recombinant HisMbTBP206, despite being abundantly expressed at 37°C could not be resolved, which would pose a major problem in protein purification.

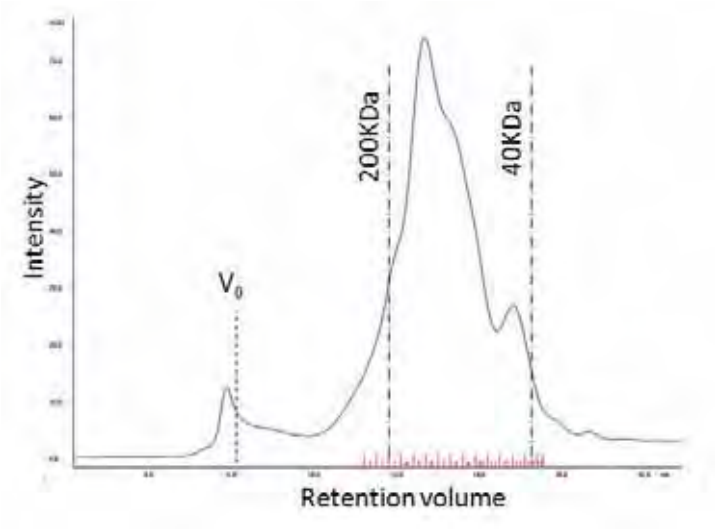


Figure 20: A typical chromatogram from the size-exclusion chromatography of His-tag purified HisMbTBP206 that was induced to expression at 37°C.

V_0 refers to the void volume of the Superdex 200 10/300GL column. The retention volumes for the 200KDa and 40KDa protein standards are marked. The column was attached to an AKTA system (GE Healthcare) with a flow rate of 0.1mL/min.

Since it was not possible to purify HisMbTBP206 that was produced at 37°C, strategies were devised to counter the undesirable low abundance of HisMbTBP206 at 15°C – 25°C (Figure 19A) instead. The unidentified protein band had an apparent molecular weight of approximately 18KDa, which seems to coincide with the theoretical molecular weight of MbTBP183, suggesting that it could also be an MbTBP. This conjecture, if correct, would mean that the 18KDa protein was translated from the recombinant HisMbTBP206 gene in *E. coli*. It also implies that the DNA sequence in the recombinant HisMbTBP206 gene can potentially be mutated to abolish the translation of the 18KDa protein in *E. coli*, thereby improving the yield of HisMbTBP206. In the event that the 18KDa protein was not MbTBP183 but rather, a protein endogenous to *E. coli* that was co-expressed with HisMbTBP206, other expression systems would have to be considered for producing HisMbTBP206.

In order to verify if the identity of the protein band was MbTBP183, a hexa-histidine tag was placed at the 3'-end of the MbTBP206 gene for over-expression in *E. coli*. Purification of two His-tagged protein bands would imply that the 18KDa protein band

that was co-expressed with MbTBP206His was translated from the same recombinant MbTBP206His gene. As expected, two protein bands of molecular weights that were consistent with that of an MbTBP206His and an MbTBP183His were purified (Figure 19B). These findings indicated that similar to *M. burtonii*, *E. coli* was able to translate the shorter MbTBP183 from a full length *tbp* gene. Therefore, the lower molecular weight protein, which was predominantly expressed at low temperatures when *E. coli* was transformed with the HisMbTBP206 gene (Figure 18A), was likely to be the shorter MbTBP183. This important finding was the basis for creating a HisMbTBP206 mutant gene, which would abolish the translation of MbTBP183 when the mutant gene was transformed into *E. coli.*

5.2.4 *E. coli* transformed with the HisMbTBP206 mutant gene expressed only the HisMbTBP206 mutant but not MbTBP183

The expression of a recombinant HisMbTBP206 posed a couple of challenges – irresolvable purification of the recombinant protein with contaminants when expressed at 37°C, and minor expression of the protein at 15°C to 20°C. In order to circumvent these two issues, the HisMbTBP206 gene was designed to contain DNA mutations that would abolish the expression of MbTBP183 in *E. coli*. Specifically, the putative Shine-Dalgarno sequence that regulates the translation of MbTBP183 was disrupted, and the AUG initiation codon for MbTBP183 was altered to a non-initiator CUG codon (Figure 21). The HisMbTBP206 mutant gene was also codon optimised [142] to maximize its expression in *E. coli*. This mutant gene was subsequently synthesized and cloned into a pJexpress404 vector (DNA 2.0).

Cloning of the HisMbTBP206 mutant gene into the pJexpress404 vector, which is controlled by a T5 promoter [180] allows for the expression of the mutant protein in any *E. coli* strain. Subsequently a pilot expression study of the recombinant HisMbTBP206 mutant was conducted using BL21 and TOP10 strains of *E. coli*. The cultures were grown and induced with IPTG at temperatures of 10°C and 20°C for each strain, to

prevent heat-induced denaturation and aggregation of the over-expressed protein at high (>20°C) temperatures.

Figure 21: Differences between the gene (and its translated sequence) for both HisMbTBP206 and the HisMbTBP206 mutant.

The first start codons are highlighted in blue. The putative ribosome-binding site for MbTBP183 is boxed. The start codon that initiates the translation of MbTBP183 is also underlined. Sequence coloured in red are the mutation introduced to abolish the translation of MbTBP183 in the HisMbTBP206 mutant.

At an induction temperature of 10°C, the TOP10 strain expressed the highest levels of recombinant protein per cell absorbance unit, as compared to the BL21 strain (Figure 22A). However, the yield of TOP10 obtained was low despite a long induction period of 48h (Figure 22B). Therefore, the amount of HisMbTBP206 mutant that could be harvested from TOP10 after 48h of protein expression would also very likely be low.

The BL21 strain, at an induction temperature of 20°C, was found to produce the highest amounts of recombinant protein per cell absorbance unit after 24 hours of induction, and registered the highest cell yield (Figure 22A and 22B). Conversely, the TOP10 strain produced a much lower cell yield than the BL21 strain (Figure 22B), which makes the TOP10 strain undesirable for recombinant protein production. Also, the mutations introduced into HisMbTBP206 to create the HisMbTBP206 mutant were able to increase the expression of the protein at low temperatures without producing MbTBP183. The recombinant HisMbTBP206 was no longer co-expressed as a minor protein species, but was solely expressed at low (20°C) temperatures instead.

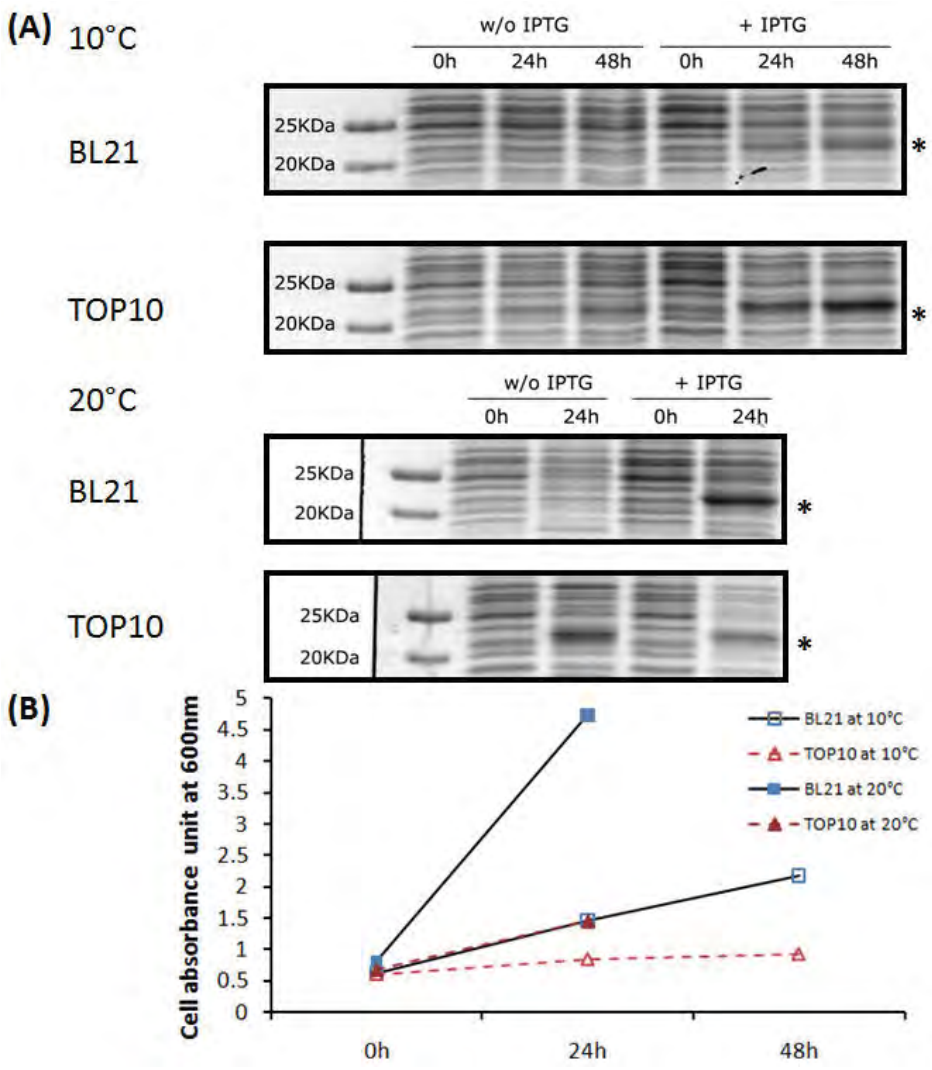


Figure 22: Optimal production of HisMbTBP206 mutant in the BL21 strain of *E. coli*, at an induction temperature of 20°C for 24 hours.

(A) SDS-PAGE expression analysis of the levels of HisMbTBP206 mutant produced at different induction temperatures and durations in the various *E. coli* host strains. Volumes corresponding to 0.75 absorbance unit of cell at 600nm were loaded in each lane. The asterisk refers to the expected position of the HisMbTBP206 mutant on SDS-PAGE. (B) Bacteria cell yield per mL of volume (y-axis), at various time points after IPTG induction of HisMbTBP206 mutant expression.

5.2.5 Recovery of soluble protein faction containing recombinant HisMbTBP isoforms

E. coli, induced to produce recombinant HisMbTBP isoforms, were incubated in cold *E. coli* lysis buffer before being subjected to lysis by a French pressure cell press (Figure 23A). The cell lysate was subsequently centrifuged and the soluble protein fraction

(supernatant) was collected. The remaining cell pellet was then washed four times with cold *E. coli* lysis buffer for further extraction of the recombinant proteins. Supernatants from these four washes were collected and analysed for recombinant protein recovery (Figure 23B).

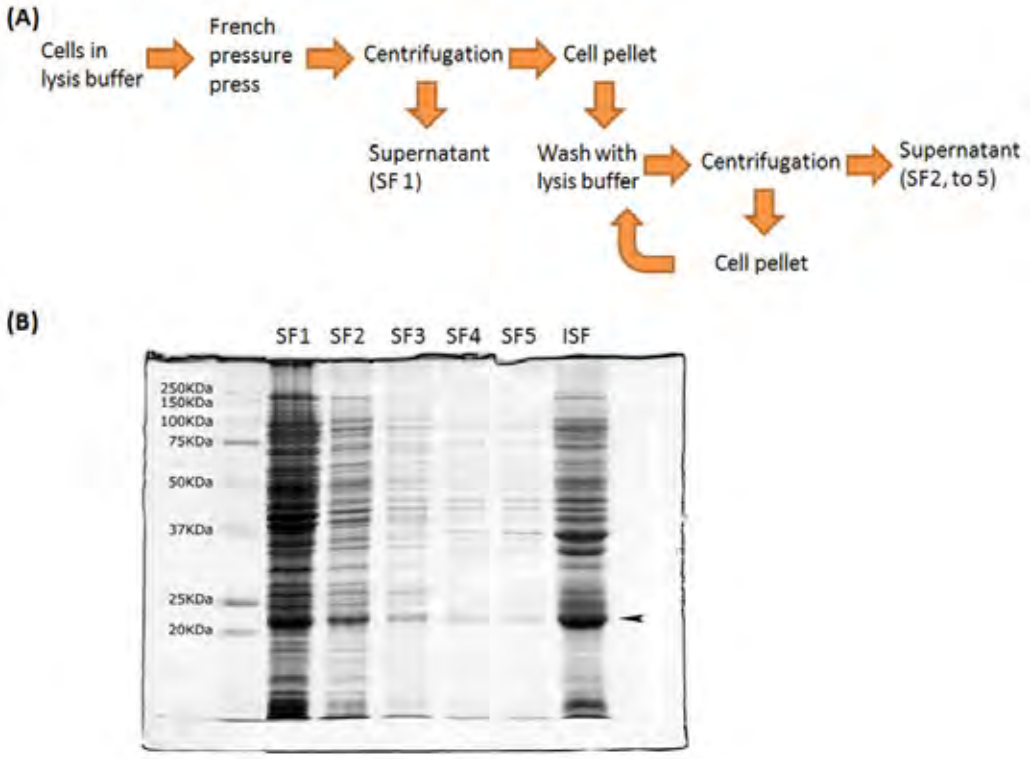


Figure 23: Optimisation of the protocol for obtaining soluble protein fractions containing recombinant proteins.

(A) A flowchart of the recovery trial used to determine the amount of washes required to maximise the efficiency of the extraction of soluble recombinant proteins from the cell lysates. 'SF' refers to soluble fraction. (B) SDS-PAGE analysis of the soluble fractions from different washes (SF1 – SF5), and the insoluble fraction (ISF; final protein pellet after the fourth wash). Equal volumes of the respective protein fraction were loaded in each lane. The arrowhead denotes the expected position of the recombinant protein.

Figure 23B shows that a reasonably fair amount of the recombinant protein could be extracted into the soluble protein fraction. The recombinant protein was present in all soluble protein fractions tested. However, it was noted that soluble protein fractions from the third (SF4) and final (SF5) wash step yielded little recombinant proteins. The

total amount of recombinant protein recovered in the soluble fractions (SF1 to 3) was approximately 40% – 50% of the total amount of recombinant protein expressed.

5.2.6 Prevalence of protein contaminants after His-tag purification

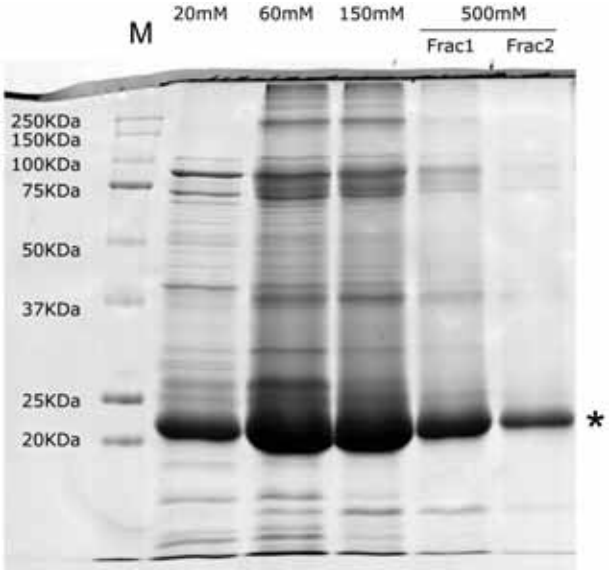


Figure 24: Presence of contaminating proteins after His-tag purification.

SDS-PAGE analysis of the different wash fractions from the His-tag purification of the HisMbTBP206 mutant. Concentrations of imidazole used are indicated on top of the respective lanes. The asterisk refers to the expected position of the HisMbTBP206 mutant protein. M refers to the protein standard molecular weight marker.

Subsequently, His-tag purification of the recovered soluble protein fractions containing the recombinant HisMbTBP206 mutant (Figure 23) was performed to purify the recombinant protein. However, the protein purification process proved to be only partially successful. Despite using a step-gradient method to elute the recombinant protein, the result was an enrichment of the recombinant HisMbTBP206 mutant rather than a complete purification of the protein (Figure 24). Even after washing the resin with 60mM of imidazole, which would remove nonspecific proteins; the protein contaminants were still found to interact strongly with the resin. As such, another wash using 150mM of imidazole was performed. Although this managed to remove a significant portion of the contaminants, a considerable amount of the target protein

was lost as well. As such, the His-tag purification step was employed as a method of protein enrichment instead of its use as a final purification step in the large scale production and purification of recombinant HisMbTBP isoforms. The enriched protein products were subsequently purified with size-exclusion chromatography.

5.2.7 Procedure adopted for the large scale expression and purification of HisMbTBP isoforms

All the results from the pilot expression studies and trials conducted in the previous sections culminated into a standard procedure adopted for the large-scale expression and purification of HisMbTBP isoforms (Material and Methods, Chapter 2.2.5). All stages within the adopted procedure, except for protein expression, were performed at 4°C.

A summary of the adopted procedure (Figure 25) were as follows:

i. The HisMbTBP206 mutant and the MbTBP183 were expressed by BL21 and Rosetta (DE3) *E.coli* strain of competent cells, respectively at an induction temperature of 20°C for an overnight period.
ii. The harvested cells were mechanically and chemically disrupted, with a French pressure cell press and lysis buffers, respectively. The resultant soluble protein fractions were used in the purification of recombinant HisMbTBPs.
iii. Recombinant proteins were enriched by the cobalt-based Talon resin (Clontech), and eluted using 500mM of imidazole.
iv. Size-exclusion chromatography, with a preparative-grade Superdex 75 column, was subsequently used to further purify the recombinant proteins. Fractions containing the purified recombinant proteins were then pooled and concentrated.
v. Finally, the purified recombinant proteins were flash-frozen in liquid nitrogen and stored at -80°C.

Both HisMbTBP206 mutant and HisMbTBP183 were highly purified at the end of the purification process, which is evident in Figure 26. On average, approximately 15mg of recombinant proteins could be purified from four litres of bacterial cell culture.

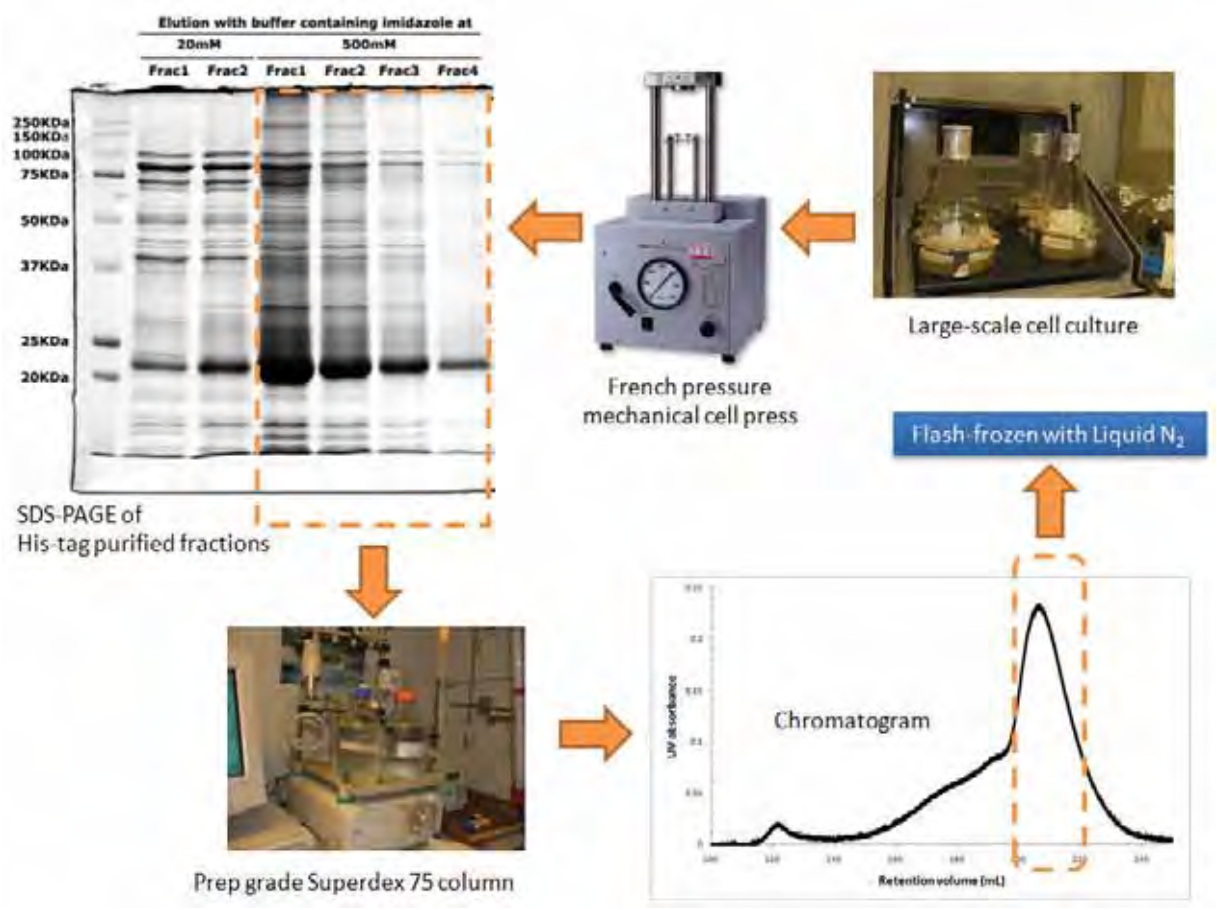


Figure 25: Schematics for the adopted procedure for large-scale expression and purification of recombinant HisMbTBP206 mutant and HisMbTBP183.

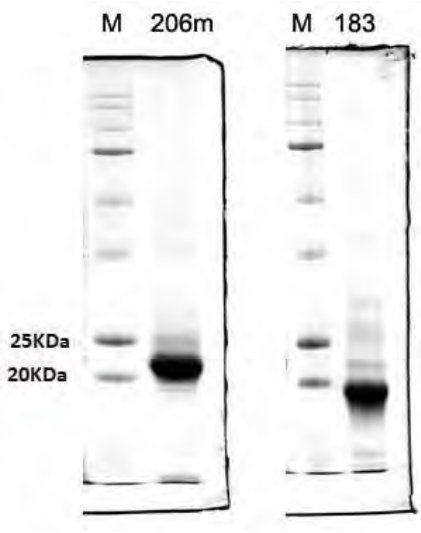


Figure 26: SDS-PAGE analysis of the final purified product of HisMbTBP206 mutant and HisMbTBP183. The HisMbTBP206 mutant was denoted as 206m, while HisMbTBP183 was denoted as 183. M refers to the protein standard marker. Five micrograms of purified protein were loaded per lane.

5.3 Discussion

5.3.1 Rationale for protocol development

The MbTBP isoforms could be extracted in two methods. The first of the two methods involves culturing *M. burtonii*, and extracting the two MbTBP isoforms from the culture. However, a problem arising from this approach is the protracted culturing period (1 – 2 months) that is required for culturing *M. burtonii*. More importantly, the fact that the two isoforms are nearly identical to each other would make separation by any chromatographic method difficult, especially when the purified product has to be functional. Both protein isoforms differ from each other in molecular weight and isoelectric point (pI) by only 2.4KDa and 0.16, respectively rendering it almost impossible to resolve these two proteins with size-exclusion and ion-exchange chromatography. A partially resolved mixture containing the two MbTBP isoforms would also complicate the interpretation of data when comparing the two proteins at a later stage.

The second method, which was preferred, was to individually express each hexa-histidine tagged MbTBP isoforms in an *E. coli* expression system before performing a

His-tag purification of the proteins. During His-tag purification of the hexa-histidine tagged MbTBP isoforms, contaminating proteins were co-purified (Figure 24). These protein contaminants are likely to be metal-chelating *E. coli* proteins such as SlyD and YodA, as these proteins are known to bind to metal-based resin [181-184] and therefore might not be easily removed through washing steps that eliminate non-specific binding proteins. In order to further purify the enriched His-tagged protein isoforms, a subsequent size-exclusion chromatography (SEC) step was included. By coupling His-tag purification with SEC, the recombinant HisMbTBP206 mutant and HisMbTBP183 could be purified to a high degree of homogeneity – which was comparable to other archaeal TBP expression studies [185-186].

5.3.2 Possible cause of the temperature-dependent shift in recombinant isoform expression

Functional psychrophilic proteins have previously been expressed in *E. coli* at 37°C [177-179]. However in other studies, expression of psychrophilic proteins have also been performed at induction temperatures of less than 37°C, ranging from 14°C to 25°C [77,187-189], possibly to prevent thermal denaturation of the proteins at 37°C [189]. Therefore, recombinant HisMbTBP isoforms were expressed at temperatures ranging from 15°C – 37°C, to determine the optimal expression temperature for production of the recombinant isoforms.

An unexpected consequence of expressing HisMbTBP206, from a HisMbTBP206 gene-containing plasmid, in *E. coli* was the co-expression of MbTBP183. The shorter MbTBP isoform was the predominant isoform at expression temperatures of less than 25°C, while HisMbTBP206 was the predominant isoform at expression temperatures of above 25°C. It is unclear why such a temperature – dependent shift in protein isoform expression occurred in the recombinant protein expression system (Figure 19), but it is possible that the secondary structure of the recombinant mRNA transcripts might have contributed to this phenomenon.

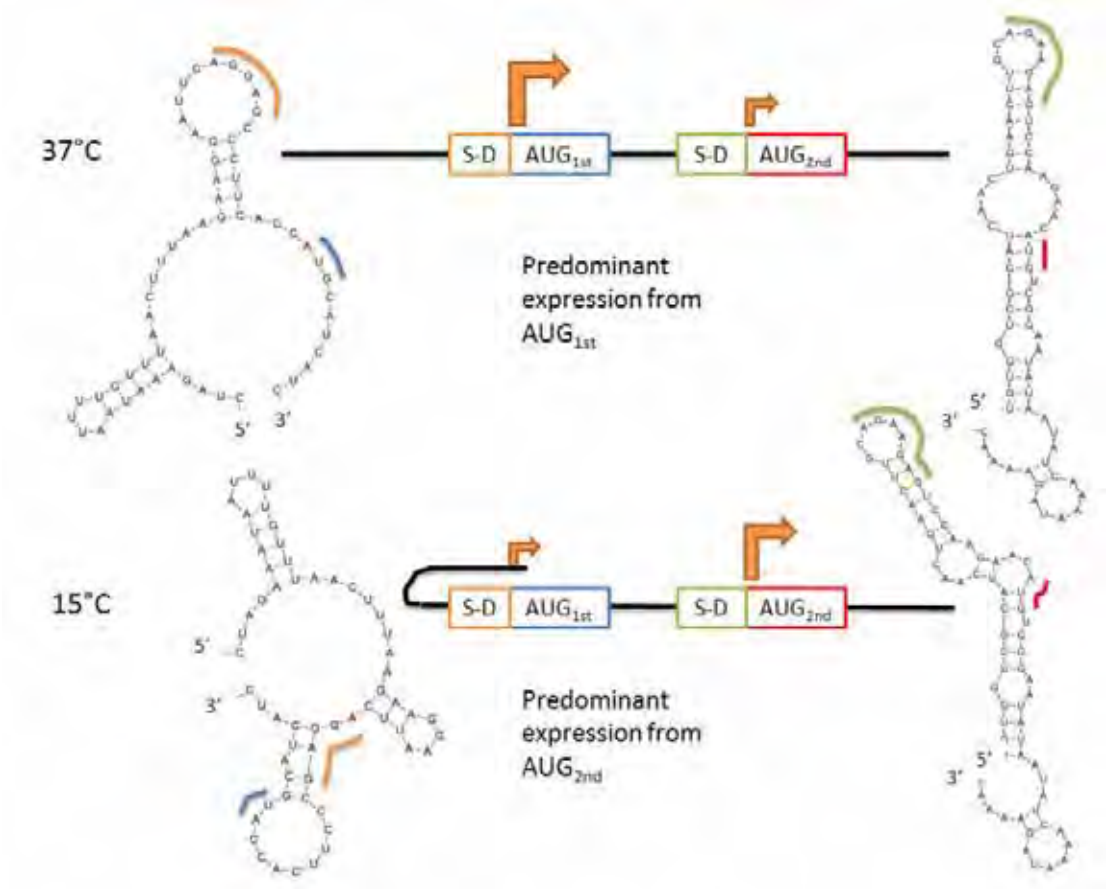


Figure 27: A schematic model showing the hypothetical rationale behind the temperature-dependent switch of protein expression.

Secondary structures from the translation initiation region (containing S-D and AUG initiation codon) of the HisMbTBP206 gene-containing pET101 plasmid were generated with the Mfold [190] program with the default settings. The positions of the S-D elements and AUG codons are marked, and colour-coded on the illustrated secondary structures. Expression (orange-coloured right-angled arrows) from the first and second AUG codon would produce HisMbTBP206 and MbTBP183, respectively.

Several studies have shown that the secondary structures of the mRNA at the Shine-Dalgarno (S-D) sites can influence the translational efficiency of a protein [191-193]. The translation efficiency of a protein was found to decrease as a result of an increased proportion of mRNAs with folded secondary structures around the S-D sites. As illustrated in Figure 27, secondary structure predictive modeling of the translation initiation elements (S-D and initiation codon) of the HisMbTBP206 gene at 37°C revealed that the elements for HisMbTBP206 were exposed and unfolded, while that of MbTBP183 was folded. Consistent with the results from the predictive modeling,

translation of the shorter MbTBP183 was found to be lower as compared to HisMbTBP206 at 37°C (Figure 19).

At 15°C, the translation initiation elements (TIE) for HisMbTBP206 were predictably folded, thereby reducing translational capability for HisMbTBP206. Conversely, the predicted structure of TIE for MbTBP183 was partially unfolded at 15°C, potentially increasing the translation of MbTBP183. Consistent with the predictive modeling results; MbTBP183 was the predominant isoform when attempts to express the HisMbTBP206 protein at 15°C were made (Figure 19). The changes in the structure of recombinant mRNA transcripts under different temperatures (Figure 27) could probably explain the temperature – dependent shift in protein isoform expression (Figure 19).

In conclusion, a protocol for the large-scale purification of both the recombinant HisMbTBP206 mutant, and HisMbTP183 was developed, after numerous pilot expression studies and protein purification trials. Using the adopted protocol, both recombinant protein isoforms were highly purified, for use in subsequent biophysical and functional studies detailed in the following chapters.

CHAPTER VI

Biophysical characterisations of recombinant MbTBP isoforms

6.1 Introduction

Biophysical characterisation allows the examination of various structural properties of a protein that give rise to its biological function. Many techniques and tools can be used to investigate the biophysical aspects of a protein. These include: size-exclusion chromatography (SEC), circular dichroism (CD), and ANS-fluorescence spectroscopy that can be used to assess the oligomerisation status, secondary structure, and hydrophobic content of a protein, respectively.

6.1.1 Chapter objectives

This chapter reports on the characterisation of the biophysical properties of both HisMbTBP183 and HisMbTBP206 mutant, through analysis with size-exclusion chromatography, circular dichroism, and ANS-fluorescence spectroscopy. In addition, the use of SEC coupled with native-Western blotting would provide an insight of the oligomerisation states of MbTBP *in vivo* and *in vitro*. Further, the structure of the RepA segment in the N-terminal region of MbTBP206 (Figure 5) would also be determined by comparing the CD-estimated structures of HisMbTBP183 and HisMbTBP206 mutant.

Additionally, since the interactions between the concave surface of a TBP and the minor groove of the promoter DNA sequence are primarily hydrophobic in nature [112-114], ANS fluorescence spectroscopy was used to study the amount of hydrophobic surfaces present on both HisMbTBP isoforms. The stability of the protein isoforms could also be evaluated by comparing changes to the hydrophobic content of the protein isoforms when heated at different temperatures.

6.2 Results

6.2.1 HisMbTBP isoforms are dimeric in solution

Each purified batch of recombinant HisMbTBP206 mutant and HisMbTBP183 (from Chapter 5) was subjected to analytical analysis on a Superdex 75 10/300GL (GE Healthcare) chromatographic column, under the MB buffer system, at approximately 23°C. The gel filtration LMW calibration kit (GE Healthcare) was used as the protein

molecular weight standards, and the standards were analysed under the same conditions (Chapter 2.2.6.6) as that of the recombinant proteins.

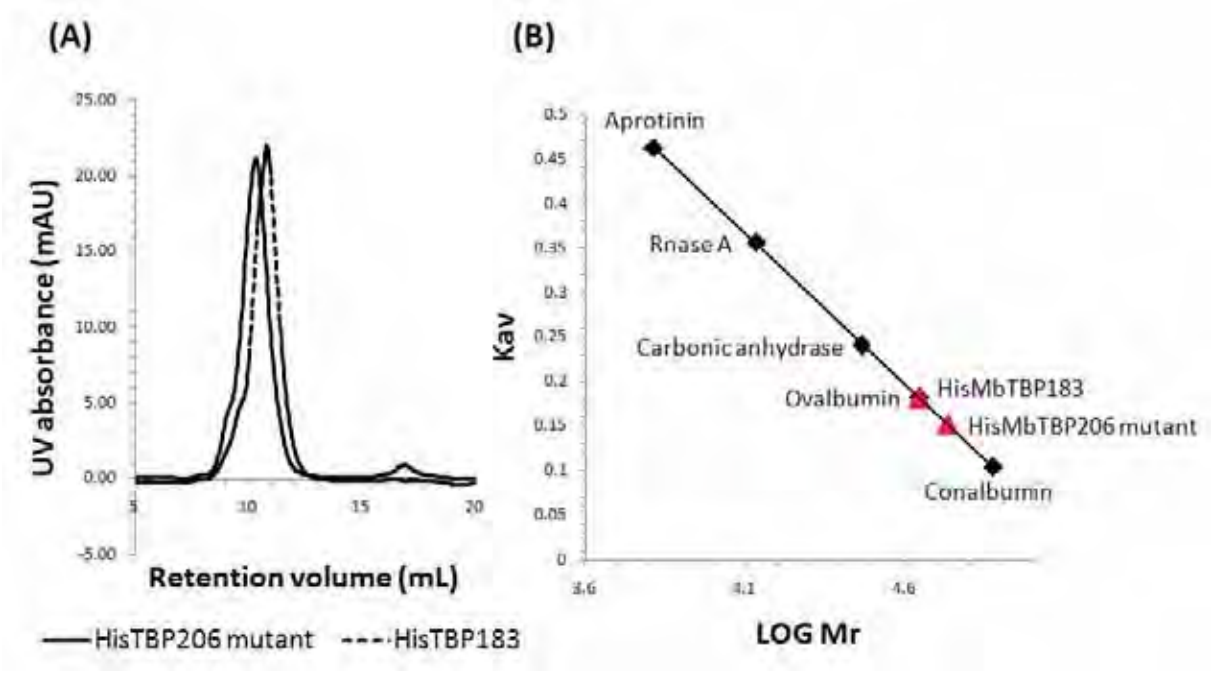


Figure 28: Analytical size-exclusion chromatography (SEC) of the recombinant MbTBP isoforms.

Two separate batches of proteins, for each isoform, were used in analytical SEC. Since data from different batches of the same isoform were virtually identical, only one data set from each protein isoform was shown. (A) A typical chromatogram for HisMbTBP206 mutant, and HisMbTBP183. (B) The calibration graph used for the determination of the molecular weight of the recombinant isoforms. The triangle and square data points refer to the recombinant isoforms, and protein standards respectively.

The molecular weights of the HisMbTBP206 mutant and the HisMbTBP183, in a phosphate-buffered solution, was determined to be 52.6kDa and 40.8KDa, respectively (Figure 28). Since the theoretical molecular weights of HisMbTBP206 mutant and HisMbTBP183 was 23.2kDa and 20.7KDa respectively, the results obtained are consistent with that expected of a protein that forms a homo-dimeric quaternary structure in aqueous solution.

6.2.2 Higher-ordered MbTBP oligomer *in vivo*

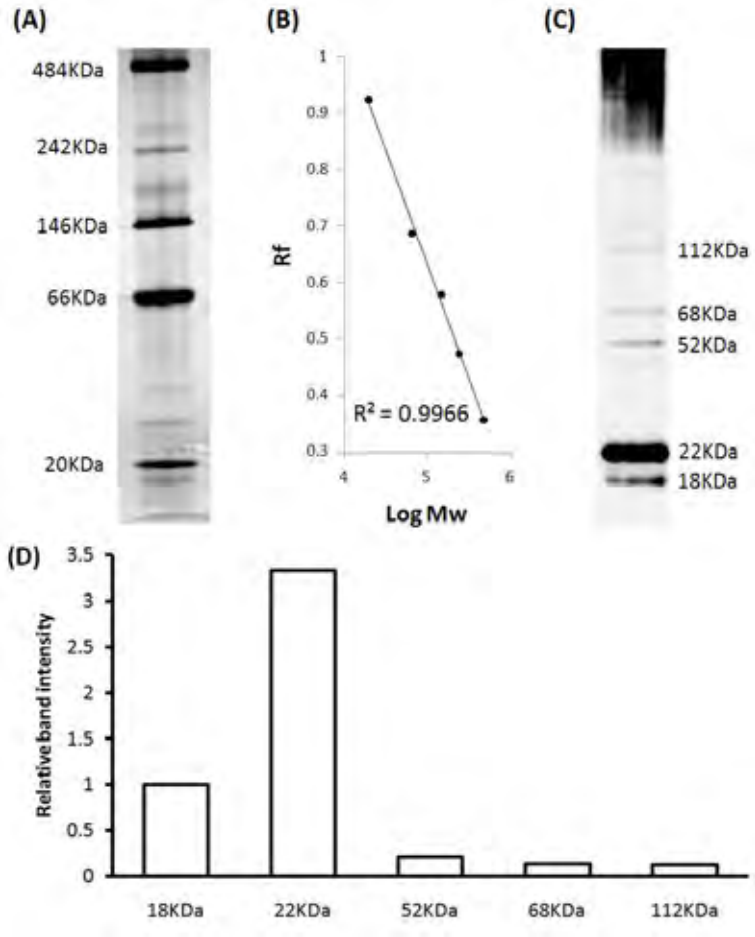


Figure 29: Native-Western analysis of MbTBP isoforms.

(A) Blot result from the native-Western analysis of the native protein marker. (B) Calibration graph of Rf versus logarithm of protein molecular weight, based on (A). Rf refers to the ratio of the distance migrated by the protein band against that of the dye front. (C) Blot result from native-Western analysis of soluble protein lysates from *M. burtonii*. The high molecular weight smear at the top of the blot is likely to be aggregated MbTBP. (D) Relative band intensities of the different MbTBP bands identified from (C), as determine by densitometry analysis of the bands.

In order to examine the oligomerisation state of MbTBP *in vivo*, a native-Western approach was used. Briefly, soluble protein lysates from *M. burtonii* were resolved using native-blue polyacrylamide gel electrophoresis, transferred onto a PVDF membrane, and probed for MbTBP using Western blotting (Chapter 2.2.6.5). A native

protein marker was used as a molecular weight standard to estimate the molecular weight of the MbTBP oligomers.

As shown in Figure 29, the estimated size of the MbTBP monomers was 22kDa and 18KDa, which is similar to the molecular weight of 22.3kDa and 19.7KDa expected of MbTBP206 and MbTBP183, respectively. The molecular weights of the higher-ordered MbTBP oligomers were estimated to be 52KDa, 68KDa, and 112KDa. It is possible that the 52KDa and the 68KDa MbTBP oligomer might have dimerised with each other to form the 112KDa oligomer MbTBP observed. As these high molecular weight MbTBP bands (52KDa, 68KDa, and 112KDa) were evident on a native-Western blot, it is clear that MbTBP oligomerises *in vivo*. Additionally, based on the relative intensity of the protein bands (Figure 29C and 29D), it is evident that the majority (90%) of the proteins remain monomeric *in vivo*. Further, between the two monomeric protein species observed, MbTBP206 was found to be approximately 3.5 times more abundant than MbTBP183 (Figure 29D).

6.2.3 MbTBP isoforms are structurally similar

Far UV circular dichroism (CD) analysis was employed to answer two questions: firstly, are the purified recombinant HisMbTBP isoforms folded; and secondly, how do the two isoforms differ in structure and stability. As such, CD spectra of the recombinant proteins at their naturally occurring (4°C) and denaturing (80°C) temperatures, as well as their melting profiles were determined, and analysed.

Comparisons that could be made between the recombinant isoforms at 4°C and 80°C were limited by the high tension voltage (HTV) generated in the far UV region of 190nm – 200nm wavelength at 80°C (Figure 30B). This is because high HTV (exceeding 600V) potentially produces artifacts and inaccurate CD data [194]. Attempts to lower the HTV below 600V, by diluting the chloride salts in the purified protein samples that could contribute to the HTV at high concentrations [143,194], were unsuccessful. As such, CD data obtained at 200nm – 260nm wavelength at both temperatures were

compared and analysed instead; since the HTV registered for both recombinant isoforms in this range was less than 600V. Analysis of the data revealed that the far UV spectra of both recombinant isoforms shifted towards zero Δε when both were heated to 80°C. Such changes to the UV spectra imply a loss of secondary structures and possible denaturation of the isoforms at 80°C (Figure 30A).

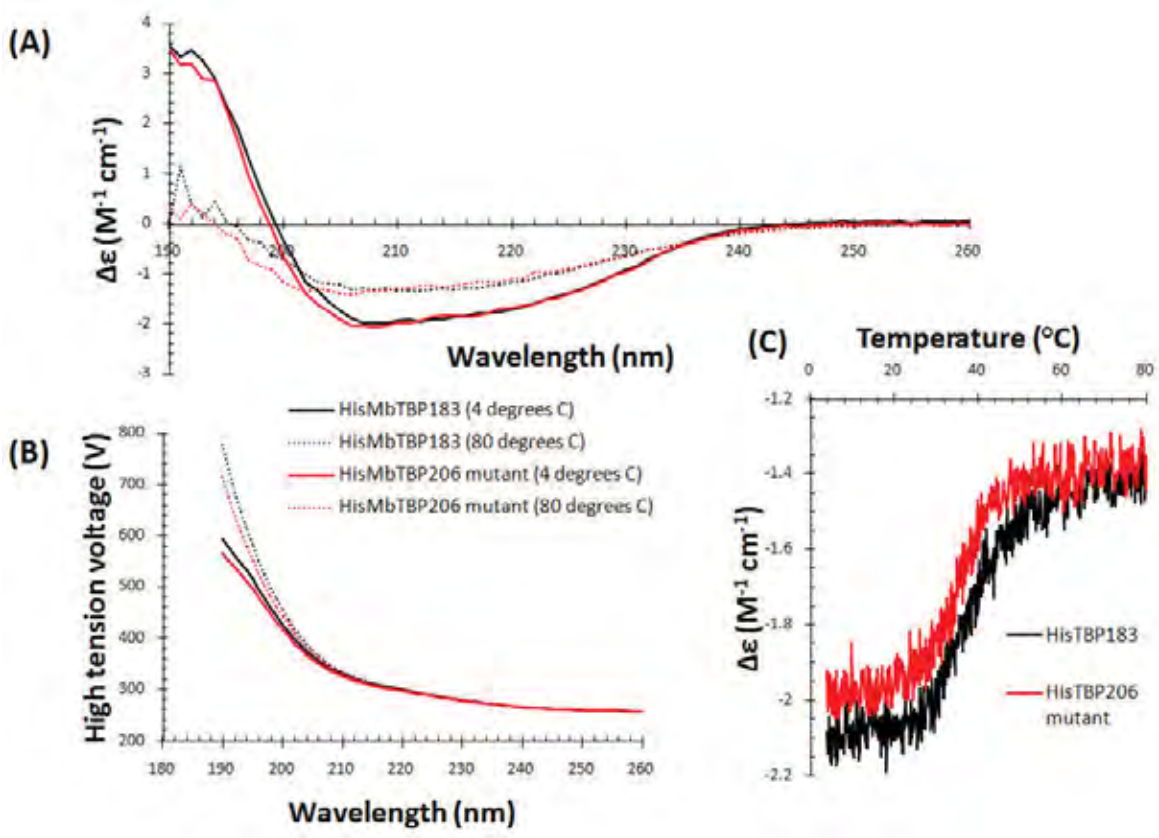


Figure 30: Circular dichroism (CD) analysis of recombinant HisMbTBP isoforms.

Two separate batches of proteins, for each isoform, were used in the CD analysis. Since data from different batches of the same isoform were virtually identical, only one data set from each protein isoform was shown. (A) Far UV CD spectra and (B) corresponding high tension voltage traces of the two recombinant HisMbTBP isoforms at 4°C and 80°C. (C) Melting profile of the recombinant HisMbTBP isoforms, at 208nm wavelength. Both recombinant HisMbTBP isoforms produced a thermal melting temperature of approximately 36°C.

The far UV CD spectra for both HisMbTBP isoforms, at 4°C, are consistent with that of a properly folded protein (Figure 30A). In order to understand the type of protein secondary structures that the spectra measurements translate into, the CD spectra of

the recombinant isoforms at 4°C were uploaded onto the online DICHROWEB [195-197] server and deconvoluted using both the CONTINLL [198-199] and CDSSTR [200] algorithms.

Secondary structure estimations of both the HisMbTBP183 and HisMbTBP206 mutant were obtained and deemed reliable, based on their NRMSD values, that were less than 0.1 for both the CONTINLL and the CDSSTR algorithms. A normalised root mean square deviation (NRMSD) is a reflection of the 'goodness of fit' parameter between experimental and calculated CD spectra by algorithms. An NRMSD value of zero constitutes a perfect fit, and values of 0.1 or lower would indicate that the deconvolution analysis was unlikely to be incorrect [194,197]. Both the CONTINLL and the CDSSTR algorithms predicted similar secondary structure estimations for both recombinant isoforms (Table 4).

The α-helix and β-sheet structural compositions obtained from the deconvoluted CD-spectra of the HisMbTBP183 protein are similar to that of other archaeal and eucaryal TBP cores determined by x-ray crystallography (Table 4 & 5). This was expected since the entire MbTBP183 protein sequence aligns closely to that of the core regions of other archaeal and eucaryal TBPs. As both the HisMbTBP206 mutant and the HisMbTBP183 have identical TBP cores, and are estimated to contain an almost identical number of residues displaying α-helix and β-sheet structures (Table 4), both isoforms appear to share the same TBP core conformation.

6.2.4 The N-terminal RepA segment in HisMbTBP206 mutant is unstructured

The main difference between the two recombinant isoforms lies in the amount of unordered structures within the proteins. Specifically, the HisMbTBP206 mutant has 14 more amino acid residues that form unordered structures in the protein compared to HisMbTBP183 (Table 4). Since the core structure of both isoforms is virtually identical, it is possible that these 14 amino acid residues might reside on the N-terminal RepA segment (Figure 5) of the HisMbTBP206 mutant, which is not present in HisMbTBP183.

The thermal melting profiles of HisMbTBP183 and HisMbTBP206 mutant were also determined over the temperature range from 4°C to 80°C. Both isoforms exhibited a highly similar thermal melting profile, and shared an estimated protein melting temperature of 36°C (Figure 30C). As thermal energy is required to destablise ordered structures such as α-helix and β-sheet; it follows that a higher temperature would be required to denature the HisMbTBP206 mutant compared to HisMbTBP183, should the RepA segment that is unique to the longer isoform be structurally ordered. However, the closely identical thermal melting temperatures of both isoforms imply that very little extra heat was needed to denature the RepA segment in the longer isoform. Therefore these observations are consistent with the proposal that the RepA segment in MbTBP206 is unordered.

	HisMbTBP183				HisMbTBP206 mutant			
Structure / Algorithm	α-helix	β-strand	Turn	Unordered	α-helix	β-strand	Turn	Unordered
CONTINLL	17% ± 0.7%	27.95% ± 0.25%	19.35% ± 0.15%	35.6% ± 0.6%	17.1% ± 0.9%	25.75% ± 0.55%	18.75% ± 0.25%	38.35% ± 0.55%
CDSSTR	12% ± 4%	32% ± 3%	20% ± 1%	35% ± 2%	12.5% ± 4.5%	29.5% ± 3.5%	20.5% ± 0.5%	38% ± 1%
Average of both algorithm	14.5%	29.975%	19.675%	35.3%	14.8%	27.625%	19.625%	38.175%
No. of Residue	28	57	37	67	32	59	42	81

Table 4: Secondary structure estimation of both HisMbTBP183 and HisMbTBP206 mutant recombinant proteins, using the CONTINLL and CDSSTR algorithm.

The percentage of secondary structures in each protein isoform was obtained from two distinct recombinant protein batches, and is displayed as mean ± standard error of mean. The numbers of residues predicted to bear the respective secondary structures were calculated by multiplying the structural percentage with the total length of either HisMbTBP183 (190 residues) or HisMbTBP206 mutant (213 residues).

PDB ID	Origin	Helical	β-sheet
3EIK	*Encephalitozoon cuniculi*	20.8%	36.5%
2Z8U	*Methanococcus jannaschii*	19.7%	29.3%
1NH2	*Saccharomyces cerevisiae*	21.1%	34.4%
1MP9	*Sulfolobus acidocaldarius*	19.2%	32.1%
1NVP	*Homo sapiens*	20.4%	35.4%
1QN5	*Arabidopsis thaliana*	20.0%	29.2%
1PCZ	*Pyrococcus woesei*	19.1%	32.8%
Average		20.0%	32.8%

Table 5: Structural composition of the core regions of various archaeal and eucaryal TBPs, derived from x-ray crystallography.

The percentage of helical and β-sheet structures was based on the number of author-approved residues from protein database (PDB) that bear the respective structure, as compared to the length of the TBP core.

6.2.5 Cell solutes reduce the effects of heat on the HisMbTBP isoforms

ANS (1-anilino-8-naphthalene sulfonate) fluorescence spectroscopy was used to probe and record the hydrophobic content on the recombinant HisMbTBP isoforms, as ANS fluoresces only in a non-aqueous environment, such as an apolar organic buffer [201], or when interacting with hydrophobic regions of a protein [202-203]. The fluorescence emission spectra of the recombinant proteins showed that the ANS molecule fluoresces intensely upon interaction with the proteins (Figure 31A). The fluorescence emission maxima of this interaction occurred at wavelengths of approximately 460nm – 465nm.

In order to estimate the amount of hydrophobic surfaces on a particular protein, the thermal melting profile of the protein was measured using ANS at 465nm wavelength, from 4°C to 50°C (Chapter 2.2.6.8). At low temperatures, proteins are structurally rigid and compact, thereby restricting the access of ANS molecules to hydrophobic surfaces

within proteins. When the temperature increases, proteins gain fluidity and are no longer compact. This therefore allows ANS molecules to interact with hydrophobic surfaces, and to emit intense fluorescence signals. As such, the destruction of protein conformation due to thermal denaturation would result in a loss of hydrophobic surfaces available to interact with ANS molecules. Consequently, ANS fluorescence signals would diminish as thermal denaturation of the protein proceeds.

Therefore, the amount of hydrophobic surfaces on a protein, at any given temperature, can be derived from data in the thermal melting profile by calculating the ratio of its ANS fluorescence intensity at that temperature to the maximum possible ANS fluorescence intensity for the protein; and expressing the ratio in terms of percentages (Figure 31B). For example, a protein that with a surface hydrophobicity of 68% at 28°C implies that the amount of surface hydrophobicity at 28°C for that protein is 68% of its maximum possible surface hydrophobicity.

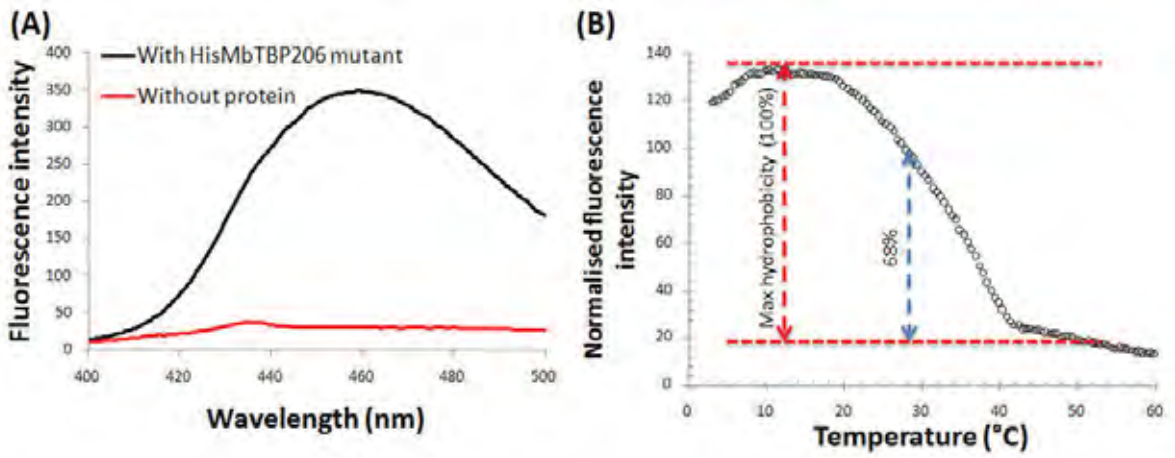


Figure 31: ANS fluorescence analysis of HisMbTBP206 mutant and HisMbTBP183.

(A) Typical ANS fluorescence emission spectra for HisMbTBP206 mutant and HisMbTBP183. (B) A typical ANS fluorescence melting profile of a HisMbTBP206 mutant. Calculation of the percentage hydrophobicity of a protein at a particular temperature was based on the fluorescence intensity measured at the chosen temperature (blue arrow) against the maximum possible hydrophobicity, marked by the highest fluorescence intensity of the protein (red arrow).

Comparison of the surface hydrophobicity of the same protein in different buffers would allow a study of the effects of various solutes on surface hydrophobicity. In this case, the melting profiles of recombinant HisMbTBPs in two different buffers, gel filtration (GF) and MB buffers were examined by ANS fluorescence spectroscopy. MB buffer is a modified GF buffer (Material and Methods, Chapter 2.1) that contains solutes such as potassium, aspartate, and glutamate, which are known to be present in *M. burtonii* [204]. Specifically, NaCl in GF buffer was replaced with 200mM of KCl, 100mM of potassium aspartate, and 100mM of potassium glutamate in MB buffer.

	HisMbTBP183		HisMbTBP206 mutant	
Buffers / Temperatures	GF	MB	GF	MB
4°C	79.6% ±3.6%	54.9% ±11.8%	83.5% ±6.4%	81.9% ±5.8%
23°C	74.7% ±8.1%	93.2% ±3.7%	71.5% ±1.2%	89.1% ±2.8%
28°C	47.6% ±11.2%	77.7% ±3.6%	41.5% ±8.5%	70.1% ±1.8%
37°C	2.0% ±2.4%	28.7% ±0.1%	0.0% ±1.6%	26.2% ±0.0%

Table 6: A comparison of the amount of hydrophobic surfaces in a recombinant HisMbTBP isoform in GF and MB buffer, at various temperatures.

The percentage is an indicator of the amount of protein surface hydrophobicity as compared to the maximum possible amount of surface hydrophobicity that a protein can display. Calculations for the percentage hydrophobicity were performed based on the description in Figure 31B. The percentage hydrophobicity was obtained from two separate protein batches of each protein isoform, and displayed as mean ± standard error of mean. In MB buffer, differences in the protein surface hydrophobicity of HisMbTBP183 at 4°C and 23°C were significant (p-value < 0.025), while those of the HisMbTBP206 mutant were not (p-value > 0.05).

Heating both HisMbTBP isoforms from 4°C to 37°C caused a loss of hydrophobic surfaces in the proteins (Table 6). However, both HisMbTBP isoforms, when placed in MB buffer, retained more than 20% hydrophobicity at 28°C and 37°C as compared to

recombinant protein isoforms incubated in GF buffer (Table 6). For example at 28°C, which is the maximum temperature at which *M. burtonii* is known to grow, the average hydrophobicity for both isoforms was approximately 41 – 48% in GF buffer. On the other hand, in a buffer that mimicked the cell solute content of *M. burtonii* (MB buffer), the average hydrophobicity of both isoforms was higher, at 70 – 78%. Therefore, these observations imply that solutes uniquely found in the MB buffer, such as aspartate and glutamate reduce the extent of heat-induced loss of surface hydrophobicity for both MbTBP isoforms at 28°C and 37°C.

Finally, it is also important to note that both HisMbTBP isoforms, in MB buffer, had approximately 89% – 91% surface hydrophobicity at 23°C, which is the temperature at which *M. burtonii* grows fastest. Since the interaction between a TBP and promoter DNA sequence is hydrophobic in nature [113-114], the high (> 85%) MbTBP surface hydrophobicity can potentially translate into an increased binding efficiency between both macromolecules, and possibly increased transcriptional efficiency of genes in *M. burtonii*, thus resulting in the rapid growth of *M. burtonii*.

6.3 Discussion

Gel filtration studies revealed that the oligomerisation status of the recombinant HisMbTBP206 mutant and HisMbTBP183 to be dimeric. This observation is in agreement with studies showing that in the absence of DNA, TBPs interact and bind to each other to form a homodimer [81,115,124,126,205-206]. Dimerisation of TBPs was previously found to be a key factor in transcription regulation in eucaryotes. Jackson-Fisher *et al*. (1999) had reported through mutational studies, that the activator-independent transcription of *lacZ* reporter gene was significantly increased in yeast that contained unstable TBP dimers [207]. Further, in the absence of DNA binding, TBP dimerisation also appears to reduce TBP inactivation [208] and degradation [207]. Therefore dimerisation of the HisMbTBP206 mutant and the HisMbTBP183 could potentially serve similar purposes. In addition, TBP dimerisation also provides a readily available reserve pool of TBPs for eliciting adaptive responses when needed.

Contrary to the gel filtration studies, native-Western analysis of *M. burtonii* protein lysates showed that the majority (approximately 90%) of MbTBP exists as monomers, with only a small proportion existing as oligomers. The discrepancy between the gel filtration and native-Western studies can be accounted for by the nature of the latter method.

Native-Western relies on a BisTris-Tricine native gel system, which unlike gel filtration, was performed in the absence of stablising solutes or agents. Therefore, the lack of stabilising solutes or agents could destabilise TBP oligomers and thus account for the large amount of monomeric MbTBP detected. Further, similar to magnesium cations, the BisTris cation could exert an additional destabilising effect on oligomeric TBP formations [209]. Such experimental conditions can potentially result in the bulk of MbTBP being monomeric, rather than dimeric as in the case of gel filtration studies. Nonetheless, since oligomeric MbTBPs were detected by both gel filtration and native-Western procedure, there is supporting evidence that MbTBPs oligomerise in *M. burtonii*.

Dissociation of TBP dimers into monomers is the rate-limiting step in a TBP-DNA binding event [208,210]. A slow dissociation rate of TBP dimers would result in a slower and less efficient transcription initiation. In an eucaryotic transcription system, dissociation of TBP dimers is facilitated by the transcriptional factor TFIIA [210]. However, no TFIIA homologue was identified in *M. burtonii* or other Archaea. Therefore, it is possible that other transcription factors might be responsible for the TBP dimer dissociation in Archaea. It would be worthwhile to identify such factors in future studies to understand the transcription initiation process in Archaea.

MbTBP183 is essentially the C-terminal core region of an MbTBP. As X-ray crystal structures of many similar eucaryal and archaeal C-terminal TBP core regions have already been previously determined [81,112,115,124-126], it would be appropriate to follow

up with X-ray crystallography studies on the full length MbTBP206 that contains the extra RepA segment at its N-terminal. However, attempts to elucidate the X-ray crystal structure of HisMbTBP206 were unsuccessful as no crystals were obtained despite using many different buffer conditions. This was unexpected since the RepA segment shares a high sequence identity with the RepB segment, where homologues of the RepB segment in other eucaryal and archaeal TBP were found to form β-strand structures (Figure 8). Therefore, RepA was predicted to form similar β-strand structures that should not interfere with the protein crystallisation process. However, the lack of crystals for HisMbTBP206 indicated that the RepA segment might not exist as an ordered structure. To verify this conclusion, CD studies were carried out on the MbTBP protein isoforms to determine the protein secondary structures present.

As expected, the CD studies indicated that the N-terminal region of the recombinant HisMbTBP206 mutant was found to be potentially disordered, which suggested the presence of flexible structures; thus it is likely that the flexible structures might have interfered with the protein crystallisation. This finding was in line with other observations. For example, Nikolov *et al*. (1992) reported that the N-terminal regions of both TBP copies in an *Arabidopsis* TBP dimer displayed different conformations, providing evidence of disordered N-terminal regions [124]. Further attempts to elucidate the structure of the long N-terminal domain of yeast TBP through X-ray crystallography were unsuccessful due to the presence of these disordered regions [125].

Current knowledge on the function of a long flexible N-terminal region in TBP is limited. In the TBP of yeast, the flexibility of the N-terminal region was found to be crucial for the autoregulatory function of TBP [121-122,211], as it formed a flexible 'lid' for the concave DNA-binding cleft. However, the N-terminal domain (containing RepA segment) of MbTBP206 was unlike that of yeast, which contained a unique tryptophan residue that could be used as a reporter of the structural plasticity of the domain in fluorescence quenching studies [121]. Thus, it was not possible to conduct such direct studies on MbTBP to investigate the dynamics of its long N-terminal region. However, given that

the RepA segment was unstructured and flexible, which makes it similar to the N-terminal domain of yeast TBP; this implies that the long N-terminal region of MbTBP206 could also potentially exert a similar autoregulatory effect on the DNA-binding function of MbTBP206.

In buffer conditions mimicking the intracellular solute content of *M. burtonii*, it was also found that differences in the protein surface hydrophobicity of HisMbTBP183 at 4°C and 23°C was significant (p-value < 0.025), while those of the HisMbTBP206 mutant were not (p-value > 0.05; Table 6). These findings imply that MbTBP206, as compared to MbTBP183 demonstrates a greater resistance towards temperature-induced loss of protein surface hydrophobicity, which is crucial for TBP-DNA interaction [113-114]. Thus, it appears that the addition of a RepA segment to MbTBP206 might serve to provide MbTBP206 with an added functional advantage over MbTBP183, by allowing the former to retain a higher protein surface hydrophobicity and thus functionality over a range of temperatures such as 4°C to 23°C compared to the latter. Having a preference for the better adapted MbTBP206 over MbTBP183 could also be the reason why *M. burtonii* was found to contain an over-abundance of the former compared to the latter (Chapter 4; Table 3).

ANS fluorescence studies showed that solutes such as potassium, aspartate and glutamate minimised the loss of hydrophobic surfaces on recombinant HisMbTBP isoforms, over the temperature range from 4°C to 28°C. In particular, potassium [212] and L-α-glutamate [204] are two main intracellular solutes in methanogenic archaea such as *M. burtonii*. *M. burtonii* also contains intracellular solutes specific to it, such as L-aspartate, which is absent in *Methanosarcina thermophila*, a closely related thermophile [204]. Therefore, it is likely that the hydrophobic DNA-binding surfaces of MbTBP are stabilised by solutes (potassium, aspartate, and glutamate) intracellular to *M. burtonii*, since the loss of MbTBP hydrophobic surfaces due to heating was greatly reduced when MbTBP was incubated in buffers containing these solutes.

Further work conducted on the EF-2 protein from *M. burtonii* demonstrated a role that these intracellular solutes play in stabilising the EF-2 protein [204,213]. Together, the work on EF-2 and MbTBP implies that intracellular solutes of *M. burtonii* stablise protein function over a wide range of growth temperatures (from -2°C to 28°C) of *M. burtonii*. These studies also highlight the importance of using a buffer system containing compatible solutes to preserve protein function.

In conclusion, both recombinant HisMbTBP isoforms were structurally folded. Both isoforms dimerised in solution, potentially creating a reservoir of readily available but dormant TBPs that can be used when required by *M. burtonii*. Further characterisation of both MbTBP isoforms by CD and ANS-fluorescence revealed differences between the two isoforms. In particular, MbTBP206 was found to contain a disordered RepA segment unique to it that was not found in MbTBP183. From the studies on protein surface hydrophobicity, it also appears that MbTBP206 was better adapted than MbTBP183 to function between 4°C and 23°C. This significant finding might account for why the RepA segment in MbTBP206 was retained throughout the evolutionary history of *M. burtonii*. Finally, the two recombinant isoforms were stabilised when placed in a buffer containing KCl, and potassium salts of glutamate and aspartate, all of which were intracellular solutes present in *M. burtonii*. This was a vital finding applicable for further functional studies of recombinant HisMbTBPs in the following chapter.

CHAPTER VII

Functional characterisations of recombinant HisMbTBP isoforms

7.1 Introduction

7.1.1 Determination of the TATA-box-binding properties of both HisMbTBP isoforms by EMSA

An electrophorectic mobility shift assay [214] (EMSA), when used in combination with radioisotope-labeled DNA, is a highly sensitive technique. It can provide both qualitative and quantitative data [214-215], and is as described in Chapter 2 and Figure 32. Essentially, solutions of protein and radiolabeled nucleic acid fragments are mixed to allow for any binding to occur between the protein and the nucleic acids. The resulting mixture is then electrophoresed under native conditions. Subsequently, the protein-nucleic acid complexes and free nucleic acids are visualised using autoradiography. The protein-nucleic acid complexes migrate at a slower rate than free nucleic acids as they have a higher molecular weight. This method provides a useful way of identifying nucleic acid-binding proteins such as archaeal TBP and transcriptional regulators [33,216-217].

Although archaeal-based cell-free transcription systems could also be used to identify archaeal TBPs [137,139,218-219], existing archaeal cell-free transcription systems are derived from mesophilic or thermophilic archaea. Therefore, these cell-free transcription assays require a relatively high temperature range of 37°C – 75°C [137,220-222] to function. This is in contrast to the low temperatures generally required for psychrophilic proteins to maintain their native protein conformation and function [69,79]. Since psychrophilic proteins would have been thermally denatured at such high temperatures, it was not feasible to determine the TATA-box binding properties of psychrophilic TBPs using these cell-free transcription assays. Instead, EMSA will be a more suitable and efficient technique.

7.1.2 Elucidating HisMbTBP-binding DNA motifs and sequences by genome-wide promoter pull-down assays and high throughput sequencing

Experiments such as fusion tag pull-down, chromatin immunoprecipitation (ChIP) and SELEX (systematic evolution of ligands by exponential enrichment) are commonly used

to study protein-DNA interactions [223-225]. In pull-down experiments, fusion-tagged proteins are used as baits to capture interacting DNA fragments (prey) from a pool of DNA. Subsequently, the bait-prey complex is purified using an immobilised affinity ligand that is specific to the fusion tag used. The nucleotide sequence of the interacting DNA fragment is subsequently determined to allow the identification of protein-interacting DNA motifs.

Alternatively, the ChIP [223] method uses antibodies to capture its protein antigen as well as the DNA partners that may be bound to the antigen. However, in a cell containing multiple protein isoforms that are virtually identical, the antibody would capture all these isoforms and their respective interacting DNA partners. Therefore, it would not be possible to assign the interacting DNA partners to their respective protein isoform. Thus an *in vitro* pull-down experiment is preferred to an *in vivo* ChIP experiment when determining the interacting DNA partners of a protein isoform. This procedure allows each individual protein isoform to be expressed separately, and be used as bait in pull-down experiments to capture the DNA partner.

SELEX [225] is another method that has also been used to identify protein-interacting DNA such as transcription factors (TF) [226-228]. However, SELEX tends to select against DNA motifs that display weak or less than optimal interactions with TF, potentially eliminating true protein-binding DNA motifs that interact weakly with proteins. Therefore, DNA sequences obtained from SELEX would not be a good representation of the actual transcription factor-binding DNA motif *in vivo* [229-230]. Instead of using SELEX, a genome-wide promoter pull-down approach, coupled with high throughput sequencing would be better suited for the identification of naturally occurring TF-binding DNA motifs and their consensus DNA sequences.

DNA sequencing is crucial in characterising protein-interacting DNAs that are obtained from pull-down, ChIP or SELEX. In particular, the Illumina/Solexa [231-232] next generation sequencing technology relies on a cloning-free DNA amplification and DNA

sequencing-by-synthesis approach to rapidly produce at least one billion base pairs of sequence data per run in 2 – 3 days, with read lengths of approximately 30bp – 40bp [232]. Additionally, a heterogeneous pool of DNA (of different DNA sequences) can be sequenced in parallel in a single run. Such high-throughput sequencing technology has since been used for whole genome and metagenomic sequencing [233-234], transcriptome sequencing [235], identification of single nucleotide polymorphism (SNP) [235], and profiling of genomic DNA methylation [236]. This well established technology is useful in the rapid sequencing of the heterogeneous pool of promoter DNAs obtained from pulled down or ChIP experiments and from there consensus DNA sequences for protein binding.

7.1.3 Chapter objectives

One of the aims of this chapter was to demonstrate the binding between the two psychrophilic HisMbTBP isoforms and TATA-box sequences by using EMSA [214]. EMSA was preferred over mesophilic/thermophilic archaeal-based cell-free transcription systems for demonstrating HisMbTBP-TATA binding because the experiment can be performed at low temperatures. It would preserve the HisMbTBP surface hydrophobicity that is crucial for DNA interaction, which would otherwise be lost at temperatures above 28°C (Chapter 6) when using mesophilic/thermophilic archaeal-based cell-free transcription systems. Additionally, binding of the TATA-box sequence by the HisMbTBP isoforms would also confirm that the recombinant isoforms produced using the methods developed in Chapter 5 were functional. Further, HisMbTBP-TATA bindings would also signify that the gene locus Mbur_1496 had been correctly annotated as a *tbp* gene of *M. burtonii*, since HisMbTBP is a hexa-histidine tagged version of the protein product of Mbur_1496.

Finally, since the anti-MbTBP antibody can detect both MbTBP183 and MbTBP206 isoforms in Western blot, it was not suitable for use in the ChIP method to obtain specific TBP-binding DNA for each MbTBP isoform. As such, pull-down experiments were used to identify interacting DNAs corresponding to each specific HisMbTBP isoform. In addition, a genome-wide promoter pull-down approach coupled with

Illumina sequencing was used to elucidate naturally occurring HisMbTBP-binding DNA motifs and consensus sequences. Subsequently, gene targets that are potentially regulated by the two MbTBP isoforms could be determined based on the identified putative promoters. Functional differences between the two MbTBP isoforms could then be inferred from the results.

7.2 Results

7.2.1 Both HisMbTBP183 and HisMbTBP206 mutant binds TATA sequence

The TATA-box binding properties of the recombinant HisMbTBP isoforms were assessed using EMSA. The recombinant isoforms were mixed at 4°C, with radiolabeled DNA oligonucleotides containing the consensus TATA sequence, which was designed based on the promoter elements of methanogenic archaea [99,217]. The TBP-bound and free DNA oligonucleotides were subsequently separated on a polyacrylamide gel and visualised by autoradiography.

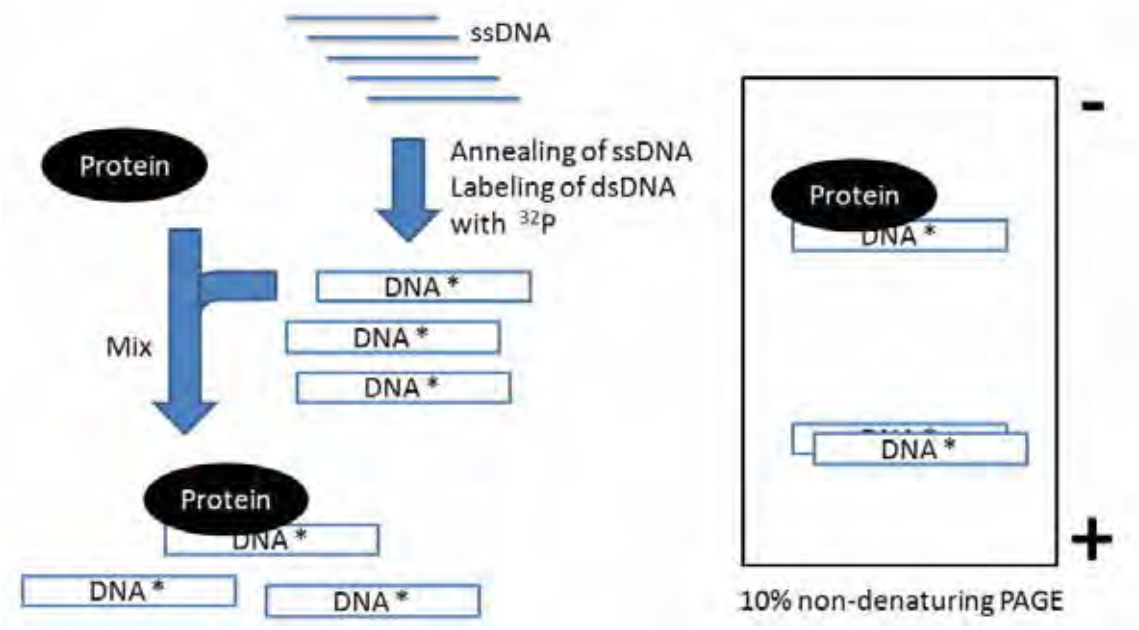


Figure 32: Schematics detailing the EMSA procedures that were used to assess the functionality of the recombinant HisMbTBP isoforms.

The binding of the recombinant HisMbTBP isoforms to radiolabeled DNA oligonucleotides yielded distinct 'gel-shifted' bands that migrated at a slower rate than the free DNA oligonucleotides (Figure 33). In addition, the gel-shifted bands were absent when the recombinant proteins were thermally denatured. The results showed that the recombinant HisMbTBP isoforms were properly folded, thus allowing them to bind DNA.

A smear was observed when the recombinant proteins were incubated with a non-specific competitor oligonucleotide (Figure 33), indicative of non-specific and weak interactions between the proteins and DNA. In the presence of a consensus TATA sequence however, the binding between the protein isoforms and DNA was highly specific resulting in a single and distinct protein-DNA band. Therefore, the binding of the recombinant HisMbTBP isoforms to DNA oligonucleotides containing the TATA sequence was specific.

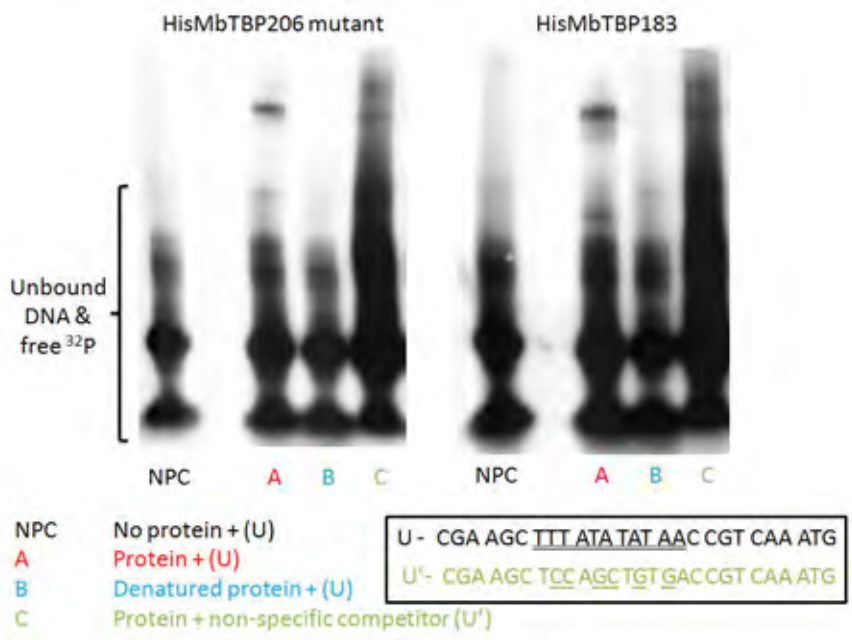


Figure 33: Recombinant HisMbTBPs bind specifically to TATA-containing DNA oligonucleotides.

Autoradiographs showing the gel-shift results of the recombinant HisMbTBP isoforms, with either DNA oligonucleotide (U) or (U'). The double underlined nucleotide sequences in (U) are consensus TATA-box sequences. The underlined nucleotide sequences in the non-specific competitor (U') are sequence variants of (U), that do not contain a consensus TATA-box sequence.

7.2.2 Recombinant HisMbTBPs bind putative promoters of *M. burtonii*

In order to determine if the recombinant HisMbTBP isoforms bind to natural promoters other than *in silico* and consensus sequence-derived putative promoters of *M. burtonii*, promoter-containing DNA sequences from *M. burtonii* were used in the EMSA analysis. After visualisation of the gel shifted bands with autoradiography, densitometry estimations of protein-bound and total DNA was performed using ImageJ software (NIH). The ratio of TBP-bound DNA was subsequently estimated semi-quantitatively (Figure 34), to indicate the DNA-binding efficiency of the protein.

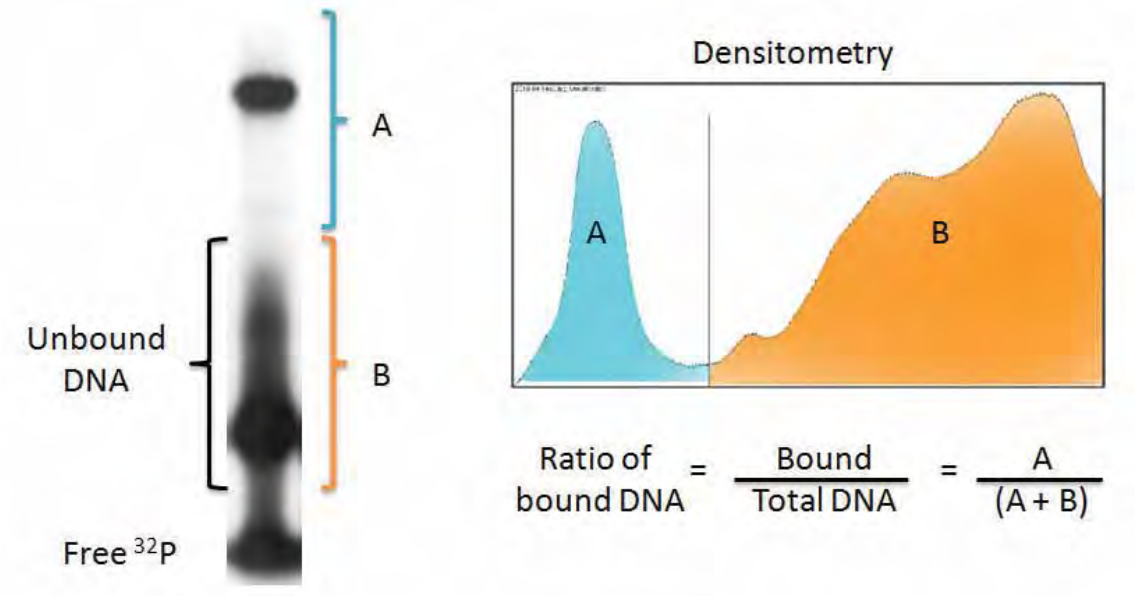


Figure 34: Semi-quantitative estimation of the ratio of bound DNA.

Densities of TBP-bound (A) and free (B) DNA were estimated from the autoradiograph. The ratio of TBP-bound DNA was calculated by dividing TBP-bound DNA by total DNA.

Putative promoters of three gene loci, namely Mbur_1494, Mbur_1950, and Mbur_2064, were selected for binding efficiency studies with the recombinant HisMbTBP isoforms. These gene loci were selected as they encode for genes that are over-expressed at either low or high temperatures in the growth range of *M. burtonii*. Additionally, putative promoters of the *tbp* gene of *M. burtonii* were also included to elucidate possible TBP-binding sites.

Semi-quantitative estimation of the TBP-bound DNA ratio (Figure 34) was used to compare the binding efficiencies of the two HisMbTBP isoforms with various putative promoter-containing DNA oligonucleotides. Control experiments yielded an average DNA-bound ratio ranging from 0.019 to 0.049, depending on the type of controls and HisMbTBP isoforms used (Table 7). This range of values was used as a baseline, in which a DNA-bound ratio of a higher value would indicate protein-DNA interaction. For example, the DNA-bound ratios of HisMbTBP206 mutant and HisMbTBP183 with DNA oligonucleotide 'U', that contains the consensus TATA-box sequence, were higher than the baseline and were thus indicative of protein-DNA interactions (Table 7).

Probe ID (5' position)	Sequences (5'–> 3')	Target gene information			N per protein	Mean (bound ratio) ± SEM (bound ratio)	
		Gene locus	4°C/23°C			HisMbTBP206 mutant	HisMbTBP183
			mRNA	Protein			
NPC (N.A)	Probe U, 1496p4	N.A			7	0.049 ± 0.010	0.030 ± 0.006
U (N.A)	CGAAGCTTTATATATAACCGTCAAATG	N.A			6	0.186± 0.019	0.191 ± 0.020
• Denatured protein	Same as probe U	N.A			2	0.047± 0.016	0.019 ± 0.006
1494p1 (-175nt)	CTGTAATGGCCTATAATTTTTTTATTC	Mbur_1494	0.24	0.49	2	0.147± 0.035	0.051± 0.018
1494p2 (-234nt)	GCCAATTTTATAGATATATGTTTATCT	Mbur_1494	0.24	0.49	4	0.174 ± 0.019	0.136 ± 0.018
1496p1 (-32nt)	TACCTTTCCTTATAATACTGAAGGTCA	Mbur_1496	1.39	2.0	2	0.266 ± 0.013	0.230± 0.006
1496p2 (-47nt)	CTGCGCTTGCGAATATACCTTTCCTTA	Mbur_1496	1.39	2.0	2	0.209 ± 0.011	0.141 ± 0.001
1496p3 (-60nt)	ATAAGTATATATACTGCGCTTGCGAAT	Mbur_1496	1.39	2.0	2	0.267 ± 0.051	0.219 ± 0.013
1496p4 (-90nt)	GTTTTTAAGTAGATATATAAAGGAAAG	Mbur_1496	1.39	2.0	4	0.301 ± 0.026	0.306 ± 0.031
1496p5 (-105nt)	AATTTCATTGATATAGTTTTTAAGTAG	Mbur_1496	1.39	2.0	2	0.175 ± 0.016	0.135 ± 0.018
1950p1 (-168nt)	TAGGACAACATTTATAAGTAGAAAGAT	Mbur_1950	1.82	2.2	2	0.226 ± 0.015	0.152 ± 0.018
2064p1 (-65nt)	AATTTGTCATTTATATTACATAAACAT	Mbur_2064	11.00	6.1	2	0.175 ± 0.025	0.107 ± 0.029
2064p2 (-142nt)	GTGTAACGTAATATATAATGTGGAGCT	Mbur_2064	11.00	6.1	2	0.277 ± 0.021	0.220 ± 0.033

Table 7: Binding efficiencies of recombinant HisMbTBP isoforms with respective putative promoter-containing DNA oligonucleotides.

The putative promoter-containing DNA oligonucleotide sequences and the gene information it represents are listed in the above table. The differential gene expression and protein abundance was extracted from Campanaro *et al.* (2010) [145] and Williams *et al.* (2010) [82], respectively. The term 'N per protein' refers to the number of technical replicates used to generate the mean bound ratio. Technical replicates were derived from two biological replicates. The term 'SEM' refers to the standard error of mean.

The recombinant HisMbTBP isoforms were able to bind to DNA oligonucleotides containing putative promoter sequences that occur naturally, other than *in silico*-derived TATA sequences, but with different efficiencies (Table 7). The binding efficiency of the HisMbTBP isoforms with DNA oligonucleotide 1496p4 was twice as great as that of 1494p2 although both DNA oligonucleotides contained the same putative TATA-box sequence 'AGATATAT' (Table 7 and 8). Since 1494p2 and 1496p4 differ in their nucleotide sequence flanking the putative TATA-box sequence 'AGATATAT', variants were created to assess the impact of the TATA flanking nucleotide sequences on the DNA binding efficiency of the TBP. In particular, a variant (1494p2c3) that was created by changing three nucleotides adjacent to the TATA-box of 1494p2 to mirror that of 1496p4 (GTT → AAA), increased the binding efficiency of both HisMbTBP isoforms by at least twice as that of 1494p2 and exceeded that of 1496p4 (Table 8). Conversely, other 1494p2 variants (1494p2c1, 1494p2c2, and 1494p2c4) that contained fewer nucleotide substitutions in the TATA flanking sequences showed only slight increases in their binding efficiencies to both HisMbTBP isoforms compared to 1494p2 (Table 8).

Probe ID • Variant ID	Sequences (5' –> 3')	N per protein	Mean (bound ratio) ± SEM (bound ratio)	
			HisMbTBP206 mutant	HisMbTBP183
1496p4	GTTTTTAAGTAGATATATAAAGGAAAG	4	0.301 ± 0.026	0.306 ± 0.031
1494p2	GCCAATTTTATAGATATATGTTTATCT	4	0.174 ± 0.019	0.136 ± 0.018
• 1494p2c1	GCCAATTTTATAGATATATATTTATCT	2	0.216 ± 0.039	0.188 ± 0.025
• 1494p2c2	GCCAATTTTATAGATATATAATTATCT	2	0.207 ± 0.049	0.196 ± 0.027
• 1494p2c3	GCCAATTTTATAGATATATAAATATCT	2	0.458 ± 0.008	0.449 ± 0.006
• 1494p2c4	GCCAATTTTGTAGATATATGTTTATCT	2	0.241 ± 0.054	0.216 ± 0.018

Table 8: Binding efficiencies of HisMbTBP isoforms with putative promoter-containing oligonucleotides 1496p4, 1494p2, and variants of 1494p2.

Sequences highlighted in red shows the region which is identical between 1496p4 and 1494p2. Nucleotide(s) changes to 1494p2 to mirror 1496p4 were underlined in the variants 1494p2c1 to 1494p2c4. The term 'N per protein' refers to the number of technical replicates used to generate the mean bound ratio. Technical replicates were derived from two biological replicates. The term 'SEM' refers to the standard error of mean.

Multiple putative promoters were found in a single 250-nucleotide region upstream of a gene (Figure 35) in *M. burtonii*. EMSA analysis presented in this section confirmed that most DNA oligonucleotides containing these putative promoters were capable of interacting with both HisMbTBP isoforms. For example, five separate DNA oligonucleotides containing putative promoters upstream of the *tbp* gene were shown to interact with both HisMbTBP isoforms (Table 7), but with different efficiencies. Additionally, certain DNA oligonucleotides such as 1494p2, 1496p4 and 2064p2 bind more strongly to HisMbTBP isoforms than others. This observation suggests that these oligonucleotides are likely to contain the main promoter for Mbur_1494, Mbur_1496, and Mbur_2064, respectively (Table 7), and that promoters upstream of a gene generally do not bind to TBPs with equal efficiency.

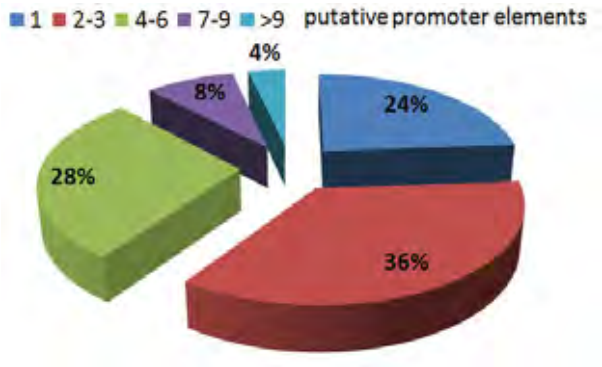


Figure 35: Percentage of putative promoter elements that are located upstream of genes found in the genome of *M. burtonii*.

The number of putative promoter elements in *M. burtonii*, which lies within 250 nucleotides upstream of each gene, was kindly generated by Campanaro S. using an algorithm to search for a consensus TATAnwTA sequence.

Finally, both HisMbTBP isoforms interacted differentially with the putative promoter-containing DNA oligonucleotides (Table 7). For example, interactions of most DNA oligonucleotides with HisMbTBP183, but not 1496p4 resulted in a lower DNA-bound ratio as compared to the HisMbTBP206 mutant. It suggests that HisMbTBP183 might potentially exhibit a higher specificity for promoter sequences.

7.2.3 Consensus TATA-box elements were identified from *M. burtonii*

The recombinant HisMbTBP isoforms were used to pull down DNA fragments from *M. burtonii* to which they were bound (see Materials and Methods chapter and Figure 36).

Genomic DNA from *M. burtonii* was first sonicated into 100 – 400bp long fragments (Figure 37) and incubated with either the HisMbTBP206 mutant or the HisMbTBP183. The protein-DNA complex was subsequently captured by cobalt-based Talon resin, while free DNA fragments were washed away. The protein-DNA complex was then eluted from the resin. DNA fragments pulled down from two technical replicates were purified and quantified. They were then pooled before being sent for high throughput Illumina-based sequencing (BGI, Hong Kong).

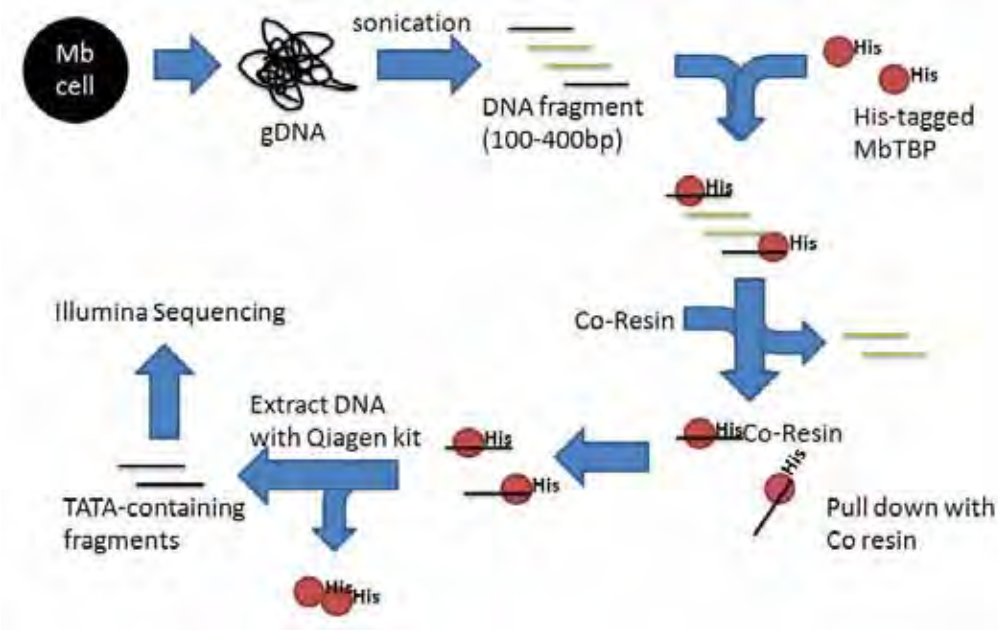


Figure 36: Schematics detailing the procedure for the genome-wide pull-down assay for TBP-binding sites using recombinant HisMbTBP isoforms.

'His' and 'Co-Resin' refers to the hexa-histidine fusion tag and cobalt-based resin, respectively.

To assess the non-specific recovery of DNA fragments, a 'no protein control' (NPC) was included as a negative control in the pull-down assay. However DNA fragments recovered from the NPC were 6 to 10 fold lower compared to both protein isoforms,

which was insufficient for downstream sequencing reactions. Therefore input DNA fragments were used as an alternative to pull-down DNA from NPC for the mock control, which was then sent for sequencing and subsequent bioinformatics analysis.

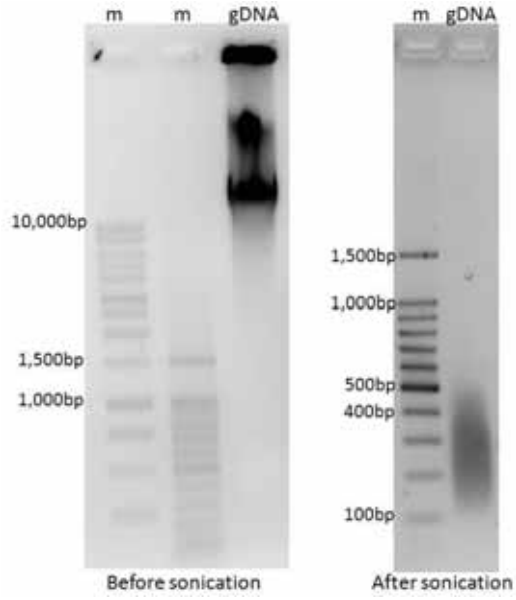


Figure 37: DNA gel electrophoresis results of intact and sonicated genomic DNA from *M. burtonii*.
Lanes with 'm' and 'gDNA' refers to DNA molecular weight ladder and genomic DNA respectively.

Fifteen million raw sequencing reads were obtained from pull-down DNA of mock control and both HisMbTBP isoforms, with each read measuring 49bp in length. Approximately 92% of the sequencing reads were deemed clean, after removal of adaptor sequences and low quality reads. Subsequently the sequencing reads were mapped onto the genome of *M. burtonii*, in sectors of 50bp in length by BGI (Hong Kong).

Sectors with less than 50 mapped reads were disregarded in subsequent analysis to minimise false discoveries of binding site-containing sectors. In order to identify genome sectors enriched with HisMbTBP isoform-binding motifs, a ratiometric analysis [144] was performed in which sequencing reads from the isoforms were divided by those of the mock control for the same sector. A significant (p-value < 0.001) read ratio would indicate that a sector contains binding sites for HisMbTBP isoforms.

Significant genome sectors were categorised into sectors specific to either HisMbTBP183 and/or HisMbTBP206 mutant. Thereafter, Gibbs motif site sampler [237] was used to identify the binding motifs shared by sectors under the same category. The consensus sequences of the motifs were displayed as WebLogo [238] sequence logos.

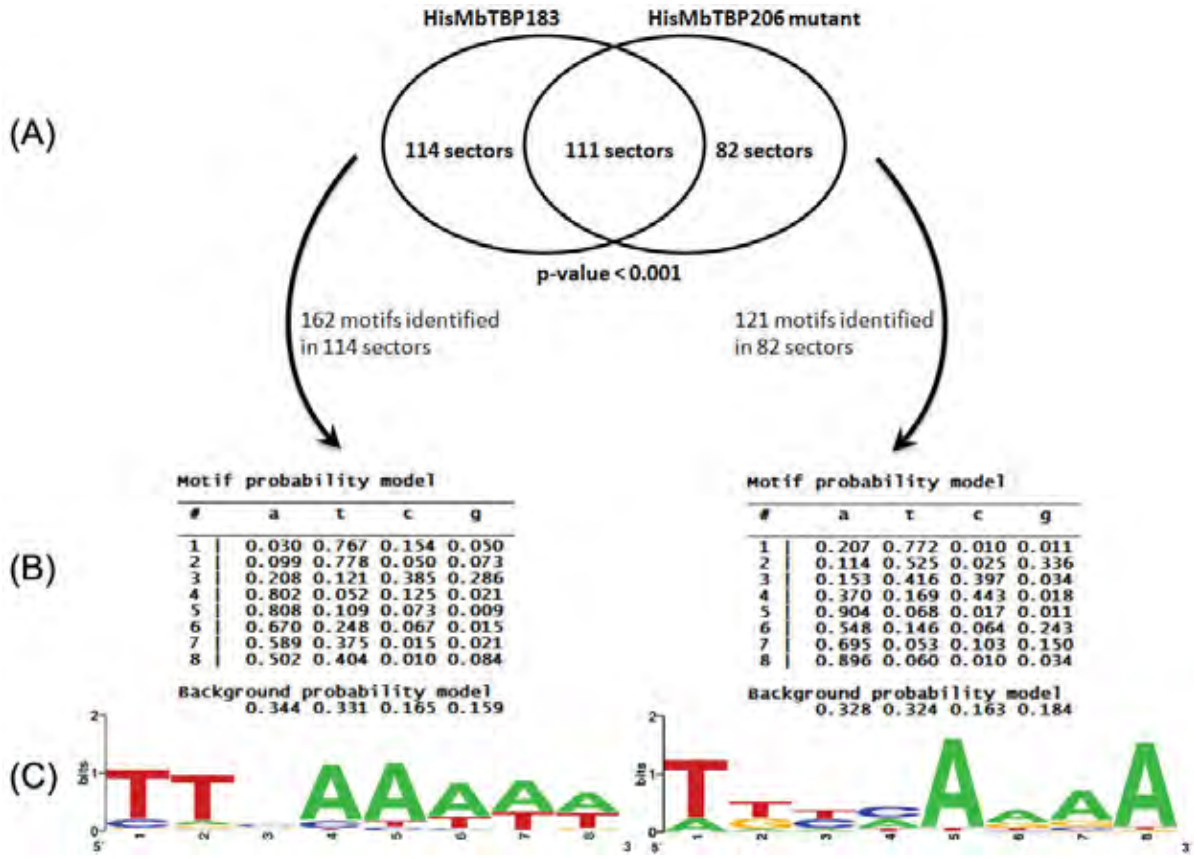


Figure 38: Identification of TBP-binding DNA motifs for HisMbTBP183 and HisMbTBP206 mutant.

(A) Genomic sectors that were significantly (p-value < 0.001) enriched with TBP-binding DNA motifs for HisMbTBP183, HisMbTBP206 mutant, or both. Each sector is 50bp in length. (B) The probability of nucleotides in each position of the TBP-binding DNA motif, and background were generated with the Gibbs motif site sampler [237]. (C) The consensus DNA sequences for HisMbTBP183- and HisMbTBP206 mutant-binding motifs were displayed, as sequence logos.

A total of 1314 sectors (p-value < 0.005) were identified as containing motifs that bind to HisMbTBP isoforms. When the selection criterion was made more stringent by decreasing the p-value, 307 sectors (p-value < 0.001) with HisMbTBP isoforms-binding

motifs were identified (Figure 38). In order to identify DNA signatures that bind strongly to MbTBP isoforms, only genome sectors with a p-value of less than 0.001 were analysed for MbTBP isoform-binding motifs.

Of the 307 genome sectors (p-value < 0.001), 114 sectors were found to be unique to the HisMbTBP183, which resulted in 162 HisMbTBP183-binding DNA motifs identified by Gibbs motif site sampler [237] (Figure 38). On the other hand, 82 sectors were found to be unique to HisMbTBP206 mutant, with 121 HisMbTBP206 mutant-binding DNA motifs identified. The DNA consensus sequence for HisMbTBP183-, and HisMbTBP206 mutant-binding DNA motifs was determined to be 'TTnAAAww' and 'TkymArAA' respectively (Figure 38). Nucleotides 'w', 'k', 'y', 'm', and 'r' refer to nucleotides 'A or T', 'G or T', 'C or T', 'A or C', and 'A or G' respectively. In general, the consensus MbTBP183-binding DNA motif was more conserved than that of MbTBP206 mutant, as determined from the height of the WebLogo [238] sequence logos.

In order to identify genes that are putatively regulated by either of the two HisMbTBP isoforms, sectors were assigned to their nearest downstream gene. The distance between a sector and its adjacent gene had to be less than 1kb for a HisMbTBP-binding DNA motif within the sector to be considered a putative promoter for its adjacent gene, since the 5' UTR of a gene can be as long as approximately 1kb in length [239-240].

A total of 95 genes were found to be potentially regulated by putative promoters of either or both HisMbTBP isoforms (Table 9 – 11). In many gene loci such as Mbur_0167, Mbur_1923 and Mbur_2027, multiple sectors containing putative promoters were found upstream of a single gene located at these loci. Of the 95 genes, 17 genes and 30 genes were potentially regulated by HisMbTBP206 mutant and HisMbTBP183 respectively, while the rest were regulated by both isoforms. Other than the genes that were annotated to have hypothetical functions, the remaining identified genes were responsible for a diverse range of functions such as transcription regulation and

the CRISPR (Clustered regularly interspaced short palindromic repeats) defense mechanism.

Gene locus tag	Protein ID (Evidence rating)	Distance to nearest putative promoter element (bp)
Mbur_0162	Transposase (ER4)	117
Mbur_0369	Oxidoreductase with FAD-binding domain (ER4)	398
Mbur_0389	ABC transporter, ATPase subunit (ER2)	492
Mbur_0390	Hypothetical protein (ER5)	21
Mbur_0501	Glycine betaine ABC transporter ATP-binding protein opuAA (EC 3.6.3.32) (ER2)	427
Mbur_0630	Carboxymuconolactone decarboxylase-domain protein (ER4)	164
Mbur_0638	Hypothetical protein (ER5)	58
Mbur_0639	PleC (non-motile and phage-resistance protein) -like multisensor signal transduction histidine kinase (ER4)	95
Mbur_0668	Hypothetical protein (ER5)	27
Mbur_0755	GPR1/FUN34/yaaH family protein (ER4)	188
Mbur_1160	Adenylosuccinate lyase (purB) (EC 4.3.2.2) (ER2)	787
Mbur_1163	Hypothetical protein (ER5)	779
Mbur_1212	DnaJ-like molecular chaperone (ER3)	184
Mbur_1458	Sulfopyruvate decarboxylase subunit alpha (comD) (EC 4.1.1.79) (ER2)	198
Mbur_1639	Transposase (ER3)	17
Mbur_1778	TatD-related deoxyribonuclease (ER2)	202
Mbur_2308	Trimethylamine methyltransferase (mttB) (EC 2.1.1.-) (ER2)	354

Table 9: Genes that are potentially regulated by HisMbTBP206 mutant.

Gene locus tag	Protein ID (Evidence rating)	Distance to nearest putative promoter element (bp)
Mbur_0075	Major facilitator superfamily transporter, large fragment (ER4)	307
Mbur_0227	Thioredoxin family protein (ER3)	146
Mbur_0275	Hypothetical protein (ER5)	359
Mbur_0341	Hypothetical protein (ER5)	240
Mbur_0342	Protein with peptidase M1 domain (ER4)	158
Mbur_0344	Hypothetical protein (ER5)	429
Mbur_0399	CheR/B methyltransferase-like protein (ER4)	202
Mbur_0401	Transposase (ER4)	444
Mbur_0444	Hypothetical protein (ER5)	349
Mbur_0445	Hypothetical protein (ER5)	102
Mbur_0459	Transposase DDE domain protein (ER4)	108
Mbur_0460	Hypothetical protein (ER5)	248
Mbur_0519	Hypothetical protein (ER5)	646
Mbur_0543	Protein of unknown function DUF255 (ER4)	729
Mbur_0583	Hypothetical protein (ER5)	104
Mbur_0637	Transposase, mutator type (ER3)	594
Mbur_0725	Glycosyl transferase, group 1 (ER3)	359
Mbur_1018	Transposase (ER4)	156
Mbur_1085	Hypothetical protein (ER5)	203
Mbur_1126	CRISPR-associated protein Cas4 containing DUF83 domain (ER4)	75
Mbur_1127	CRISPR-associated protein Cas1 containing DUF48 domain (ER4)	517
Mbur_1252	Hypothetical protein (ER5)	211
Mbur_1637	Hypothetical protein (ER5)	601
Mbur_1695	Sodium/hydrogen antiporter (ER3)	407
Mbur_1714	Hypothetical protein (ER5)	424
Mbur_1868	Deoxyhypusine synthase (ER3)	125
Mbur_1919	Hypothetical protein (ER5)	323
Mbur_2160	Transposase (ER3)	94
Mbur_2216	Protein with DUF1328 domain (ER4)	245
Mbur_2251	Transposase (ER4)	64

Table 10: Genes that are potentially regulated by HisMbTBP183.

Gene locus tag	Protein ID (Evidence rating)	Distance to nearest putative promoter element (bp)
Mbur_0070	Amino acid transporter domain protein (ER4)	299
Mbur_0071	Hypothetical protein (ER5)	174
Mbur_0167	Hypothetical protein (ER5)	157
Mbur_0169	Hypothetical protein (ER5)	180
Mbur_0296	Hypothetical protein (ER5)	205
Mbur_0298	Protein of unknown function DUF11 (ER4)	740
Mbur_0314	Cadherin domain protein (ER4)	122
Mbur_0316	Aconitate hydratase 2 (EC 4.2.1.3) (acnB) (ER2)	394
Mbur_0318	PRC-barrel domain protein (ER4)	848
Mbur_0321	Nicotinate phosphoribosyltransferase (pncB) (EC 2.4.2.11) (ER2)	704
Mbur_0402	Hypothetical protein (ER5)	215
Mbur_0403	DDE superfamily, IS4 transposase (ER4)	192
Mbur_0446	DDE superfamily, IS4 transposase (ER4)	17
Mbur_0457	Threonine synthase (EC 4.2.3.1) (ER2)	279
Mbur_0458	Hypothetical protein (ER5)	20
Mbur_0491	Atypical ABC-ATPase SufC (ER2)	189
Mbur_0493	Protein of unknown function DUF1622 (ER4)	27
Mbur_0500	Iron complex/Vitamin B12 ABC transporter substrate binding protein (ER3)	478
Mbur_0504	Metallophosphoesterase domain protein (ER4)	167
Mbur_0505	Hypothetical protein (ER5)	174
Mbur_0575	Hypothetical protein (ER5)	435
Mbur_0711	(R)-citramalate synthase (cimA) (EC 2.3.3.-) (ER2)	248
Mbur_0762	Protein containing N-terminal Pyruvate flavodoxin/ferredoxin oxidoreductase region, C-terminal TPP-binding region and ferredoxin domain (ER4)	891
Mbur_0794	Hypothetical protein (ER5)	48
Mbur_0865	Magnesium/cobalt chelatase-domain containing protein (ER4)	579
Mbur_0900	Hydantoinase/oxoprolinase (ER3)	109
Mbur_1123	CRISPR-associated protein (CT1132 family) with DUF694 domain (ER4)	182
Mbur_1139	Hypothetical protein (ER5)	839
Mbur_1140	Transcriptional regulator (ER4)	140
Mbur_1557	Response regulator receiver (ER4)	252
Mbur_1643	Cell surface glycoprotein (S-layer protein) with copper-binding domain (ER4)	544
Mbur_1666	Transcriptional regulator (ER3)	88
Mbur_1732	L-tyrosine decarboxylase (EC 4.1.1.25) (ER2)	180
Mbur_1733	ORC complex protein Cdc6/Orc1 (ER2)	418
Mbur_1767	FAD-dependent pyridine nucleotide-disulphide oxidoreductase-like protein (ER4)	298
Mbur_1768	Hypothetical protein (ER5)	95
Mbur_1798	Hypothetical protein (ER5)	473
Mbur_1801	Imidazole glycerol phosphate synthase subunit hisH (EC 2.4.2.-) (ER2)	146
Mbur_1923	Transposase (ER3)	42
Mbur_2027	Glycosyl transferase, group I (ER3)	121
Mbur_2058	ABC transporter, ATPase subunit (ER3)	25
Mbur_2170	Transposase (ER3)	727
Mbur_2173	Transposase (ER3)	81
Mbur_2175	PBS lyase HEAT-like repeat protein (ER4)	501
Mbur_2199	Chaperone ClpB/Hsp104 (ER2)	244
Mbur_2290	Transposase (ER4)	25
Mbur_2292	Amino acid/polyamine transporter I (ER4)	958

Table 11: Genes that are potentially regulated by both HisMbTBP183 and HisMbTBP206 mutant.

7.3 Discussion

Both HisMbTBP isoforms were found to exist in a folded conformation, which allowed them to be functional and to interact specifically with TATA-boxes. Subsequently, a genome-wide method was developed to identify the TBP-binding DNA for each HisMbTBP isoform. Consensus TBP-binding DNA motifs were then elucidated from the

TBP-binding DNA, in order to provide a better understanding of the transcription initiation process of *M. burtonii*.

The genome-wide method was developed based on the premise that a promoter-containing DNA, when bound by a HisMbTBP, could be pulled down by immobilised metal-affinity chromatography (IMAC) [174-175]. This method would allow for promoter-containing DNA to be pulled down from the genome of *M. burtonii* rapidly, and easily. This method of obtaining HisMbTBP-binding DNA is also preferred as there would be no interference to the HisMbTBP-DNA interaction from other molecular components such as transcriptional repressors, since only HisMbTBPs and genomic DNA would interact. Subsequent high throughput sequencing (Illumina) [231-232,241] of the pulled down DNA would also allow for large amounts of pulled down DNA to be sequenced without the need for tedious large scale cloning.

7.3.1 Consensus MbTBP-binding DNA motifs

The consensus sequence of the HisMbTBP183-binding DNA motif is unlike that of a typical TATA-box from methanogenic archaea. Specifically, the consensus sequence of the former contains thymine- and adenine-rich clusters, while the latter generally contains thymine and adenine in alternation [99,242]. In addition, the consensus sequence of the HisMbTBP206 mutant-binding motif was not as conserved as that of its shorter counterpart, except at nucleotide positions 1, 5, and 8 (Figure 38). Therefore since HisMbTBP183 recognises a more conserved DNA motif, it suggests an increased specificity in promoter sequence recognition by HisMbTBP183 than its longer isoform.

The substitution of a trinucleotide 'GTT' with an adenine-rich cluster in oligonucleotide 1494p2 created a TATA-box element with the sequence 'TATATAAA'. This variant, together with oligonucleotide 1496p4 that also contains the 'TATATAAA' sequence, was found to exhibit a high binding efficiency with MbTBP isoforms. On the other hand, substitution of a single or double nucleotide bases with adenine in 1494p2 generated the TATA-box elements 'TATATATT' and 'TATATAAT' respectively, that possessed half

the binding efficiencies compared to that of the 'TATATAAA' sequence to TBP (Table 8). These nucleotide substitution studies potentially support the notion that adenine-rich clusters in the 3' half of a TATA-box element improves the promoter-TBP binding efficiency.

7.3.2 Genes putatively regulated by MbTBP183 and/or MbTBP206

The distance between a HisMbTBP-binding DNA motif and its nearest downstream gene can be used to determine if a HisMbTBP-binding DNA motif is a putative promoter. This region would also include the 5' UTR of the genes. The 5' UTRs of eucaryal, archaeal and bacterial genes were previously determined to be approximately as long as 1kb [239], 0.5kb [243], and 1kb [240,244] in length, respectively. Therefore, it appears that the upper limit in the length of a 5' UTR in organisms is approximately 1kb. Additionally, long 5' untranslated regions (5' UTR) in gene transcripts have been associated with transcriptional and translational regulation [245-251]. Taking into account the potential importance of long 5' UTR in gene regulation, only HisMbTBP-binding DNA motifs that were approximately 1kb upstream of their nearest gene were considered as putative promoters. As such, a gene located 1kb downstream of a putative promoter would be potentially regulated by the respective HisMbTBP isoforms that interacted with the gene promoter upstream of it.

Both isoforms of HisMbTBP potentially regulate distinct as well as similar genes in *M. burtonii*. However there was no apparent common association between the genes within either group, other than the fact that they were potentially regulated by the same HisMbTBP isoform. Moreover since both HisMbTBP isoforms differ in their N-terminal, it is apparent that differences in their regulatory function might be attributed by the N-terminal regions of the proteins. The extra RepA segment (Figure 5) found in the N-terminal of HisMbTBP206 mutant could possibly affect the consensus promoter sequence recognised by the longer isoform, and thus decrease its specificity in promoter recognition and binding.

Approximately 4% of the total protein-coding genes in *M. burtonii* were found to be potentially regulated by the two HisMbTBP isoforms. This could be a consequence of the stringent criterion (p-value < 0.001) used in the identification of genomic sectors that contain HisMbTBP-binding motifs. In addition, about half of the identified HisMbTBP-binding DNA motifs were found to be located more than 1kb away from any gene. Therefore not all HisMbTBP-binding DNA motifs were putative promoters, which could potentially regulate downstream genes. Coupled with the fact that multiple putative promoters were found upstream of the same gene in *M. burtonii* (Chapter 7.2.3 and Table 7), the number of protein-coding genes that could be identified as putatively regulated by MbTBP isoforms will be lower, as compared to that of having a single putative promoter regulating a gene.

It is unclear why some HisMbTBP-binding DNA motifs are situated at a significant distance away from their downstream gene. A possible explanation might be that these putative promoters potentially transcribe nearby non-coding RNAs (ncRNAs) rather than distant downstream protein-coding genes in *M. burtonii*. Non-coding RNAs are present in all the three domains of life [242-243,252-254], and have been implicated in transcriptional and translational regulations of many genes [254-256], as well as chromatin remodeling for transcriptional activation [257]. Therefore, apart from gene regulation through direct interaction with a promoter sequence, TBPs from *M. burtonii* can also potentially increase its regulatory capacity indirectly through the ncRNAs.

In conclusion, this chapter describes the first investigation of the functional characteristics of a psychrophilic TBP that was found to exist in two isoforms in *M. burtonii*. Additionally, results obtained from the genome-wide identification of TBP-binding DNA motifs also provided insight into the differences between the two isoforms of MbTBP, in terms of the DNA sequences and genes that they bind to and regulate, respectively.

CHAPTER VIII

Concluding remarks

8.1 N-terminal sequence repeats may be unique to psychrophilic methanogens of Ace Lake

The work described in this thesis represents the first known attempt to describe the TBPs of a psychrophilic archaeon. Of particular relevance is that, an N-terminal sequence repeat (NTSR) was identified within MbTBP206, one of the two TBPs found in the psychrophilic *M. burtonii*. MbTBP206 contains a pair of sequence repeats (RepA and RepB) in its N-terminal region. With the exception of MbTBP, no other NTSR-containing TBPs were found within the 84 archaeal TBPs currently in the Integrated Microbial Genome (IMG) [149] database.

However, preliminary data from the recently completed analysis of the draft genome of another psychrophilic archaeon, *Methanogenium fridigum* (Webster, J., 2010, personal communication) revealed the existence of two *tbp* open reading frames (ORF). Translation from the two ORFs would produce two TBPs of 222 (MfTBP222) and 185 (MfTBP185) amino acids in length, respectively (Figure 39A).

Figure 39: N-terminal regions and sequence repeats of TBPs from *M. burtonii* and *M. frigidum*.

(A) Alignment of the N-terminal region of TBPs from *M. burtonii* (MbTBP206 and MbTBP183), and *M. fridigum* (MfTBP222 and MfTBP185). The sequence repeats are highlighted in blue and red boxes for *M. burtonii* and *M. fridigum*, respectively. (*) denotes identical alignment of residue; (:) denotes strong conservation of physico-chemical properties of mismatch residues; (.) denotes weak conservation of physico-chemical properties of mismatch residues. (B) Weblogo sequence display for the NTSRs of MbTBP206 and MfTBP222. The alignment result for both NTSRs is indicated in the boxed insert while the consensus sequence found in all the NTSRs is highlighted in the dotted box.

Similar to MbTBP206, MfTBP222 also contains an NTSR. The NTSRs of both MbTBP206 and MfTBP222 share a consensus amino acid sequence of 'KIE(E/N)' (Figure 39B). Previously, these residues were found to bind DNA (Figure 8) in the S1 β-strand structure of TBP cores [112-114]. Thus it is possible that in psychrophilic methanogens, such as *M. burtonii* and *M. frigidum*, these residues in the NTSRs play a role in transcription initiation by TBPs through interacting with DNA.

The presence of NTSRs in TBPs from both *M. burtonii* and *M. fridigum* implies a putative association between NTSRs with psychrophilic methanogens of Ace Lake in Antarctica, since both archaea were isolated from this source. It would be valuable to search for NTSRs in TBPs from the metagenome and metaproteome of Ace Lake microbial communities to verify such a possible association. Screening psychrophilic archaea from locations other than Antarctica would also address the issue of whether NTSR-containing TBPs are a general feature of psychrophilic archaea. If that is the case, NTSRs could well represent a cold-adaption strategy employed by archaeal TBPs to ensure sufficient transcriptional capability in cold environments.

8.2 Sequence repeats in NTSR may alter the binding specificity of TBPs to TATA-box elements

Since the MbTBP183 protein makes up the core of MbTBP206, it was anticipated that the consensus TATA-box element for MbTBP183-binding would be preserved in MbTBP206 as well. However, this was not the case. MbTBP206 appears to recognise a different consensus TATA-box element, which implied that the extra RepA in the NTSR of MbTBP206 altered the specificity of its MbTBP183 core (Figure 40). Further, RepA contains residues identical to previously established DNA-interacting residues found in the S1 β-strand structure of TBP cores [112-114]. Therefore, additional studies of these putative DNA-interacting residues on RepA would be needed to determine whether they influence the specificity of TBP-TATA-box recognition and DNA binding.

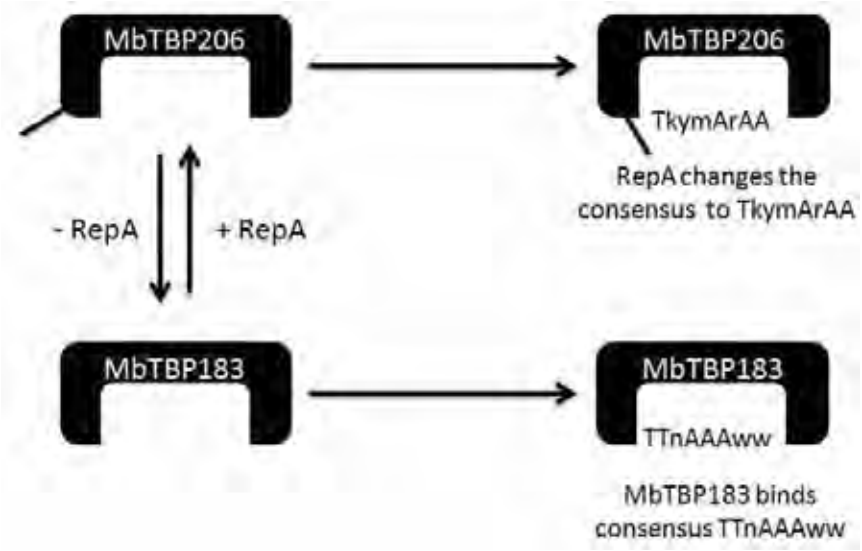


Figure 40: RepA relaxes the specificity of MbTBP206 for TATA-box element.

The TATA-box elements 'TTnAAAww' and 'TkymArAA' are the natural consensus TATA-box elements specific for MbTBP183 and MbTBP206, respectively.

The addition of the RepA segment to the N-terminus of MbTBP183 forms MbTBP206. This pairing is analogous to that of a TBP-associated factor (TAF)-TBP interaction that is specific to eucaryotes. Eucaryotes produce TAFs, which interact with TBP to regulate promoter selectivity and transcription initiation of genes [258-260]. Similarly, RepA potentially interacts with the TBP core to regulate promoter specificity and gene transcription. Therefore the RepA segment probably serves as an archaeal TAF, although homologues of eucaryotic TAFs have not been found in archaea [109,261]. However, unlike the non-covalent TAF-TBP interactions in eucaryotes, the RepA segment is covalently bound to the MbTBP183 core of MbTBP206. As such, MbTBP206 resembles an MbTBP183 that is permanently regulated by a TAF.

The RepA repeat forms part of the NTSR of MbTBP206, and appeared to alter the binding specificity of the protein to TATA-box elements. Since MfTBP222 of *M. frigidum* also contains individual repeats in its NTSR, it would be interesting to compare it with a NTSR-lacking MfTBP222 mutant to determine if its sequence repeats also influenced the binding specificity of MfTBP (*M. frigidum* TBP) to TATA-box elements. However, the fastidious nature of *M. frigidum* makes it a difficult archaea to

culture [75]. Thus elucidating the binding specificity of MfTBPs to genomic promoters through either genome-wide promoter pull-down or chromatin IP would be challenging. As such, any MfTBP-promoter specificity studies at this stage would be limited to SELEX experiments, by generating DNA sequences that bind optimally to MfTBPs instead.

8.3 Future work and conclusion

This study also established the foundation for the future development of a psychrophilic cell-free transcription system for archaea, which would be useful in studying the functions of psychrophilic transcriptional regulators. Many putative transcriptional regulators have been identified to be cold-adapted and/or growth substrate-specific in the psychrophilic *M. burtonii* [67,82], with most of their mechanisms of action unknown. Therefore, a psychrophilic cell-free archaeal transcription assay would allow an in-depth examination of how such putative psychrophilic transcriptional regulators modulate the process of gene transcription in *M. burtonii*. In particular, the recombinant MbTBPs produced, using the methods described in Chapter 5, could be used to develop such an assay.

In conclusion, this thesis has documented unprecedented evidence that implicated two TBP isoforms in the transcription initiation of genes in *M. burtonii*. Both isoforms display differential recognition and binding specificities for TATA-box sequences, highlighting the unique role that each TBP isoform has in gene regulation in *M. burtonii*.

CHAPTER IX

References

1 Sapp, J. The prokaryote-eukaryote dichotomy: meanings and mythology. *Microbiol Mol Biol Rev* **69**, 292-305 (2005).

2 Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *J Mol Biol* **13**, 373-398 (1965).

3 Fox, G. E., Pechman, K. R. & Woese, C. R. Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *Int J Syst Bacteriol* **27**, 44-57 (1977).

4 Sogin, S. J., Sogin, M. L. & Woese, C. R. Phylogenetic measurement in procaryotes by primary structural characterization. *J Mol Evol* **1**, 173-184 (1971).

5 Woese, C. R. *et al.* Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**, 83-86 (1975).

6 Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S. & Woese, C. R. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci U S A* **74**, 4537-4541 (1977).

7 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).

8 Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**, 4576-4579 (1990).

9 Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63-67 (2002).

10 Burggraf, S., Heyder, P. & Eis, N. A pivotal Archaea group. *Nature* **385**, 780 (1997).

11 Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci U S A* **93**, 9188-9193 (1996).

12 Brochier-Armanet, C., Boussau, B., Gribaldo, S. & Forterre, P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* **6**, 245-252 (2008).

13 Cavicchioli, R. Archaea - timeline of the third domain. *Nat Rev Microbiol* (2010).

14 Koga, Y., Nishihara, M., Morii, H. & Akagawa-Matsushita, M. Ether polar lipids of methanogenic bacteria: structures, comparative aspects, and biosyntheses. *Microbiol Rev* **57**, 164-182 (1993).

15 van de Vossenberg, J. L., Driessen, A. J. & Konings, W. N. The essence of being extremophilic: the role of the unique archaeal membrane lipids. *Extremophiles* **2**, 163-170 (1998).

16 Kandler, O. & Konig, H. Cell wall polymers in Archaea (Archaebacteria). *Cell Mol Life Sci* **54**, 305-308 (1998).

17 Claus, H. *et al.* Molecular organization of selected prokaryotic S-layer proteins. *Can J Microbiol* **51**, 731-743 (2005).

18 Sara, M. & Sleytr, U. B. S-Layer proteins. *J Bacteriol* **182**, 859-868 (2000).

19 Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B. E. & Mrazek, J. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. *Theor Popul Biol* **61**, 367-390 (2002).

20 Cavicchioli, R., Curmi, P. M., Saunders, N. & Thomas, T. Pathogenic archaea: do they exist? *Bioessays* **25**, 1119-1128 (2003).

21 Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**, 456-502 (1997).

22 Charlebois, R. L., Schalkwyk, L. C., Hofman, J. D. & Doolittle, W. F. Detailed physical map and set of overlapping clones covering the genome of the archaebacterium Haloferax volcanii DS2. *J Mol Biol* **222**, 509-524 (1991).

23 Ng, W. V. *et al.* Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res* **8**, 1131-1141 (1998).

24 Keeling, P. J., Charlebois, R. L. & Doolittle, W. F. Archaebacterial genomes: eubacterial form and eukaryotic content. *Curr Opin Genet Dev* **4**, 816-822 (1994).

25 Zillig, W., Stetter, K. O. & Janekovic, D. DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius. *Eur J Biochem* **96**, 597-604 (1979).

26 Lupas, A. *et al.* Eubacterial proteasomes. *Mol Biol Rep* **24**, 125-131 (1997).

27 Bell, S. D. & Jackson, S. P. Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol* **6**, 222-228 (1998).

28 Hausner, W., Frey, G. & Thomm, M. Control regions of an archaeal gene. A TATA box and an initiator element promote cell-free transcription of the tRNA(Val) gene of Methanococcus vannielii. *J Mol Biol* **222**, 495-508 (1991).

29 Reiter, W. D., Hudepohl, U. & Zillig, W. Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. *Proc Natl Acad Sci U S A* **87**, 9509-9513 (1990).

30 Ouzounis, C. & Sander, C. TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes. *Cell* **71**, 189-190 (1992).

31 Hanzelka, B. L., Darcy, T. J. & Reeve, J. N. TFE, an archaeal transcription factor in Methanobacterium thermoautotrophicum related to eucaryal transcription factor TFIIEalpha. *J Bacteriol* **183**, 1813-1818 (2001).

32 Langer, D., Hain, J., Thuriaux, P. & Zillig, W. Transcription in archaea: similarity to that in eucarya. *Proc Natl Acad Sci U S A* **92**, 5768-5772 (1995).

33 Dahlke, I. & Thomm, M. A Pyrococcus homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment. *Nucleic Acids Res* **30**, 701-710 (2002).

34 Bell, S. D., Cairns, S. S., Robson, R. L. & Jackson, S. P. Transcriptional regulation of an archaeal operon in vivo and in vitro. *Mol Cell* **4**, 971-982 (1999).

35 Dennis, P. P. Ancient ciphers: translation in Archaea. *Cell* **89**, 1007-1010 (1997).

36 Zepp Falz, K. *et al.* Vertical distribution of methanogens in the anoxic sediment of Rotsee (Switzerland). *Appl Environ Microbiol* **65**, 2402-2408 (1999).

37 Simankova, M. V. *et al.* Methanosarcina lacustris sp. nov., a new psychrotolerant methanogenic archaeon from anoxic lake sediments. *Syst Appl Microbiol* **24**, 362-367 (2001).

38 Kublanov, I. V., Bidjieva, S., Mardanov, A. V. & Bonch-Osmolovskaya, E. A. Desulfurococcus kamchatkensis sp. nov., a novel hyperthermophilic protein-

degrading archaeon isolated from a Kamchatka hot spring. *Int J Syst Evol Microbiol* **59**, 1743-1747 (2009).

39 Reigstad, L. J., Jorgensen, S. L. & Schleper, C. Diversity and abundance of Korarchaeota in terrestrial hot springs of Iceland and Kamchatka. *ISME J* **4**, 346-356 (2010).

40 Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci U S A* **91**, 1609-1613 (1994).

41 Gonzalez, J. M. *et al.* Pyrococcus horikoshii sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. *Extremophiles* **2**, 123-130 (1998).

42 Jeanthon, C. *et al.* Methanococcus infernus sp. nov., a novel hyperthermophilic lithotrophic methanogen isolated from a deep-sea hydrothermal vent. *Int J Syst Bacteriol* **48 Pt 3**, 913-919 (1998).

43 Walsby, A. E. A square bacterium. *Nature* **283**, 69-71 (1980).

44 Bolhuis, H., Poele, E. M. & Rodriguez-Valera, F. Isolation and cultivation of Walsby's square archaeon. *Environ Microbiol* **6**, 1287-1291 (2004).

45 Baliga, N. S. *et al.* Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. *Genome Res* **14**, 2221-2234 (2004).

46 Bruneel, O. *et al.* Archaeal diversity in a Fe-As rich acid mine drainage at Carnoules (France). *Extremophiles* **12**, 563-571 (2008).

47 Edwards, K. J., Bond, P. L., Gihring, T. M. & Banfield, J. F. An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. *Science* **287**, 1796-1799 (2000).

48 Karner, M. B., DeLong, E. F. & Karl, D. M. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**, 507-510 (2001).

49 Franzmann, P. D., Springer, N., Ludwig, W., Conway De Macario, E. & Rohde, M. A methanogenic archaeon from Ace Lake, Antarctica: Methanococcoides burtonii sp. nov. *System Appl Microbiol* **15**, 573-581 (1992).

50 Franzmann, P. D. *et al.* Methanogenium frigidum sp. nov., a psychrophilic, H2-using methanogen from Ace Lake, Antarctica. *Int J Syst Bacteriol* **47**, 1068-1072 (1997).

51 Zhang, L. M., Wang, M., Prosser, J. I., Zheng, Y. M. & He, J. Z. Altitude ammonia-oxidizing bacteria and archaea in soils of Mount Everest. *FEMS Microbiol Ecol* **70**, 52-61 (2009).

52 DeLong, E. F. Everything in moderation: archaea as 'non-extremophiles'. *Curr Opin Genet Dev* **8**, 649-654 (1998).

53 Hongoh, Y. Diversity and genomes of uncultured microbial symbionts in the termite gut. *Biosci Biotechnol Biochem* **74**, 1145-1151 (2010).

54 Skillman, L. C., Evans, P. N., Strompl, C. & Joblin, K. N. 16S rDNA directed PCR primers and detection of methanogens in the bovine rumen. *Lett Appl Microbiol* **42**, 222-228 (2006).

55 Dridi, B., Henry, M., El Khechine, A., Raoult, D. & Drancourt, M. High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. *PLoS One* **4**, e7063 (2009).

56 DeLong, E. F. & Pace, N. R. Environmental diversity of bacteria and archaea. *Syst Biol* **50**, 470-478 (2001).

57 Francis, C. A., Beman, J. M. & Kuypers, M. M. New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J* **1**, 19-27 (2007).

58 Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543-546 (2005).

59 Galperin, M. Y. Using archaeal genomics to fight global warming and clostridia to fight cancer. *Environmental Microbiology* **9**, 279-286 (2007).

60 Lu, Y. & Conrad, R. In situ stable isotope probing of methanogenic archaea in the rice rhizosphere. *Science* **309**, 1088-1090 (2005).

61 Erkel, C. *et al.* Retrieval of first genome data for rice cluster I methanogens by a combination of cultivation and molecular techniques. *FEMS Microbiology Ecology* **53**, 187-204 (2005).

62 Sakai, S. *et al.* Methanocella paludicola gen. nov., sp. nov., a methane-producing archaeon, the first isolate of the lineage 'Rice Cluster I', and proposal of the new archaeal order Methanocellales ord. nov. *Int J Syst Evol Microbiol* **58**, 929-936 (2008).

63 Bapteste, E., Brochier, C. & Boucher, Y. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* **1**, 353-363 (2005).

64 Luo, H. *et al.* Gene order phylogeny and the evolution of methanogens. *PLoS One* **4**, e6069 (2009).

65 Ferry, J. G. How to make a living by exhaling methane. *Annu Rev Microbiol* **64**, 453-473 (2010).

66 Jetten, M. S. M., Stams, A. J. M. & Zehnder, A. J. B. Methanogenesis from acetate: a comparison of the acetate metabolism in Methanothrix soehngenii and Methanosarcina spp. *FEMS Microbiology Letters* **88**, 181-197 (1992).

67 Williams, T. J. *et al.* Global proteomic analysis of the insoluble, soluble, and supernatant fractions of the psychrophilic archaeon Methanococcoides burtonii. Part II: the effect of different methylated growth substrates. *J Proteome Res* **9**, 653-663 (2010).

68 Cavicchioli, R. Cold-adapted archaea. *Nat Rev Microbiol* **4**, 331-343 (2006).

69 Feller, G. & Gerday, C. Psychrophilic enzymes: hot topics in cold adaptation. *Nat Rev Microbiol* **1**, 200-208 (2003).

70 Li, L., Kato, C. & Horikoshi, K. Microbial Diversity in Sediments Collected from the Deepest Cold-Seep Area, the Japan Trench. *Mar Biotechnol (NY)* **1**, 391-400 (1999).

71 Simon, C., Wiezer, A., Strittmatter, A. W. & Daniel, R. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl Environ Microbiol* **75**, 7519-7526 (2009).

72 Gibson, J. A. *et al.* Unsaturated diether lipids in the psychrotrophic archaeon Halorubrum lacusprofundi. *Syst Appl Microbiol* **28**, 19-26 (2005).

73 Nichols, D. S. *et al.* Cold adaptation in the Antarctic Archaeon Methanococcoides burtonii involves membrane lipid unsaturation. *J Bacteriol* **186**, 8508-8515 (2004).

74 Lim, J., Thomas, T. & Cavicchioli, R. Low temperature regulated DEAD-box RNA helicase from the Antarctic archaeon, Methanococcoides burtonii. *J Mol Biol* **297**, 553-567 (2000).

75 Giaquinto, L. *et al.* Structure and function of cold shock proteins in archaea. *J Bacteriol* **189**, 5738-5748 (2007).

76 Fields, P. A. & Somero, G. N. Hot spots in cold adaptation: localized increases in conformational flexibility in lactate dehydrogenase A4 orthologs of Antarctic notothenioid fishes. *Proc Natl Acad Sci U S A* **95**, 11476-11481 (1998).

77 Thomas, T. & Cavicchioli, R. Effect of temperature on stability and activity of elongation factor 2 proteins from Antarctic and thermophilic methanogens. *J Bacteriol* **182**, 1328-1332 (2000).

78 Georlette, D. *et al.* Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic, and thermophilic DNA ligases. *J Biol Chem* **278**, 37015-37023 (2003).

79 Siddiqui, K. S. & Cavicchioli, R. Cold-adapted enzymes. *Annu Rev Biochem* **75**, 403-433 (2006).

80 Saunders, N. F. *et al.* Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. *Genome Res* **13**, 1580-1588 (2003).

81 Koike, H. *et al.* Origins of protein stability revealed by comparing crystal structures of TATA binding proteins. *Structure* **12**, 157-168 (2004).

82 Williams, T. J. *et al.* Global proteomic analysis of the insoluble, soluble, and supernatant fractions of the psychrophilic archaeon Methanococcoides burtonii. Part I: the effect of growth temperature. *J Proteome Res* **9**, 640-652 (2010).

83 Goodchild, A., Raftery, M., Saunders, N. F., Guilhaus, M. & Cavicchioli, R. Cold adaptation of the Antarctic archaeon, Methanococcoides burtonii assessed by proteomics using ICAT. *J Proteome Res* **4**, 473-480 (2005).

84 Goodchild, A. *et al.* A proteomic determination of cold adaptation in the Antarctic archaeon, Methanococcoides burtonii. *Mol Microbiol* **53**, 309-321 (2004).

85 Lyimo, T. J., Pol, A., Jetten, M. S. & den Camp, H. J. Diversity of methanogenic archaea in a mangrove sediment and isolation of a new Methanococcoides strain. *FEMS Microbiol Lett* **291**, 247-253 (2009).

86 Sowers, K. R. & Ferry, J. G. Isolation and Characterization of a Methylotrophic Marine Methanogen, Methanococcoides methylutens gen. nov., sp. nov. *Appl Environ Microbiol* **45**, 684-690 (1983).

87 Singh, N., Kendall, M. M., Liu, Y. & Boone, D. R. Isolation and characterization of methylotrophic methanogens from anoxic marine sediments in Skan Bay, Alaska: description of Methanococcoides alaskense sp. nov., and emended description of Methanosarcina baltica. *Int J Syst Evol Microbiol* **55**, 2531-2538 (2005).

88 Allen, M. A. *et al.* The genome sequence of the psychrophilic archaeon, Methanococcoides burtonii: the role of genome evolution in cold adaptation. *ISME J* **3**, 1012-1035 (2009).

89 Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **5**, e1000605 (2009).

90 Saunders, N. F. *et al.* Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon Methanococcoides burtonii. *J Proteome Res* **4**, 464-472 (2005).

91 Macfarlane, W. M. Demystified.... Transcription. *Mol Pathol* **53**, 1-7 (2000).

92 Bell, S. D. Archaeal transcriptional regulation--variation on a bacterial theme? *Trends Microbiol* **13**, 262-265 (2005).

93 Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**, 327-335 (1996).

94 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).

95 Barolo, S. & Posakony, J. W. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev* **16**, 1167-1181 (2002).

96 Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186-192 (2009).

97 Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205 (2009).

98 Browning, D. F. & Busby, S. J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57-65 (2004).

99 Soppa, J. Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. *Mol Microbiol* **31**, 1589-1592 (1999).

100 Qureshi, S. A., Khoo, B., Baumann, P. & Jackson, S. P. Molecular cloning of the transcription factor TFIIB homolog from Sulfolobus shibatae. *Proc Natl Acad Sci U S A* **92**, 6077-6081 (1995).

101 Rowlands, T., Baumann, P. & Jackson, S. P. The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. *Science* **264**, 1326-1329 (1994).

102 Leffers, H., Gropp, F., Lottspeich, F., Zillig, W. & Garrett, R. A. Sequence, organization, transcription and evolution of RNA polymerase subunit genes from the archaebacterial extreme halophiles Halobacterium halobium and Halococcus morrhuae. *J Mol Biol* **206**, 1-17 (1989).

103 Berghofer, B. *et al.* Relatedness of archaebacterial RNA polymerase core subunits to their eubacterial and eukaryotic equivalents. *Nucleic Acids Res* **16**, 8113-8128 (1988).

104 Puhler, G., Lottspeich, F. & Zillig, W. Organization and nucleotide sequence of the genes encoding the large subunits A, B and C of the DNA-dependent RNA polymerase of the archaebacterium Sulfolobus acidocaldarius. *Nucleic Acids Res* **17**, 4517-4534 (1989).

105 Soppa, J. Transcription initiation in Archaea: facts, factors and future aspects. *Mol Microbiol* **31**, 1295-1305 (1999).

106 Bell, S. D., Kosa, P. L., Sigler, P. B. & Jackson, S. P. Orientation of the transcription preinitiation complex in archaea. *Proc Natl Acad Sci U S A* **96**, 13662-13667 (1999).

107 Wolner, B. S. & Gralla, J. D. TATA-flanking sequences influence the rate and stability of TATA-binding protein and TFIIB binding. *J Biol Chem* **276**, 6260-6266 (2001).

108 Bell, S. D., Brinkman, A. B., van der Oost, J. & Jackson, S. P. The archaeal TFIIEalpha homologue facilitates transcription initiation by enhancing TATA-box recognition. *EMBO Rep* **2**, 133-138 (2001).

109 Kyrpides, N. C. & Ouzounis, C. A. Transcription in archaea. *Proc Natl Acad Sci U S A* **96**, 8545-8550 (1999).

110 Bell, S. D. & Jackson, S. P. Mechanism of autoregulation by an archaeal transcriptional repressor. *J Biol Chem* **275**, 31624-31629 (2000).

111 Starr, D. B. & Hawley, D. K. TFIID binds in the minor groove of the TATA box. *Cell* **67**, 1231-1240 (1991).

112 Kosa, P. F., Ghosh, G., DeDecker, B. S. & Sigler, P. B. The 2.1-A crystal structure of an archaeal preinitiation complex: TATA-box-binding protein/transcription factor (II)B core/TATA-box. *Proc Natl Acad Sci U S A* **94**, 6042-6047 (1997).

113 Kim, J. L., Nikolov, D. B. & Burley, S. K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520-527 (1993).

114 Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512-520 (1993).

115 Adachi, N., Senda, M., Natsume, R., Senda, T. & Horikoshi, M. Crystal structure of Methanococcus jannaschii TATA box-binding protein. *Genes Cells* **13**, 1127-1140 (2008).

116 Rashid, N., Morikawa, M. & Imanaka, T. An abnormally acidic TATA-binding protein from a hyperthermophilic archaeon. *Gene* **166**, 139-143 (1995).

117 Adachi, N. *et al.* Purification, crystallization and preliminary X-ray analysis of Methanococcus jannaschii TATA box-binding protein (TBP). *Acta Crystallogr D Biol Crystallogr* **60**, 2328-2331 (2004).

118 Wu, Y. R. *et al.* Analysis of polyglutamine-coding repeats in the TATA-binding protein in different neurodegenerative diseases. *J Neural Transm* **112**, 539-546 (2005).

119 Reid, S. J., Whittaker, D. J., Greenwood, D. & Snell, R. G. A splice variant of the TATA-box binding protein encoding the polyglutamine-containing N-terminal domain that accumulates in Alzheimer's disease. *Brain Res* **1268**, 190-199 (2009).

120 Hobbs, N. K., Bondareva, A. A., Barnett, S., Capecchi, M. R. & Schmidt, E. E. Removing the vertebrate-specific TBP N terminus disrupts placental beta2m-dependent interactions with the maternal immune system. *Cell* **110**, 43-54 (2002).

121 Khrapunov, S. & Brenowitz, M. Influence of the N-terminal domain and divalent cations on self-association and DNA binding by the Saccharomyces cerevisiae TATA binding protein. *Biochemistry* **46**, 4876-4887 (2007).

122 Gupta, S. *et al.* DNA and protein footprinting analysis of the modulation of DNA binding by the N-terminal domain of the Saccharomyces cerevisiae TATA binding protein. *Biochemistry* **46**, 9886-9898 (2007).

123 Delgadillo, R. F., Whittington, J. E., Parkhurst, L. K. & Parkhurst, L. J. The TATA-Binding Protein Core Domain in Solution Variably Bends TATA Sequences via a Three-Step Binding Mechanism. *Biochemistry* (2009).

124 Nikolov, D. B. *et al.* Crystal structure of TFIID TATA-box binding protein. *Nature* **360**, 40-46 (1992).

125 Chasman, D. I., Flaherty, K. M., Sharp, P. A. & Kornberg, R. D. Crystal structure of yeast TATA-binding protein and model for interaction with DNA. *Proc Natl Acad Sci U S A* **90**, 8174-8178 (1993).

126 DeDecker, B. S. *et al.* The crystal structure of a hyperthermophilic archaeal TATA-box binding protein. *J Mol Biol* **264**, 1072-1084 (1996).

127 Wu, J., Parkhurst, K. M., Powell, R. M., Brenowitz, M. & Parkhurst, L. J. DNA bends in TATA-binding protein-TATA complexes in solution are DNA sequence-dependent. *J Biol Chem* **276**, 14614-14622 (2001).

128 Qian, X., Strahs, D. & Schlick, T. Dynamic simulations of 13 TATA variants refine kinetic hypotheses of sequence/activity relationships. *J Mol Biol* **308**, 681-703 (2001).

129 Whittington, J. E. *et al.* TATA-binding protein recognition and bending of a consensus promoter are protein species dependent. *Biochemistry* **47**, 7264-7273 (2008).

130 Powell, R. M., Parkhurst, K. M., Brenowitz, M. & Parkhurst, L. J. Marked stepwise differences within a common kinetic mechanism characterize TATA-binding protein interactions with two consensus promoters. *J Biol Chem* **276**, 29782-29791 (2001).

131 Parkhurst, K. M., Richards, R. M., Brenowitz, M. & Parkhurst, L. J. Intermediate species possessing bent DNA are present along the pathway to formation of a final TBP-TATA complex. *J Mol Biol* **289**, 1327-1341 (1999).

132 Wolner, B. S. & Gralla, J. D. Roles for non-TATA core promoter sequences in transcription and factor binding. *Mol Cell Biol* **20**, 3608-3615 (2000).

133 Faiger, H., Ivanchenko, M., Cohen, I. & Haran, T. E. TBP flanking sequences: asymmetry of binding, long-range effects and consensus sequences. *Nucleic Acids Res* **34**, 104-119 (2006).

134 De Francisci, D. *et al.* The RNA polymerase subunits E/F from the Antarctic archaeon Methanococcoides burtonii bind to specific species of mRNA. *Environ Microbiol* (2010).

135 Roovers, M., Hethke, C., Legrain, C., Thomm, M. & Glansdorff, N. Isolation of the gene encoding Pyrococcus furiosus ornithine carbamoyltransferase and study of its expression profile in vivo and in vitro. *Eur J Biochem* **247**, 1038-1045 (1997).

136 Gohl, H. P., Hausner, W. & Thomm, M. Cell-free transcription of the nifH1 gene of Methanococcus thermolithotrophicus indicates that promoters of archaeal nif genes share basic features with the methanogen consensus promoter. *Mol Gen Genet* **231**, 286-295 (1992).

137 Thomsen, J. *et al.* The basal transcription factors TBP and TFB from the mesophilic archaeon Methanosarcina mazeii: structure and conformational changes upon interaction with stress-gene promoters. *J Mol Biol* **309**, 589-603 (2001).

138 Hausner, W., Lange, U. & Musfeldt, M. Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J Biol Chem* **275**, 12393-12399 (2000).

139 Hausner, W. & Thomm, M. The translation product of the presumptive Thermococcus celer TATA-binding protein sequence is a transcription factor related in structure and function to Methanococcus transcription factor B. *J Biol Chem* **270**, 17649-17651 (1995).

140 Keese, A. M., Schut, G. J., Ouhammouch, M., Adams, M. W. & Thomm, M. Genome-wide identification of targets for the archaeal heat shock regulator phr by cell-free transcription of genomic DNA. *J Bacteriol* **192**, 1292-1298 (2010).

141 Gill, S. C. & von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**, 319-326 (1989).

142 Welch, M. *et al.* Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One* **4**, e7002 (2009).

143 Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**, 2876-2890 (2006).

144 Illumina. ChIP-Seq Data Analysis. *Technical Note: Illumina Sequencing* (2007).

145 Campanaro, S. *et al.* Temperature-dependent global gene expression in the Antarctic archaeon Methanococcoides burtonii. *Environ Microbiol* (2010).

146 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

147 Hickey, A. J., Conway de Macario, E. & Macario, A. J. Transcription in the archaea: basal factors, regulation, and stress-gene expression. *Crit Rev Biochem Mol Biol* **37**, 537-599 (2002).

148 Sartorius-Neef, S. & Pfeifer, F. In vivo studies on putative Shine-Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. *Mol Microbiol* **51**, 579-588 (2004).

149 Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res* **34**, D344-348 (2006).

150 Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201 (2006).

151 Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460 (2007).

152 Daugherty, M. A., Brenowitz, M. & Fried, M. G. Participation of the amino-terminal domain in the self-association of the full-length yeast TATA binding protein. *Biochemistry* **39**, 4869-4880 (2000).

153 Khrapunov, S., Pastor, N. & Brenowitz, M. Solution structural studies of the Saccharomyces cerevisiae TATA binding protein (TBP). *Biochemistry* **41**, 9559-9571 (2002).

154 Perez-Howard, G. M., Weil, P. A. & Beechem, J. M. Yeast TATA binding protein interaction with DNA: fluorescence determination of oligomeric state, equilibrium binding, on-rate, and dissociation kinetics. *Biochemistry* **34**, 8005-8017 (1995).

155 Reddy, P. & Hahn, S. Dominant negative mutations in yeast TFIID define a bipartite DNA-binding region. *Cell* **65**, 349-357 (1991).

156 Horikoshi, M., Yamamoto, T., Ohkuma, Y., Weil, P. A. & Roeder, R. G. Analysis of structure-function relationships of yeast TATA box binding factor TFIID. *Cell* **61**, 1171-1178 (1990).

157 Cormack, B. P., Strubin, M., Ponticelli, A. S. & Struhl, K. Functional differences between yeast and human TFIID are localized to the highly conserved region. *Cell* **65**, 341-348 (1991).

158 Gribaldo, S. & Brochier-Armanet, C. The origin and evolution of Archaea: a state of the art. *Philos Trans R Soc Lond B Biol Sci* **361**, 1007-1022 (2006).

159 Preston, C. M., Wu, K. Y., Molinski, T. F. & DeLong, E. F. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc Natl Acad Sci U S A* **93**, 6241-6246 (1996).

160 Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. & Fukatsu, T. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A* **99**, 14280-14285 (2002).

161 Nikoh, N. & Nakabachi, A. Aphids acquired symbiotic genes via lateral gene transfer. *BMC Biol* **7**, 12 (2009).

162 Podar, M. *et al.* A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biol* **9**, R158 (2008).

163 Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**, 605-618 (2008).

164 Burnette, W. N. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal Biochem* **112**, 195-203 (1981).

165 Alegria-Schaffer, A., Lodge, A. & Vattem, K. Performing and optimizing Western blots with an emphasis on chemiluminescent detection. *Methods Enzymol* **463**, 573-599 (2009).

166 Goodchild, A., Raftery, M., Saunders, N. F., Guilhaus, M. & Cavicchioli, R. Biology of the cold adapted archaeon, Methanococcoides burtonii determined

by proteomics using liquid chromatography-tandem mass spectrometry. *J Proteome Res* **3**, 1164-1176 (2004).

167 Campanaro, S. *et al.* Thermal regulation of gene and operon expression in the Antarctic archaeon Methanococcoides burtonii. *Manuscript submitted* (2010).

168 Reichlen, M. J., Murakami, K. S. & Ferry, J. G. Functional analysis of the three TATA binding protein homologs in Methanosarcina acetivorans. *J Bacteriol* **192**, 1511-1517 (2010).

169 Coker, J. A. & DasSarma, S. Genetic and transcriptomic analysis of transcription factor genes in the model halophilic Archaeon: coordinate action of TbpD and TfbA. *BMC Genet* **8**, 61 (2007).

170 Jana, S. & Deb, J. K. Strategies for efficient production of heterologous proteins in Escherichia coli. *Appl Microbiol Biotechnol* **67**, 289-298 (2005).

171 Weickert, M. J., Doherty, D. H., Best, E. A. & Olins, P. O. Optimization of heterologous protein production in Escherichia coli. *Curr Opin Biotechnol* **7**, 494-499 (1996).

172 Francis, D. M. & Page, R. Strategies to optimize protein expression in E. coli. *Curr Protoc Protein Sci* **Chapter 5**, Unit 5 24 21-29 (2010).

173 Samuelson, J. C. Recent developments in difficult protein expression: a guide to E. coli strains, promoters, and relevant host mutations. *Methods Mol Biol* **705**, 195-209 (2011).

174 Porath, J., Carlsson, J., Olsson, I. & Belfrage, G. Metal chelate affinity chromatography, a new approach to protein fractionation. *Nature* **258**, 598-599 (1975).

175 Hochuli, E., Dobeli, H. & Schacher, A. New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *J Chromatogr* **411**, 177-184 (1987).

176 Hagel, L. Gel-filtration chromatography. *Curr Protoc Protein Sci* **Chapter 8**, Unit8.3 (2001).

177 Lee, M. S., Kim, G. A., Seo, M. S., Lee, J. H. & Kwon, S. T. Characterization of heat-labile uracil-DNA glycosylase from Psychrobacter sp. HJ147 and its

application to the polymerase chain reaction. *Biotechnol Appl Biochem* **52**, 167-175 (2009).

178 Jeon, J. H., Kim, J. T., Kang, S. G., Lee, J. H. & Kim, S. J. Characterization and its potential application of two esterases derived from the arctic sediment metagenome. *Mar Biotechnol (NY)* **11**, 307-316 (2009).

179 Di Rocco, G. *et al.* Cloning, expression and physicochemical characterization of a di-heme cytochrome c (4) from the psychrophilic bacterium Pseudoalteromonas haloplanktis TAC 125. *J Biol Inorg Chem* **13**, 789-799 (2008).

180 Gentz, R. & Bujard, H. Promoters recognized by Escherichia coli RNA polymerase selected by function: highly efficient promoters from bacteriophage T5. *J Bacteriol* **164**, 70-77 (1985).

181 Mukherjee, S., Shukla, A. & Guptasarma, P. Single-step purification of a protein-folding catalyst, the SlyD peptidyl prolyl isomerase (PPI), from cytoplasmic extracts of Escherichia coli. *Biotechnol Appl Biochem* **37**, 183-186, doi:10.1042/ (2003).

182 Parsy, C. B., Chapman, C. J., Barnes, A. C., Robertson, J. F. & Murray, A. Two-step method to isolate target recombinant protein from co-purified bacterial contaminant SlyD after immobilised metal affinity chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* **853**, 314-319 (2007).

183 Bolanos-Garcia, V. M. & Davies, O. R. Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochim Biophys Acta* **1760**, 1304-1313 (2006).

184 David, G., Blondeau, K., Schiltz, M., Penel, S. & Lewit-Bentley, A. YodA from Escherichia coli is a metal-binding, lipocalin-like protein. *J Biol Chem* **278**, 43728-43735 (2003).

185 Kopitz, A., Soppa, J., Krejtschi, C. & Hauser, K. Differential stability of TATA box binding proteins from archaea with different optimal growth temperatures. *Spectrochim Acta A Mol Biomol Spectrosc* **73**, 799-804 (2009).

186 Soppa, J. & Link, T. A. The TATA-box-binding protein (TBP) of Halobacterium salinarum. Cloning of the tbp gene, heterologous production of TBP and folding of TBP into a native conformation. *Eur J Biochem* **249**, 318-324 (1997).

187 Lu, Z., Chen, W., Liu, R., Hu, X. & Ding, Y. A novel method for high-level production of psychrophilic TAB5 alkaline phosphatase. *Protein Expr Purif* **74**, 217-222 (2010).

188 Yang, X., Lin, X., Fan, T., Bian, J. & Huang, X. Cloning and expression of lipP, a gene encoding a cold-adapted lipase from Moritella sp.2-5-10-1. *Curr Microbiol* **56**, 194-198 (2008).

189 Watanabe, S., Yasutake, Y., Tanaka, I. & Takada, Y. Elucidation of stability determinants of cold-adapted monomeric isocitrate dehydrogenase from a psychrophilic bacterium, Colwellia maris, by construction of chimeric enzymes. *Microbiology* **151**, 1083-1094 (2005).

190 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).

191 de Smit, M. H. & van Duin, J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci U S A* **87**, 7668-7672 (1990).

192 de Smit, M. H. & van Duin, J. Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *J Mol Biol* **244**, 144-150 (1994).

193 Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255-258 (2009).

194 Kelly, S. M., Jess, T. J. & Price, N. C. How to study proteins by circular dichroism. *Biochim Biophys Acta* **1751**, 119-139 (2005).

195 Whitmore, L. & Wallace, B. A. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers* **89**, 392-400 (2008).

196 Lobley, A., Whitmore, L. & Wallace, B. A. DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics* **18**, 211-212 (2002).

197 Whitmore, L. & Wallace, B. A. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res* **32**, W668-673 (2004).

198 Provencher, S. W. & Glockner, J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry* **20**, 33-37 (1981).

199 van Stokkum, I. H., Spoelder, H. J., Bloemendal, M., van Grondelle, R. & Groen, F. C. Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal Biochem* **191**, 110-118 (1990).

200 Compton, L. A. & Johnson, W. C., Jr. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal Biochem* **155**, 155-167 (1986).

201 Hawe, A., Sutter, M. & Jiskoot, W. Extrinsic fluorescent dyes as tools for protein characterization. *Pharm Res* **25**, 1487-1499 (2008).

202 Stryer, L. The interaction of a naphthalene dye with apomyoglobin and apohemoglobin. A fluorescent probe of non-polar binding sites. *J Mol Biol* **13**, 482-495 (1965).

203 Russo, S. F. & Chen, W.-C. Fluorescent probe studies of haptoglobin type 2-1. *Physiol Chem Phys* **8**, 229-236 (1976).

204 Thomas, T., Kumar, N. & Cavicchioli, R. Effects of ribosomes and intracellular solutes on activities and stabilities of elongation factor 2 proteins from psychrotolerant and thermophilic methanogens. *J Bacteriol* **183**, 1974-1982 (2001).

205 Nikolov, D. B. & Burley, S. K. 2.1 A resolution refined structure of a TATA box-binding protein (TBP). *Nat Struct Biol* **1**, 621-637 (1994).

206 Coleman, R. A., Taggart, A. K., Benjamin, L. R. & Pugh, B. F. Dimerization of the TATA binding protein. *J Biol Chem* **270**, 13842-13849 (1995).

207 Jackson-Fisher, A. J., Chitikila, C., Mitra, M. & Pugh, B. F. A role for TBP dimerization in preventing unregulated gene expression. *Mol Cell* **3**, 717-727 (1999).

208 Jackson-Fisher, A. J. *et al.* Dimer dissociation and thermosensitivity kinetics of the Saccharomyces cerevisiae and human TATA binding proteins. *Biochemistry* **38**, 11340-11348 (1999).

209 Vanathi, P., Mishra, A. K. & Bhargava, P. Regulation of activity of the yeast TATA-binding protein through intra-molecular interactions. *J Biosci* **28**, 413-421 (2003).

210 Coleman, R. A., Taggart, A. K., Burma, S., Chicca, J. J., 2nd & Pugh, B. F. TFIIA regulates TBP and TFIID dimers. *Mol Cell* **4**, 451-457 (1999).

211 Rashidzadeh, H., Khrapunov, S., Chance, M. R. & Brenowitz, M. Solution structure and interdomain interactions of the Saccharomyces cerevisiae "TATA binding protein" (TBP) probed by radiolytic protein footprinting. *Biochemistry* **42**, 3655-3665 (2003).

212 Robertson, D. E. & Roberts, M. F. Organic osmolytes in methanogenic archaebacteria. *Biofactors* **3**, 1-9 (1991).

213 Siddiqui, K. S., Cavicchioli, R. & Thomas, T. Thermodynamic activation properties of elongation factor 2 (EF-2) proteins from psychrotolerant and thermophilic Archaea. *Extremophiles* **6**, 143-150 (2002).

214 Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2**, 1849-1861 (2007).

215 Fried, M. G. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis* **10**, 366-376 (1989).

216 Weidenbach, K., Ehlers, C., Kock, J. & Schmitz, R. A. NrpRII mediates contacts between NrpRI and general transcription factors in the archaeon Methanosarcina mazei Go1. *FEBS J* **277**, 4398-4411 (2010).

217 De Biase, A., Macario, A. J. & Conway de Macario, E. Effect of heat stress on promoter binding by transcription factors in the cytosol of the archaeon Methanosarcina mazeii. *Gene* **282**, 189-197 (2002).

218 Qureshi, S. A., Baumann, P., Rowlands, T., Khoo, B. & Jackson, S. P. Cloning and functional analysis of the TATA binding protein from Sulfolobus shibatae. *Nucleic Acids Res* **23**, 1775-1781 (1995).

219 Hausner, W., Wettach, J., Hethke, C. & Thomm, M. Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. *J Biol Chem* **271**, 30144-30148 (1996).

220 Hethke, C., Geerling, A. C., Hausner, W., de Vos, W. M. & Thomm, M. A cell-free transcription system for the hyperthermophilic archaeon Pyrococcus furiosus. *Nucleic Acids Res* **24**, 2369-2376 (1996).

221 Hethke, C., Bergerat, A., Hausner, W., Forterre, P. & Thomm, M. Cell-free transcription at 95 degrees: thermostability of transcriptional components and DNA topology requirements of Pyrococcus transcription. *Genetics* **152**, 1325-1333 (1999).

222 Frey, G., Thomm, M., Brudigam, B., Gohl, H. P. & Hausner, W. An archaebacterial cell-free transcription system. The expression of tRNA genes from Methanococcus vannielii is mediated by a transcription factor. *Nucleic Acids Res* **18**, 1361-1367 (1990).

223 Collas, P. The current state of chromatin immunoprecipitation. *Mol Biotechnol* **45**, 87-100 (2010).

224 He, A. & Pu, W. T. Genome-wide location analysis by pull down of in vivo biotinylated transcription factors. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21 20 (2010).

225 Sun, S. Technology evaluation: SELEX, Gilead Sciences Inc. *Curr Opin Mol Ther* **2**, 100-105 (2000).

226 Cribb, P., Esteban, L., Trochine, A., Girardini, J. & Serra, E. Trypanosoma cruzi TBP shows preference for C/G-rich DNA sequences in vitro. *Exp Parasitol* **124**, 346-349 (2010).

227 Perez, C. A., Ott, J., Mays, D. J. & Pietenpol, J. A. p63 consensus DNA-binding site: identification, analysis and application into a p63MH algorithm. *Oncogene* **26**, 7363-7370 (2007).

228 Sevilimedu, A., Shi, H. & Lis, J. T. TFIIB aptamers inhibit transcription by perturbing PIC formation at distinct stages. *Nucleic Acids Res* **36**, 3118-3127 (2008).

229 Gaal, T. *et al.* Promoter recognition and discrimination by EsigmaS RNA polymerase. *Mol Microbiol* **42**, 939-954 (2001).

230 Ouhammouch, M. & Geiduschek, E. P. A thermostable platform for transcriptional regulation: the DNA-binding properties of two Lrp homologs from the hyperthermophilic archaeon Methanococcus jannaschii. *EMBO J* **20**, 146-156 (2001).

231 Bennett, S. Solexa Ltd. *Pharmacogenomics* **5**, 433-438 (2004).

232 Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255-264 (2008).

233 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).

234 Kim, M. Y. *et al.* Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc Natl Acad Sci U S A* **107**, 22032-22037 (2010).

235 Liu, S. *et al.* Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC Genomics* **12**, 53 (2011).

236 Bibikova, M. & Fan, J. B. Genome-wide DNA methylation profiling. *Wiley Interdiscip Rev Syst Biol Med* **2**, 210-223 (2010).

237 Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-214 (1993).

238 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

239 Sehgal, A., Briggs, J., Rinehart-Kim, J., Basso, J. & Bos, T. J. The chicken c-Jun 5' untranslated region directs translation by internal initiation. *Oncogene* **19**, 2836-2845 (2000).

240 Ogura, M. & Tanaka, T. The Bacillus subtilis late competence operon comE is transcriptionally regulated by yutB and under post-transcription initiation control by comN (yrzD). *J Bacteriol* **191**, 949-958 (2009).

241 Bennett, S. T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373-382 (2005).

242 Li, E., Reich, C. I. & Olsen, G. J. A whole-genome approach to identifying protein binding sites: promoters in Methanocaldococcus (Methanococcus) jannaschii. *Nucleic Acids Res* **36**, 6948-6958 (2008).

243 Jager, D. *et al.* Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* **106**, 21878-21882 (2009).

244 Hahn, J., Inamine, G., Kozlov, Y. & Dubnau, D. Characterization of comE, a late competence operon of Bacillus subtilis required for the binding and uptake of transforming DNA. *Mol Microbiol* **10**, 99-111 (1993).

245 Ding, N., Zheng, Y., Wu, Q. & Mao, X. The 5' untranslated region of fruA mRNA is required for translational enhancement of FruA synthesis during Myxococcus xanthus development. *Arch Microbiol* **189**, 279-288 (2008).

246 Fang, L., Xia, B. & Inouye, M. Transcription of cspA, the gene for the major cold-shock protein of Escherichia coli, is negatively regulated at 37 degrees C by the 5'-untranslated region of its mRNA. *FEMS Microbiol Lett* **176**, 39-43 (1999).

247 Fang, L., Hou, Y. & Inouye, M. Role of the cold-box region in the 5' untranslated region of the cspA mRNA in its transient expression at low temperature in Escherichia coli. *J Bacteriol* **180**, 90-95 (1998).

248 Zhao, X. *et al.* The optional long 5'-untranslated region of human ACAT1 mRNAs impairs the production of ACAT1 protein by promoting its mRNA decay. *Acta Biochim Biophys Sin (Shanghai)* **41**, 30-41 (2009).

249 Jovcic, B., Bertani, I., Venturi, V., Topisirovic, L. & Kojic, M. 5' Untranslated region of the Pseudomonas putida WCS358 stationary phase sigma factor rpoS mRNA is involved in RpoS translational regulation. *J Microbiol* **46**, 56-61 (2008).

250 Wang, J. T. *et al.* Characterization of Epstein-Barr virus BGLF4 kinase expression control at the transcriptional and translational levels. *J Gen Virol* **91**, 2186-2196 (2010).

251 Myers, S. J., Huang, Y., Genetta, T. & Dingledine, R. Inhibition of glutamate receptor 2 translation by a polymorphic repeat sequence in the 5'-untranslated leaders. *J Neurosci* **24**, 3489-3499 (2004).

252 Kumar, R. *et al.* Identification of novel non-coding small RNAs from Streptococcus pneumoniae TIGR4 using high-resolution genome tiling arrays. *BMC Genomics* **11**, 350 (2010).

253 Jiang, R. P. *et al.* Identification of four novel small non-coding RNAs from Xanthomonas campestris pathovar campestris. *BMC Genomics* **11**, 316 (2010).

254 Tang, T. H. *et al.* Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol* **55**, 469-481 (2005).

255 Altuvia, S., Zhang, A., Argaman, L., Tiwari, A. & Storz, G. The Escherichia coli OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J* **17**, 6069-6075 (1998).

256 Espinoza, C. A., Allen, T. A., Hieb, A. R., Kugel, J. F. & Goodrich, J. A. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol* **11**, 822-829 (2004).

257 Hirota, K. *et al.* Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**, 130-134 (2008).

258 Davidson, I., Kobi, D., Fadloun, A. & Mengus, G. New insights into TAFs as regulators of cell cycle and signaling pathways. *Cell Cycle* **4**, 1486-1490 (2005).

259 Green, M. R. TBP-associated factors (TAFIIs): multiple, selective transcriptional mediators in common complexes. *Trends Biochem Sci* **25**, 59-63 (2000).

260 Pugh, B. F. Control of gene expression through regulation of the TATA-binding protein. *Gene* **255**, 1-14 (2000).

261 Qureshi, S. A., Bell, S. D. & Jackson, S. P. Factor requirements for transcription in the Archaeon Sulfolobus shibatae. *EMBO J* **16**, 2927-2936 (1997).

Appendix

Appendix A

List of TBP gene locus tags with their corresponding TBP length and archaeal genome source.

IMG Locus tag	TBP length	Genome
AF0373	183aa	Archaeoglobus fulgidus DSM 4304
APE1862	222aa	Aeropyrum pernix K1
CENSYa_0092	186aa	Cenarchaeum symbiosum A
CENSYa_1178	182aa	Cenarchaeum symbiosum A
CENSYa_1234	186aa	Cenarchaeum symbiosum A
Cmaq_1206	194aa	Caldivirga maquilingensis IC-167
FaciDRAFT_1431	184aa	Ferroplasma acidarmanus Fer1
Hbut_0934	202aa	Hyperthermus butylicus DSM 5456
HlacDRAFT_0660	185aa	Halorubrum lacusprofundi ATCC 49239
HlacDRAFT_1728	186aa	Halorubrum lacusprofundi ATCC 49239
HQ2527A	186aa	Haloquadratum walsbyi DSM 16790
HQ3410A	186aa	Haloquadratum walsbyi DSM 16790
Igni_0989	207aa	Ignicoccus hospitalis KIN4/I
Kcr_1466	211aa	Candidatus Korarchaeum cryptofilum OPF8
MA0179	185aa	Methanosarcina acetivorans C2A
MA0278	185aa	Methanosarcina acetivorans C2A
MA4331	183aa	Methanosarcina acetivorans C2A
Maeo_0186	181aa	Methanococcus aeolicus Nankai-3
Mbar_A0595	183aa	Methanosarcina barkeri fusaro
Mbar_A1062	184aa	Methanosarcina barkeri fusaro
Mboo_0323	184aa	Candidatus Methanoregula boonei 6A8
Mboo_0916	186aa	Candidatus Methanoregula boonei 6A8
Mboo_1680	184aa	Candidatus Methanoregula boonei 6A8
Mbur_1496	206aa	Methanococcoides burtonii DSM 6242
Memar_1252	193aa	Methanoculleus marisnigri JR1
Memar_1388	187aa	Methanoculleus marisnigri JR1
Mevan_1270	181aa	Methanococcus vannielii SB
Mhun_0568	186aa	Methanospirillum hungatei JF-1
Mhun_0593	192aa	Methanospirillum hungatei JF-1
MJ0507	183aa	Methanocaldococcus jannaschii DSM 2661
MK0839	185aa	Methanopyrus kandleri AV19
Mlab_0440	188aa	Methanocorpusculum labreanum Z
Mlab_0447	189aa	Methanocorpusculum labreanum Z
MM1027	183aa	Methanosarcina mazei Go1
MM1028	183aa	Methanosarcina mazei Go1
MM2184	185aa	Methanosarcina mazei Go2
MmarC5_0612	174aa	Methanococcus maripaludis C5
MmarC5_1416	181aa	Methanococcus maripaludis C5
MmarC6_0696	181aa	Methanococcus maripaludis C6
MmarC7_1260	181aa	Methanococcus maripaludis C7
MMP0257	189aa	Methanococcus maripaludis S2
Msed_1728	176aa	Metallosphaera sedula DSM 5348

Appendix A (Continued)

IMG Locus tag	TBP length	Genome
Msm_0720	181aa	Methanobrevibacter smithii ATCC 35061
Msp_1097	183aa	Methanosphaera stadtmanae DSM 3091
MTH1627	181aa	Methanothermobacter thermautotrophicus Delta H
Mthe_0152	183aa	Methanosaeta thermophila PT
MvolDRAFT_0951	181aa	Methanococcus voltae A3
NEQ039	217aa	Nanoarchaeum equitans Kin4-M
Nmar_0598	186aa	Nitrosopumilus maritimus SCM1
Nmar_1519	187aa	Nitrosopumilus maritimus SCM1
NP1064A	187aa	Natronomonas pharaonis DSM 2160
OE4146F	186aa	Halobacterium salinarum R1
OE7186F	181aa	Halobacterium salinarum R1
OE7214R	186aa	Halobacterium salinarum R1
OE8017F	181aa	Halobacterium salinarum R1
OE8046R	186aa	Halobacterium salinarum R1
PAB1726	191aa	Pyrococcus abyssi GE5
PAE2164	199aa	Pyrobaculum aerophilum IM2
Pars_0806	201aa	Pyrobaculum arsenaticum DSM 13514
Pcal_1078	199aa	Pyrobaculum calidifontis JCM 11548
PF1295	191aa	Pyrococcus furiosus DSM 3638
PH1009	191aa	Pyrococcus horikoshii OT3
Pisl_0807	199aa	Pyrobaculum islandicum DSM 4184
PTO0506	184aa	Picrophilus torridus DSM 9790
RCIX1238	188aa	Methanocella sp. RC-I
rrnAC0681	186aa	Haloarcula marismortui ATCC 43049
Saci_1336	197aa	Sulfolobus acidocaldarius DSM 639
Smar_1322	196aa	Staphylothermus marinus F1
SSO0951	181aa	Sulfolobus solfataricus P2
ST1275	198aa	Sulfolobus tokodaii 7
Ta0199	184aa	Thermoplasma acidophilum DSM 1728
TK0132	190aa	Thermococcus kodakaraensis KOD1
Tneu_1352	197aa	Thermoproteus neutrophilus V24Sta
Tpen_0293	142aa	Thermofilum pendens Hrk 5
TVG1467237	184aa	Thermoplasma volcanium GSS1
VNG2243G	186aa	Halobacterium sp. NRC-1
VNG6050G	186aa	Halobacterium sp. NRC-1
VNG6140G	181aa	Halobacterium sp. NRC-1
VNG6438G	186aa	Halobacterium sp. NRC-1
VNG6476G	186aa	Halobacterium sp. NRC-1
VNG7038	186aa	Halobacterium sp. NRC-1
VNG7100	181aa	Halobacterium sp. NRC-1
VNG7114	186aa	Halobacterium sp. NRC-1
VNG7171	186aa	Halobacterium sp. NRC-1

Appendix B

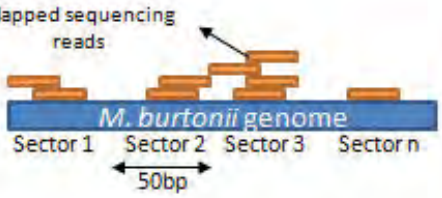
List of TBP gene locus tags with their corresponding TBP length and eucaryal genome source.

IMG Locus tag	TBP length	Genome
AT3G13445	200aa	Arabidopsis thaliana Columbia
ATEG_04542	263aa	Aspergillus terreus NIH2624
Dmel_CG9874	353aa	Drosophila melanogaster
T20B12.2	340aa	Caenorhabditis elegans
Tb10.61.0330	262aa	Trypanosoma brucei TREU927
YER148W	240aa	Saccharomyces cerevisiae
Accession number	TBP length	Genome
NP_001105318	200aa	Zea mays
NP_001098077	336aa	Pan troglodytes
NP_956390	302aa	Danio rerio
NP_990434	302aa	Gallus gallus
NP_001004198	318aa	Rattus norvegicus
XP_001704121	200aa	Giardia lamblia ATCC 50803
XP_001351620	327aa	Plasmodium falciparum 3D7
AAH16476	260aa	Mus musculus
NP_003185	339aa	Homo sapiens

Appendix C

Procedure used to determine whether a sector contains sequences that interact significantly with MbTBP.

Sector (every 50bp)	Mock	HisMbTBP206 mutant	HisMbTBP183
1st	16	20	23
2nd	66	116	78
3rd	212	254	318
n	a	b	c



Eliminate sectors containing less than 50 mapped reads

Sector	Mock	HisMbTBP206 mutant	HisMbTBP183
2nd	66	116	78
3rd	212	254	318
n	a	b	c

$$\text{Read ratio} = \frac{\text{reads (protein isoform)}}{\text{reads (mock)}}$$

Calculate read ratio

Sector	Mock	HisMbTBP206 mutant	HisMbTBP206 mutant read ratio	HisMbTBP183	HisMbTBP183 read ratio
2nd	66	116	1.75	78	1.18
3rd	212	254	1.19	318	1.5
n	a	b	b/a	c	c/a

Obtain mean and standard deviations of read ratios for each isoform

Sector	HisMbTBP206 mutant read ratio	HisMbTBP183 read ratio
2nd	1.75	1.18
3rd	1.19	1.5
n	b/a	c/a
Mean (read ratio)	0.992	0.986
Standard deviation (read ratio)	0.101	0.128

Calculate the ratio z-score of all sectors

Sector	HisMbTBP206 mutant read ratio	HisMbTBP206 mutant Ratio Z-score	HisMbTBP183 read ratio	HisMbTBP183 Ratio Z-score
2nd	1.75	7.5	1.18	1.5
3rd	1.19	1.96	1.5	4.02
n	b/a	[(b/a)-0.992]/0.101	c/a	[(c/a)-0.986]/0.128
Mean (read ratio)	0.992	-	0.986	-
Standard deviation (read ratio)	0.101	-	0.128	-

Z-score > 3.5
p-value < 0.001