NTNU
Norwegian University of
Science and Technology

NMR-Structural Analysis of Proteins from Oil Reservoir Metagenomes

Hilde Opdahl Erlandsen

Biotechnology (5 year)

Submission date:	October 2015
Supervisor:	Martin Frank Hohmann-Marriott, IBT
Co-supervisor:	Finn L. Aachmann, IBT Anna Lewin, SINTEF Alexander Wentzel, SINTEF

Norwegian University of Science and Technology
Department of Biotechnology

Acknowledgements

This master thesis was carried out at the Norwegian University of Science and Technology (NTNU), at the department of Biotechnology. The experiment and analyses took place at the molecular biology laboratories and NMR laboratory. The thesis was accomplished and written under guidance of Finn L. Aachmann, Anna Lewin, Finn Drabløs and Alexander Wentzel.

I would like to thank my supervisors for their excellent feedback and guidance in the laboratory and in the writing process; your help has been invaluable in improving the overall quality of the thesis. I would also like to thank Gerd Inger Sætrom for her help and guidance during the heterologous expression and purification of the proteins. My fellow lab-coworkers also deserve warm thanks for their helpfulness and the pleasant time spent in the laboratory.

I also want to thank my dear friends and family for their encouragement and support. Last, but not least, I want to thank my amazing boyfriend who I do not think I could have done without.

Abstract

Microorganisms are found ubiquitously around the world, where they survive as a result of their specialized biochemical reactions and enzymes. In the extreme environment of deep sea oil reservoirs, the microbes have to withstand difficulties such as endo- and exogenous toxins, as well as physical extremes (e.g. high temperature and pressure). The aim of this study was to identify putative novel enzymes suitable for NMR spectroscopy (by use of bioinformatic tools), from a translated metagenomic sequence database made of microorganisms obtained from such a reservoir. The proteins were to be of well characterized protein families, and to be studied by thermostability as well as NMR analysis to examine the proteins stability, integrity, purity and fold. This is important to further the understanding of their adaptations to high temperature and pressure, compared to their mesophilic counterparts. Identified proteins were to be heterologously expressed (with and without a His-tag) and purified prior to characterization with NMR spectroscopy, and thermo-exclusion analysis. Five candidates were found in the metagenomic database, namely Ars1, Ars2, Ars3, Glx1 and Glx2 – of the ArsC and GlxI protein families. The three candidates Ars1, Ars2 and Glx1 (the two latter with a His-tag) were heterologously expressed in both isotope labeled and non-isotope labeled versions. The Glx1-His protein seemed to be thermostable up to 60 °C, but the NMR results showed no indication of structure. The NMR spectra for Ars1 and Ars2-His show that these candidates seem to be folded and may therefore have potential for more detailed NMR-based structural analysis. In addition, Ars1 appear to be thermostable while Ars2-His does not. Thus only the Ars1 candidate might be used to provide more insight to the thermophilic and piezophilic characteristics of these polyextreme proteins.

Sammendrag

Mikroorganismer finnes overalt på kloden, og deres overlevelse er et resultat av deres spesialiserte enzymer og biokjemiske reaksjoner. Mikrober som lever i oljebrønner på dyphavet utsettes for et ekstremt fysisk miljø (blant annet høy temperatur og høyt trykk), og må i tillegg takle utfordringer som endo- og eksogene giftstoffer. Målet med denne studien var å identifisere potensielt nye enzymer som er egnet for NMR-spektroskopi (ved hjelp av bioinformatiske verktøy), fra en translatert metagenomisk sekvensdatabase laget fra mikroorganismer som er tatt fra en slik oljebrønn. Termostabiliteten til proteinene ble studert, i tillegg til at NMR analyse ble benyttet til å undersøke proteinenes stabilitet, integritet, renhet og folding. Et av kriteriene var at kandidatene måtte tilhøre godt kjente proteinfamilier, slik at de skulle kunne sammenlignes med sine mesofile slektninger. Dette er viktig for å se hvilke tilpasninger som har skjedd for at proteinene i oljebrønnene skal tåle høy temperatur og høyt trykk. Identifiserte proteiner ble uttrykt heterologt (med og uten His-tag) og renset før karakterisering med NMR-spektroskopi, og termo-ekskluderingsanalyse. Følgende fem kandidater ble funnet i den metagenomiske databasen: Ars1, Ars2, Ars3, Glx1 og Glx2 som er medlemmer av proteinfamiliene ArsC og GlxI. Tre kandidater (Ars1, Ars2 og Glx1 – de to sistnevnte med His-tag) ble uttrykt i både isotopmerkede og umerkede versjoner. Glx1-His viste termostabilitet opptil 60 °C men NMR resultatene viste ingen strukturell folding. NMR spektrene for Ars1 og Ars2-His viste begge å ha noe strukturell folding og kan derfor kunne brukes til mer detaljerte NMR baserte strukturanalyser. Kun Ars1 virker å være termostabil av disse to, og er dermed den eneste kandidaten som kan brukes til å gi mer innsikt til de termofile og piezofile tilpasningene hos disse polyekstreme proteinene

.

Table of contents

Abbreviations

amp	–	Ampicillin
ArsC	–	Arsenate reductase
BLAST	–	Basic local alignment search tool
BME	–	2-Mercaptoethanol
Dam	–	DNA adenine methylation
EDTA	–	Ethylendiaminetetra acetic acid
Glx	–	Glyoxalase
Grx	–	Glutaredoxin
HEPES	–	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid
HSQC	–	Heteronuclear single quantum coherence
IPTG	–	Isopropyl β-D-1-thiogalactopyranoside
LB	–	Lysogeny broth
LBA	–	LB agar
MEM	–	Minimum essential medium
mH_2O	–	Milli-Q water
MWCO	–	Molecular weight cut-off
MOPS	–	4-Morpholinepropanesulfonic acid
Mrx	–	Mycoredoxin
NCS	–	Norwegian continental shelf
NMR	–	Nuclear magnetic resonance
Ori	–	Origin of replication
PDB	–	Protein databank
PTPase	–	Low molecular weight protein tyrosine phosphatase
SDS-PAGE	–	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
smH_2O	–	Sterile mH_2O
SOC	–	Super optimal broth
TAE	–	Tris-acetate EDTA
TE	–	Tris-EDTA
TFB	–	Transformation buffer
Tris	–	Tris-(hydromethyl) aminomethane
Trx	–	Thioredoxin
WEB	–	Wash and equilibrium buffer

1 Introduction

1.1 Background

Microorganisms exist all across the globe, from high up in the troposphere to deep within the Earth's crust (Madigan et al., 2009, DeLeon-Rodriguez et al., 2013). This is possible due to their highly adapted biochemical reactions that provide metabolic diversity and the ability to withstand chemical and physical environmental stresses (Heulin et al., 2012). Metabolism of ordinarily safe compounds can produce endogenous[1] toxins as metabolites. Exogenous toxins, like heavy metals, that are present in the surrounding environment can also provide challenges for the organisms. In order to survive such toxins, the cells use various biochemical systems to maintain vital equilibrium by detoxification of the toxic substances involved, mainly by keeping the external environment out (Rothschild and Mancinelli, 2001). The systems are dependent on specialized enzymes to catalyze the biochemical reactions involved in the process. Without these enzymes the reactions could take a long time to occur spontaneously in the cell environment (Wolfenden, 2006)

1.1.1 Arsenate reductase

One example of keeping the environment out is shown with the arsenate reductase (ArsC - not to be confused with arsenate respiratory reductase[2]), which is part of a system that provides arsenic resistance in cells. Arsenic is an exotoxic metalloid[3] that is abundant in the Earth's crust, where it is mainly present as the oxyanions of arsenate or arsenite (Messens and Silver, 2006). It has been found in ground water in more than thirty countries around the world, affecting more than 100 million people in total (Mukherjee et al., 2008). Arsenate has the same chemical properties as phosphate and can substitute for phosphate in chemical compounds such as ATP (Hughes, 2002). However, the As-O bond is not a stable bond like the P-O bond is, and will spontaneously hydrolyze causing depletion of ATP in the cell. Arsenite is more toxic than arsenate in that it reacts strongly with thiol groups (a sulfhydryl group that is commonly found in enzymes) so encounters with arsenite may disrupt enzyme activity (Hughes, 2002). The principal mechanism for arsenate detoxification involves reduction of intracellular arsenate to arsenite by arsenate reductase (Luis et al., 2012). Despite

[1] Endogenous means that the toxin is produced inside the cell.
[2] Are large membrane bound enzymes involved in the metabolism of arsenic and not detoxification
[3] Its properties are a mixture of the properties of metals and nonmetals

arsenite being more toxic than arsenate, it is more readily exported out of the cell or into vacuoles by arsenate efflux permeases (Mukhopadhyay and Rosen, 2002).

At least three independent families of the arsenate reductase enzyme have been found that provide arsenic resistance in prokaryotes (Messens and Silver, 2006, Ordóñez et al., 2009). The different families are categorized by which redox system they utilize to provide reduction activity (Villadangos et al., 2011). All of the families are small cytosolic redox enzymes that work in sequential reaction with three different thiolate nucleophiles in a redox cascade (Messens and Silver, 2006, Ordóñez et al., 2009). To date, three reduction systems are found to be involved with arsenate reductases in prokaryotes, namely the thioredoxin (Trx - prevent oxidative stress), glutaredoxin and mycoredoxin systems (Grx and Mrx respectively – both prevent disulfide stress). Each thiol/disulphide system includes the named nucleophile and their respective reductase with the same given name (e.g. Trx and Trx reductase) (Messens and Silver, 2006, Ordóñez et al., 2009).

The Trx-coupled arsenate reductase (EC 1.20.4.4) is most excessively studied in the *Staphylococcus aureus* where it is found as a small (~15 kDa) monomeric enzyme (Messens and Silver, 2006). A larger (~57 kDa - each subunit is ~23 kDa) homodimeric version of this enzyme has been found in *Corynebacterium glutamicum* where the subunit contains a three-helical bundle added at the N-terminal (Villadangos et al., 2011). Both these enzymes comprise a low molecular weight phosphatase (PTPase) fold with a four-stranded parallel β-sheet and three α-helixes (Zegers et al., 2001, Villadangos et al., 2011). They also have three conserved catalytic cysteines (Cys-10, -82 and -89 in *S. aureus*, and Cys-88, -162 and -166 in *C. glutamicum*) that are essential for arsenate reductase activity, one of which is located in the characteristic P-loop (CX_5R) in the N-terminal, which binds to the oxyanion (Villadangos et al., 2011, Messens et al., 1999). The ArsC from *S. aureus* is the only ArsC found to use a metal ion (potassium) as a cofactor, providing higher ArsC activity (Lah et al., 2003). In *C. glutamicum* Trx-ArsC the charged nitrogen group at the end of the carbon chain ($N^{\zeta}H^{+}$) of Lys-144 is located in this space (Villadangos et al., 2011), while in *Bacillus subtilis* Trx-ArsC yet another modification has been found to engage metal independency (Roos et al., 2006). Some of these enzymes also possess PTPase activity, seen as the catalytic site of PTPase is conserved in ArsC (Zegers et al., 2001, Su et al., 1994).

The Grx-associated arsenate reductase (EC 1.20.4.1) is also a small (~16 kDa in *Escherichia coli*) monomeric enzyme with five α-helixes and a four stranded mixed β-sheet, one parallel and three anti parallel strands (Stevens et al., 1999). Most Grx-ArsC shows little sequence similarity with the other ArsC groups except for some preserved residues of the P-loop motif (CX_nR), but it is more similar to glutaredoxin and other glutathione-binding proteins (Shi et al., 1998). The Grx-coupled ArsC only contains only one conserved catalytic cysteine that is located in the P-loop on the N-terminal end (Liu et al., 1995). However, a Grx-dependent ArsC from *Synechocystis* sp. have the same LMW-PTPase fold as the *S. aureus* Trx-coupled ArsC, and encompass low PTPase activity (Li et al., 2003). This enzyme is a small homodimer (~30 kDa – each subunit is ~15 kDa) with three nucleophilic cysteines (Cys-11, -80, and -82), though these enzymes cannot use the Trx reduction system (Li et al., 2003, Yu et al., 2011).

The Mrx-dependent arsenate reductase (EC 2.8.4.2, also known as arsenate-mycothiol transferase) is a novel arsenate reductase family found together with the Trx-dependent homodimeric ArsC in the *C. glutamicum,* and contain the two protein members ArsC1 and ArsC2 (Ordóñez et al., 2009). These proteins are also small (~14 and 15 kDa) monomers with a PTPase I fold (unknown if they possess the activity), which only has one catalytic cysteine located in the conserved P-loop of the two enzymes (Villadangos et al., 2011). The remaining two nucleophiles needed for reduction is located on the mycothiol and mycoredoxin (Ordóñez et al., 2009). The metal-binding pocket analogous of *S. aureus* is occupied by the charged $N^{\zeta}H^{+}$ of the conserved Lys-64 (Villadangos et al., 2011).

Due to a high sequence similarity between these different ArsC families they are thought to have emerged independently by convergent evolution, thus they are considered to be analogous to each other (Messens and Silver, 2006). Over the last three decades the ArsC enzymes has been heavily focused on. Thorough knowledge of arsenic metabolism and detoxification is important for developing efficient and selective arsenic bioremediation[4] tools (Del Giudice et al., 2013). Engineered *C. glutamicum* strains have already shown potential as a bioremediation tool, due to their ability to accumulate large amounts of arsenic (Villadangos et al., 2014). This way the toxic arsenite (which would usually be pumped out into the environment) can be efficiently removed from drinking and crop waters.

[4] To remove or neutralize hazardous substances by the use of microorganisms

1.1.2 Glyoxalase I

When the toxic compound is produced inside the cell it is not readily kept out of the cell. Glyoxalase I (GlxI - EC 4.4.1.5, also known as lactoylglutathione lyase or hemithioacetal isomerase) is a metalloenzyme that regulates the endotoxic metabolite methylglyoxal. This compound is primarily produced during glycolysis, but is also found as an exotoxic pollutant which is known to inhibit cell growth (Cooper, 1984, Maeta et al., 2004, Kalapos, 1999). GlxI is part of the ubiquitous glyoxalase system that degrades α-oxoaldehydes (mainly methylglyoxal) into α-hydroxyacids (Thornalley, 1990). The system consists of two enzymes, namely glyoxalase I and glyoxalase II (GlxII – EC 3.1.2.6) (Racker, 1951). In most bacteria, methylglyoxal and glutathione forms an adduct (methylglyoxal-glutathione hemithioacetal) in a non-enzymatic reaction, which is then isomerized into S-D-lactoylglutathione by GlxI (Larsson, 1983, Mannervik, 1980). The S-D-lactoylglutathione is then converted into D-lactate by GlxII, which also reduces the glutathione from the hemithioacetal so it can continue to bind methylglyoxal (Dolphin et al., 1989).

The highly conserved structures of GlxI from different organisms suggests that they may have arisen by divergent evolution from a common ancestor, into several homologous proteins (Thornalley, 2003). The GlxI protein family is divided into the Ni^{2+}- or Zn^{2+}-dependent GlxI where in prokaryotes, both groups are found to be homodimeric (Suttisansanee and Honek, 2011). Enzymes of both groups have been found capable of using several types of metal ions, but they have the highest activity with the given ion (Clugston et al., 1998, Saint-Jean et al., 1998). In addition, the Ni^{2+}-dependent enzymes is inhibited by Zn^{2+} (Clugston et al., 1998). The bacterial GlxI is usually of the small Ni^{2+}-dependent group, where each subunit (∼130 amino acids long) contains two unrelated βαβββ domains which is connected through a flexible loop (Suttisansanee and Honek, 2011). Two different types of dimeric conformations have been found in this group of enzymes. Generally the monomers are joined to each other through the βαβββ domains in a hetero conformation (N-terminal domain bind to the C-terminal domain), creating two active sites in the dimer interface (Figure 1.1, panel B). However, in a newly discovered GlxI from *Clostridium acetobutylicum* the monomers are connected in a back-to-back conformation, where one active site is located within each of the monomers (Figure 1.1, panel A) (Suttisansanee et al., 2011). The Zn^{2+}-dependent enzymes consist of larger subunits at about 180 amino acids in length which is found as the βαβββ connected dimer (Suttisansanee and Honek, 2011). For all the GlxI enzymes, an active site consists of a catalytic pocket with a metal ion bound to the bottom.

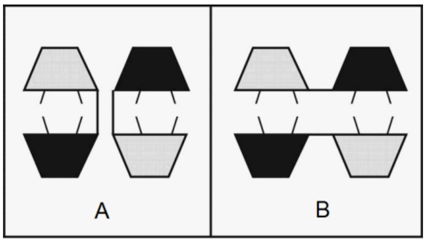


Figure 1.1: Illustration of the two dimeric conformations of GlxI found in prokaryotes (Suttisansanee et al., 2011). The back-to-back dimer (A) and the βαβββ connected dimer (B).

The prokaryotic GlxI enzyme has been a well studied enzyme for several reasons. For instance, the amino acid sequence and structure of GlxI from several pathological bacteria has been found to contain certain differences from the human GlxI (Clugston and Honek, 2000). These differences are believed to be critical for GlxI activity and may therefore be targeted when developing selective inhibitors of the bacterial enzymes (Clugston and Honek, 2000). In addition, an overexpression of the glyoxalase system has shown to provide heavy-metal and saline resistance in transgenic tobacco plants, providing an effective strategy for stress tolerance in crop plants (Singla-Pareek et al., 2006, Singla-Pareek et al., 2003).

1.1.3 Extreme and polyextreme protein adaptations

In some cases it is impossible to keep the environment out of the cell, particularly when dealing with physical extremes such as high temperature and pressure (Rothschild and Mancinelli, 2001). Organisms that thrive in one extreme setting are called extremophiles, while organisms that thrive in more than one extreme condition are called polyextremophiles. Through evolution, microbes have developed several protein adaptations to withstand these different extremes.

Thermophiles are microbes that thrive in 50-70 °C, while hyperthermophiles can grow at temperatures up to 105 °C (Reed et al., 2013). At such high temperatures, mesophilic proteins without the required adaptations will denaturize – undergo permanent unfolding where the hydrophobic core is exposed, causing the proteins to aggregate (Tomazic and Klibanov, 1988). Several protein adaptations are found to contribute to maintaining protein fold and function in high temperatures. Many thermostable proteins are found to contain other quaternary structures than seen in their mesophilic counterparts, for example some mesophilic monomers can be found as thermophilic oligomers (Reed et al., 2013). This oligomerization is thought to increase the rigidity of each subunit and endorse a tighter packaging of the hydrophobic core, thus reducing the exposure of hydrophobic residues to the solvent (Vieille and Zeikus, 2001). An increased amount of disulfide bridging between cysteine residues have been shown to increase stability in thermophilic proteins, as well as being important in multimerization by increasing the rigidity and interlocking adjacent chains between monomeric subunits respectively (Cacciapuoti et al., 2012, Boutz et al., 2007). There is also an increased amount of salt-bridging found in hyperthermophilic enzymes, which unlike the mesophilic enzymes are stabilized by the interactions due to the higher temperature (Karshikoff and Ladenstein, 2001). In addition, thermostable proteins can have more charged residues on their surface that provides salt-dependent stability, while the core contains a higher amount of large hydrophobic amino acids (Liu et al., 2012, Reed et al., 2013).

Piezophilies thrive under extremely high hydrostatic pressure and are often found in polyextreme environments in combination with high or low temperature (Reed et al., 2013). When high pressure is applied to proteins, water molecules are inclined to penetrate the hydrophobic core of the proteins and disrupt their structure (Tanaka et al., 2000). Therefore, the piezophilic proteins contain a higher amount of small polar amino acids, which provides a pressure stable, densely packed core (Di Giulio, 2005). The piezophilic adaptations seem to be less significant for polyextreme proteins than the thermophilic adaptations (Hay et al., 2009).

1.1.4 Employment of extremophilic or polyextremophilic proteins

It seems that subtle changes in the amino acid sequence that alter the charge, hydrophobicity and structure of a protein, are a common adaptation for proteins to maintain active in extreme environments (Reed et al., 2013). These adapted proteins (just a few examples are mentioned here) are highly valued in the industry where several applications have been found for them. For instance, industrial reactions can often be favored at high temperatures or pressures, where the chance of bacterial contamination is lower and the product yield is higher, such as in food and drug applications (Unsworth et al., 2007, Simonato et al., 2006). By using thermophilic or piezophilic proteins in such reactions the operation costs could be reduced by avoiding a constant enzyme replacement due to denaturation (Unsworth et al., 2007, Simonato et al., 2006). There is still a lot to learn about the different extreme adaptations in proteins, and understanding these adaptations is greatly desirable to science. With understanding comes knowledge, which can be used to engineer and utilize proteins that are capable to function in extreme environments. There are many possible fields of application, including industry, environment and biotechnology (Reed et al., 2013).

SINTEF, in collaboration with Statoil and NTNU, conducted a project where microbial samples were extracted from two oil reservoirs (Well I and II) located on the Norwegian Continental Shelf (NCS). These oil reservoirs are located about 2.5 km below the sea surface and can be considered as polyextreme environments based on their high temperature (~85 °C) and high pressure (~250 bar), salinity, and the possible presence of heavy metals and other toxic compounds (Lewin et al., 2014). Two specialized pressure flasks were used for sampling to avoid loss of genome due to rapid pressure changes, and thorough care was taken to avoid contamination (Kotlar et al., 2011). A translated metagenomic sequence database was made from the two samples by 454 pyrosequencing of the isolated DNA (~60 µg in total) and initial bioinformatic analyses. The resulting database contains 32.6 Mbp of assembled contigs with a high DNA sequence coverage (Kotlar et al., 2011, Lewin et al., 2014), providing a large supply of unstudied polyextreme proteins.

1.2 Gene mining

In order to locate novel proteins in a translated metagenomic sequence database, a process known as gene mining can be used. This is a form of sequence-based screening, which detects novel proteins (in e.g. a translated metagenomic database) based on the amino acid sequences of well characterized proteins (found in a protein database such as Swiss-Prot) by the use of alignment algorithms (Streit et al., 2004). The Basic Local Alignment Search Tool (BLAST) is widely used to approximate alignment similarity between a query sequence and sequences from amino acid or nucleic acid databases (Altschul et al., 1990). It can be applied in several ways, such as gene alignment with protein sequences, protein-protein alignment, and gene-gene alignment (Altschul et al., 1990).

The putative novel proteins that are found need to be processed and biochemically tested to verify their protein activity (Streit et al., 2004). Analysis by the use of nuclear magnetic resonance (NMR) spectroscopy can also be performed on the proteins, which can provide insight to their structure and functions (Kwan et al., 2011). Novel proteins from extreme environments can be difficult and expensive to cultivate in a laboratory, because the microbes containing them might depend on such an environment to grow (Rothschild and Mancinelli, 2001). Therefore it is proven more beneficial to use heterologous expression by means of recombinant DNA technology. There are several aspects to consider when expressing recombinant proteins. The most beneficial gene codons, hosts, vectors, recombination- and purification techniques should be established for each particular protein.

1.3 DNA recombination and heterologous expression

A protein sequence retrieved from a metagenomic database can be synthesized into a nucleic acid in several ways. In fact, there are multiple companies that have specialized in doing this for others (e.g. GenScript). When sequencing the DNA, it is favorable to optimize the codon sequence by which host organism it should be expressed in, as to make sure the tRNA pool in the host can translate the protein without stalling (Burgess-Brown et al., 2008, Kane, 1995). In DNA recombination it is important to produce large quantities of the target gene, since multiple attempts can be necessary in order to succeed. Therefore, following the target gene sequencing it should be inserted into a cloning vector used in quantification of recombinant genes. This vector is likely to possess an origin of replication (Ori – specific for the host organism to be used) which provides a high-copy number.

An expression vector would need to possess a promoter with a gene regulatory system. The regulatory system is necessary to prevent the host organisms from expressing the protein until they have been quantified to a satisfactory level. This is to make sure that a large amount of product is achieved, in case the recombinant protein is toxic to the cell (Studier and Moffatt, 1986). An example is the T7 promoter (from bacteriophage T7) that has been put under the control of the Lac repressor in the pET expression system (Studier and Moffatt, 1986, Tabor, 2001). The repression system contains a repressor (encoded by *LacI*) that binds to the *LacO* operon, and can be inactivated by isopropyl β-D-1-thiogalactopyranoside (IPTG) – an analog to allolactose. However, this particular vector requires a specific bacterial host that contain a recombinant T7 polymerase encoded in the genome, that is regulated by the same means as the target gene (Studier and Moffatt, 1986).

All vector systems have a way to screen the host organisms to identify the ones possessing the vector (selective markers). A common way to do this is by adding antibiotic resistance genes into the vector, such as amp^R that encodes β-lactamase (provides ampicillin resistance) (Sutcliffe, 1978). This is then complimented with the use of the specific antibiotic in the medium, resulting in selective growth of the organisms containing the vector (Hershfield et al., 1974). The target gene is inserted into the required vector by the use of type II restriction endonucleases, which binds to and digests the DNA at specific recognition sites leaving either blunt or sticky-ends (e.g. with HindIII and EcoRI respectively) (Smith and Welcox, 1970, Hedgpeth et al., 1972). DNA ligase can then join matching DNA ends together (Zimmerman et al., 1967). Some endonucleases are inhibited by certain DNA methylations, which protect the host DNA against the specific enzyme as the use the enzyme as a defense mechanism against foreign viral DNA (Raleigh and Wilson, 1986, Arber, 1965). However, DNA methylation is also used in a various other cellular functions, such as Dam methylation, which stall one round of replication in bacteria, until the first round is complete (Von Freiesleben et al., 2000), will also block restriction cleavage of DNA at the specific site.

1.3.1 Host organisms

The host organism that is employed for quantification of a target vector can be manipulated in several ways; such as to make sure of high transformation efficiency, and/or to remove a specific methylase in order to use a certain restriction enzymes on the DNA. In protein expression, the host can be favored with mutations that provide a higher protein yield, such as knockout of proteases that are involved in recombinant degradation (Phillips et al., 1984) and additional tRNA genes to prevent a halt in translation (Kurland and Gallant, 1996). The selection of taxa in a host organism used for expression can depend on which organism the target protein is retrieved from, since posttranslational processes available in this particular organism can be needed in order to produce a functional protein (Terpe, 2006). The most abundantly used host organism is *E. coli*, because they have well studied biochemical reactions, are easy to grow in the laboratory, and has a short replication time (Terpe, 2006).

1.3.2 Protein isolation

The purification of target proteins from the cell culture can be done by making use of the characteristics of the specific protein. Such as, thermophilic proteins will continue to stay solvent when treated with heat, while the non-thermophilic proteins will aggregate (Tomazic and Klibanov, 1988). Other characteristics that can be used are certain traits of individual residues, such as the nickel-binding ability of histidine that can be disrupted by imidazole (Hoffmann and Roeder, 1991). This specific trait has been utilized by tagging target proteins with His-oligomers, which then is easily isolated in a column with nickel covered resin (Yip and Hutchens, 1992). Similar characteristics are utilized when separating or isolating DNA, where the DNA possess a negative charge that can interact with certain resins (such as silica or magnetic beads) in a column (Berensmeier, 2006, Melzak et al., 1996), and move DNA through a meshed network of agar polysaccharides toward the positive charge when electrical current is applied (Thorne, 1966). The different sized DNA fragments will separate on the gel since the movement rate through the gel is dependent on fragment size (Thorne, 1966). Proteins can be separated in the same manner on a polyacrylamide gel (PAG) after being denaturized and applied a negative charge with sodium dodecyl sulfate (SDS) – in a technique called SDS-PAGE (E for electrophoresis) (Shapiro et al., 1967). The gels need to be stained in order for the DNA and proteins to show, and a standard (a sample of pre-determined fragment sizes and concentrations) can be added to determine the molecular-weight and quantum of the DNA fragment or protein (Weber and Osborn, 1969).

1.4 NMR spectroscopy

Sufficiently isolated proteins can be studied by NMR spectroscopy. Signals in NMR spectra arise from transitions made by (atomic) nuclei between different excitation states in a magnetic field (Kwan et al., 2011). When electromagnetic radiation at resonance frequency is applied, signals can be observed. The resonance signals created by an NMR active nucleus (e.g. ^{1}H – proton, ^{13}C and ^{15}N) depend on the surrounding magnetic environment created by other nuclei, which among other things is based on their electron negativity, as well as their distance and position to the signaling nucleus. These small differences are called chemical shifts (differences) and are shown in parts per million (ppm – corrects for the differences that scale with NMR magnet size) in a one or several dimensional spectrum. Thus, signals arising from one specific isotopic nucleus can be distinguished. In addition, the peak intensity is proportional to the number of isotopic nuclei. In such spectra, each specific nuclei environment will provide a specific signal which in the case of small molecules, can be used to find the molecular structure. Macro molecules such as proteins, give rise to complex 1D spectra where numerous of these signals may overlap, thus making it impossible to establish detailed information regarding each nuclei (Kwan et al., 2011).

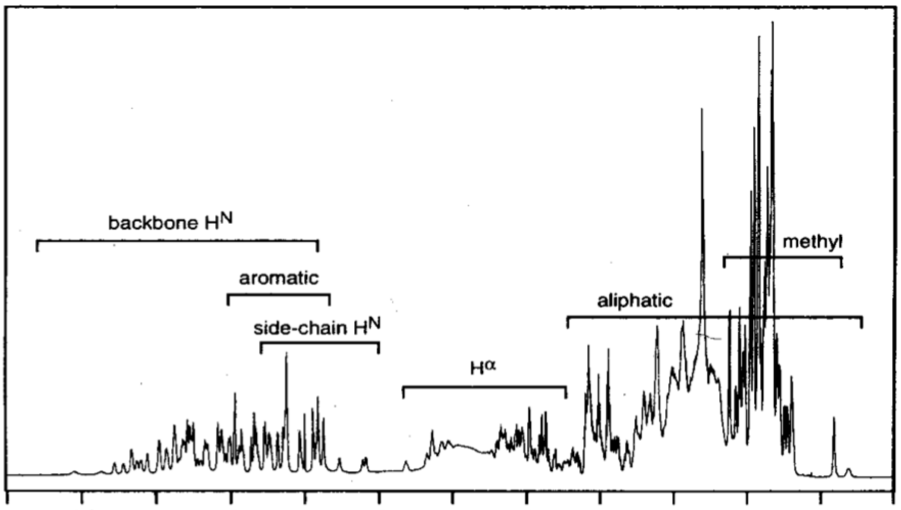


Figure 1.2: Proton chemical shift positions of chemical groups in the ubiquitin of 8.5 kDa (Cavanagh, 1996).

However, some tendencies have been found in the occurrence of signals belonging to certain chemical groups in a protein. In a 1D proton spectrum (Figure 1.2), protons coupled to carbons (except aromatic protons) have a low frequency and tend to appear between -1-6 ppm, while protons coupled to the more electronegative amides have a higher frequency and tend to appear in the 6-11 ppm region of the spectrum (Groß and Kalbitzer, 1988). These two regions can be divided into several sub-regions of carbon or amide connected protons (H^C and H^N respectively), such as methyl proton, α-proton – H^α, side-chain H^N, and backbone H^N regions (Figure 1.2) (Cavanagh, 1996, Groß and Kalbitzer, 1988). Since several of these regions overlap it is difficult to pinpoint the signals in each region to a specific chemical group. However, the H^α region can further be divided into sub-regions of lower or higher frequencies, based on whether the backbone H^C groups are arranged in α-helixes or β-sheets respectively. These frequencies of β-sheet protons can more readily be distinguished from other signals (found between 5-6 ppm), thus providing some information of the proteins secondary structure (Kwan et al., 2011).

1D spectrum of NMR cannot be used to determine the structure of folded proteins, but it is possible to transfer magnetization from one nucleus to another (homo- or heteronucleus) (Kwan et al., 2011). This produces signals that link the frequencies of two or several nuclei (if several transfers are made) in a 2D or multidimensional spectrum respectively. In order to use this technique for proteins, they must be enriched with carbon and/or nitrogen NMR active isotopes (^{13}C and ^{15}N instead of ^{12}C and ^{14}N). This can be done by growing the host organism in a medium enriched with these isotopes. A 2D spectrum that correlates protons with an heteronucleus is called heteronuclear single quantum coherence spectrum (HSQC). The two cross-sections (one horizontal and one vertical) through the correlated 2D spectrum show a 1D spectrum for each of the two NMR active nuclei (Kwan et al., 2011). The 2D {1H-15N} HSQC spectra – also referred to as the protein fingerprint – show a correlation peak for each pair of covalently bonded ^1H-^{15}N nuclei in the protein backbone, and for some of the side-chains who contain NH groups, such as Asn and Gln (Bodenhausen and Ruben, 1980). Analysis of the 2D {1H-15N} HSQC spectrum can establish if the protein is suitable for more detailed NMR-based structural analysis (Kwan et al., 2011).

Both the 1D and 2D spectra can give a relatively good indication of the order and stability of a protein. A nucleus in an unfolded or randomly coiled protein does not have one distinct environment as in a folded protein, providing an average NMR signal of all the different environments (Kwan et al., 2011). The environments of many nuclei are not as distinct as in a folded protein, and thus they tend to cluster around 8 and 1 ppm for the H^N and methyl protons respectively (Wishart et al., 1995). The averaged signals from an unfolded protein tend to provide broad peaks, while well folded proteins tend to have sharp narrow peaks. However, the shape of the signals may also depend on other factors, such as the strength of the magnetic field that is applied, temperature used, size and shape of the protein, and which buffer the protein prefers in order to fold properly (Kwan et al., 2011). Higher magnetic strength and temperature, in addition to smaller size and more globular shape on the proteins, each provide a higher resolution on the NMR signals (sharper and narrowed peaks) (Kwan et al., 2011).

2 Aim of study

The aim of this study was to identify and analyze putative novel enzymes (suitable for structural NMR spectroscopy) from a metagenomic sequence database made from microorganisms native to two oil reservoirs on the NCS. The proteins were to be identified using different bioinformatic tools. A thermal test would be used to determine thermostability while NMR analyses would determine the proteins stability, integrity, purity and fold. This is important to gain further insight regarding their adaptations to high temperature and pressure, compared to their mesophilic counterparts. Prior to analyzes, the respective genes were to be modified with a His-tag, heterologously expressed (both with and without isotope labeling) and purified to homogeneity satisfactory for NMR analysis (preferably 90 % or higher).

3 Materials and methods

3.1 Candidate selection

Protein candidates suitable for NMR spectroscopy were selected from 546,238 manually annotated and reviewed proteins in the Swiss-Prot database, using the website UniProt knowledgebase.

Selection criteria

Certain criteria had to be fulfilled for the proteins to be considered as relevant candidates for the subsequent NMR study. Key search criteria used in this search were protein size, their localization in the cell, and to some extent protein classification and physical traits.
They needed to be about 50-150 amino acids in size (molecular weight < 20 kDa – the smaller protein the easier it is to get high resolution) and not be dependent on strongly paramagnetic ions (such as Cd^{2+} or Cu^{2+}) as cofactors, since the NMR signals of nuclei near such metals are difficult to observe (Otting, 2010). In order to simplify the isolation process, the proteins should be soluble and not be embedded in the cell membrane. Also, the enzyme queries should be verified functional monomeric or homodimeric proteins (to ensure functional proteins) who are not members of the superfamily of proteases due to the risk of proteolytic cleavage within the sample, which provide broad averaged NMR signals. These criteria were fulfilled by using the advanced search function in UniProt (Appendix A) and manual selection among the search results.

Protein alignment

The protein sequences were then used as a query to search for similar proteins in a metagenomic library of 52,415 annotated contigs. This was done by using the tblastn algorithm in BLAST (National center for biotechnology information – NCBI), which allows a translated nucleotide database to be searched using a protein query, thereby providing the name of contigs containing the genes of interest and the reading frame of the gene.

The full nucleotide sequences of the contigs were then located in the database by using the contig names as queries in a regular expression based search in the text editor TextPad (Helios Software Solutions). The contig sequences were applied into ORF finder (NCBI) to locate the full amino acid sequence of the candidates. The correct ORF was found by looking at ORFs with the same reading frame and size of the candidate gene which were obtained from the tblastn search. If several ORFs of this size were found in the same reading frame, a closer look at the given sequence was performed to find the candidate sequence. ORF finder is a graphical analysis tool which identifies all open reading frames in a nucleic acid sequence by the use of standard or alternative genetic codes. In addition, ORF finder translates the ORFs into amino acid sequences.

The amino acid sequences were then used to search for similar proteins (identity >30 % is considered as similar) in the nr database (a set of several protein databases) by using the blastp algorithm in BLAST. This particular algorithm searches a protein database using an amino acid query providing a list of similar proteins and information about them. The information of these proteins was studied to make sure that the proteins enclose the final criteria for relevant candidates; that they resemble intact functional proteins with already well established protein activity assays. Finally the candidates selected were used as queries in the protein databank (PDB – a database of proteins with known structure) by using the blastp algorithm in BLAST. This was done in order to find which protein group within their family they resembles the most.

Multiple sequence alignments were made in Clustal Omega (EMBL-EBI) and annotated with Jalview 2.8.2 (University of Dundee), while pairwise alignments were made directly in Jalview 2.8.2. All programs used in this thesis were conducted with their default settings.

3.2 Bacterial strains and vectors

Table 3.1: The properties and sources of the different bacterial strains and vectors used in the thesis, in addition to the sizes of the different genes and their modified derivates used. For the gene sequences for each candidate and their orientations in pUC57 see appendix C.

Bacterial strain or plasmid	Properties	Source
Strains		
E. coli, DH5-α	fhuA2 lac(del)U169 phoA glnV44 Φ80' lacZ(del)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17; Cloning strain	Invitrogen
Dam$^-$/Dcm$^-$ competent *E. coli*	ara-14 leuB6 fhuA31 lacY1 tsx78 glnV44 galK2 galT22 mcrA dcm-6 hisG4 rfbD1 R(zgb210::Tn10) Tet^S endA1 rspL136 (Str^R) dam13::Tn9 (Cam^R) xylA-5 mtl-1 thi-1 mcrB1 hsdR2; Cloning strain	New England Biolabs - NEB
E. coli BL21-CodonPlus (DE3)-RIPL	*E. coli* B F$^-$ ompT hsdS(r_B^- m_B^-) dcm$^+$ Tet^R gal λ(DE3) endA Hte [argU proL Cam^R] [argU ileY leuW Strep/Spec^R]; Production strain	Stratagene
Vectors		
pUC57-"Candidate"	*amp^R*, pMB1 derived Ori, Blue/white screen compatible with insert of a candidate gene fragment in EcoRV site; Cloning vector	GenScript
pUC57-"Candidate"-His	Derivative of pUC57-"Candidate" with a 6 bp DNA fragment removed by BclI and BglII and thus destructing a TGA stop codon; Cloning vector	This study
pOD1	*amp^R*, pBR322 Ori, *Rop*, T7 promoter under LacI regulation with *SelW* (266 bp) inserted in NdeI-XhoI sites; Expression vector	Another project (Dikiy et al., 2007)
pET21-"Candidate"	Derivative of pOD1, with a candidate gene fragment from pUC57-"Candidate" subcloned into the NdeI-XhoI sites; Expression vector	This study
pET21-"Candidate"-His	Derivative of pOD1, with a His-tagged candidate gene fragment from pUC57-"Candidate"-His subcloned into the NdeI-XhoI sites; Expression vector	This study

Candidate genes	Without His-tag (bp)	With His-tag (bp)
Ars1	484	478
Ars2	469	463
Ars3	476	-
Glx1	440	434
Glx2	415	-

3.3 Vector construction

When not specified otherwise the centrifuge used in this thesis was a Eppendorf Centrifuge 5415R with F-45-24-11 rotor for samples in 1.5 ml Eppendorf tubes and tubes from different kits, and a Eppendorf Centrifuge 5804R with F-34-6.38 rotor for samples in 10 or 50 ml tubes.

Media prepared

All media were mixed with a proper amount of Milli-Q water (mH_2O) and autoclaved. The LBA was cooled down to 50 °C after autoclavation before amp (100 µg/ml, AppliChem) was added and the agar plates were made (20-25 ml/dish).

Table 3.2: The compositions of the different media prepared for vector construction: Lysogeny broth (LB), LB agar (LBA), super optimal broth (SOC) and Psi broth.

LB:	10 g/l	Tryptone (OXOID)	**LBA:**	10 g/l	Tryptone
	5 g/l	Yeast extract (OXOID)		5 g/l	Yeast extract
	10 g/l	Sodium chloride (VWR)		10 g/l	Sodium chloride
				15 g/l	Agar (OXOID)
SOC:	20 g/l	Tryptone	**Psi:**	20 g/l	Tryptone
	5 g/l	Yeast extract		5 g/l	Yeast extract
	0.5 g/l	Sodium chloride		10.24 g/l	Magnesium sulfide heptahydrate
	0.186 g/l	Potassium chloride (Merck)			
	3.6 g/l	D-glucose (VWR)			
	4.8 g/l	Magnesium sulfide heptahydrate (VWR)			

Buffers and loading dye prepared

All solutions were made with mH_2O. All buffers (except TAE) were filtrated using Stericup™ Millipore Express™PLUS 0.22 µm filters (EMD Millipore) before use. The pH of TFB I was adjusted with acetic acid (VWR), while the remaining buffers were adjusted with sodium hydroxide (VWR).

Table 3.3: The compositions of the DNA loading dye (DLD) and different buffers used for vector construction: Transformation buffer (TFB) I and II, Tris-acetate-EDTA (TAE) and Tris-EDTA (TE).

TFB I:	30 mM	Potassium acetate (Merck)	**TFB II:**	10 mM	MOPS* (Sigma Aldrich)
pH 5.8	100 mM	Rubidium chloride (AMRESCO)	pH 5.8	10 mM	Rubidium chloride
	10 mM	Calsium chloride (VWR)		75 mM	Calsium chloride
	50 mM	Manganese(II) chloride tetrahydrate		15 %	Glycerol
	15 %	Glycerol (VWR)			
TAE:	40 mM	Tris** (VWR)	**TE:**	10 mM	Tris
pH 8.0	20 mM	Acetic acid	pH 8.0	1 mM	EDTA
	1 mM	EDTA*** (VWR)			
DLD:	60 mM	EDTA			
	10 mM	Tris-HCl (Sigma Aldrich)			
	0.03 %	Bromophenol blue (Sigma Aldrich)			
	0.03 %	Xylene cyanol (Sigma Aldrich)			
	60 %	Glycerol			

* 4-Morpholinepropanesulfonic acid, ** Tris-(hydroxymethyl)aminomethane, ***Ethylenediaminetetraacetic acid.

DNA digestion

DNA digestion with the use of a single restriction enzyme was achieved by mixing 10-50 µl of DNA with 0.2-1 µl of endonuclease (NEB) and a 1:10 dilution of the appropriate endonuclease reaction buffer (NEB) to total reaction mixture. To bring the reaction mixture to a certain volume, sterile mH_2O (smH_2O) was added as the extra liquid if needed. Enzymes (endonucleases and ligases) were always the last to be added to reaction mixtures to ensure that it was not exposed to extreme conditions. The reaction mixture was then incubated for 1-2 hrs at mutual optimum temperature.

DNA double digestion was achieved by mixing 10-50 µl DNA with 0.2-1 µl of each specific endonuclease and a 1:10 dilution of a mutually shared endonuclease reaction buffer to total reaction mixture (one providing the most activity for both enzymes). To bring the reaction mixture to a certain volume, smH_2O was added as the extra liquid if needed. The reaction mixture was incubated for 1-2 hrs at mutual optimum temperature or 1.5 hrs at each optimum temperature if a mutual optimum temperature did not exist.
The appropriate restriction enzymes were found using the program Clone Manager Suite 6.0 (Sci-Ed Software).

DNA separation and isolation

Agarose gel electrophoresis with GelGreen™ dye (50 µl/l gel, VWR) was used to separate DNA fragments by loading them together with an approximate ratio of 1:8 DNA loading dye to DNA into wells on the agarose gel (0.8 %) and 2 µl of ready to use O'GeneRuler™ 1 kb DNA standard (250-10,000 bp, Thermo Scientific) was loaded into separate wells to be used as DNA ladder. The gel was run in TAE buffer at 60-100 V for 0.5-2 hrs until the loading dye had run approximately ¾ of the gel. For visualization and data footage of the DNA fragments on gel, Gel Doc™ XR+ Imager (BioRad) was used.

For extraction of DNA from the gel, the wanted DNA fragments were cut out of the gel and purified with QIAquick Gel Extraction Kit with Centrifuge Protocol (QIAGEN), where 30 µl of sterile Milli-Q water (smH_2O) was used to elute the DNA fragments. When gel electrophoresis was not applied after restriction digestion, QIAquick® PCR Purification Kit Centrifuge Protocol (QIAGEN) was used to isolate the DNA from restriction enzymes and their respective buffer.

DNA ligation

Re-ligation was achieved by mixing 30 µl of DNA with 1 µl T4 DNA ligase (NEB) and 3.5 µl of T4 DNA ligase reaction buffer (NEB). smH$_2$O (0.5 µl) was added to bring the total reaction volume up to 35 µl. The ligation mixture was incubated at 8 °C over night or 2 hrs in room temperature. The re-ligated plasmids were then transformed and propagated in the *E. coli* DH5-α strain, which has several features that make it useful for recombinant DNA methods (Table 3.1).

Ligation of genes into the target vector pET21 was achieved by mixing approximately 1:3 of vector to insert DNA, with 1 µl T4 DNA ligase and a dilution of 1:9 T4 DNA ligase reaction buffer to the total reaction mixture of 20-35 µl. The DNA ratio was calculated by plotting the size of the two fragments and the mass of vector DNA to be used into the ligation calculator of NEBioCalculator™ v1.2.1 (NEB). To bring the reaction mixture to a certain volume, smH$_2$O was added as the extra liquid if needed. The ligated expression vectors were transformed and propagated in the BL21 cells, which enable efficient high-level expression of heterologous proteins (Table 3.1).

Preparation of RbCl cells

A preculture of DH5-α and BL21 cells was made by growing the cells in LB (5 ml) with amp (100 µg/ml) over night at 37 °C. Psi-medium (100 ml) was inoculated with 1 % preculture and incubated at 37 °C until OD$_{600}$ ~0.410. The Psi medium is used in order to increase cell density. The cells were then incubated on ice for 15 minutes before they were centrifuged down (5 minutes at 4500 rpm) in 50 ml sterile tubes. The supernatant was discarded before the pellet was resuspended in cold TFBI (40 ml) and incubated on ice for 5 minutes. The cells were then centrifuged down a second time (5 minutes at 4500 rpm) and the supernatant discarded before the cells were carefully resuspended in cold TFBII (3 ml) on ice. Aliquots of 100 µl per tube (on ice) were then quick-frozen in liquid nitrogen and stored at -80 °C.

Plasmid amplification

Amplification of plasmids was achieved by propagation in transformed bacteria. The bacteria were transformed by the use of heat shock, where aliquot of chemical competent (RbCl) cells or dam$^-$/dcm$^-$ competent *E. coli* were thawed on ice and DNA (2-5 µl) was added. The reaction mixture was incubated on ice for 30 minutes followed by incubation at 42 °C for 25 or 45 sec and then on ice for 2 or 5 minutes for the dam$^-$/dcm$^-$ competent *E. coli and* chemical competent cells respectively. SOC-medium (250 µl) was then added before the cells were incubated at 37°C for 1 hour and plated (50 and 200 µl) onto LBA plates with amp (100 µg/ml). The plates were incubated at 37 °C over night. For long time storage, a bacterial glycerol stock was made by mixing 500 µl of glycerol solution (50 %) and 500 µl overnight culture of the bacteria transformed with the desired plasmid.

Isolation of plasmids from bacteria

Preculture (5-100 ml) was harvested by centrifugation (5-10 minutes at 5,000 rpm). Several batches with the same plasmid could be made and purified separately before they were eluted into the same tube. The supernatant was removed and the plasmids were isolated from the cell pellets using WizardPlus SV Minipreps DNA Purification System (Promega), or NucleoBond® Xtra Midi (Macherey-Nagel) for plasmids with low copy number (used to obtain the pET21 backbone). For long time storage, 11 µl of TE buffer was added per 100 µl of eluted DNA. The DNA concentrations were measured using the NanoDrop 1000 Spectrophotometer.

3.4 Protein expression and purification

Media prepared

The 2xLB and base solution of M9 (marked in gray) were mixed with the proper amount of mH_2O and autoclaved. Stock solutions of D-glucose, $MgSO_4$, trace metals and MEM vitamins were sterile filtrated through a Steriflip™ and added together with the Bioexpress Cell Growth Media to the base solution right before use.

Table 3.4: The compositions of the media used in protein expression (in addition to LB).

2xLB:	20 g/l	Tryptone	**M9:**	6 g/l	Disodium phosphate (Merck)	Base
	10 g/l	Yeast extract		3 g/l	Monopotassium phosphate (Merck)	
	10 g/l	Sodium chloride		0.5 g/l	Sodium chloride	
				1 g/l	Ammonium sulfate($^{15}N_2$) (CIL*)	
				2 g/l	D-glucose	
				60.2 g/l	Magnesium sulfate heptahydrate	
				20 ml/l	Trace metal solution**	
				6.6/20 ml/l	Bioexpress Cell Growth Media (^{15}N)(CIL)	
				10 ml/l	MEM*** Vitamin Solution (Gibco)	

*Cambridge Isotope Laboratories, **Contains zinc sulfate (0.1 g/l, Sigma Aldrich), manganese (II) sulfate (0.8 g/l, Sigma Aldrich), iron(II)sulfate (0.5 g/l, Sigma Aldrich), copper(II) sulfate (0.1 g/l, Sigma Aldrich) and calcium chloride (1 g/l, Sigma Aldrich) in mH_2O. ***Minimum Essential Medium.

Buffers and loading dye prepared

The resuspension buffer, WEB, and protein loading dye were made with mH_2O, the NMR buffer was made with a 90:10 dilution of H_2O to D_2O (deuterium oxide, CIL), while the rest were made with WEB. The pH of all buffers were adjusted with HCl (Chiron) and sodium hydroxide, before being sterile filtrated using Stericup™ Millipore Express™PLUS 0.22 µm filters.

Table 3.5: The compositions of the protein loading dye (PLD) and different buffers used for protein expression and purification; Resuspension buffer (Res), Wash and equilibration buffer (WEB), Regeneration buffer (Reg), Elution buffer (Elu), Wash buffer (Wash), and NMR buffer (NMR).

Res:	25 mM	Monosodium phosphate (Merck)	**WEB:**	25 mM	Monosodium phosphate
pH 7.0	25 mM	Disodium phosphate	pH 7.0	25 mM	Disodium phosphate
	50 mM	Sodium chloride		300 mM	Sodium chloride
	0.05 %	Triton X-100 (Sigma Aldrich)			
Reg:	500 mM	Imidazole (Merck)	**Elu:**	200 mM	Imidazole
pH 7.0			pH 7.0		
Wash:	2/5/10/15/20 mM	Imidazole	**NMR:**	50 mM	Tris
pH 7.0			pH 7.5	50 mM	Sodium chloride
PLD:	715 mM	BME* (Sigma Aldrich)			
	50 mM	Tris-HCl			
	0.25 %	Bromophenol blue			
	10 %	SDS (Sigma Aldrich)			
	50 %	Glycerol			

* 2-Mercaptoethanol

Expression test

Preculture (1 %) of transformed *E. coli* BL21 clones were inoculated into 25 ml of LB medium. The two sets of culture for each target gene were incubated at 37 °C for 3 hours before IPTG (100 µg/ml, VWR) were added to one of the two sets, in order to induce gene expression. The induced and non-induced clones were then incubated for 3 more hours before being harvested, which was done by centrifuging 1 ml of culture (5,000 rpm for 5 minutes) and discarding the supernatant. The cell pellets were resuspended in NMR buffer (100 µl) before being sonicated (Branson sonifier 250) in 1.5 ml Eppendorf tubes placed in ice water, using a Double step Microtip, an output control of 3.5, and a duty cycle of 40 % for 1 minute. The lysed cell suspensions were centrifuged (13,200 rpm for 30 minutes) and the supernatant were then carefully transferred to a separate tube using a pipette. The sonication lysate and cell debris pellet from induced and non-induced cells, were then analyzed for target proteins.

Expression of proteins

500 ml of 2xLB or M9 medium (Aalborg) containing amp (100 µg/ml) were inoculated with preculture (1 %) of Bl21 clones, before incubation at 37 °C until OD_{600} ~0.6-0.8. The culture was then incubated on ice for 5 minutes before it was induced with IPTG (100 µg/ml) and incubated at 16 °C overnight. The cells were harvested at 4 °C by centrifugation for 5 minutes

at 5500 rpm (Sorvall, SLA-1500), the supernatant was discarded and the pellet was stored at either 4 °C or -20 °C (if not used directly).

The cells were lysed by sonication, where the cells were first resuspended on ice in ~10-15 ml of Resuspension buffer with a ½ tablet of protease inhibitor complex with EDTA (Roche) before being sonicated. The sonication was performed in a 50 ml tube placed in ice water using a Step Horn together with a 3 mm Tapered Microtip, an output control of 3.5, and a duty cycle of 35 % for 10 minutes. The lysate was then centrifuged at 14,000 rpm (~23,000 x G at 4 °C) for 45 minutes (Sorvall, SS-34 rotor), and the supernatant transferred to a 50 ml tube and stored at 4 °C. The protein content from sonication lysate and sonication cell debris pellet were then analyzed for target proteins.

His-tag isolation

Lysate containing His-tagged target proteins were filtered through a 0.22 μm Steriflip™ filter (EMD Millipore) using a syringe (10 ml). A 20 ml gravity flow column (Bio-Rad) with cross-section was set up with 2 ml PerfectPro Ni-NTA Superflow resin (VWR). The column was equilibrated by washing it with 40-50 ml of WEB, which was discarded before the lysate was applied to it and incubated for 20 minutes (flipped every 5 minutes). The lysate flow-through was discharged into a 50 ml tube and the resin was washed with 10 ml of WEB collected in the same tube.

The resin was then washed in two steps with wash buffer (20 mM imidazole), the first with 10 ml and the other with 5 ml, collected in separate tubes. The proteins were also eluted in two steps; one with 10 ml elution buffer (200 mM imidazole) that was discharged in one tube, and the second with 5 ml elution buffer that was discharged into a second tube before 2 ml of regeneration buffer was used and collected in same tube. The proteins were at all times kept at low temperatures (around 4 °C).

The column was regenerated using 10-15 ml of regeneration buffer that was discarded before it was washed 2-3 times with 40-50 ml of WEB. When stored (<1 week) the column was flushed with 20-40 ml of mH_2O and stored with 10-20 ml of mH_2O. Optimized versions of the protocol were made using 2, 3, 5, 10 and 20 mM imidazole in the wash steps, and target proteins were retrieved from high imidazole containing buffer by dialysis using Spectra/Por® 1 dialysis membrane (6-8 molecular weight cut-off – MWCO, Spectrum Labs) in WEB (20-40 ml buffer per ml protein sample) over night. The protein content in each collected flow-through was then analyzed by SDS-PAGE.

Analysis of protein content

The protein expression and purification was analyzed by running up to 10 µl of samples on SDS-PAGE. Loading dye (equal amount as supernatant to supernatant samples and 100 µl to pellets) was applied to the different samples, vortexed and centrifuged (14,500 rpm for 60 seconds, Eppendorf MiniSpisPlus 5453 with F-45-12-11 rotor) before they were incubated at 95 °C for 5 minutes. Samples with cell debris were then centrifuged for 10 minutes at 14,500 rpm (same centrifuge) before 10-20 µl of the top liquid was applied to a ClearPAGE™ SDS-Gel 12 % (12 or 17 wells, CBS Scientific).

Samples taken from supernatants were centrifuged down shortly (same centrifuge and speed) before 5-20 µl was added to the gel. Precision Plus Protein™ All blue standards (BioRad) (6-7 µl) was used as protein ladder and the gel was run in ClearPAGE™ SDS Run Buffer Non-reducing (CBS Scientific) at 130 V for 75-85 minutes to separate the proteins. The protein bands were visualized by incubating the gel in InstantBlue™ (Expedeon) for 1 hour and washed thoroughly with water.

3.5 Protein characterization

Buffers prepared

The buffers were made with a 90:10 dilution of H_2O to D_2O. All buffers were pH adjusted with hydrogen chloride and sodium hydroxide, before being sterile filtrated using Stericup™ Millipore Express™PLUS 0.22 µm filters.

Table 3.6: The composition of the different NMR buffers (pH 5.0 and 6.9) used for NMR spectroscopy.

NMR:	25 mM	Disodium phosphate	**NMR:**	20 mM	HEPES* (VWR)
pH 5.0	10/25 mM	Sodium chloride	pH 6.9	40 mM	Sodium chloride

* 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid

Thermostability analysis and thermo-exclusion purification

The thermo stability of target proteins was analyzed by incubating lysate or purified samples (200 µl) at 60 and 80 °C at different time intervals. Parallels were made for each time interval and each sample were placed on ice after the heat treatment and centrifuged at 13,200 rpm for 5 minutes to spin down the aggregated proteins.

Isotope labeled proteins without His-tag were purified by incubation of the lysate (~10-15 ml) at 80 °C for 3 hrs. The aggregated non-thermostable proteins were centrifuged down at 5,000 rpm for 30 minutes in 4 °C (Eppendorf Centrifuge 5430R, F-35-6-30 rotor). The supernatant was carefully transferred to a syringe (10 ml) and filtered through a Steriflip™ into a new sterile tube (50 ml). The protein content in the sample was analyzed by SDS-PAGE and the purification process was repeated until accepted purity was acquired.

NMR analysis

Prior to analysis the target proteins were transferred and concentrated in an NMR buffer with pH 5.0 or 6.9 (25 mM phosphate buffer, or 20 mM HEPES buffer respectively). This was achieved by using VivaSpin 6 spin columns (10 kDa MWCO, GE healthcare) with centrifugation (5,000 rpm for up to 2 hrs at a time) until reaching a volume of 400-500 µl. All recording processes were performed by Finn L. Aachmann, where the NMR spectra were recorded at 298 K on a Burker Advance 600 MHz spectrometer equipped with 5 mm Z-gradient CP-TCI(H/C/N) probe using BRUKER TopSpin 2.1 software.

4 Results

The protein sequences of the enzyme candidates studied in this thesis were found by data mining a translated metagenomic sequence database of annotated contigs, created from polyextremophile microbes in oil reservoirs. The target genes were modified with a His-tag and heterologously expressed in *E. coli* K12 strains before their structure and thermostability was analyzed with NMR spectroscopy and thermo-exclusion study respectively.

4.1 Candidate selection

Enzyme classes to be used for further study were chosen by searching the Swiss-Prot database, containing over 500,000 annotated and reviewed proteins. The advanced search (Appendix A) narrowed the proteins down to approximately 1,200 hits which were analyzed manually to find interesting candidates among them. The amino acid sequences of about 200 proteins were then used as queries to find similar proteins amongst 52,415 contigs in the translated metagenomic sequence library (Figure 4.1).

Approximately 20 of the queries were found to have similar proteins encoded in the metagenomic sequence library, which showed to have 5 to 100 hits each, providing over 1,000 potential candidates. To limit the number of candidates, only 1 to 15 of the best aligned sequences (alignment score >40) were chosen for further study, providing approximately 150 possible candidates.

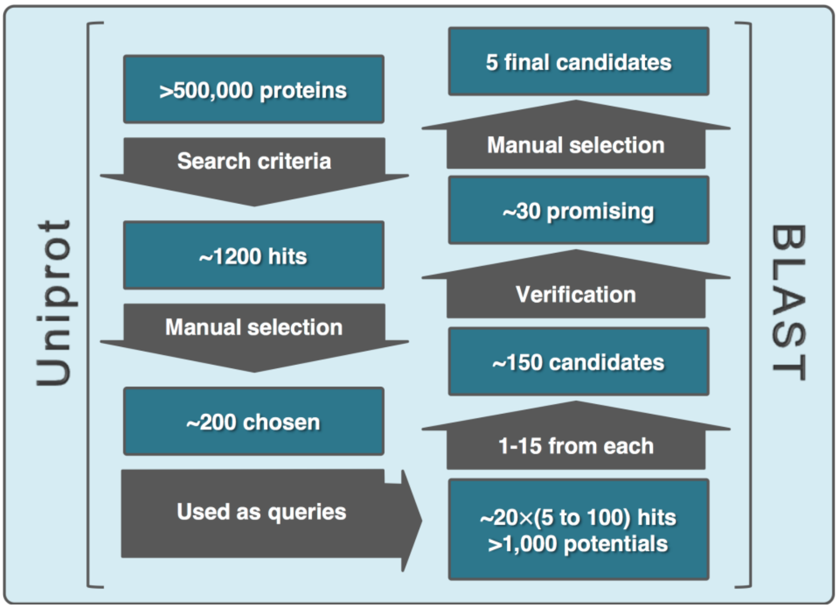


Figure 4.1: A summary of the results (rectangles) after each step (arrows) applied during the data mining, including which program produced the different results.

The contigs of these 150 candidates were analyzed with ORF finder to locate the translated genes which were then used as queries for a BLAST search in the nr database. This was done to make sure the proteins were of the right class and resembles known proteins with full length. This step provided approximately 30 promising candidates which were used to manually select the final 5 candidates, where three were predicted to be arsenate reductases and two glyoxalases. Finally each of the five candidates were aligned with proteins in the PDB database with BLAST, to establish which of the well studied ArsC and GlxI groups they resemble the most.

4.1.1 Arsenate reductase

Thirteen potential arsenate reductase proteins were located in the translated metagenomic sequence library by using the 129 amino acids long sequence of arsenate-mycothiol transferase ArsC2 of *C. glutamicum* (UniProt ID P0DKS7) as a query. Of the thirteen, three final arsenate reductase candidates, named Ars1, Ars2 and Ars3, showed to have an analogy to the query (Appendix B), and have the qualities needed for NMR analysis.

Results from the analysis in ORF finder and BLAST alignment in the nr database (Appendix B) confirmed the lengths of candidate Ars2 and Ars3 but showed that candidate Ars1 probably was missing a small part of the C-terminal region. However, due to a high similarity (99 % identity to the full sequence) to an ArsC protein from *Thermococcus litoralis*, the C-terminal sequence of this protein was used to substitute the missing amino acid sequence on the Ars1 candidate (Figure 4.2, boxed in green). With that in mind, candidate Ars1 might have the wrong sequence and thus might provide flawed results (e.g. no activity and/or wrong structure). Of the PDB registered proteins, Ars1 showed highest resemblance to the homodimeric Trx-dependent ArsC from *C. glutamicum* (PDB ID 3T38_A) with 34 % identity to 96 % of the query. Candidate Ars2 and Ars3 were shown to resemble the *B. subtilis* Trx-coupled ArsC (PDB ID 1JL3_A) with 49 and 50 % identity to 93 and 95 % of the query respectively.

An multiple sequence alignment (Figure 4.2) was used to get an idea of which class of ArsC enzymes the candidates might belong to, and to show their similarity to the PDB search results. The alignment was made from protein sequences of each class (Trx, Mrx and Grx[5]) verified by experiments on protein level and from the PDB search results, together with the

[5] Only the Grx with PTPase fold and activity was used since the other Grx-ArsC has a sequence identity of <20 to the other ArsC groups.

three candidate sequences and the UniProt query used to find them. The annotations of the ArsC and PTPase active sites, and the metal-binding site are based on studies of the *S. aureus* ArsC (Lah et al., 2003, Zegers et al., 2001, Su et al., 1994, Messens et al., 2002, Messens et al., 1999), the residues believed to yield metal-binding independency are based on studies from *B. subtilis* (Roos et al., 2006) and *C. glutamicum* ArsC (Villadangos et al., 2011), while the nucleophilic cysteines are from studies done on each separate protein (Bennett et al., 2001, Li et al., 2003, Villadangos et al., 2011, Ordóñez et al., 2009).

The results showed that the metal-binding residues are only partially conserved among most of the ArsC groups (Figure 4.2, boxed in blue and annotated with ▲). However, the residues thought to be important in yielding metal-binding independency (Figure 4.2, boxed in magenta and annotated with ▲) align well in the different groups of ArsC, meaning that all the proteins (except *S. aureus* and candidate Ars3) has at least one of the two alterations (His-62→Gln and Asp-65→Lys relative to *S. aureus*) found in proteins that are independent of metal-binding. Conversely, candidate Ars3 does not have the fully conserved metal-binding site from *S. aureus*.

The PTPase active site aligns well in all the groups, except for the Mrx-ArsC from *C. glutamicum*. The ArsC active site however, does not align very well between the different groups. First, the Arg-16 of *S. aureus* is part of both the PTPase and ArsC active site and thus, the alteration in the Mrx-ArsC apply for the ArsC active site as well. Second, the nucleophilic cysteines (Figure 4.2, highlighted in blue) in the C-terminal do not align between the different groups and might be used to separate the groups from each other (further studies are needed to confirm this). The cysteines of candidate Ars2 and Ars3 are aligned with the monomeric Trx-ArsC from *S. aureus* and *B. subtilis* (Figure 4.2, highlighted in blue and annotated with ▲). As for the cysteines of candidate Ars1 are closer in resemblance to the homodimeric Trx-ArsC from *C. glutamicum* though they are not aligned, indicating that the multiple alignment could be improved by introducing two additional gaps and removing another (Figure 4.2, boxed in cyan). This fault in alignment might be because of a high gap penalty.

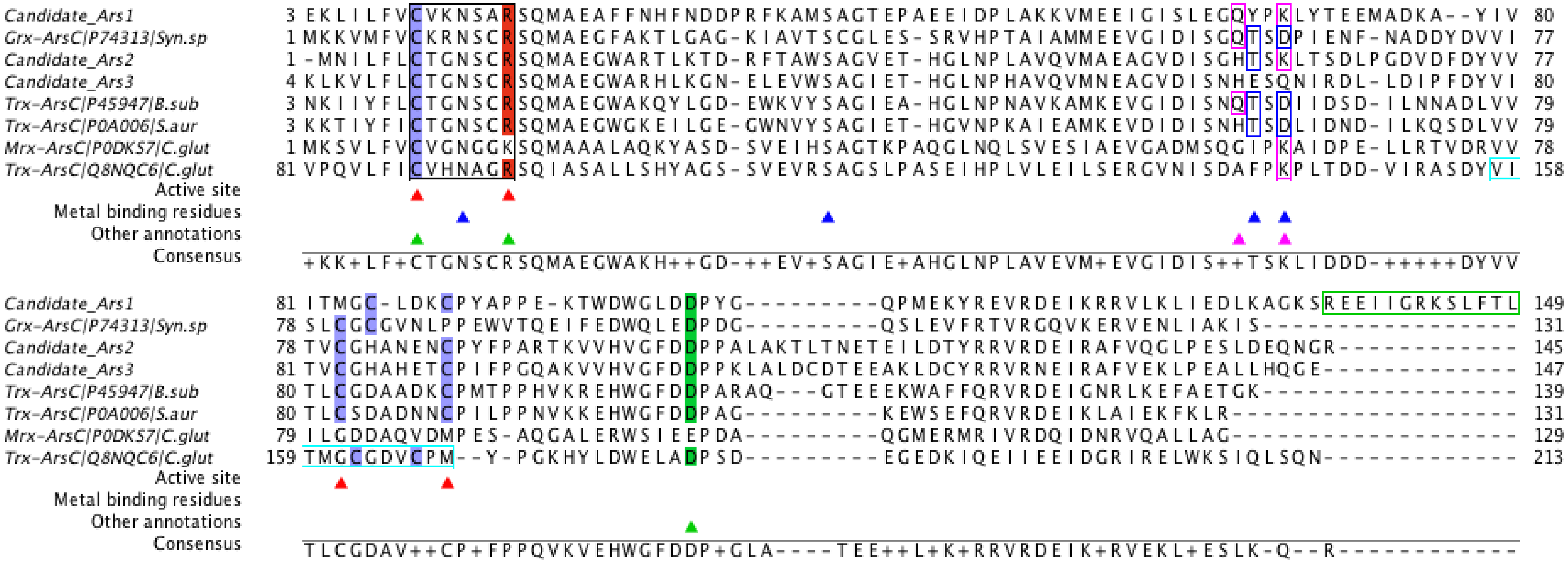

Candidate_Ars1 3 EKLILFVCVKNSARSQMAEAFFNHFNDDPRFKAMSAGTEPAEEIDPLAKKVMEEIGISLEGQYPKLYTEEMADKA--YIV 80
Grx-ArsC|P74313|Syn.sp 1 MKKVMFVCKRNSCRSQMAEGFAKTLGAG-KIAVTSCGLES-SRVHPTAIAMMEEVGIDISGQTSDPIENF-NADDYDVVI 77
Candidate_Ars2 1 -MNILFLCTGNSCRSQMAEGWARTLKTD-RFTAWSAGVET-HGLNPLAVQVMAEAGVDISGHTSKLTSDLPGDVDFDYVV 77
Candidate_Ars3 4 KLKVLFLCTGNSCRSQMAEGWARHLKGN-ELEVWSAGIET-HGLNPHAVQVMNEAGVDISNHESQNIRDL-LDIPFDYVI 80
Trx-ArsC|P45947|B.sub 3 NKIIYFLCTGNSCRSQMAEGWAKQYLGD-EWKVYSAGIEA-HGLNPNAVKAMKEVGIDISNQTSDIIDSD-ILNNADLVV 79
Trx-ArsC|P0A006|S.aur 3 KKTIYFICTGNSCRSQMAEGWGKEILGE-GWNVYSAGIET-HGVNPKAIEAMKEVDIDISNHTSDLIDND-ILKQSDLVV 79
Mrx-ArsC|P0DKS7|C.glut 1 MKSVLFVCVGNGGKSQMAAALAQKYASD-SVEIHSAGTKPAQGLNQLSVESIAEVGADMSQGIPKAIDPE-LLRTVDRVV 78
Trx-ArsC|Q8NQC6|C.glut 81 VPQVLFICVHNAGRSQIASALLSHYAGS-SVEVRSAGSLPASEIHPLVLEILSERGVNISDAFPKPLTDD-VIRASDYVI 158
Active site
Metal binding residues
Other annotations
Consensus +KK+LF+CTGNSCRSQMAEGWAKH++GD-++EV+SAGIE+AHGLNPLAVEVM+EVGIDIS++TSKLIDDD-+++++DYVV
Candidate_Ars1 81 ITMGC-LDKCPYAPPE-KTWDWGLDDPYG---------QPMEKYREVRDEIKRRVLKLIEDLKAGKSREEIIGRKSLFTL 149
Grx-ArsC|P74313|Syn.sp 78 SLCGCGVNLPPEWVTQEIFEDWQLEDPDG---------QSLEVFRTVRGQVKERVENLIAKIS----------------- 131
Candidate_Ars2 78 TVCGHANENCPYFPARTKVVHVGFDDPPALAKTLTNETEILDTYRRVRDEIRAFVQGLPESLDEQNGR------------ 145
Candidate_Ars3 81 TVCGHAHETCPIFPGQAKVVHVGFDDPPKLALDCDTEEAKLDCYRRVRNEIRAFVEKLPEALLHQGE------------- 147
Trx-ArsC|P45947|B.sub 80 TLCGDAADKCPMTPPHVKREHWGFDDPARAQ---GTEEEKWAFFQRVRDEIGNRLKEFAETGK----------------- 139
Trx-ArsC|P0A006|S.aur 80 TLCSDADNNCPILPPNVKKEHWGFDDPAG---------KEWSEFQRVRDEIKLAIEKFKLR------------------- 131
Mrx-ArsC|P0DKS7|C.glut 79 ILGDDAQVDMPES-AQGALERWSIEEPDA---------QGMERMRIVRDQIDNRVQALLAG------------------- 129
Trx-ArsC|Q8NQC6|C.glut 159 TMGCGDVCPM--Y-PGKHYLDWELADPSD---------EGEDKIQEIIEEIDGRIRELWKSIQLSQN------------- 213
Active site
Metal binding residues
Other annotations
Consensus TLCGDAV++CP+FPPQVKVEHWGFDDP+GLA----TEE++L+K+RRVRDEIK+RVEKL+ESLK-Q--R------------

Figure 4.2: An annotated multiple sequence alignment made of selected verified ArsC enzymes together with the three putative ArsC candidates (Ars1, Ars2 and Ars3) and the Mrx-ArsC query from *C. glutamicum* (UniProt ID P0DKS7) used to find the candidates. The multiple alignment is made using Clustal Omega and the annotation is done in Jalview 2.8.2. The selected ArsC sequences are from *B. subtilis* (UniProt ID P45947) – Trx-dependent with the highest identity to both candidate Ars2 and Ars3 of the PDB registered proteins, *C. glutamicum* (UniProt ID Q8NQC6) – homodimeric Trx-dependent with the highest resemblance to candidate Ars1 in the PDB database, *S. aureus* (UniProt ID P0A006) – Trx-dependent which uses K^+ as a cofactor, and *Synechocystis* sp. (UniProt ID P74313) – Grx-dependent with a PTPase fold and activity. The first part of the alignment was removed since most of this part only contains information regarding the Trx-ArsC from *C. glutamicum*.

Residues important for activity and for metal-binding (as seen in the *S. aureus* ArsC) are annotated with ▲ (highlighted in red where mismatches occur), and ▲ (boxed in blue where mismatches occur) respectively. The nucleophilic cysteines of each ArsC enzyme and possible nucleophilic cysteines of each candidate are highlighted in blue. Residues believed to be important for PTPase activity are annotated with ▲ where the preserved residues are highlighted in red or green when mismatches occur. Residues believed to yield metal-binding independency are annotated with ▲ and boxed with magenta. The characteristic P-loop recognized from PTPases is boxed in black. A consensus sequence is shown underneath the alignment. The sequence of candidate Ars1 boxed in green is added from the *T. litoralis* ArsC. The sequence of the Trx-ArsC from *C. glutamicum* boxed in cyan can be aligned with the Ars1 candidate by moving the sequence one step to the right and introducing two gaps (one gap in front of the sequence and the other between the two cysteines).

4.1.2 Glyoxalase 1

Six potential glyoxalase 1 proteins were found to be encoded in the translated metagenomic sequence library by using the 135 amino acid long sequence of lactoylglutathione lyase 1 from *Haemophilus influenzae* (UniProt ID P44638) as a query. Two turned out to be less than 20 kDa in size with homology to the query (Appendix B), thus both were used as the final glyoxalase candidates named Glx1 and Glx2. The analysis in ORF finder and BLAST alignments in the nr database (Appendix B) showed that both Glx1 and Glx2 matched the lengths of other similar GlxI enzymes, indicating that they are of full length. Of the PDB registered proteins, candidate Glx1 showed most resemblance to the Ni^{2+} binding GlxI from *E. coli* (PDB ID 1F9Z_A) with 57 % identity to 94 % of the query, while Glx2 showed a greater resemblance to a GlxI from *Pseudomonas aeruginosa* PAO1 (PDB ID 4MTQ_A) with 34 % identity to 96 % of the query.

An annotated multiple sequence alignment was made to estimate which group of GlxI the candidates might belong to and to show their resemblance to the PDB registered GlxI (Figure 4.3). Sequences of verified proteins (from experiments on protein level) from the Ni^{2+}- and Zn^{2+}-dependent GlxI group were used to make the alignment, including examples of the back-to-back and the βαβββ connected dimers that exist for the Ni^{2+}-GlxI. In addition, sequences of the proteins from the PDB database with the highest resemblance to the candidates were applied in the alignment together with the two candidates and the *H. influenzae* query that were used to find the candidates. The annotations were applied from the studies done on the βαβββ-motif bound *E. coli* Ni^{2+}-GlxI (He et al., 2000, Suttisansanee and Honek, 2011).

The results showed that all the proteins had fully conserved metal-binding sites (Figure 4.3, annotated with ▲). However, the active site is not well conserved between the three different types of GlxI (annotated with @ and ▲), suggesting they depend on other residues for the active site in Zn^{2+}-GlxI and back-to-back oriented Ni^{2+}-GlxI. Candidate Glx1 has almost a fully conserved active site (amino acid wise) as found in the *E. coli* GlxI. The exception is Val-103 in *E. coli* which is replaced by Met-110 in candidate Glx1, where both residues contain a hydrophobic side chain. Candidate Glx2 on the other hand, only possess two of the conserved active site residues, though a somewhat better alignment is achieved if parts of the aligned sequence is manually altered (Figure 4.3, boxed in cyan). The remaining sites for the active site are replaced by unrelated residues (not with the same properties as the conserved ones) especially in the C-terminal domain, indicating that this protein might also depend on other residues for activity than the GlxI from *E. coli*.

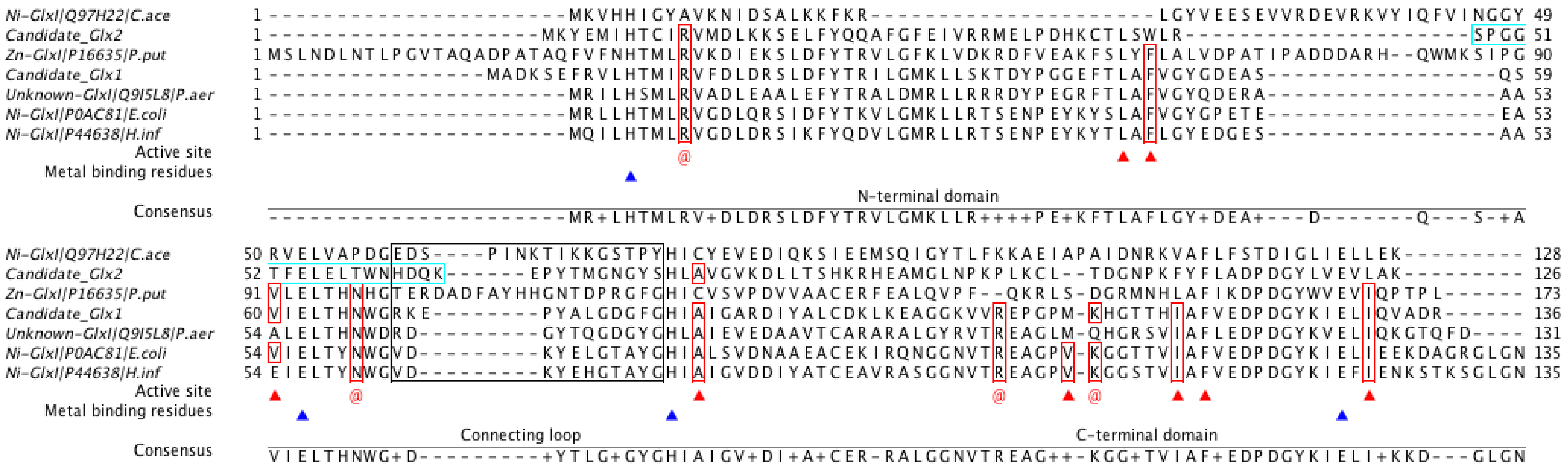

Ni-GlxI/Q97H22/C.ace 1 ----------------------MKVHHIGYAVKNIDSALKKFKR---------------------LGYVEESEVVRDEVRKVYIQFVINGGY 49
Candidate_Glx2 1 --------------------MKYEMIHTCIRVMDLKKSELFYQQAFGFEIVRRMELPDHKCTLSWLR---------------------SPGG 51
Zn-GlxI/P16635/P.put 1 MSLNDLNTLPGVTAQADPATAQFVFNHTMLRVKDIEKSLDFYTRVLGFKLVDKRDFVEAKFSLYFLALVDPATIPADDDARH--QWMKSIPG 90
Candidate_Glx1 1 ----------------MADKSEFRVLHTMIRVFDLDRSLDFYTRILGMKLLSKTDYPGGEFTLAFVGYGDEAS-----------------QS 59
Unknown-GlxI/Q9I5L8/P.aer 1 ----------------------MRILHSMLRVADLEAALEFYTRALDMRLLRRRDYPEGRFTLAFVGYQDERA-----------------AA 53
Ni-GlxI/P0AC81/E.coli 1 ----------------------MRLLHTMLRVGDLQRSIDFYTKVLGMKLLRTSENPEYKYSLAFVGYGPETE-----------------EA 53
Ni-GlxI/P44638/H.inf 1 ----------------------MQILHTMLRVGDLDRSIKFYQDVLGMRLLRTSENPEYKYTLAFLGYEDGES-----------------AA 53
Active site
@
Metal binding residues
N-terminal domain
Consensus ----------------------MR+LHTMLRV+DLDRSLDFYTRVLGMKLLR++++PE+KFTLAFLGY+DEA+---D--------Q---S-+A
Ni-GlxI/Q97H22/C.ace 50 RVELVAPDGEDS----PINKTIKKGSTPYHICYEVEDIQKSIEEMSQIGYTLFKKAEIAPAIDNRKVAFLFSTDIGLIELLEK--------- 128
Candidate_Glx2 52 TFELELTWNHDQK------EPYTMGNGYSHLAVGVKDLLTSHKRHEAMGLNPKPLKCL--TDGNPKFYFLADPDGYLVEVLAK--------- 126
Zn-GlxI/P16635/P.put 91 VLELTHNHGTERDADFAYHHGNTDPRGFGHICVSVPDVVAACERFEALQVPF--QKRLS-DGRMNHLAFIKDPDGYWVEVIQPTPL------ 173
Candidate_Glx1 60 VIELTHNWGRKE--------PYALGDGFGHIAIGARDIYALCDKLKEAGGKVVREPGPM-KHGTTHIAFVEDPDGYKIELIQVADR------ 136
Unknown-GlxI/Q9I5L8/P.aer 54 ALELTHNWDRD---------GYTQGDGYGHLAIEVEDAAVTCARARALGYRVTREAGLM-QHGRSVIAFLEDPDGYKVELIQKGTQFD---- 131
Ni-GlxI/P0AC81/E.coli 54 VIELTYNWGVD---------KYELGTAYGHIALSVDNAAEACEKIRQNGGNVTREAGPV-KGGTTVIAFVEDPDGYKIELIEEKDAGRGLGN 135
Ni-GlxI/P44638/H.inf 54 EIELTYNWGVD---------KYEHGTAYGHIAIGVDDIYATCEAVRASGGNVTREAGPV-KGGSTVIAFVEDPDGYKIEFIENKSTKSGLGN 135
Active site
@ @ @
Metal binding residues
Connecting loop
C-terminal domain
Consensus VIELTHNWG+D---------+YTLG+GYGHIAIGV+DI+A+CER-RALGGNVTREAG++-KGG+TVIAF+EDPDGYKIELI+KKD---GLGN

Figure 4.3: An annotated multiple sequence alignment of selected verified GlxI proteins together with the GlxI query from *H. influenzae* (UniProt ID P44638) used, and the two putative GlxI candidates, Glx1 and Glx2. The multiple alignment is made using Clustal Omega and the annotation is done in Jalview 2.8.2. The selected GlxI sequences are from *C. acetobutylicum* (UniProt ID Q97H22) – Ni^{2+} dependent with the back-to-back dimer conformation, *E. coli* (UniProt ID P0AC81) – Ni^{2+} dependent with the βαβββ domain binding conformation which were found to be the PDB registered GlxI with the highest resemblance to candidate Glx1, *P. aeruginosa* (UniProt ID Q9I5L8) – Unknown cofactor and conformation which had the highest identity to candidate Glx2 of the PDB registered GlxI enzymes, and *Pseudomonas putida* (UniProt ID P16635) – Zn^{2+} dependent with βαβββ domain binding conformation.

The active site is annotated with @ and ▲ showing the residues that constitute the hemithioacetal binding site and the catalytic pocket respectively, as found in the Ni^{2+}-GlxI from *E. coli* (preserved residues are marked with red boxes where mismatches occur). The metal binding residues are annotated with ▲. In addition, the flexible connective loop in between the N- and C-terminal domains is annotated with a black box, and a consensus sequence is shown underneath the alignment. The sequence from candidate Glx2 that is boxed in cyan can be moved two steps to the left in order for the Asn-60 to align with the active site.

4.2 Vector construction

4.2.1 Cloning vector

A “genetic framework” (i.e. additional nucleotides flanking the gene) was designed around the target genes (Appendix C). This provides the genes with the stop codon TGA flanked by restriction sites for BclI and BglII, and the code for a 6*His-tag downstream of the stop codon, with a TAA stop codon on its end. The BclI and BglII restriction sites provide the translated proteins with one to two additional amino acids (Arg and/or Ser) in the C-terminal end, but allow the option of expressing the gene in frame with a His-tag (due to the sequence compatibility) by removing the stop codon. In addition, the genetic framework contains restriction sites for NdeI and XhoI flanking the inserted gene with protective nucleotides bordering the framework. The target gene sequences were codon-optimized for expression in *E. coli*, chemically synthesized, and inserted into an EcoRV restriction site in cloning vector pUC57 (by GenScript).

4.2.2 His-tag modification

For the His-tag modification (i.e. bringing the His-tag in frame with the target gene, Figure 4.4, top panel) the cloning vector of each plasmid (except for Ars1) was amplified in dam$^-$/dcm$^-$ competent *E. coli* to obtain unmethylated BclI restriction sites. Afterwards, the cloning vectors were modified by double digestion using BclI and BglII. The digestion was expected to generate two fragments, one at about 3,000 bp and the other at 6 bp (which would not show on the gel). If partial digestion, these plasmids were expected to show the same size as the fully digested plasmids since they only differ with six bp in length. The results for all candidates showed a single band at about 3,000 bp, and the fragments were isolated and purified from the gel before they were re-ligated using T4 ligase.

The modified cloning vectors were then enriched before verification (Figure 4.4, center panel). The enrichment of correctly modified plasmids involved digesting them with BglII to linearize unmodified plasmids, and then transforming them into DH5-α cells. This should increase the number of clones carrying modified plasmids compared to unmodified ones (since the restriction site of BglII is removed in modified plasmids), and thus reduce the total number of colonies. The results showed a decreased number of transformants for all the modified candidates (Table 4.1).

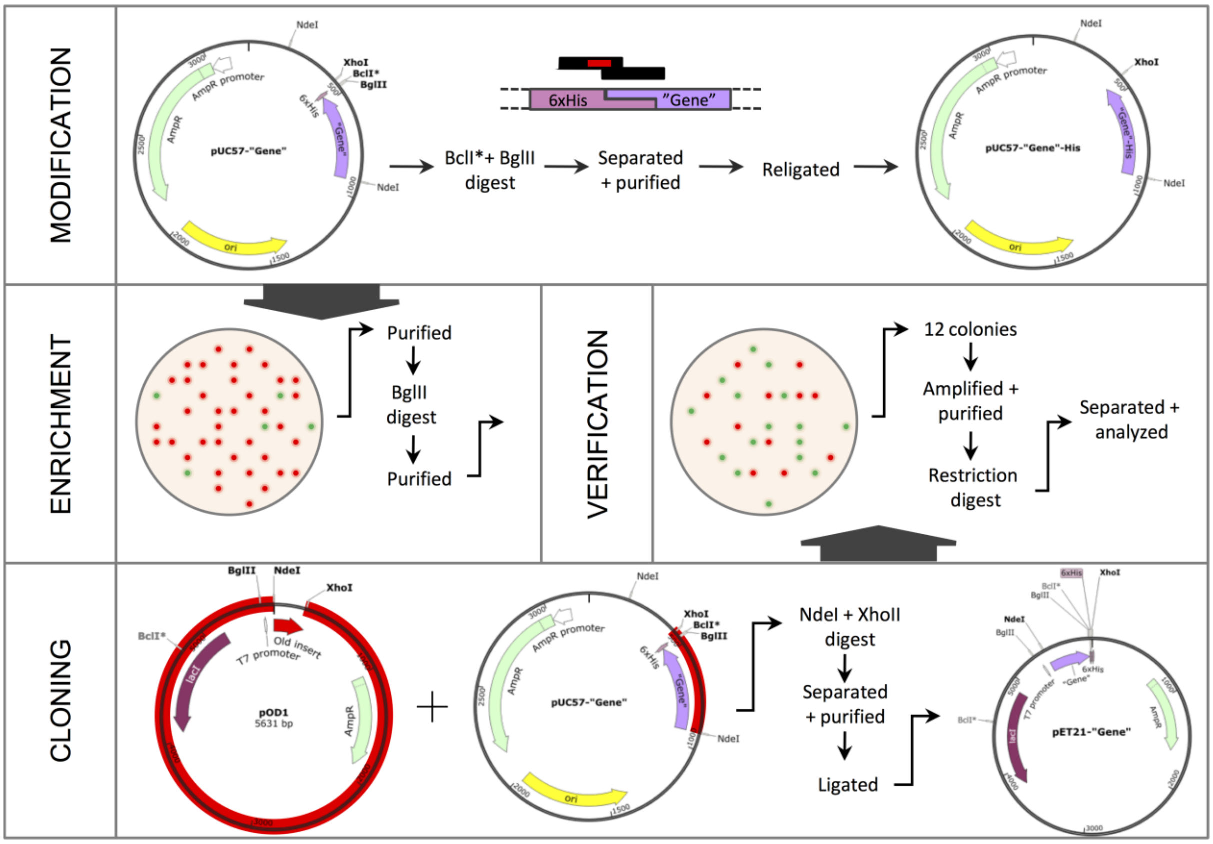


Figure 4.4: An illustration of the genetic framework for the vectors (made in SnapGene®Viewer 2.5), and the various steps involved in modification, cloning and verification of the plasmids.

MODIFICATION: The red mark on the fragment withdrawn from pUC57-"Target gene" illustrates the stop codon TGA, which, when removed allow the 6×His-tag sequence to be expressed in frame with the target gene.

ENRICHMENT AND VERIFICATION: The red and green dots illustrates transformed colonies on agar plates, with unmodified (in red) or correctly modified (in green) plasmids (the amount and proportion of the red and green colonies are not necessarily applicable). Only the His-tag modification required enrichment of the correctly modified clones.

CLONING: The pET21 vector fragment (to the left) and the target gene fragment (shown here as unmodified) are marked in red. The digest of pOD1 and pUC57-"Target gene" was done in separate reaction mixtures. The direction of the "Target gene" and 6*His-tag in pUC57 is reversed for some of the candidates.

Table 4.1: The number of transformed colonies counted for each of the candidates before and after enrichment of the correctly modified clones. The results are taken from plates plated with 50 µl of transformation mix.

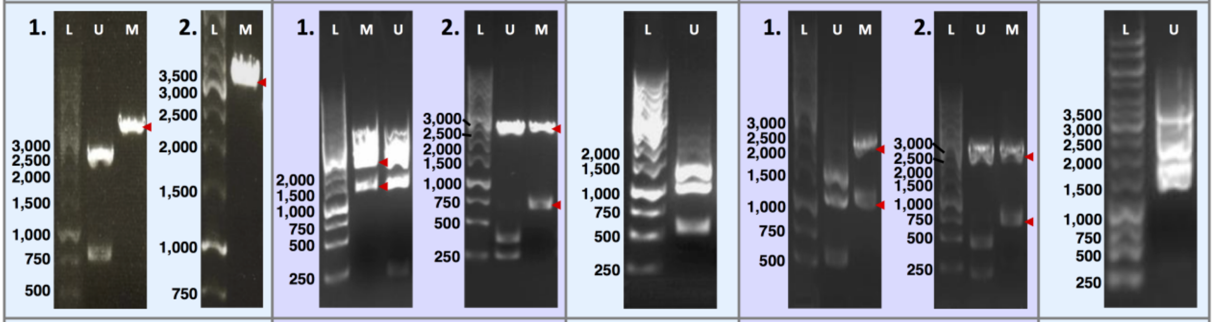
Target vector	Before enrichment	After enrichment
pUC57-Ars1-His	TCTC*	96
pUC57-Ars2-His	TCTC	100
pUC57-Ars3-His	TCTC	24
pUC57-Glx1-His	TCTC	200
pUC57-Glx2-His	TCTC	>200

*Too confluent to count

Constructs of pUC57-target genes with His-tag were verified by restriction enzyme digestion, where the plasmids of twelve clones (for each candidate) were double digested using BglII together with another appropriate endonuclease (Figure 4.5). Two fragments were expected for the unmodified pUC57-Ars1, and three fragments were expected for the remaining four unmodified candidate vectors (See Figure 4.5 for specified lengths), because a BglII restriction site would still be in place. As for the modified plasmids, the BglII restriction site is removed, hence digestion is expected to result in one less fragment compared to their unmodified counterpart, indicating that the TGA stop codon was removed and the His-tag is in frame with the target gene. Unclear results were verified a second time, either with the same enzymes or using a new combination.

The gel analysis showed that at least one colony for three candidates (Ars1, Ars2 and Glx1) indicated to contain His-tagged target genes (Figure 4.5, panel A.1, B.1 and D.1 - red cursors mark correct fragments), which were verified by a second digestion (Figure 4.5, panel A.2, B.2, and D.2). All constructs tested for the Ars3 candidate showed the digestion pattern of unmodified plasmids, while the ones for the Glx2 candidate showed ambiguous results that were interpreted as star activity at the time (Figure 4.5, panel C and E respectively). The constructs from Ars3 and Glx2 were tested again with similar results (Appendix E, Figure E.4 and E.8 respectively). This resulted in eight target genes; five without a His-tag and three with a His-tag, which were proceeded with and cloned into the pET21 expression vector.

A.		B.		C.		D.		E.	
pUC57-Ars1		pUC57-Ars2		pUC57-Ars3		pUC57-Glx1		pUC57-Glx2	



	PciI		BglI (1)		BglI		BglI (1)		BglI
M	3,203	M	1,118 + 2,070	M	1,118 + 2,075	M	1,118 + 2,039	M	1,118 + 2,016
U	859 + 2,350	U	219 + 1,118 + 1,857			U	590 + 1,118 + 1,455		
			NdeI (2)	U	626 + 1,118 + 1,455		PvuII (2)	U	219 + 1,118 + 1,803
		M	716 + 2,472			M	793 + 2,364		
		U	286 + 436 + 2,472			U	266 + 533 + 2,364		

Figure 4.5: Selected results from His-tag verification for each candidate, with the fragments expected for modified (M) and unmodified (U) candidates when digested with BglII and their respective endonucleases (listed underneath each specific candidate). One modified and/or one unmodified digestion pattern is shown for each candidate. The fragments of a successfully modified candidate are emphasized by red cursors. The specified values for the O'GeneRuler™ 1 kb DNA standard (L) are given in bp. The full gel pictures showing all twelve selections of each target gene can be found in the appendix E.

4.2.3 Expression vector

The cloning (Figure 4.4, base panel) was done by double digestion of pOD1 and each of the pUC57- target gene constructs, using NdeI and XhoI. The digestion pattern of pOD1 was expected to show two DNA fragments, the *SelW* gene fragment (266 bp) and the pET21 vector backbone (5,365 bp) when run in agarose gel. Each pUC57-target gene construct was expected to show three DNA fragments, the target gene fragment (See Table 4.2 for specific lengths) and two pieces of the pUC57 backbone (250 and 2,472 bp). The results showed two bands for pOD1 and three bands of approximately correct lengths for each candidate (Figure 4.6). The target gene fragments (red) and pET21 vector backbone (green) were isolated and purified from the gel, before each target gene was separately ligated into linearized pET21 vectors using T4 DNA ligase.

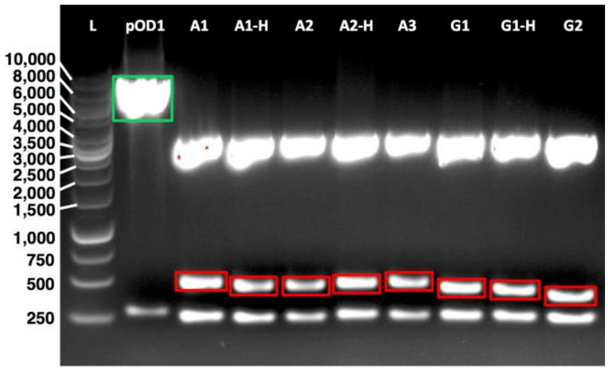


Figure 4.6: The resulting digestion pattern of pOD1 and each candidate cloning vector using NdeI and XhoI. The specified values for the O'GeneRuler™ 1 kb DNA standard (L) are given in bp. The pET21 backbone fragment of pOD1 is boxed in green and the target genes to be inserted into pET21 in red.

Table 4.2: The specific lengths of each target gene fragment excised from pUC57 vectors

Target gene	Length (bp)
Ars1	484
Ars1-His	478
Ars2	469
Ars2-His	463
Ars3	476
Glx1	440
Glx1-His	434
Glx2	415

The constructs of expression vector with target gene were verified by restriction digestion (Figure 4.4, center panel). Twelve selected clones of each construct were digested with appropriate endonucleases (Figure 4.7). A digestion pattern with two or three bands was expected from digestion of the pET21-target gene constructs, while one, two or four bands were expected for the pOD1 re-ligations (for details, see Figure 4.7). Ambiguous results were verified up to three times, either with the same enzymes or using a new combination of enzymes.

Nearly all the Ars1 and Ars1-His clones and less than half of the other six candidates showed the desired digestion pattern (Appendix F). However, some digestion patterns had a displacement of the bands on the gel, where the digested DNA fragments had larger sizes than expected (when compared to the DNA ladder). This has been an ongoing problem on several other projects in the laboratory.

The first verification of the pET21-Ars2 construct showed signs of partial digestion (Figure 4.7, panel C.1), and a second verification was made with a longer incubation time that showed the expected digestion pattern (Figure 4.7, panel C.2). In addition, the two first pET21-Glx1-His construct verifications showed inconclusive results (Appendix F, Figure F.7). However, by comparing the ambiguous gel pictures a "probable consensus" could be derived for some of the clones. Two constructs were selected for a third verification, one construct which was thought to contain the target gene and one that was thought to be a re-ligated pOD1 (to be used as a negative control). The digestion pattern of the third verification confirmed the perception of these constructs (Figure 4.7, panel G), indicating that the Glx1-His target gene was successfully inserted into the pET21 vector.

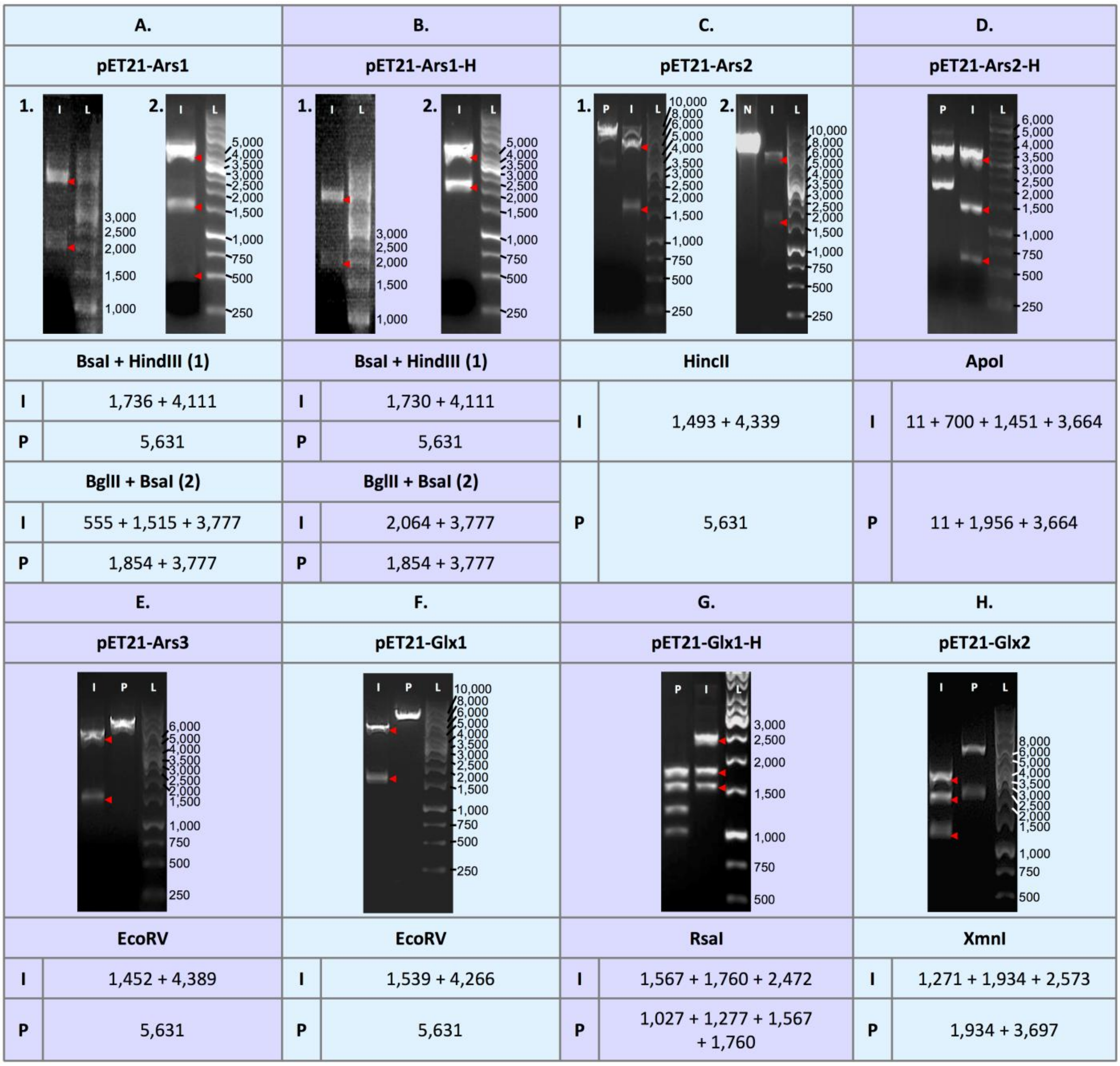
Since the Ars1-His target gene was not amplified in dam$^-$/dcm$^-$ competent *E. coli* cells prior to His-tag modification, the Ars1 and Ars1-His constructs were digested a second time using a different enzyme (BsaI) with BglII (expected to produce three bands for Ars1 and two bands for Ars1-His), to make sure the single band obtained in the first two modifications was not due to partial digestion. The results showed two bands at approximately the right size for the Ars1-His (Figure 4.7, panel B.2) which were clearly distinct from the Ars1 results (Figure 4.7, panel A.2), indicating that the BglII restriction site is removed in Ars1-His and that the first His-tag verifications were correct.

Hence, all eight target genes were successfully inserted into pET21, which were then introduced into *E. coli* BL21 (DE3) cells by transformation. The tagged and untagged target genes of Ars1, Ars2 and Glx1 were chosen for further heterologous expression.

4.3 Heterologous expression and His-tag purification

4.3.1 Expression test of candidates

Expression of the target genes in *E. coli* BL21 (DE3) were tested by comparing IPTG induced cell cultures against non-induced cell cultures (25 ml). This should show an additional band on SDS-PAGE for the induced cell cultures, with a molecular weight that corresponds to the sizes expected for the respective target proteins. In addition, most of the target proteins are preferably shown in the sonication lysate and not in the cell debris pellet, since insoluble proteins are more difficult to process.



A.		B.		C.		D.	
pET21-Ars1		pET21-Ars1-H		pET21-Ars2		pET21-Ars2-H	
BsaI + HindIII (1)		BsaI + HindIII (1)		HincII		ApoI	
I	1,736 + 4,111	I	1,730 + 4,111	I	1,493 + 4,339	I	11 + 700 + 1,451 + 3,664
P	5,631	P	5,631				
BglII + BsaI (2)		BglII + BsaI (2)		P	5,631	P	11 + 1,956 + 3,664
I	555 + 1,515 + 3,777	I	2,064 + 3,777				
P	1,854 + 3,777	P	1,854 + 3,777				

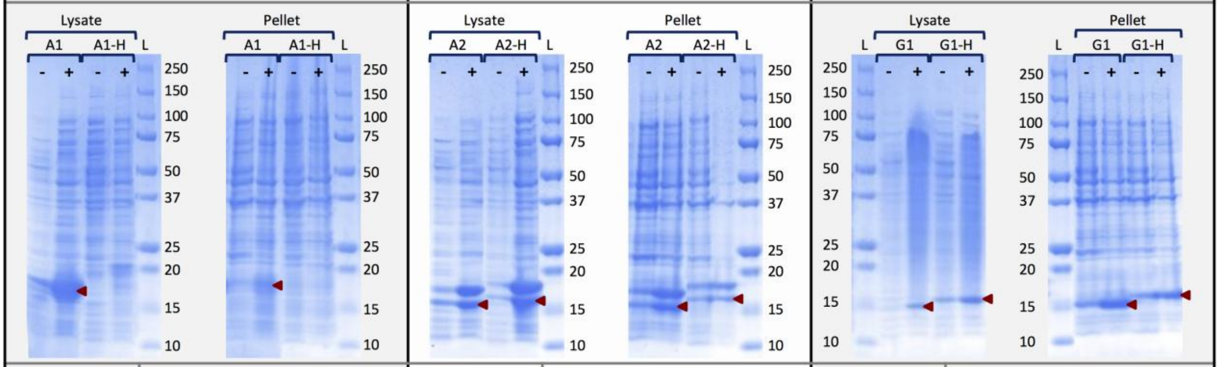
E.		F.		G.		H.	
pET21-Ars3		pET21-Glx1		pET21-Glx1-H		pET21-Glx2	
EcoRV		EcoRV		RsaI		XmnI	
I	1,452 + 4,389	I	1,539 + 4,266	I	1,567 + 1,760 + 2,472	I	1,271 + 1,934 + 2,573
P	5,631	P	5,631	P	1,027 + 1,277 + 1,567 + 1,760	P	1,934 + 3,697

Figure 4.7: Selected results from verification of the eight pET21 constructs with target gene insert. The expected fragments for correct construct (I) and re-ligated pOD1 (P) when digested with the respective endonucleases is listed underneath each specific candidate. The resulting digestion pattern of a correct construct (denoted with red marks) and a re-ligated pOD1 (except for Ars1 and Ars1-His) is shown for each candidate verification. In addition, a negative control (N) is shown in the second verification of Ars2 where pOD1 is digested with HincII. The specified values of the O'GeneRuler™ 1 kb DNA standard (L) are given in bp. The full gel pictures showing all twelve selections for each candidate can be found in appendix F.

The expression of all target genes, except Ars1-His, showed an additional band on SDS-PAGE with the expected molecular weight, indicating that these target genes were expressed in the induced cells (Figure 4.8). Some of the non-induced samples also seem to have a strong expression of the target gene. This could be caused by leak expression and/or an overflow from the neighboring wells during the loading of samples to the SDS-PAGE. All of the successfully expressed target proteins seem primarily to be located in the lysate, indicating that the proteins are soluble. In addition, the SDS-PAGE results for the Ars2 and Ars2-His protein showed a doublet band, where the smallest band seemed to have the expected molecular weight while the other band was around 2 kDa larger.

Only one variant of each protein was needed for the characterization. The untagged Ars1 target protein was chosen since it was the only version of Ars1 that showed to be expressed. Both Ars2 and Glx1 were expressed with and without the His-tag, where the size difference between the two versions (tagged and untagged) is clearly seen on the SDS-PAGE. The His-tag variant of the proteins was preferred as it allows a faster purification and higher protein purity than what can be achieved by the thermo-exclusion purification.

A.	B.	C.
Ars1 and Ars1-His	Ars2 and Ars2-His	Glx1 and Glx1-His



Ars1	17.46 kDa	Ars2	16.05 kDa	Glx1	15.30 kDa
Ars1-His	18.28 kDa	Ars2-His	16.87 kDa	Glx1-His	16.12 kDa

Figure 4.8: SDS-PAGE for the expression test performed on the six selected candidates with their expected size listed underneath. The candidates were expressed both with and without a His-tag. Samples of sonication lysate and cell debris pellet from induced (+) and non-induced (-) cells were analyzed. Red indicators denote the expected molecular weight of expressed target genes. The specified values for the Precision Plus Protein™ All blue ladder (L) are given in kDa.

4.3.2 Expression of target genes and His-tag isolation

The target genes were first expressed in unlabeled 2xLB medium to for thermo characterization. Prior to the NMR analysis the target proteins were isotope labeled using M9 medium with ^{15}N nitrogen source for the protein expression. This allows easy assessment of the protein stability, integrity, purity and fold.

Non-isotope labeled

His-tagged Ars2 and Glx1 were purified using the standard protocol for His-tag purification. The SDS-PAGE results for both candidates showed that the majority of the host cell proteins were located in the flow-through of lysate and wash steps, and that some of the target proteins were eluated in the first of the two elution steps (Figure 4.9, panel A.1 and B.1). However, the main part of both target proteins was located in the flow-through of lysate and the two wash steps. This could indicate that the proteins did not bind sufficiently to the column, and that a lower concentration of imidazole should be used to wash the host cell proteins out.

Prior to the second purification, the fractions that contained target proteins were dialyzed to remove imidazole. In this attempt, the washing was done in four steps using four different concentrations of imidazole (2, 5, 10 and 15 mM), instead of 20 mM imidazole. The elution flow-through in this second attempt of both candidates, showed a higher concentration of target proteins in the first step relative to the corresponding flow-through in the first attempt, where both candidates seem to have up to 90 % homogeneity (Figure 4.9, panel A.2 and B.2). The flow-through of wash buffer containing 2 and 5 mM imidazole contained the smallest amount of target proteins, indicating that a concentration of 2-5 mM imidazole should be sufficient for washing the column.

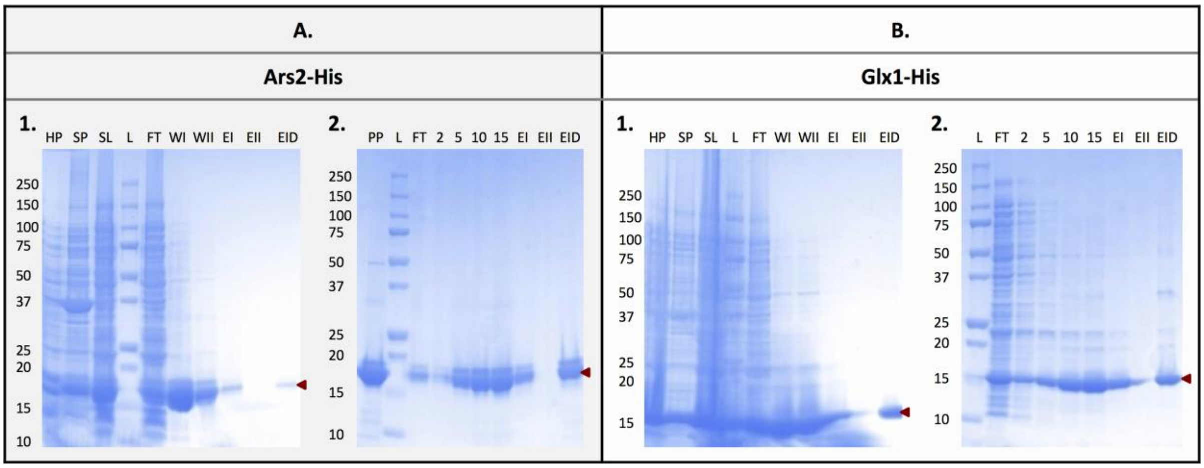


Figure 4.9: SDS-PAGE of Ars1-His and Glx1-His expression, and their His-tag purification. For the expression, samples of the harvest pellet (HP), the sonication pellet (SP), and the sonication lysate (SL) are shown for both candidates on their first gel. In addition, the gel shows the flow-through of the lysate (FT), the two wash steps (WI and WII), the two elution steps (EI and EII), and a double amount of the first elution step (EID) for both candidates during His-tag purification.

The other gels show the second attempt of purification with the same samples as in the first except for the two wash steps which has been replaced with four steps of washing using the concentrations 2, 5, 10 and 15 mM of imidazole. Red cursors indicate the molecular weights of the target proteins, and the specified values for the Precision Plus Protein™ All blue ladder (L) are given in kDa.

During dialysis of the His-tagged Ars2 protein sample a white precipitation was formed. A sample of the precipitate was analyzed by SDS-PAGE to determine if it contained any target protein. The SDS-PAGE for the precipitate showed a band with an undesirably high concentration of Ars2-His-tagged proteins (Figure 4.9, panel A.2). However, the elution result showed that there still was a significant amount of target proteins left to be used for thermo characterization. In addition, the Ars2-His-tagged protein results showed the same doublet band as in the expression test.

Isotope labeled

Ars2-His and Glx1-His were purified with an optimized His-tag protocol. The results showed a minimal amount of the target proteins in the wash buffer, while the most were located in the elution buffer with over 90 % homogeneity (Figure 4.10). The amount of expressed ^{15}N-labeled proteins seems to be lower compared to the unlabeled proteins. In addition, the doublet band also appeared for the isotope labeled Ars2-His-tagged protein.

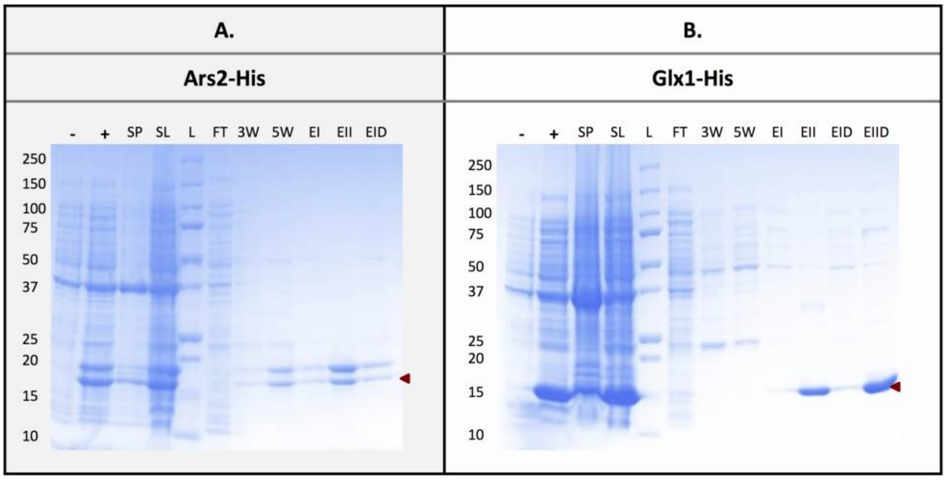

A.
Ars2-His
- + SP SL L FT 3W 5W EI EII EID
250
150
100
75
50
37
25
20
15
10
B.
Glx1-His
- + SP SL L FT 3W 5W EI EII EID EIID
250
150
100
75
50
37
25
20
15
10

Figure 4.10: SDS-PAGE of isotope labeled Ars1-His and Glx1-His expression and their His-tag purification. From the expression of both candidates, harvest pellets from induced (+) and non-induced (-) cells are shown. In addition, the sonication pellet (SP) and the sonication lysate (SL) is added.

From the purification, flow-through of the lysate (FT), the two wash steps with 3 (3W) and 5 (5W) mM imidazole, the two elution steps (EI and EII), and a double amount of the first elution step (EID) for both candidates during His-tag purification is added. In addition, for Glx1-His a double amount of the second elution flow-through (EIID) is added. Red markers indicate the molecular weights of the candidates, and the specified values for the Precision Plus Protein™ All blue ladder (L) are given in kDa.

4.4 Protein characterization

4.4.1 Thermostability analysis

Thermo characterization was performed on lysate from non- isotope labeled samples of Ars1 lysate, and purified Ars2-His and Glx1-His. This was done by incubating samples of each candidate at 60 and 80 °C at different time intervals (from 5 to 180 minutes), with a parallel of samples for each interval. For thermostable proteins no significant decrease of target proteins were expected over time, indicating that the proteins are still soluble. The amount of soluble protein after heat treatment was evaluated by SDS-PAGE. The characterization of

Ars1 was analyzed to see if the protein could be purified to homogeneity of preferably 90 % using thermo-exclusion, to be used for the isotope labeled Ars1 prior to NMR spectroscopy.

Thermo characterization

The SDS-PAGE of Ars2-His showed a significant decrease of target proteins at all intervals for both temperatures (Figure 4.11, panel C and D), indicating that it is aggregated and not thermostable at the given conditions. The opposite was observed for Ars1 and Glx1-His, where Ars1 showed to remain in solution at both of the temperatures (Figure 4.11, panel A and B.1) while Glx1-His were precipitated at 80 °C (Figure 4.11, panel E). The amount of target proteins showed no significant precipitation during the experiment indicating that Ars1 and Glx1 (at 60 °C) may be thermostable (further studies are necessary to say if they are folded and active). The SDS-PAGE of Ars1 showed a decline in the amount of host cell proteins. The decline was gradually stronger with increased incubation time and higher temperature (Figure 4.11, panel A and B.1), indicating that the Ars1 protein can be purified by thermo-exclusion.

However, the accomplished reduction of host proteins was not adequate for purification to NMR analysis. A second purification was therefore set up, using a new set of Ars1 samples at 80 °C with prolonged incubation intervals, was set up to see if a higher purity of the target protein could be reached. The results showed a significantly stronger reduction of host proteins when compared to the previous attempt, and almost no reduction of the target protein (Figure 4.11, panel B.2). The amount of target protein encompasses approximately 90 % purity, indicating that the purification of Ars1 by thermo-exclusion can be sufficient for NMR spectroscopy. A distinct band at about twice the molecular weight (approximately 35 kDa) of Ars1 showed a relatively high endurance to the increased temperature compared to the host proteins, which may indicate homodimerization of Ars1.

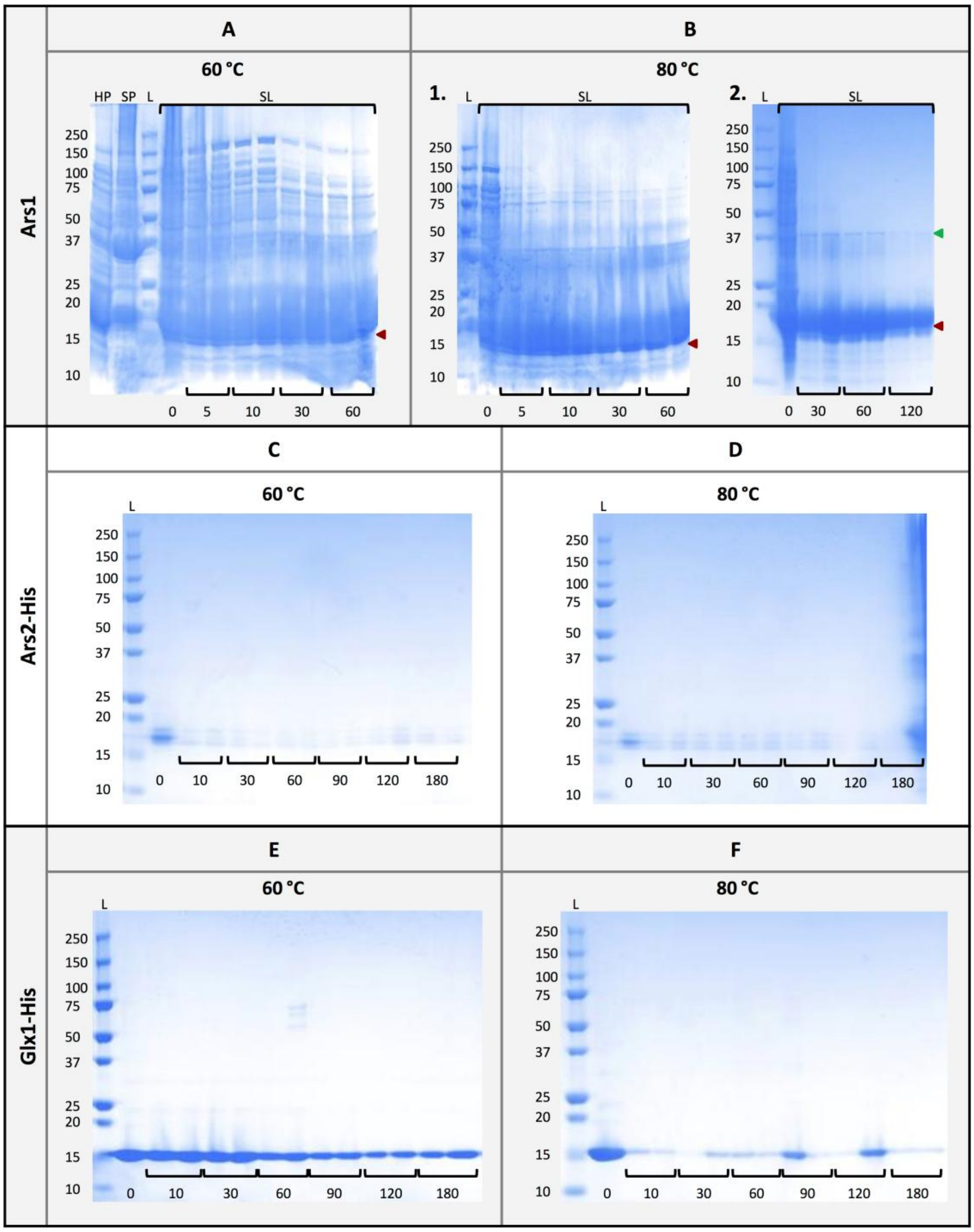


Figure 4.11: SDS-PAGE for Ars1 expression and thermo purification, together with purified Ars2-His and Glx1-His thermo test results. From the Ars1 expression, samples of the harvest pellet (HP), the sonication pellet (SP) and the sonication lysate (SL) are shown. For the purification and thermo testing of each candidate, samples of the lysate are shown with a parallel for each incubation interval at 60 °C (panel A, C and E) and 80 °C (panel B1, B2, D and F). A negative control of the lysate with no heat treatment (0 minutes of incubation) was made for each candidate.

Red cursors denote the molecular weight of candidate Ars1 while the green indicate the weight of a potential homodimer of Ars1. The high amount of proteins observed for the 180 minute parallel at 80 °C for Ars2-His (panel D) are contamination from the adjacent well with a nonrelated sample. The specified values for the Precision Plus Protein™ All blue ladder (L) are given in kDa.

Thermo purification of isotope labeled Ars1

Isotope labeled Ars1 was purified from the cell lysate by incubation at 80 °C for a total of nine hours. The results showed a gradually more precipitation of soluble host cell proteins, while the level of target protein did not show significant precipitation. After nine hours a purity of less than 90 % was achieved for the Ars1 protein (Figure 4.12, panel C), though further isolation was not attempted, to prevent significant loss of target protein. As for the non-isotope labeled Ars1, an extra band was seen at 37 kDa, indicating homodimerization.

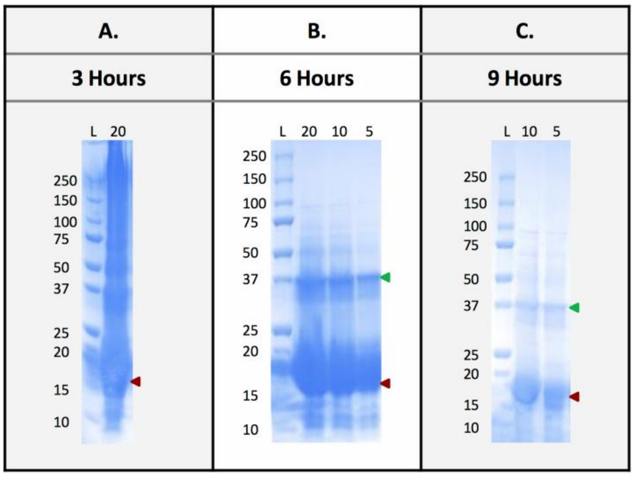


Figure 4.12: SDS-PAGE of isotope labeled Ars1 thermo purification, showing samples from the sonication lysate after 3, 6 and 9 hours of incubation at 80 °C. 20, 10 and/or 5 µl of sample was loaded on to the gel following each incubation period. Red markers indicate the molecular weight of the isotope labeled Ars1 while the green denote potential homodimers of the candidate. The specified values for the Precision Plus Protein™ All blue standards ladder (L) are given in (kDa).

4.4.2 Structural analysis by NMR

Two samples of Ars1 (pH 5.0 and pH 6.9), one sample of Ars2 (pH 5.0) and one sample of Glx1 (pH 5.0) isotope labeled proteins were prepared for NMR spectroscopy. Both a 1D proton and a 2D {1H-15N} spectrum were made for each protein sample.

The NMR 1D spectra showed that both samples of Ars1 had good dispersion (especially in the methyl region) with relatively well defined peaks compared to the other two candidates (Figure 4.13 and Figure 4.15Figure 4.15, black arrows), indicating that the proteins in the sample are folded. In addition, the Ars1 spectra show distinct peaks between 5-6 ppm in the H^{α} region (Figure 4.13 and Figure 4.15, red brackets), suggesting that the protein contains β-sheet fold. The 2D {1H-15N} HSQC spectra of the Ars1 samples also show relatively good dispersion of chemical shifts, though the line-width of several peaks are large with low resolution, especially between 7.2-8.8 ppm and 7.0-9.2 ppm of the proton dimension, and 115-125 ppm and 110-128 ppm in the amide dimension, in pH 6.9 and pH 5.0 respectively (Figure 4.14 and Figure 4.16). This may indicate a poorly structured part in the protein or protein degradation.

The Ars2-His showed a 1D spectrum with approximately the same dispersion as Ars1 but with much broader and less defined peaks and over all lower resolution (Figure 4.17), indicating that the proteins are partially folded. The 2D {1H-15N} HSQC spectrum of Ars2 showed a good dispersion of chemical shifts with relatively well defined peaks. However, some signals around 8 ppm in the proton dimension and 120 ppm in the amide dimension, showed broader signals whit lower resolution (Figure 4.18). As for the Ars1, this might be a sign of degradation or a non-structured part in the protein.

For Glx1, it can clearly be seen that the proton signals cluster within the regions typical for unfolded and partially folded proteins in the 1D spectrum as well as 2D {1H-15N} HSQC spectrum respectively (Figure 4.19 and Figure 4.20). In addition, the signals in both spectra were characterized by broad ill-defined peaks and the amide dimension in the 2D {1H-15N} HSQC spectrum also had poor dispersion. This, together with the other observations indicates that the protein is non-structured.

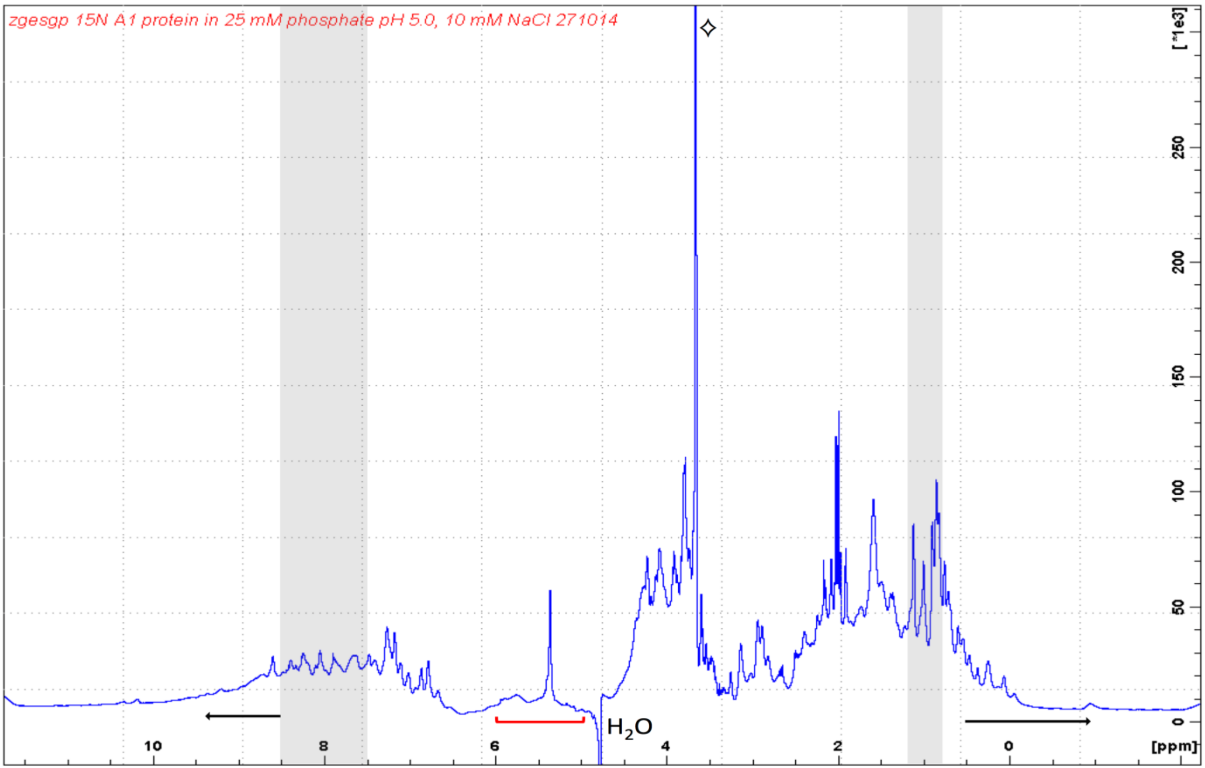


Figure 4.13: 1D proton spectrum of isotope ^{15}N-labeled Ars1-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 10 mM NaCl recorded at 298 K. The two peaks (◇) at 3.6-4.0 ppm are of unknown origin. The areas in grey denote where the H^{N} and methyl proton signals from non-structured proteins with random coil tend to be located, the black arrows designate the dispersion indicating a tertiary structure, and the red bracket signify the H$^{\alpha}$ region that correlates with β-sheets.

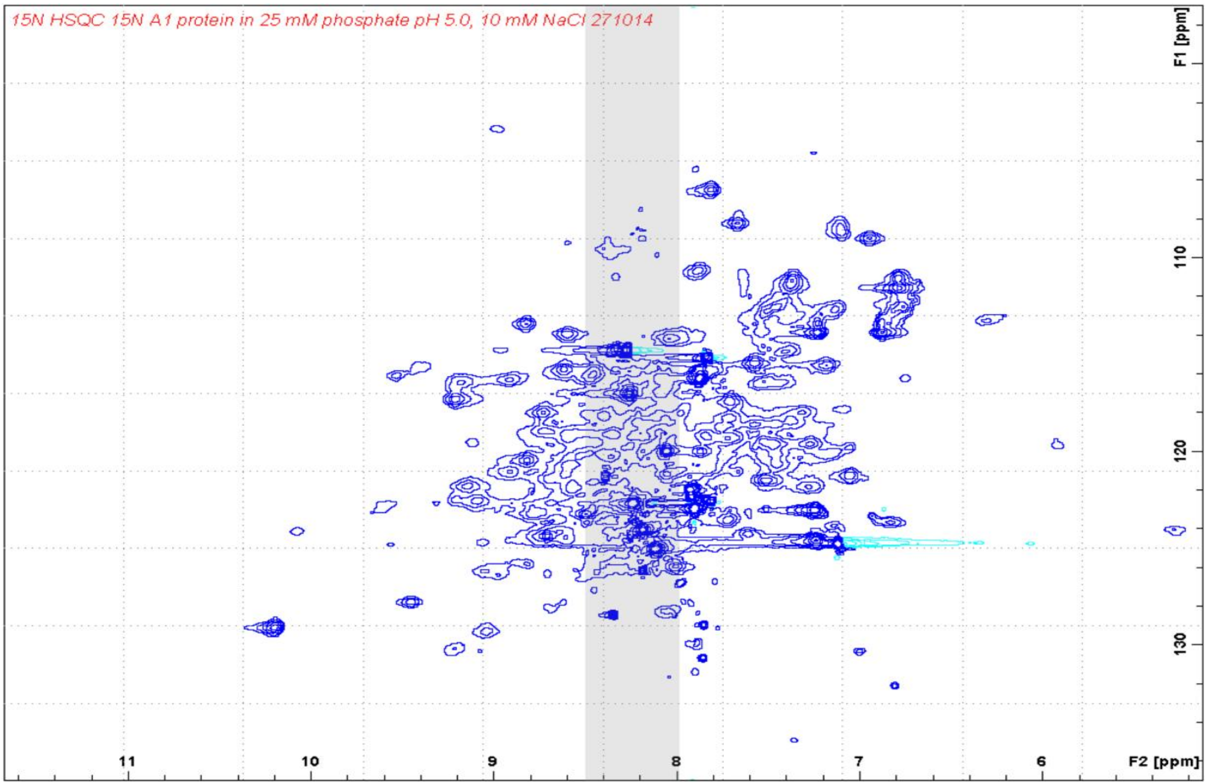


Figure 4.14: 2D {1H-15N}-HSQC spectrum of isotope ^{15}N-labeled Ars1-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 10 mM NaCl recorded at 298 K. The area in grey indicates the region in the proton spectrum were H^{N} signals from partially folded proteins tend to be located.

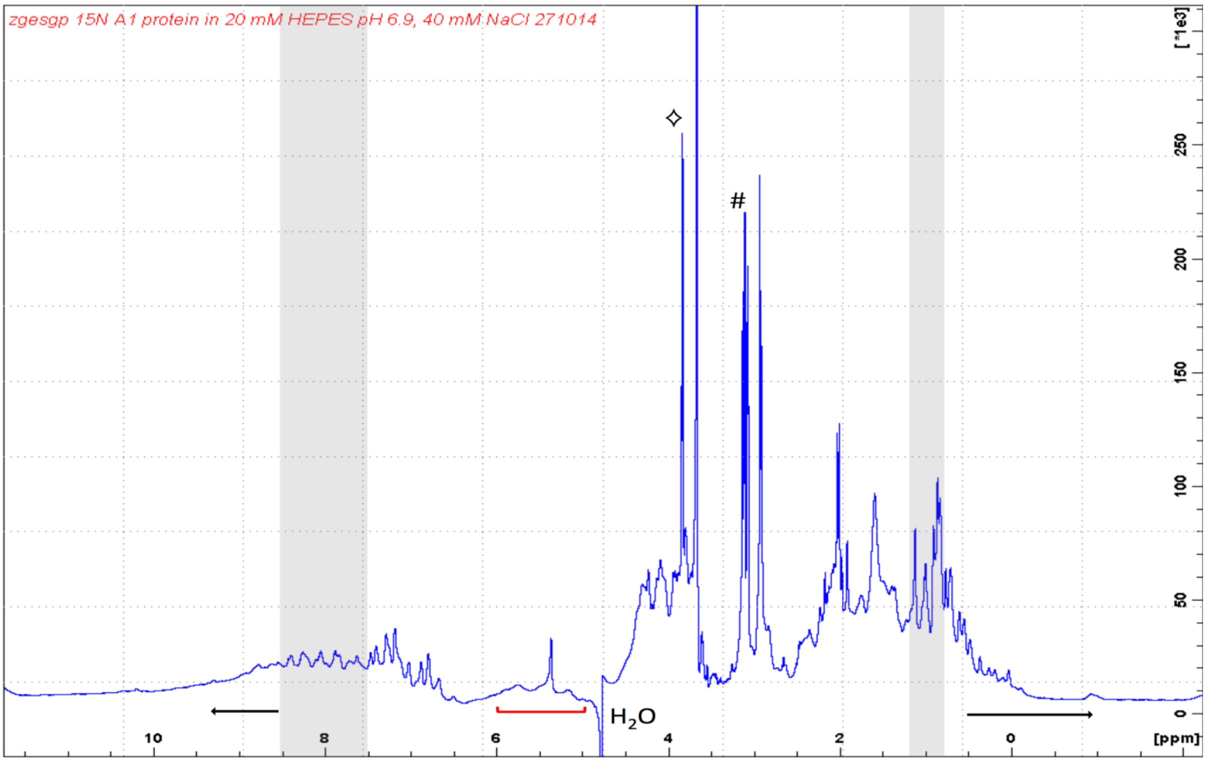


Figure 4.15: 1D proton spectrum of isotope ^{15}N-labeled Ars1-protein in (90:10) H2O:D2O 20 mM HEPES buffer pH 6.9 with 40 mM NaCl recorded at 298 K. The two peaks (✧) at 3.6-4.0 ppm are of unknown origin, while the two peaks (#) around 3 ppm correlates with HEPES from the buffer. The areas in grey denote where the H^N and methyl proton signals from non-structured proteins with random coil tend to be located, the black arrows designate the dispersion that indicate a tertiary structure, and the red bracket signify the H^α signals that correlates with β-sheets.

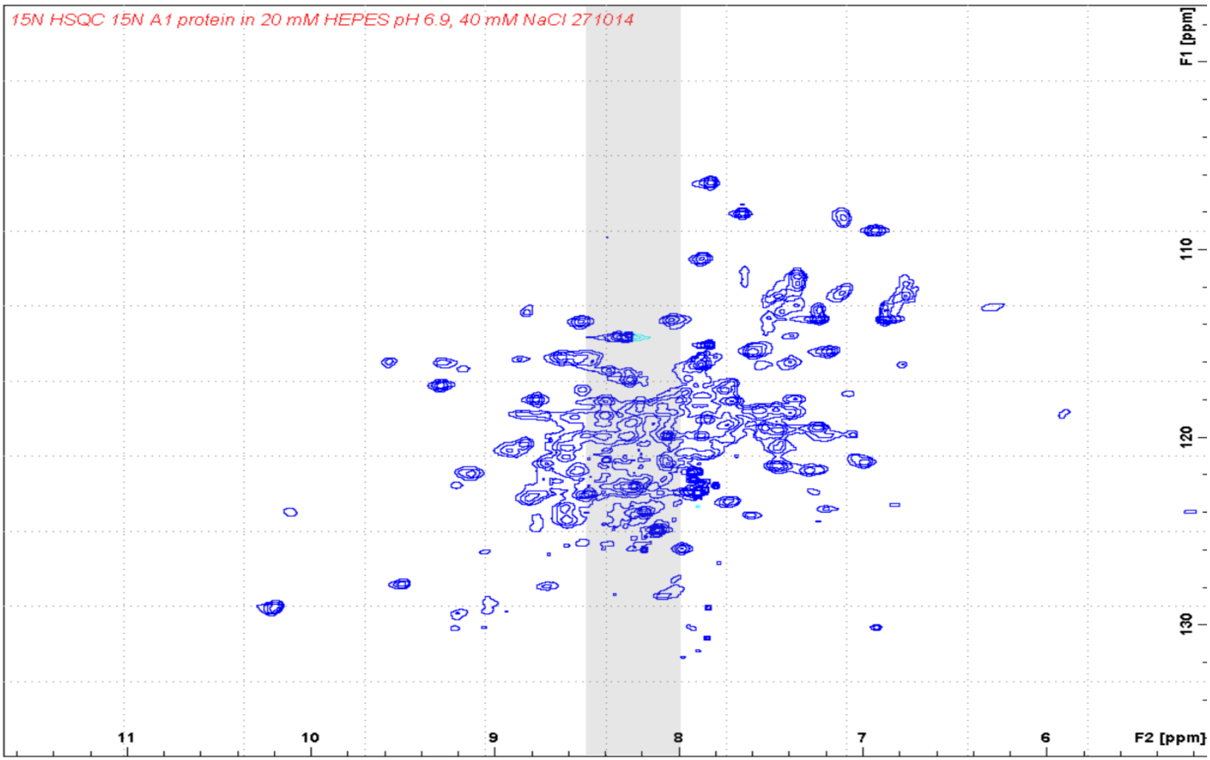


Figure 4.16: 2D {1H-15N}-HSQC spectrum of isotope ^{15}N-labeled Ars1-protein in (90:10) H2O:D2O 40 mM HEPES buffer pH 6.9 with 10 mM NaCl recorded at 298 K. The area in grey indicates the region in the proton spectrum were H^N signals from partially folded proteins tend to be located.

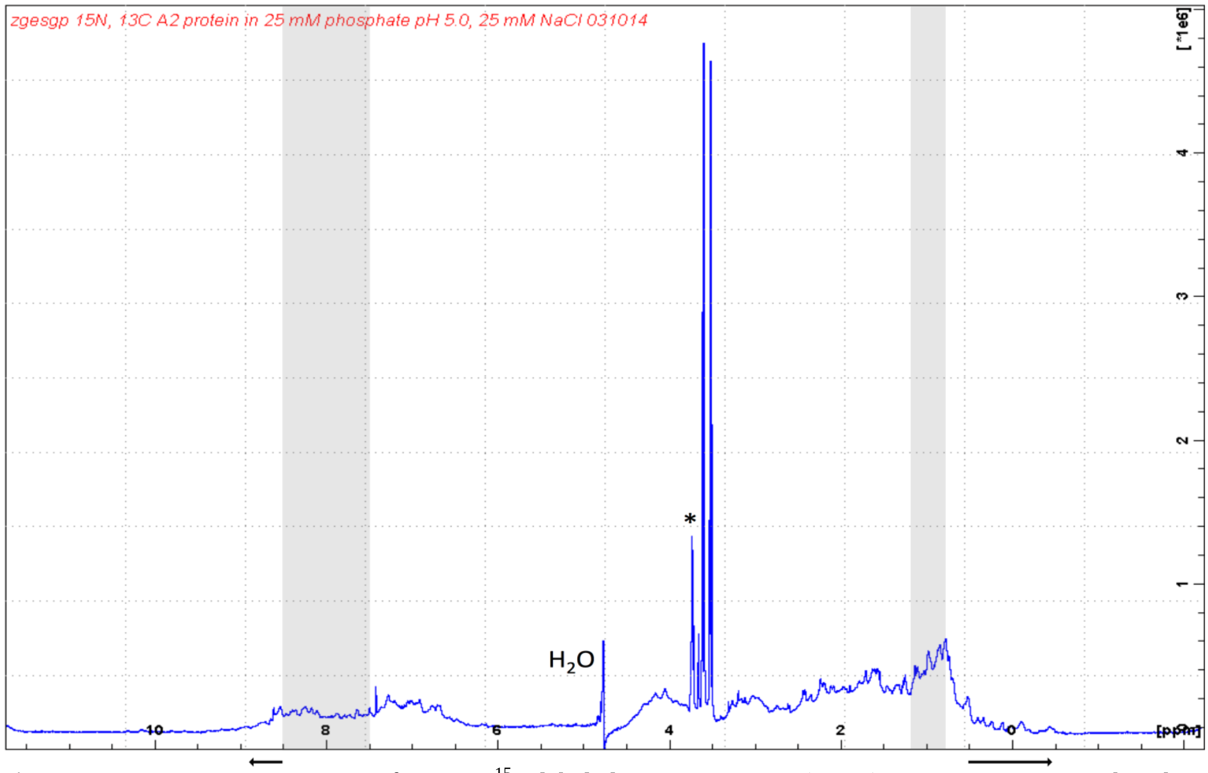


Figure 4.17: 1D proton spectrum of isotope ^{15}N-labeled Ars2-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 25mM NaCl recorded at 298 K. The three peaks (*) at 3.5-3.8 ppm are thought to be glycerol from VivaSpin column. The areas in grey indicate where the H^{N} and methyl proton signals from non-structured proteins with random coil tend to be located, and the black arrows designate the dispersion which correlates to a tertiary structure.

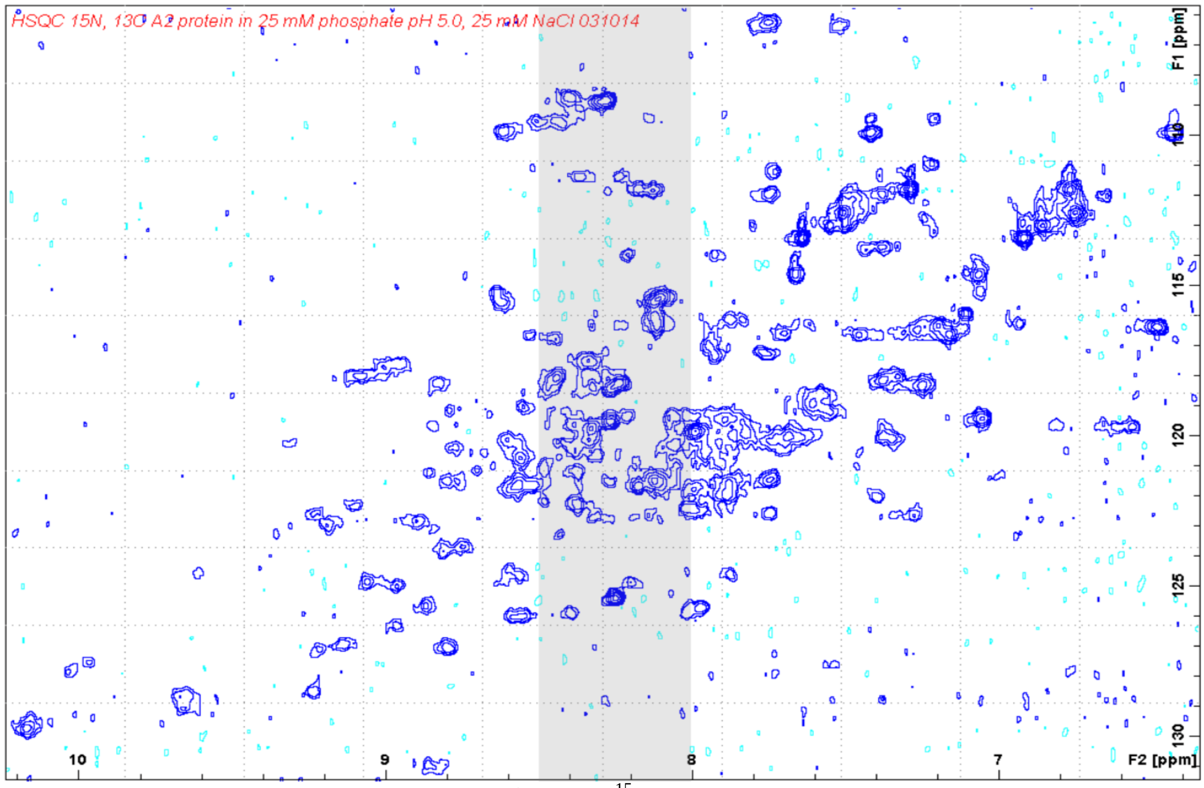


Figure 4.18: 2D {1H-15N}-HSQC spectrum of isotope ^{15}N-labeled Ars2-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 25mM NaCl recorded at 298 K. The area in grey indicates the region in the proton spectrum were H^{N} signals from partially folded proteins tend to be located.

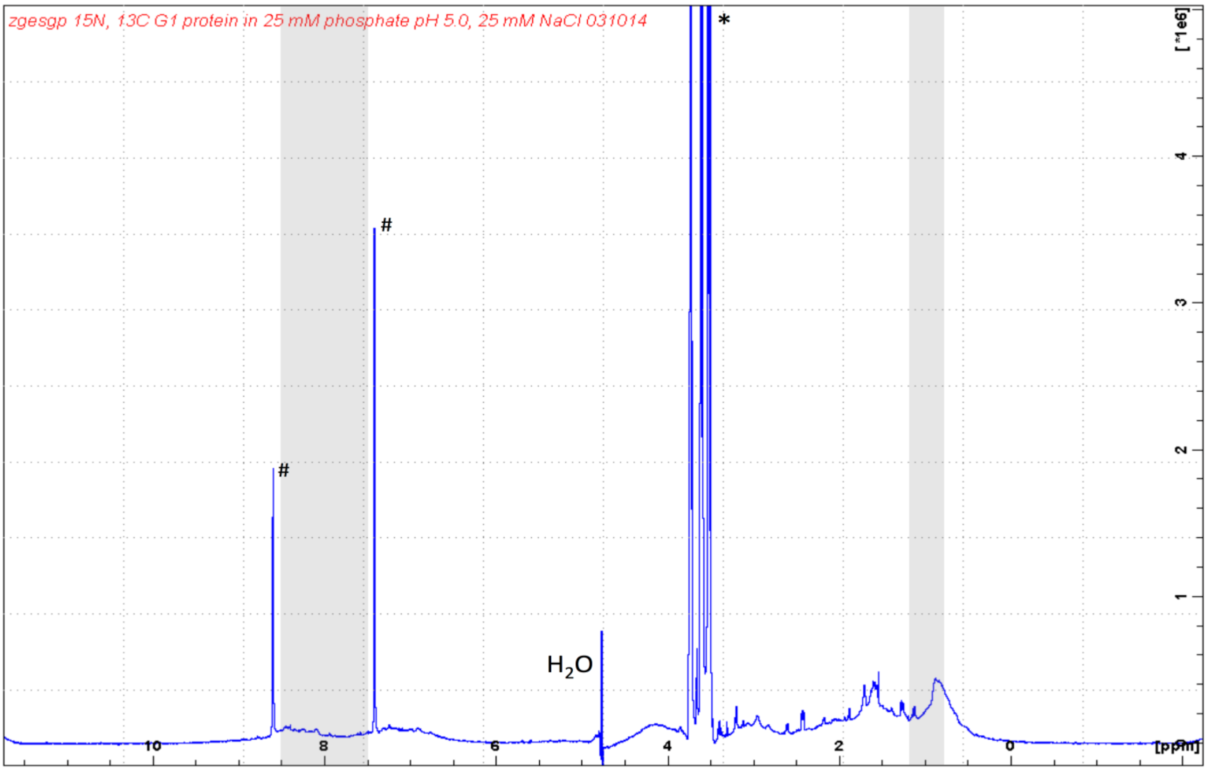


Figure 4.19: 1D proton spectrum of isotope ^{15}N-labeled Glx1-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 25mM NaCl recorded at 298 K. The three peaks (*) at 3.5-3.8 ppm are thought to be glycerol from VivaSpin column, while the two peaks (#) at ~7.5 and 8.5 ppm corresponds to signals from imidazole remnants from the His-tag isolation. The areas in grey indicate where the H^{N} and methyl proton signals from non-structured proteins with random coil tend to be located.

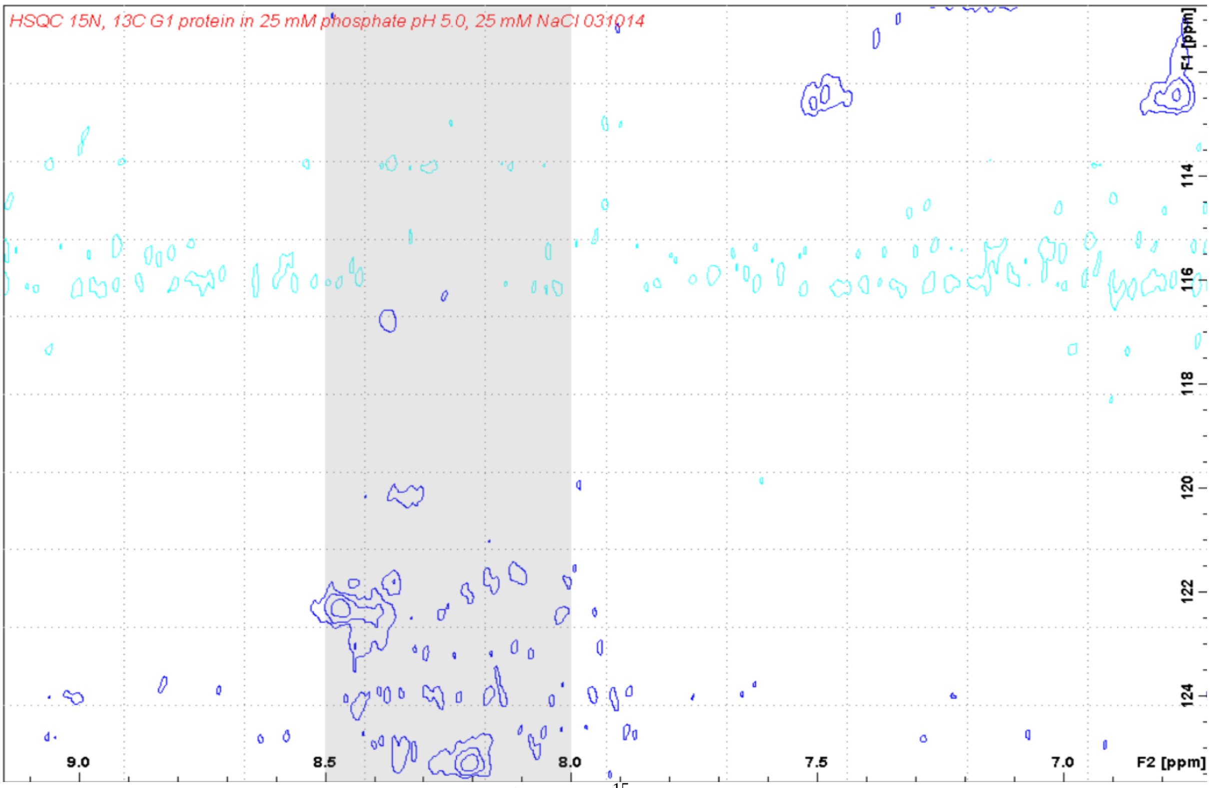


Figure 4.20: 2D {1H-15N}-HSQC spectrum of isotope ^{15}N-labeled Glx1-protein in (90:10) H2O:D2O 25 mM phosphate buffer pH 5.0 with 25mM NaCl recorded at 298 K. The area in grey indicates the region in the proton spectrum were H^{N} signals from partially folded proteins tend to be located.

5 Discussion

The aim of this study was to discover polyextreme enzymes in a translated metagenomic sequence database, made of microbial samples collected from oil reservoirs. The proteins were to be of well characterized protein families which should be suitable for structural NMR analysis, with the purpose of studying the proteins stability, integrity, purity and fold compared to their mesophilic counterparts. In addition, the thermostability of the en proteins was to be analyzed. Prior to analyzes, the respective genes were to be modified with a His-tag for purification, heterologously expressed and purified to homogeneity satisfactory for NMR analysis.

5.1 Candidate selection

The queries used

From the search in UniProt the amino acid sequence of Mrx-dependent ArsC2 from *C. glutamicum* and GlxI from *H. influenzae*, were found suitable as queries to find similar novel proteins in the translated metagenomic sequence database that were appropriate for NMR analysis. The Mrx-ArsC is a novel family of ArsC proteins containing monomeric enzymes of approximately 14 kDa, which has not been shown to bind any metal ions as cofactors but has the charged $N^{\zeta}H^{+}$ of Lys-64 to substitute for the K^{+} binding that is found in the *S. aureus* ArsC and positively affects the activity (Ordóñez et al., 2009, Villadangos et al., 2011, Lah et al., 2003). The GlxI from *H. influenzae* is thought to be a homodimeric Ni^{2+} binding GlxI due to the short sequence of the subunit (135 amino acids long), which is consistent with the length of other Ni^{2+} binding GlxI such as the ones from *E. coli*, *Yersinia pestis* and *Neisseria meningitides* (Clugston et al., 1998, Sukdeo et al., 2004). This ion is not considered to be a strongly paramagnetic species and can easily be used for NMR analysis (Kwan et al., 2011). In addition, both these proteins are of soluble cytosolic enzyme families with no known proteolytic activity, thus consistent with the criteria needed to find candidates suitable for NMR analysis (Ordóñez et al., 2009, Villadangos et al., 2011, Sukdeo et al., 2004, Kwan et al., 2011). Five candidates were selected for further processing in order to have several options.

The ArsC candidates

Analysis of the multiple alignment for the three ArsC candidates together with selected proteins of different ArsC groups, indicate that they all contain a PTPase fold and therefore might have PTPase activity. This is based on the observation that all three candidates contain the fully conserved PTPase active site Asn-13, Arg-16 and Asp-105 of *S. aureus* which is found to have PTPase activity (Zegers et al., 2001). The analysis also implies that they are related to the Trx-dependent ArsC. Candidate Ars2 and Ars3, shown to have the highest resemblance to *B. subtilis* ArsC in the PDB database, both have all their catalytic cysteines and P-loop totally conserved with the monomeric Trx-ArsC of *B. subtilis* and *S. aureus* (Messens et al., 1999, Messens et al., 2002, Bennett et al., 2001). In addition, candidate Ars1 shows the highest resemblance to the homodimeric Trx-ArsC from *C. glutamicum* both from the PDB alignment an from the positions of their catalytic cysteines (Villadangos et al., 2011). However, the finding of the PTPase-like Grx-ArsC from *Synechocystis* sp. which also encompasses three catalytic cysteines (though with slightly different positions of the C-terminal cysteines), and has over 30 % identity to Trx-enzymes (Li et al., 2003), means that it cannot be said with certainty that the candidates are Trx-ArsC based on sequence alignment alone. On the other hand, it can be said with certainty that they are not dependent on Mrx such as the ArsC query that was used to find the candidates. This enzyme only has one catalytic cysteine and had the lowest resemblance toward the candidates of all the ArsC families. Even though the candidates are from a different group of ArsC than the UniProt query that was used to find them, they still qualify for NMR spectroscopy. This is based on the fact that both the Trx-ArsC and the PTPase-like Grx-ArsC possess the same characteristics as the Mrx-ArsC, except for the K^+-binding found in *S. aureus* Trx-ArsC which is not fully conserved in any of the candidates (Kwan et al., 2011, Yu et al., 2011, Ordóñez et al., 2009, Villadangos et al., 2011, Ji et al., 1994).

The GlxI candidates

The multiple alignment of different verified kinds of GlxI enzymes together with the putative GlxI candidates indicate that the candidates are related to the Ni^{2+}-GlxI with the βαβββ-binding conformation. At least two observations in the alignment support this hypothesis. The short lengths of the candidates are consistent with other Ni^{2+} binding GlxI found, where the longer sequence of Zn^{2+} binding GlxI is necessary to obtain a octahedral geometry (Figure 5.1, A) of the Zn ion complex (He et al., 2000). This conformation of the bound ion has been found to be crucial for activity in both the Zn^{2+}- and Ni^{2+}-GlxI, binding two water molecules which are thought to have a key function in the catalytic mechanism. Conversely, when the shorter form of GlxI binds Zn ions a trigonal bipyramidal conformation is formed (Figure 5.1, B), which binds only one water molecule and thus renders the enzyme inactive. In addition to protein length, the multiple alignment shows that both of the candidates has a higher identity with the active site of enzymes with a βαβββ-binding conformation, rather than the enzyme with the back-to-back conformation. For candidate Glx1, the PDB alignment also supports this hypothesis where the *E. coli* GlxI was found to be the most similar to candidate Glx1.

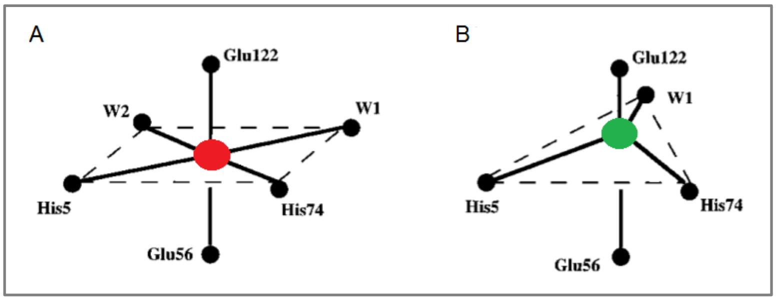


Figure 5.1: Metal binding conformations in *E. coli* GlxI, with Ni^{2+} (in red) bound in a octahedral conformation (A) and Zn^{2+} (in green) bound in a trigonal bipyramidal conformation (B) (He et al., 2000). The water ligands are labeled W1 and/or W2.

5.2 DNA recombination and heterologous expression

As a result of the His-tag modification, the following candidates Ars1, Ars2 and Glx1 showed to have a His-tag. This modification was also confirmed by SDS-PAGE results of the expression test, where the additional 0.8 kDa of the 6*His-tag is clearly seen on the tagged proteins when compared to their non-tagged counterparts as well as from the His-tag purification. However, no expression target gene was detected for candidate Ars1-His. This could be caused by too low expression, degradation of the protein in the cell, or unsuccessful expression. Seeing as the untagged version of Ars1 was expressed in high amounts, there is no obvious reason why the expression of the His-tagged version did not work. The same goes for degradation of the protein since both Ars1 and Ars1-His was produced under the same conditions. It is therefore likely that the target gene Ars1-His failed to be expressed, and a plausible reason are mutations in the gene or gene region. This hypothesis can be tested by sequencing the Ars1-His target gene and aligning it with the original sequence retrieved from GenScript (Appendix C).

A total of five of the candidate genes that was selected for expression seemed to be successfully expressed in the BL21 (DE3) cells. However, the Ars2 and Ars2-His showed an unexpected doublet band on SDS-PAGE, whereof the extra band was around 2 kDa larger than expected. Because none of the other candidates show sign of a doublet band, it can be assumed that the cause for this band must be due to the characteristics of the gene or gene product of Ars2 or Ars2-His (gene sequence or protein) or interactions between it and the surrounding pET21 backbone. In addition, since the unexpected band was the larger band, it rules out any form of proteolytic cleavage, since this would produce shorter fragments than expected.

A plausible explanation to the doublet band can be cysteine oxidation (Krondiris and Sideris, 2002). Oxidation of cysteine residues forms loops in the protein, causing an altered migration rate of the protein in the gel, leading to a doublet band. However, measures have been taken to reduce the disulfide bonds in the protein samples by adding BME. According to Sigma-Aldrich procedures a concentration of 5 % BME (~700 mM) is usually sufficient to in SDS-PAGE (http://www.sigmaaldrich.com/catalog/product/aldrich/m6250, 09.16.2015), and a concentration of approximately 350 and 700 mM has been used for these samples. However, if oxidation occurs, it might be that a longer incubation time or higher concentration of BME is required. Another reducing agent could be tested if BME is not sufficient for this protein, such as dithiothreitol that can reduce disulfides quantitatively (Cleland, 1964).

Another event that can cause doublet bands is translational read-through of the stop codon, which causes additional amino acids to be added to the protein. At least three different mechanisms can cause this event based on the gene sequence alone. One of which is by translational missense error which is generally caused by tRNA competition (Kramer and Farabaugh, 2007). The other two mechanisms is by codon context effects 5' of the stop codon where the chemical properties of nascent amino acids can effect translational termination, and by the characteristics of the stop codon used which determine translational termination efficiency (Mottagui-Tabar et al., 1994). However, none of these mechanisms are obvious to have occurred for the Ars2 candidates. The first mechanism should be prevented by the optimization of the Ars2 gene done by GenScript which reduces the risk of tRNA competition. In addition, if the doublet band was a result of codon effects, stop codon used or a combination of these two, one could expect to observe similar doublet band for the Glx1-His candidate. This is because the Glx1-His target gene has the exact same codon composition as the Ars2-His in the 3' of the gene (6*His-tag + TAA). If for some reason the doublet band is caused by read-through, it would be important to verify and remove the problem. A way to avoid this is by adding an additional stop codon neighboring to the original one (MacBeath and Kast, 1998).

5.3 Characterization

The ArsC candidates

Results from the thermo purification and thermo exclusion test of Ars1 and Ars2-His respectively showed that Ars1 seems to be thermostable while Ars2-His does not. In addition, the SDS-PAGE after thermo purification of Ars1 showed an additional band that was about twice the size of Ars1. The size of this band fits well with a homodimer of the protein. If candidate Ars1 is shown to form a natural dimer it would be the first small ArsC discovered to be homodimeric. It is not unlikely that such an event has taken place in order for this protein to become thermostable, seeing as oligomerization is a common adaption to thermostability (Reed et al., 2013). However, this might be an artificial dimerization caused by the SDS, which has been shown capable of inducing non-natural oligomerization of some proteins when employed in high concentrations (Bitan et al., 2005).

There can be several reasons for Ars2-His not to show thermostability. It might be that the host cell from the oil well does not express the Ars2 protein, rendering it a mere remnant from its initial mesophilic relative. A study done of the metagenomic libraries from the two oil wells has shown that the evolution of these organisms is probably more than 20-30 times slower than on the surface (Lewin et al., 2014), which might explain why the gene can still be found as a coding element, rather than being removed or turned into a pseudogene. Another reason might be that the protein is dependent on specific physical conditions that are not optimal during testing, such as a specific pH or cofactor. For instance, many proteins are dependent on salt bridges to maintain thermostability, and these bonds (created by charged residues) are highly dependent on pH (Reed et al., 2013, Anderson et al., 1990).

The 2D {1H-15N} HSQC spectra of Ars1 and Ars2 showed to have an area of indistinct signals centered in the 8 ppm region of their proton spectra and in the 120 ppm region of their amide spectra. Certain events may be the cause of this, for example if the sample is highly contaminated or if degradation of the target proteins is occurring in the sample (as mentioned earlier). However, these signals might also correspond to the P-loop region of the ArsC protein which can have no fixed conformation. By interacting with specific oxyanions, this particular region will be fixed to a certain conformation and the signals may be observable in a 2D {1H-15N} HSQC spectrum (Messens et al., 2002, Zegers et al., 2001). Both candidates show potential to be analyzed further with NMR spectroscopy. However, the findings of this study cannot be taken as evidence that the candidates found are novel ArsC enzymes, activity need to be verified to see if this is the case.

The GlxI candidate

The thermo exclusion test for candidate Glx1 indicated that it was thermostable at 60 °C, but not at 80 °C. This possible thermostability would have to be verified with activity analysis since no further information was provided by the NMR results itself. Further work would be determined based on whether the protein has a functional relevant form or not.

5.4 Future work

SDS-PAGE analysis of the Ars2-His should be done, where several incubation intervals and different concentrations of BME (and/or dithiothreitol) in Ars2-His samples is prepared, to see if the doublet band can be removed. Both the Ars1 and Ars2-His protein samples should be optimized for NMR spectroscopy to see if it is possible to achieve a better resolution. The proteins should be tested with higher pH values, different salts and salt concentrations, and their purification procedures should be optimized to obtain higher protein purity. Optimized Ars2-His samples should be tested for thermostability, in case a better protein fold is needed to obtain thermostability. All three candidates should be tested for their respective activity. In addition, since both Ars1 and Ars2-His has shown to contain PTPase like fold they should both be tested for PTPase activity, which can be done by a *p*-nitrophenyl phosphate assay.

6 Conclusion

Five putative novel candidates (Ars1, Ars2, Ars3, Glx1 and Glx2 of the ArsC and GlxI protein families respectively) were identified and selected from a metagenomic sequence database, which was made of microbial samples from two oil reservoirs taken on the NCS. Various bioinformatic tools were used to select the candidates based on certain criteria to make them suited for NMR analysis. Ars2 and Glx1 were successfully modified with a His-tag and heterologously expressed together with the un-tagged Ars1 in both isotope labeled and non-isotope labeled versions. The Glx1-His protein seemed to be thermostable up to 60 °C, but the NMR results showed no indication of structure. This suggests that an activity assay is needed to establish whether this candidate should be studied further. The NMR spectra for Ars1 and Ars2-His show that they seem to have a structural fold and may therefore have potential for more detailed NMR-based structural analysis. In addition, Ars1 appear to be thermostable while Ars2-His does not. Thus only the Ars1 candidate might be used to provide insight to the thermophilic and piezophilic characteristics of these polyextreme proteins. However, both these candidates require further analysis to confirm whether or not they are novel ArsC enzymes, and to determine which group of ArsC proteins they belong to in order to compare to them to their mesophilic counterparts.

References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

ANDERSON, D. E., BECKTEL, W. J. & DAHLQUIST, F. W. 1990. pH-Induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry,* 29**,** 2403-2408.

ARBER, W. 1965. Host-Controlled Modification of Bacteriophage. *Annual Review of Microbiology,* 19**,** 365-378.

BENNETT, M. S., GUAN, Z., LAURBERG, M. & SU, X.-D. 2001. Bacillus subtilis arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. *Proceedings of the National Academy of Sciences,* 98**,** 13577-13582.

BERENSMEIER, S. 2006. Magnetic particles for the separation and purification of nucleic acids. *Applied Microbiology and Biotechnology,* 73**,** 495-504.

BITAN, G., FRADINGER, E. A., SPRING, S. M. & TEPLOW, D. B. 2005. Neurotoxic protein oligomers—what you see is not always what you get. *Amyloid,* 12**,** 88-95.

BODENHAUSEN, G. & RUBEN, D. J. 1980. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters,* 69**,** 185-189.

BOUTZ, D. R., CASCIO, D., WHITELEGGE, J., PERRY, L. J. & YEATES, T. O. 2007. Discovery of a Thermophilic Protein Complex Stabilized by Topologically Interlinked Chains. *Journal of Molecular Biology,* 368**,** 1332-1344.

BURGESS-BROWN, N. A., SHARMA, S., SOBOTT, F., LOENARZ, C., OPPERMANN, U. & GILEADI, O. 2008. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expression and Purification,* 59**,** 94-102.

CACCIAPUOTI, G., FUCCIO, F., PETRACCONE, L., DEL VECCHIO, P. & PORCELLI, M. 2012. Role of disulfide bonds in conformational stability and folding of 5′-deoxy-5′-methylthioadenosine phosphorylase II from the hyperthermophilic archaeon Sulfolobus solfataricus. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics,* 1824**,** 1136-1143.

CAVANAGH, J. 1996. *Protein NMR spectroscopy: principles and practice,* San Diego, Academic Press.

CLELAND, W. W. 1964. Dithiothreitol, a New Protective Reagent for SH Groups*. *Biochemistry,* 3**,** 480-482.

CLUGSTON, S. L., BARNARD, J. F. J., KINACH, R., MIEDEMA, D., RUMAN, R., DAUB, E. & HONEK, J. F. 1998. Overproduction and Characterization of a Dimeric Non-Zinc Glyoxalase I from Escherichia coli: Evidence for Optimal Activation by Nickel Ions. *Biochemistry,* 37**,** 8754-8763.

CLUGSTON, S. L. & HONEK, J. F. 2000. Identification of sequences encoding the detoxification metalloisomerase glyoxalase I in microbial genomes from several pathogenic organisms. *Journal of Molecular Evolution,* 50**,** 491-495.

COOPER, R. A. 1984. Metabolism of Methylglyoxal in Microorganisms. *Annual Review of Microbiology,* 38**,** 49-68.

DEL GIUDICE, I., LIMAURO, D., PEDONE, E., BARTOLUCCI, S. & FIORENTINO, G. 2013. A novel arsenate reductase from the bacterium Thermus thermophilus HB27: Its role in arsenic detoxification. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics,* 1834**,** 2071-2079.

DELEON-RODRIGUEZ, N., LATHEM, T. L., RODRIGUEZ-R, L. M., BARAZESH, J. M., ANDERSON, B. E., BEYERSDORF, A. J., ZIEMBA, L. D., BERGIN, M., NENES, A. & KONSTANTINIDIS, K. T. 2013. Microbiome of the upper troposphere: Species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proceedings of the National Academy of Sciences,* 110**,** 2575-2580.

DI GIULIO, M. 2005. A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene,* 346**,** 1-6.

DIKIY, A., NOVOSELOV, S. V., FOMENKO, D. E., SENGUPTA, A., CARLSON, B. A., CERNY, R. L., GINALSKI, K., GRISHIN, N. V., HATFIELD, D. L. & GLADYSHEV, V. N. 2007. SelT, SelW, SelH, and Rdx12: Genomics and Molecular Insights into the Functions of Selenoproteins of a Novel Thioredoxin-like Family†. *Biochemistry,* 46**,** 6871-6882.

DOLPHIN, D., AVRAMOVIC, O. & POULSON, R. 1989. *Glutathione: chemical, biochemical, and medical aspects, Part A (Coenzymes and Cofactors),* New York, Wiley.

GROß, K.-H. & KALBITZER, H. R. 1988. Distribution of chemical shifts in 1H nuclear magnetic resonance spectra of proteins. *Journal of magnetic resonance* 76**,** 87-99.

HAY, S., EVANS, R. M., LEVY, C., LOVERIDGE, E. J., WANG, X., LEYS, D., ALLEMANN, R. K. & SCRUTTON, N. S. 2009. Are the Catalytic Properties of Enzymes from Piezophilic Organisms Pressure Adapted? *ChemBioChem,* 10**,** 2348-2353.

HE, M. M., CLUGSTON, S. L., HONEK, J. F. & MATTHEWS, B. W. 2000. Determination of the Structure of Escherichia coli Glyoxalase I Suggests a Structural Basis for Differential Metal Activation†. *Biochemistry,* 39**,** 8719-8727.

HEDGPETH, J., GOODMAN, H. M. & BOYER, H. W. 1972. DNA Nucleotide Sequence Restricted by the RI Endonuclease. *Proceedings of the National Academy of Sciences of the United States of America,* 69**,** 3448-3452.

HERSHFIELD, V., BOYER, H. W., YANOFSKY, C., LOVETT, M. A. & HELINSKI, D. R. 1974. Plasmid ColE1 as a Molecular Vehicle for Cloning and Amplification of DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 71**,** 3455-3459.

HEULIN, T., DE LUCA, G., BARAKAT, M., DE GROOT, A., BLANCHARD, L., ORTET, P. & ACHOUAK, W. 2012. Bacterial adaptation to hot and dry deserts. *In:* STAN-LOTTER, H. & FENDRIHAN, S. (eds.) *Adaption of Microbial Life to Environmental Extremes.* Springer Vienna.

HOFFMANN, A. & ROEDER, R. G. 1991. Purification of his-tagged proteins in non-denaturing conditions suggests a convenient method for protein interaction studies. *Nucleic Acids Research,* 19**,** 6337-6338.

HUGHES, M. F. 2002. Arsenic toxicity and potential mechanisms of action. *Toxicology Letters,* 133**,** 1-16.

JI, G., GARBER, E. A. E., ARMES, L. G., CHEN, C.-M., FUCHS, J. A. & SILVER, S. 1994. Arsenate Reductase of Staphylococcus aureus Plasmid pI258. *Biochemistry,* 33**,** 7294-7299.

KALAPOS, M. P. 1999. Methylglyoxal in living organisms: Chemistry, biochemistry, toxicology and biological implications. *Toxicology Letters,* 110**,** 145-175.

KANE, J. F. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Current Opinion in Biotechnology,* 6**,** 494-500.

KARSHIKOFF, A. & LADENSTEIN, R. 2001. Ion pairs and the thermotolerance of proteins from hyperthermophiles: a ‘traffic rule’ for hot roads. *Trends in Biochemical Sciences,* 26**,** 550-557.

KOTLAR, H. K., LEWIN, A., JOHANSEN, J., THRONE-HOLST, M., HAVERKAMP, T., MARKUSSEN, S., WINNBERG, A., RINGROSE, P., AAKVIK, T., RYENG, E., JAKOBSEN, K., DRABLØS, F. & VALLA, S. 2011. High coverage sequencing of DNA from microorganisms living in an oil reservoir 2.5 kilometres subsurface. *Environmental Microbiology Reports,* 3**,** 674-681.

KRAMER, E. B. & FARABAUGH, P. J. 2007. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA,* 13**,** 87-96.

KRONDIRIS, J. V. & SIDERIS, D. C. 2002. Intramolecular disulfide bonding is essential for betanodavirus coat protein conformation. *Journal of General Virology,* 83**,** 2211-2214.

KURLAND, C. & GALLANT, J. 1996. Errors of heterologous protein expression. *Current Opinion in Biotechnology,* 7**,** 489-493.

KWAN, A. H., MOBLI, M., GOOLEY, P. R., KING, G. F. & MACKAY, J. P. 2011. Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal,* 278**,** 687-703.

LAH, N., LAH, J., ZEGERS, I., WYNS, L. & MESSENS, J. 2003. Specific Potassium Binding Stabilizes pI258 Arsenate Reductase from Staphylococcus aureus. *Journal of Biological Chemistry,* 278**,** 24673-24679.

LARSSON, A. 1983. *Functions of glutathione: biochemical, physiological, toxicological, and clinical aspects,* New York, Raven Press.

LEWIN, A., JOHANSEN, J., WENTZEL, A., KOTLAR, H. K., DRABLØS, F. & VALLA, S. 2014. The microbial communities in two apparently physically separated deep subsurface oil reservoirs show extensive DNA sequence similarities. *Environmental Microbiology,* 16**,** 545-558.

LI, R., HAILE, J. D. & KENNELLY, P. J. 2003. An Arsenate Reductase from Synechocystis sp. Strain PCC 6803 Exhibits a Novel Combination of Catalytic Characteristics. *Journal of Bacteriology,* 185**,** 6780-6789.

LIU, J., GLADYSHEVA, T. B., LEE, L. & ROSEN, B. P. 1995. Identification of an Essential Cysteinyl Residue in the ArsC Arsenate Reductase of Plasmid R773. *Biochemistry,* 34**,** 13472-13476.

LIU, Y.-F., ZHANG, N., LIU, X., WANG, X., WANG, Z.-X., CHEN, Y., YAO, H.-W., GE, M. & PAN, X.-M. 2012. Molecular Mechanism Underlying the Interaction of Typical Sac10b Family Proteins with DNA. *PLoS ONE,* 7**,** e34986.

LUIS, M. M., BARRY, R. & JORIS, M. 2012. The arsenic stress defense mechanism of Corynebaterium glutamicum revealed. *Understanding the Geological and Medical Interface of Arsenic - As 2012.* CRC Press.

MACBEATH, G. & KAST, P. 1998. UGA read-through artifacts - when popular gene expression systems need a pATCH. *Biotechniques,* 24**,** 789-94.

MADIGAN, M. T., MARTINKO, J. M., DUNLAP, P. V. & CLARK, D. P. 2009. *Brock Biology of Microorganisms* San Francisco, Pearson Benjamin Cummings.

MAETA, K., IZAWA, S., OKAZAKI, S., KUGE, S. & INOUE, Y. 2004. Activity of the Yap1 Transcription Factor in Saccharomyces cerevisiae Is Modulated by Methylglyoxal, a Metabolite Derived from Glycolysis. *Molecular and Cellular Biology,* 24**,** 8753-8764.

MANNERVIK, B. 1980. Chapter 14 - Glyoxalase I. *In:* JAKOBY, W. B. (ed.) *Enzymatic Basis of Detoxication.* Academic Press.

MELZAK, K. A., SHERWOOD, C. S., TURNER, R. F. B. & HAYNES, C. A. 1996. Driving Forces for DNA Adsorption to Silica in Perchlorate Solutions. *Journal of Colloid and Interface Science,* 181**,** 635-644.

MESSENS, J., HAYBURN, G., DESMYTER, A., LAUS, G. & WYNS, L. 1999. The Essential Catalytic Redox Couple in Arsenate Reductase from Staphylococcus aureus. *Biochemistry,* 38**,** 16857-16865.

MESSENS, J., MARTINS, J. C., VAN BELLE, K., BROSENS, E., DESMYTER, A., DE GIETER, M., WIERUSZESKI, J.-M., WILLEM, R., WYNS, L. & ZEGERS, I. 2002. All intermediates of the arsenate reductase mechanism, including an intramolecular dynamic disulfide cascade. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 8506-8511.

MESSENS, J. & SILVER, S. 2006. Arsenate Reduction: Thiol Cascade Chemistry with Convergent Evolution. *Journal of Molecular Biology,* 362**,** 1-17.

MOTTAGUI-TABAR, S., BJÖRNSSON, A. & ISAKSSON, L. A. 1994. The second to last amino acid in the nascent peptide as a codon context determinant. *The EMBO Journal,* 13**,** 249-257.

MUKHERJEE, A., BHATTACHARYA, P., SAVAGE, K., FOSTER, A. & BUNDSCHUH, J. 2008. Distribution of geogenic arsenic in hydrologic systems: Controls and challenges. *Journal of Contaminant Hydrology,* 99**,** 1-7.

MUKHOPADHYAY, R. & ROSEN, B. P. 2002. Arsenate reductases in prokaryotes and eukaryotes. *Environmental Health Perspectives,* 110**,** 745-748.

ORDÓÑEZ, E., VAN BELLE, K., ROOS, G., DE GALAN, S., LETEK, M., GIL, J. A., WYNS, L., MATEOS, L. M. & MESSENS, J. 2009. Arsenate Reductase, Mycothiol, and Mycoredoxin Concert Thiol/Disulfide Exchange. *Journal of Biological Chemistry,* 284**,** 15107-15116.

OTTING, G. 2010. Protein NMR Using Paramagnetic Ions. *Annual Review of Biophysics,* 39**,** 387-405.

PHILLIPS, T. A., VANBOGELEN, R. A. & NEIDHARDT, F. C. 1984. lon gene product of Escherichia coli is a heat-shock protein. *Journal of Bacteriology,* 159**,** 283-287.

RACKER, E. 1951. The Mechanism of Action of Glyoxalase. *Journal of Biological Chemistry,* 190**,** 685-696.

RALEIGH, E. A. & WILSON, G. 1986. Escherichia coli K-12 restricts DNA containing 5-methylcytosine. *Proceedings of the National Academy of Sciences of the United States of America,* 83**,** 9070-9074.

REED, C. J., LEWIS, H., TREJO, E., WINSTON, V. & EVILIA, C. 2013. Protein Adaptations in Archaeal Extremophiles. *Archaea,* 2013**,** 14.

ROOS, G., BUTS, L., VAN BELLE, K., BROSENS, E., GEERLINGS, P., LORIS, R., WYNS, L. & MESSENS, J. 2006. Interplay Between Ion Binding and Catalysis in the Thioredoxin-coupled Arsenate Reductase Family. *Journal of Molecular Biology,* 360**,** 826-838.

ROTHSCHILD, L. J. & MANCINELLI, R. L. 2001. Life in extreme environments. *Nature,* 409**,** 1092-1101.

SAINT-JEAN, A. P., PHILLIPS, K. R., CREIGHTON, D. J. & STONE, M. J. 1998. Active Monomeric and Dimeric Forms of Pseudomonas putida Glyoxalase I: Evidence for 3D Domain Swapping. *Biochemistry,* 37**,** 10345-10353.

SHAPIRO, A. L., VIÑUELA, E. & MAIZEL, J. V., JR. 1967. Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochemical and Biophysical Research Communications,* 28**,** 815-820.

SHI, L., POTTS, M. & KENNELLY, P. J. 1998. The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. *FEMS Microbiology Reviews,* 22**,** 229-253.

SIMONATO, F., CAMPANARO, S., LAURO, F. M., VEZZI, A., D'ANGELO, M., VITULO, N., VALLE, G. & BARTLETT, D. H. 2006. Piezophilic adaptation: a genomic point of view. *Journal of Biotechnology,* 126**,** 11-25.

SINGLA-PAREEK, S. L., REDDY, M. K. & SOPORY, S. K. 2003. Genetic engineering of the glyoxalase pathway in tobacco leads to enhanced salinity tolerance. *Proceedings of the National Academy of Sciences,* 100**,** 14672-14677.

SINGLA-PAREEK, S. L., YADAV, S. K., PAREEK, A., REDDY, M. K. & SOPORY, S. K. 2006. Transgenic Tobacco Overexpressing Glyoxalase Pathway Enzymes Grow and Set Viable Seeds in Zinc-Spiked Soils. *Plant Physiology,* 140**,** 613-623.

SMITH, H. O. & WELCOX, K. W. 1970. A Restriction enzyme from Hemophilus influenzae: I. Purification and general properties. *Journal of Molecular Biology,* 51**,** 379-391.

STEVENS, S. Y., HU, W., GLADYSHEVA, T., ROSEN, B. P., ZUIDERWEG, E. R. P. & LEE, L. 1999. Secondary Structure and Fold Homology of the ArsC Protein from the Escherichia coli Arsenic Resistance Plasmid R773. *Biochemistry,* 38**,** 10178-10186.

STREIT, W. R., DANIEL, R. & JAEGER, K.-E. 2004. Prospecting for biocatalysts and drugs in the genomes of non-cultured microorganisms. *Current Opinion in Biotechnology,* 15**,** 285-290.

STUDIER, F. W. & MOFFATT, B. A. 1986. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology,* 189**,** 113-130.

SU, X.-D., TADDEI, N., STEFANI, M., RAMPONI, G. & NORDLUND, P. 1994. The crystal structure of a low-molecular-weight phosphotyrosine protein phosphatase. *Nature,* 370**,** 575-578.

SUKDEO, N., CLUGSTON, SUSAN L., DAUB, E. & HONEK, JOHN F. 2004. Distinct classes of glyoxalase I: metal specificity of the Yersinia pestis, Pseudomonas aeruginosa and Neisseria meningitidis enzymes. *Biochemical Journal,* 384**,** 111-117.

SUTCLIFFE, J. G. 1978. Nucleotide sequence of the ampicillin resistance gene of Escherichia coli plasmid pBR322. *Proceedings of the National Academy of Sciences of the United States of America,* 75**,** 3737-3741.

SUTTISANSANEE, U. & HONEK, J. F. 2011. Bacterial glyoxalase enzymes. *Seminars in Cell & Developmental Biology,* 22**,** 285-292.

SUTTISANSANEE, U., LAU, K., LAGISHETTY, S., RAO, K. N., SWAMINATHAN, S., SAUDER, J. M., BURLEY, S. K. & HONEK, J. F. 2011. Structural Variation in Bacterial Glyoxalase I Enzymes: INVESTIGATION OF THE METALLOENZYME GLYOXALASE I FROM CLOSTRIDIUM ACETOBUTYLICUM. *The Journal of Biological Chemistry,* 286**,** 38367-38374.

TABOR, S. 2001. Expression Using the T7 RNA Polymerase/Promoter System. *Current Protocols in Molecular Biology.* John Wiley & Sons, Inc.

TANAKA, N., IKEDA, C., KANAORI, K., HIRAGA, K., KONNO, T. & KUNUGI, S. 2000. Pressure Effect on the Conformational Fluctuation of Apomyoglobin in the Native State. *Biochemistry,* 39**,** 12063-12068.

TERPE, K. 2006. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology,* 72**,** 211-222.

THORNALLEY, P. J. 1990. The glyoxalase system: new developments towards functional characterization of a metabolic pathway fundamental to biological life. *Biochemical Journal,* 269**,** 1-11.

THORNALLEY, P. J. 2003. Glyoxalase I - structure, function and a critical role in the enzymatic defence against glycation. *Biochemical Society Transactions* 31**,** 1343–1348.

THORNE, H. V. 1966. Electrophoretic separation of polyoma virus DNA from host cell DNA. *Virology,* 29**,** 234-239.

TOMAZIC, S. J. & KLIBANOV, A. M. 1988. Mechanisms of irreversible thermal inactivation of Bacillus alpha-amylases. *Journal of Biological Chemistry,* 263**,** 3086-3091.

UNSWORTH, L. D., VAN DER OOST, J. & KOUTSOPOULOS, S. 2007. Hyperthermophilic enzymes − stability, activity and implementation strategies for high temperature applications. *FEBS Journal,* 274**,** 4044-4056.

VIEILLE, C. & ZEIKUS, G. J. 2001. Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews,* 65**,** 1-43.

VILLADANGOS, A., ORDÓÑEZ, E., PEDRE, B., MESSENS, J., GIL, J. & MATEOS, L. 2014. Engineered coryneform bacteria as a bio-tool for arsenic remediation. *Applied Microbiology and Biotechnology,* 98**,** 10143-10152.

VILLADANGOS, A. F., VAN BELLE, K., WAHNI, K., TAMU DUFE, V., FREITAS, S., NUR, H., DE GALAN, S., GIL, J. A., COLLET, J.-F., MATEOS, L. M. & MESSENS, J. 2011. Corynebacterium glutamicum survives arsenic stress with arsenate reductases coupled to two distinct redox mechanisms. *Molecular Microbiology,* 82**,** 998-1014.

VON FREIESLEBEN, U., KREKLING, M. A., HANSEN, F. G. & LØBNER-OLESEN, A. 2000. *The eclipse period of Escherichia coli*.

WEBER, K. & OSBORN, M. 1969. The Reliability of Molecular Weight Determinations by Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis. *Journal of Biological Chemistry,* 244**,** 4406-4412.

WISHART, D., BIGAM, C., HOLM, A., HODGES, R. & SYKES, B. 1995. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of Biomolecular NMR,* 5**,** 67-81.

WOLFENDEN, R. 2006. Degrees of Difficulty of Water-Consuming Reactions in the Absence of Enzymes. *Chemical Reviews,* 106**,** 3379-3396.

YIP, T.-T. & HUTCHENS, T. W. 1992. Immobilized Metal Ion Affinity Chromatography. *In:* KENNEY, A. & FOWELL, S. (eds.) *Practical Protein Chromatography.* Humana Press.

YU, C., XIA, B. & JIN, C. 2011. 1H, 13C and 15N resonance assignments of the arsenate reductase from Synechocystis sp. strain PCC 6803. *Biomolecular NMR Assignments,* 5**,** 85-87.

ZEGERS, I., MARTINS, J. C., WILLEM, R., WYNS, L. & MESSENS, J. 2001. Arsenate reductase from S. aureus plasmid pI258 is a phosphatase drafted for redox duty. *Nat Struct Mol Biol,* 8**,** 843-847.

ZIMMERMAN, S. B., LITTLE, J. W., OSHINSKY, C. K. & GELLERT, M. 1967. Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide. *Proceedings of the National Academy of Sciences of the United States of America,* 57**,** 1841-1848.

Appendix A. Advanced search criteria used in UniProt

(length:[50 TO 150] NOT taxonomy:viruses NOT taxonomy:Eukaryota NOT name:uncharacterized NOT name:putative NOT name:hypothetical NOT name:probable AND existence:"evidence at protein level" NOT name:transmembrane NOT name:subunit NOT annotation:(type:function protease) NOT annotation:(type:ptm) NOT annotation:(type:signal) AND precursor:no) AND reviewed:yes NOT go:"DNA binding" NOT name:"ribosomal protein" NOT name:UPF NOT name:ferredoxin

- Length was set to 50-150 amino acids to make sure the proteins was < 20 kDa
- It was easiest to exclude viruses and eukaryotes to make sure it was prokaryotes (both bacteria and archaea)
- Words as uncharacterized, putative and hypothetical and probable were excluded from the search to greater the chance of finding functional candidate proteins of the right class
- Existence was set to evidence at protein level to greater the chance of finding functional candidate proteins of the right class
- Transmembrane proteins were excluded from the search since solid state NMR are preferred to study such enzymes and fluid state NMR was to be used for this thesis
- Subunits, proteins dependent on post-translational modification (ptm), precursors and proteins with signal peptides were excluded from the search to greater the chances of functional proteins
- Reviewed proteins were preferred to greater the chance of functional proteins
- Protease proteins were excluded from the search since their activity could destroy the NMR samples
- DNA binding proteins, ribosomal proteins UPF were excluded from the search due to wrong type of activity

Appendix B. Specific bioinformatic results for each candidate

Ars1

The first candidate of arsenate reductase was found to be encoded on a small (1,692 bp long) contig derived from Well II that was annotated as arsenical-resistance membrane protein from *Thermococcus kodakarensis*. The candidate aligned with the UniProt query arsenate-mycothiol transferase had a 33 % identity to 97 % of the query sequence, meaning that the two proteins are far from identical but they can still be analogous. The gene was found as the only ORF in the -1 frame, located against one end of the contig and was shown to lack a stop codon, suggesting that the gene may be truncated.

When verifying the candidate, the blastp results showed six hits with an alignment score above 200 and several others with lower scores. Four of these had over 90 % identity and all of them had sequences that were at least 13 amino acids longer than the candidate sequence, also indicating that the candidate is likely truncated. The best hit was from *Thermococcus litoralis* (Figure B.1) with 99 % identity to the full length of the query. Since this reference protein was so similar to the candidate, the remaining sequence length of this protein was used to substitute the missing amino acid sequence on the Ars1 candidate, resulting in:

MEEKLILFVCVKNSARSQMAEAFFNHFNDDPRFKAMSAGTEPAEEIDPLAKKVMEEIGISLEGQYPKLYTEEMADK
AYIVITMGCLDKCPYAPPEKTWDWGLDDPYGQPMEKYREVRDEIKRRVLKLIEDLKAGKSREEIIGRKSLFTL

The letters in blue indicate the sequence which was added from the reference protein.

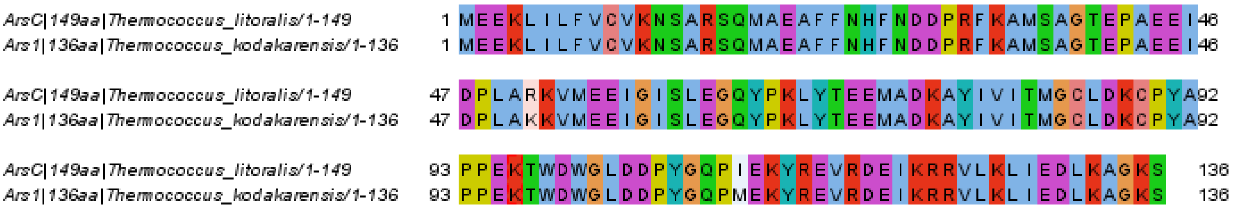

ArsC|149aa|Thermococcus_litoralis/1-149 1 MEEKLILFVCVKNSARSQMAEAFFNHFNDDPRFKAMSAGTEPAEEI 46
Ars1|136aa|Thermococcus_kodakarensis/1-136 1 MEEKLILFVCVKNSARSQMAEAFFNHFNDDPRFKAMSAGTEPAEEI 46
ArsC|149aa|Thermococcus_litoralis/1-149 47 DPLARKVMEEIGISLEGQYPKLYTEEMADKAYIVITMGCLDKCPYA 92
Ars1|136aa|Thermococcus_kodakarensis/1-136 47 DPLAKKVMEEIGISLEGQYPKLYTEEMADKAYIVITMGCLDKCPYA 92
ArsC|149aa|Thermococcus_litoralis/1-149 93 PPEKTWDWGLDDPYGQPIEKYREVRDEIKRRVLKLIEDLKAGKS 136
Ars1|136aa|Thermococcus_kodakarensis/1-136 93 PPEKTWDWGLDDPYGQPMEKYREVRDEIKRRVLKLIEDLKAGKS 136

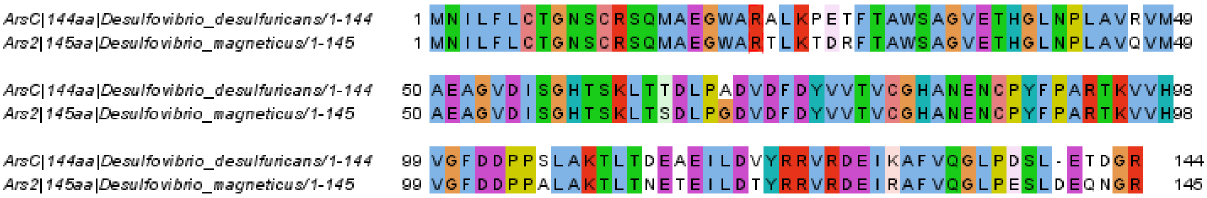
Figure B.1: A pairwise alignment made in Jalview 2.8.2 of the likely truncated amino acid sequence of Ars1 obtained from the contig together with the sequence of arsenate reductase from *T. litoralis*. The colors of the alignment indicate the nature of the amino acid; hydrophobic (red), acidic (blue), basic (magenta), or adhered to polar uncharged hydroxyl, sulfhydryl or amine side chains together with glycine (green). The color intensity fades as the alignments become less conserved.

Ars2

The second arsenate reductase candidate was found to be encoded on a large (10,011 bp long) contig derived from Well I that was annotated as putative thiosulfate reductase from *Desulfovibrio magneticus* RS-1. The candidate sequence aligned with the UniProt query had 31 % identity to 96 % of the complete query sequence, which also indicate that the proteins may contain similar activity. The candidate was positioned in frame -1 which contained six ORFs.

When verifying the candidate, the blastp results showed nineteen hits with an alignment score above or equal to 200 bits. Only one of these had query coverage of 100 %, and none had over 90 % identity. Most of the best hits had the same length as the candidate or shorter, indicating that the candidate amino acid sequence is likely intact. The best hit was from *Desulfovibrio desulfuricans* with 89 % identity to the full length of the query (Figure B.2). The resulting amino acid sequence for Ars2 retrieved from the contig was:

MNILFLCTGNSCRSQMAEGWARTLKTDRFTAWSAGVETHGLNPLAVQVMAEAGVDISGHTSKLTSDLPGDVDFD
YVVTVCGHANENCPYFPARTKVVHVGFDDPPALAKTLTNETEILDTYRRVRDEIRAFVQGLPESLDEQNGR



```
ArsC|144aa|Desulfovibrio_desulfuricans/1-144    1 MNILFLCTGNSCRSQMAEGWARALKPETFTAWSAGVETHGLNPLAVRVM49
Ars2|145aa|Desulfovibrio_magneticus/1-145       1 MNILFLCTGNSCRSQMAEGWARTLKTDRFTAWSAGVETHGLNPLAVQVM49

ArsC|144aa|Desulfovibrio_desulfuricans/1-144   50 AEAGVDISGHTSKLTTDLPADVDFDYVVTVCGHANENCPYFPARTKVVH98
Ars2|145aa|Desulfovibrio_magneticus/1-145      50 AEAGVDISGHTSKLTSDLPGDVDFDYVVTVCGHANENCPYFPARTKVVH98

ArsC|144aa|Desulfovibrio_desulfuricans/1-144   99 VGFDDPPSLAKTLTDEAEILDVYRRVRDEIKAFVQGLPDSL-ETDGR   144
Ars2|145aa|Desulfovibrio_magneticus/1-145      99 VGFDDPPALAKTLTNETEILDTYRRVRDEIRAFVQGLPESLDEQNGR   145
```

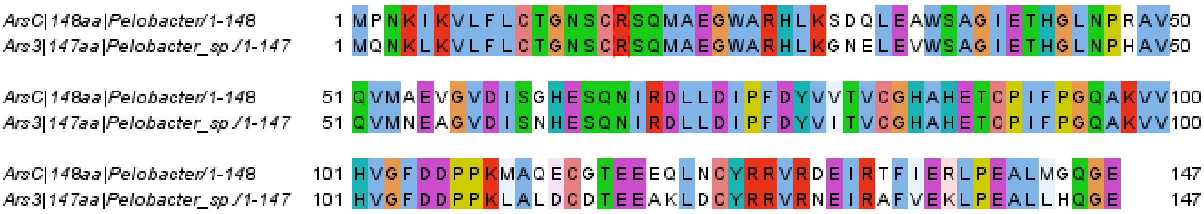
Figure B.2: A pairwise alignment in Jalview 2.8.2 of the amino acid sequence of candidate Ars2 and the arsenate reductase sequence from *D. desulfuricans*. The colors of the alignment indicate the nature of the amino acid; hydrophobic (red), acidic (blue), basic (magenta), or adhered to polar uncharged hydroxyl, sulfhydryl or amine side chains together with glycine (green). The color intensity fades as the alignments become less conserved.

Ars3

The third candidate of arsenate reductase was found to be encoded in the translated metagenomic sequences of both Well I and 2. The corresponding gene sequence was retrieved from a medium sized (6,570 bp long) contig from Well II that was annotated to encode GTPase and sulfate adenylate transferase subunit 1 from *Pelobacter* sp. The candidate in an alignment with the arsenate-mycothiol transferase had 34 % identity to 96 % of the full query sequence, which may indicate that the proteins have a similar function. The gene was located in frame -3 which enclosed nine ORFs.

The blastp results showed eighteen hits with an alignment score above or equal to 200 bits. Only one of these had query coverage of 100 % and none had over 90 % identity. Most of the hits were about the same length as the candidate, which may indicate a full sequence. The best hit was from *Pelobacter carbinolicus* with an 84 % identity to the total query sequence (Figure B.3). The amino acid sequence for Ars3 retrieved from the contig was:

MQNKLKVLFLCTGNSCRSQMAEGWARHLKGNELEVWSAGIETHGLNPHAVQVMNEAGVDISNHESQNIRDLLDIPFDYVITVCGHAHETCPIFPGQAKVVHVGFDDPPKLALDCDTEEAKLDCYRRVRNEIRAFVEKLPEALLHQGE



```
ArsC|148aa|Pelobacter/1-148     1 MPNKIKVLFLCTGNSCRSQMAEGWARHLKSDQLEAWSAGIETHGLNPRAV50
Ars3|147aa|Pelobacter_sp./1-147 1 MQNKLKVLFLCTGNSCRSQMAEGWARHLKGNELEVWSAGIETHGLNPHAV50

ArsC|148aa|Pelobacter/1-148     51 QVMAEVGVDISGHESQNIRDLLDIPFDYVVTVCGHAHETCPIFPGQAKVV100
Ars3|147aa|Pelobacter_sp./1-147 51 QVMNEAGVDISNHESQNIRDLLDIPFDYVITVCGHAHETCPIFPGQAKVV100

ArsC|148aa|Pelobacter/1-148     101 HVGFDDPPKMAQECGTEEEQLNCYRRVRDEIRTFIERLPEALMGQGE   147
Ars3|147aa|Pelobacter_sp./1-147 101 HVGFDDPPKLALDCDTEEAKLDCYRRVRNEIRAFVEKLPEALLHQGE   147
```

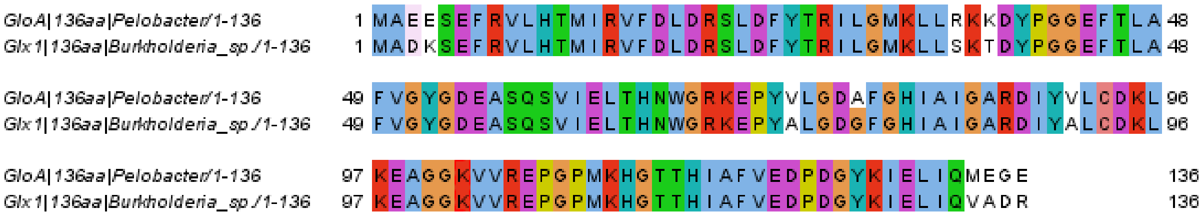
Figure B.3: A pairwise alignment in Jalview 2.8.2 of the amino acid sequence of the Ars3 candidate and the arsenate reductase sequence from *Pelobacter* sp. The colors of the alignment indicate the nature of the amino acid; hydrophobic (red), acidic (blue), basic (magenta), or adhered to polar uncharged hydroxyl, sulfhydryl or amine side chains together with glycine (green). The color intensity fades as the alignments become less conserved.

Glx1

The first candidate of glyoxalase 1 was found to be encoded in the translated metagenomes of both Well I and 2. The corresponding gene sequence was extracted from a large (32,034 bp long) contig from Well II that were annotated as a type I restriction-modification system, R subunit from *Burkholderia* sp. The candidate sequence in an alignment with the sequence of lactoylglutathione lyase 1 found in UniProt showed 56 % identity to 92 % of the complete query sequence, which may indicate homology. The gene was positioned in frame +2 which contained thirty-five ORFs.

The verification showed only one hit with an alignment score above 200 and the rest had a score of 185 or larger. Most of these hits were about the same length or shorter than the candidate sequence, indicating that the Glx1 gene sequence is likely complete. The best hit was from *Pelobacter carbinolicus* with 94 % identity to 97 % of the full query sequence (Figure B.4). The resulting amino acid sequence of Glx1 retrieved from the contig was:

MADKSEFRVLHTMIRVFDLDRSLDFYTRILGMKLLSKTDYPGGEFTLAFVGYGDEASQSVIELTHNWGRKEPYALGD
GFGHIAIGARDIYALCDKLKEAGGKVVREPGPMKHGTTHIAFVEDPDGYKIELIQVADR



```
GloA|136aa|Pelobacter/1-136        1 MAEESEFRVLHTMIRVFDLDRSLDFYTRILGMKLLRKKDYPGGEFTLA 48
Glx1|136aa|Burkholderia_sp./1-136  1 MADKSEFRVLHTMIRVFDLDRSLDFYTRILGMKLLSKTDYPGGEFTLA 48

GloA|136aa|Pelobacter/1-136       49 FVGYGDEASQSVIELTHNWGRKEPYVLGDAFGHIAIGARDIYVLCDKL 96
Glx1|136aa|Burkholderia_sp./1-136 49 FVGYGDEASQSVIELTHNWGRKEPYALGDGFGHIAIGARDIYALCDKL 96

GloA|136aa|Pelobacter/1-136       97 KEAGGKVVREPGPMKHGTTHIAFVEDPDGYKIELIQMEGE        136
Glx1|136aa|Burkholderia_sp./1-136 97 KEAGGKVVREPGPMKHGTTHIAFVEDPDGYKIELIQVADR        136
```

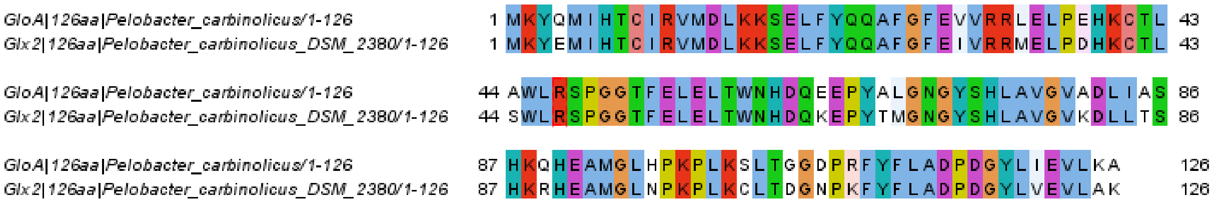
Figure B.4: A pairwise alignment in Jalview 2.8.2 of the amino acid sequence of Glx1 and the lactoylglutathione lyase 1 sequence from *P. carbinolicus*. The colors of the alignment indicate the nature of the amino acid; hydrophobic (red), acidic (blue), basic (magenta), or adhered to polar uncharged hydroxyl, sulfhydryl or amine side chains together with glycine (green). The color intensity fades as the alignments become less conserved.

Glx2

The second glyoxalase 1 candidate was also found to be encoded in the translated metagenomic data derived from Well I and 2. The gene sequence was retrieved from a large (50,843 bp long) contig from Well II that were annotated as hypothetical protein Pcar_1724 from *Pelobacter carbinolicus* DSM 2380. The sequence alignment with the UniProt query showed 35 % identity to 91 % of the full query sequence, which may place it in the same protein class as the query. The gene was positioned in frame -1 which encompassed fifty-seven ORFs.

The verification showed only one hit with an alignment score above 200 and the rest from 124 to 192 bits. Most hits had about the same length as the candidate, indicating a complete amino acid sequence. The best hit was, as Glx1, *P. carbinolicus* with 85 % identity to 98 % of the entire query sequence (Figure B.5). The fully retrieved amino acid sequence from the contig was:

MKYEMIHTCIRVMDLKKSELFYQQAFGFEIVRRMELPDHKCTLSWLRSPGGTFELELTWNHDQKEPYTMGNGYSH
LAVGVKDLLTSHKRHEAMGLNPKPLKCLTDGNPKFYFLADPDGYLVEVLAK



```
GloA|126aa|Pelobacter_carbinolicus/1-126            1 MKYQMIHTCIRVMDLKKSELFYQQAFGFEVVRRLELPEHKCTL  43
Glx2|126aa|Pelobacter_carbinolicus_DSM_2380/1-126   1 MKYEMIHTCIRVMDLKKSELFYQQAFGFEIVRRMELPDHKCTL  43

GloA|126aa|Pelobacter_carbinolicus/1-126           44 AWLRSPGGTFELELTWNHDQEEPYALGNGYSHLAVGVADLIAS  86
Glx2|126aa|Pelobacter_carbinolicus_DSM_2380/1-126  44 SWLRSPGGTFELELTWNHDQKEPYTMGNGYSHLAVGVKDLLTS  86

GloA|126aa|Pelobacter_carbinolicus/1-126           87 HKQHEAMGLHPKPLKSLTGGDPRFYFLADPDGYLIEVLKA    126
Glx2|126aa|Pelobacter_carbinolicus_DSM_2380/1-126  87 HKRHEAMGLNPKPLKCLTDGNPKFYFLADPDGYLVEVLAK    126
```

Figure B.5: A pairwise alignment in Jalview 2.8.2 of the Glx2 amino acid sequence and the lactoylglutathione lyase 1 sequence from *P. carbinolicus*. The colors of the alignment indicate the nature of the amino acid; hydrophobic (red), acidic (blue), basic (magenta), or adhered to polar uncharged hydroxyl, sulfhydryl or amine side chains together with glycine (green). The color intensity fades as the alignments become less conserved.

Appendix C. Genetic framework and optimized genes

____ – Start or stop codon

* – Cut site

Genetic framework:

🟦 – BclI binding site

🟥 – BglII binding site

🟪 – NdeI binding site

🟩 – XhoI binding site

■ – 6*His-tag

⬛ – Protective bases

Ars1

CCAATCA*T**ATG**GAAGAAAAACTGATCCTGTTCGTCTGTGTGAAAAACTCGGCCCGTTCGCA
AATGGCGGAAGCGTTCTTTAACCACTTTAATGATGACCCGCGTTTTAAAGCGATGAGTGCCG
GCACCGAACCGGCGGAAGAAATTGATCCGCTGGCCAAAAAAGTCATGGAAGAAATTGGCATC
AGCCTGGAAGGCCAGTATCCGAAACTGTACACCGAAGAAATGGCAGATAAAGCTTACATCGT
TATCACGATGGGTTGCCTGGACAAATGTCCGTACGCACCGCCGGAAAAAACCTGGGATTGGG
GCCTGGATGACCCGTATGGTCAACCGATGGAAAAATACCGTGAAGTGCGCGACGAAATTAAA
CGTCGCGTTCTGAAACTGATCGAAGACCTGAAAGCCGGCAAAAGCCGTGAAGAAATTATCGG
TCGCAAATCTCTGTTCACGCTGA*GATCT**T*GA**TCACATCATCATCATCATCAT**TAA**C*TCG
AGATTGG

Same orientation as *LacZ* when inserted in pUC57

Ars2

CCAATCA*T**ATG**AATATCCTGTTCCTGTGTACGGGTAACTCCTGTCGCTCGCAAATGGCTGA
AGGCTGGGCACGCACGCTGAAAACCGACCGCTTTACCGCGTGGTCCGCCGGCGTTGAAACGC
ATGGTCTGAACCCGCTGGCAGTTCAGGTCATGGCGGAAGCCGGCGTCGACATTAGCGGTCAC
ACCTCTAAACTGACGAGTGATCTGCCGGGCGATGTGGACTTTGATTATGTGGTTACCGTTTG
CGGTCATGCAAACGAAAATTGTCCGTACTTTCCGGCTCGTACCAAAGTCGTGCACGTGGGTT
TCGATGACCCGCCGGCACTGGCTAAAACCCTGACGAACGAAACCGAAATTCTGGACACGTAT
CGTCGCGTCCGTGATGAAATCCGTGCATTCGTGCAGGGTCTGCCGGAAAGCCTGGATGAACA
AAATGGTA*GATCT**T*GA**TCACATCATCATCATCATCAT**TAA**C*TCGAGATTGG

Same orientation as *LacZ* when inserted in pUC57

Ars3

```
TCA*TATGCAAAATAAACTGAAAGTGCTGTTCCTGTGTACGGGCAACTCCTGTCGCTCGCAG
ATGGCAGAAGGCTGGGCTCGTCACCTGAAAGGCAACGAACTGGAAGTCTGGAGTGCAGGCAT
CGAAACCCATGGTCTGAACCCGCACGCGGTCCAGGTGATGAATGAAGCCGGTGTTGATATCA
GCAACCATGAATCTCAAAACATTCGTGATCTGCTGGACATCCCGTTTGACTATGTTATTACC
GTCTGCGGCCATGCACACGAAACGTGTCCGATCTTTCCGGGCCAGGCTAAAGTGGTTCACGT
GGGTTTCGATGACCCGCCGAAACTGGCACTGGATTGCGACACGGAAGAAGCTAAACTGGATT
GTTACCGTCGCGTGCGTAATGAAATTCGCGCGTTCGTTGAAAAACTGCCGGAAGCCCTGCTG
CATCAAGGTGAAA*GATCTT*GATCACATCATCATCATCATCATTAAC*TCGAGATTGG
```

Opposite orientation of *LacZ* when inserted in pUC57

Glx1

```
TCA*TATGGCTGATAAAAGTGAATTTCGCGTGCTGCATACGATGATTCGCGTCTTCGACCTG
GACCGCTCTCTGGATTTCTACACCCGCATTCTGGGTATGAAACTGCTGAGTAAAACCGATTA
TCCGGGCGGTGAATTTACGCTGGCATTCGTCGGCTACGGTGACGAAGCTAGCCAGTCTGTGA
TTGAACTGACCCATAACTGGGGTCGTAAAGAACCGTATGCACTGGGCGATGGTTTTGGCCAC
ATTGCGATCGGCGCCCGCGATATCTACGCGCTGTGCGACAAACTGAAAGAAGCCGGCGGTAA
AGTGGTTCGTGAACCGGGTCCGATGAAACATGGCACCACGCACATTGCATTCGTTGAAGATC
CGGACGGCTATAAAATTGAACTGATCCAAGTCGCTGATA*GATCTT*GATCACATCATCATC
ATCATCATTAAC*TCGAGATTGG
```

Opposite orientation of *LacZ* when inserted in pUC57

Glx2

```
CCAATCA*TATGAAATACGAAATGATCCACACCTGCATCCGCGTTATGGACCTGAAAAAATC
TGAACTGTTCTACCAGCAAGCGTTTGGCTTTGAAATTGTTCGTCGCATGGAACTGCCGGATC
ATAAATGCACCCTGAGCTGGCTGCGTTCTCCGGGCGGTACCTTTGAACTGGAACTGACGTGG
AACCACGATCAGAAAGAACCGTATACCATGGGCAATGGTTACAGTCACCTGGCAGTGGGTGT
TAAAGACCTGCTGACGAGCCATAAACGCCACGAAGCGATGGGCCTGAACCCGAAACCGCTGA
AATGTCTGACGGACGGTAATCCGAAATTTTATTTCCTGGCGGATCCGGACGGCTACCTGGTC
GAAGTGCTGGCCAAAA*GATCTT*GATCACATCATCATCATCATCATTAAC*TCGAGATTGG
```

Same orientation as *LacZ* when inserted in pUC57

Appendix D. Ladders used for gel electrophoresis and SDS-PAGE

DNA ladder:

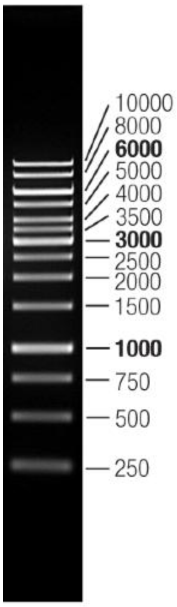


Figure D.1: A picture of the expected band pattern for the O'GeneRuler 1kb DNA ladder used for this thesis with the length for each DNA fragment in the pattern given in bp.

Protein ladder:

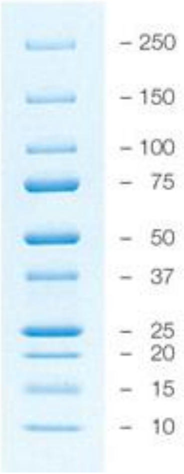


Figure D.2: A picture of the expected band pattern for the Precision Plus Protein™ All blue ladder used for this thesis with the length for each protein size in the pattern given in kDa.

Appendix E. His-tag modification

Ars1

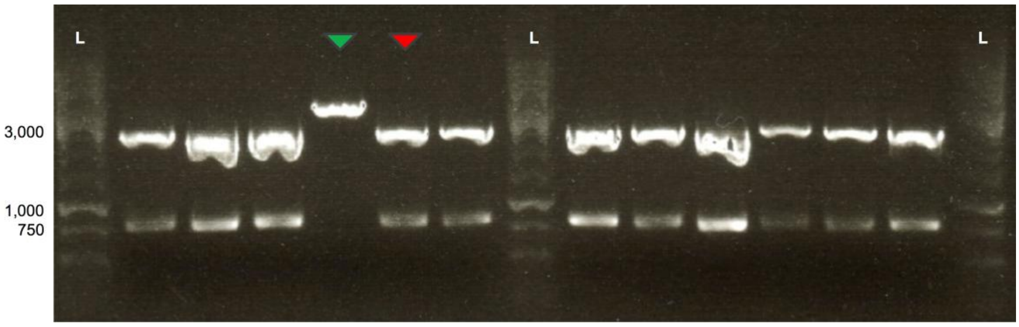


Figure E.1: The gel results for all twelve clones of pUC57-Ars1 after digestion with BglII and PciI. Modified plasmids are expected to contain one fragment of 3,203 bp while unmodified are expected to contain two fragments of 859 and 2,350 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (modified and unmodified respectively) shown in the results.

Ars2

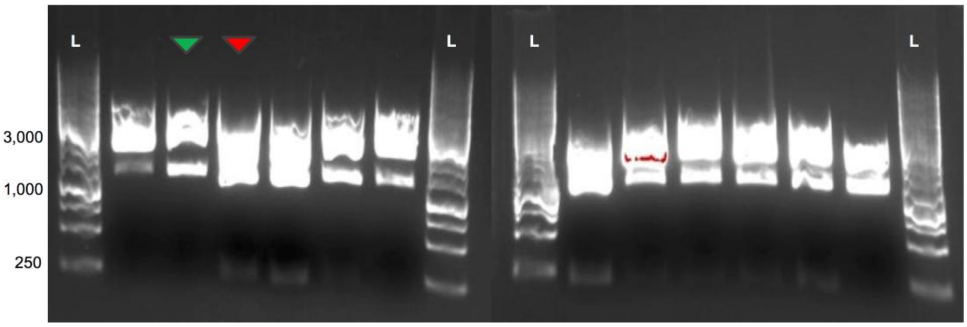


Figure E.2: The gel results for all twelve clones of pUC57-Ars2 after digestion with BglII and BglI. Modified plasmids are expected to contain two fragments of 1,118 and 2,070 bp while unmodified are expected to contain three fragments of 219, 1,118 and 1,857 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (modified and unmodified respectively) shown in the results.

Ars3

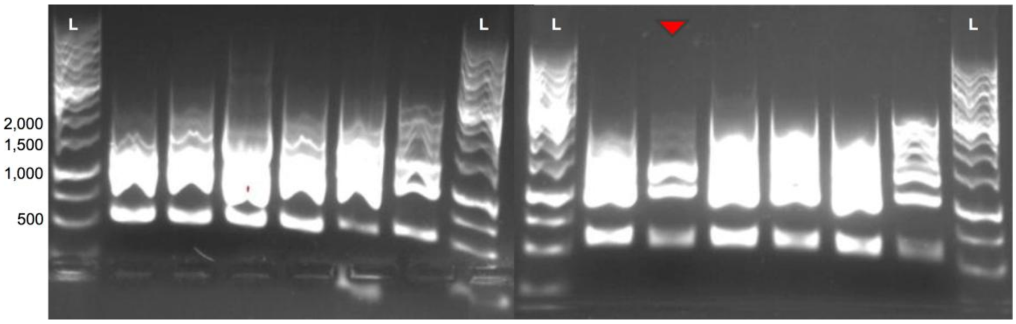


Figure E.3: The gel results for all twelve clones of pUC57-Ars3 after digestion with BglII and BglI. Modified plasmids are expected to contain two fragments of 1,118 and 2,075 bp while unmodified are expected to contain three fragments of 626, 1,118 and 1,455 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The red marker indicates the unmodified vector shown in the results.

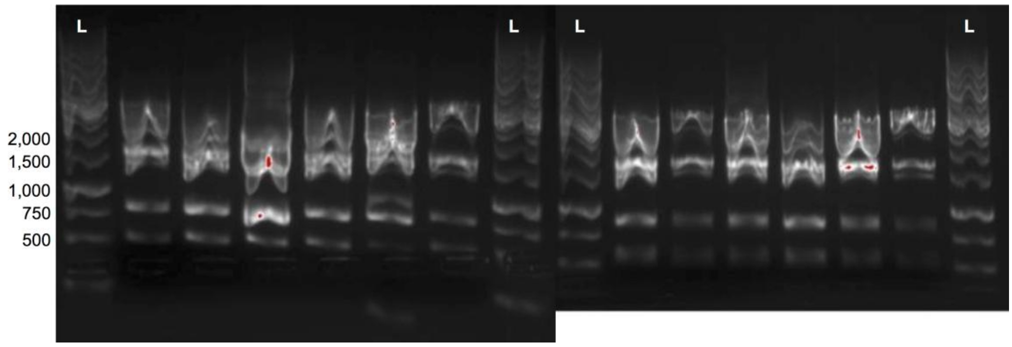


Figure E.4: The gel results for all twelve clones of pUC57-Ars3 after digestion with BglII and PvuI in the second verification. Modified plasmids are expected to contain two fragments of 896 and 2,297 bp while unmodified are expected to contain three fragments of 598, 896 and 1,705 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp.

Glx1

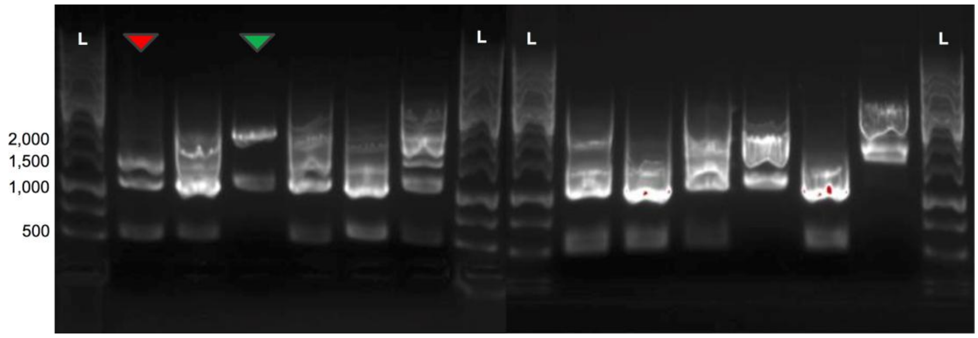


Figure E.5: The gel results for all twelve clones of pUC57-Glx1 after digestion with BglII and BglI. Modified plasmids are expected to contain two fragments of 1,118 and 2,039 bp while unmodified are expected to contain three fragments of 590, 1,118 and 1,455 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (modified and unmodified respectively) shown in the results.

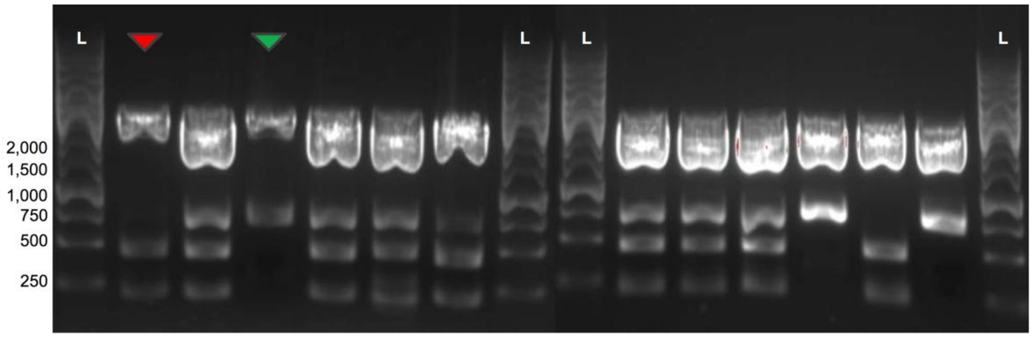


Figure E.6: The gel results for all twelve clones of pUC57-Glx1 after digestion with BglII and PvuII in the second verification. Modified plasmids are expected to contain two fragments of 793 and 2,364 bp (red markers) while unmodified are expected to contain three fragments of 266, 533 and 2,364 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (modified and unmodified respectively) shown in the results.

Glx2

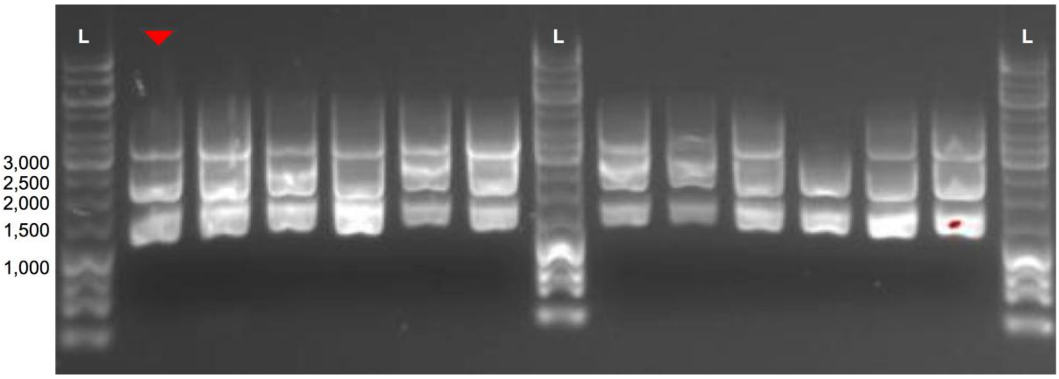


Figure E.7: The gel results for all twelve clones of pUC57-Glx2 after digestion with BglII and BglI. Modified plasmids are expected to contain two fragments of 1,118 and 2,016 bp while unmodified are expected to contain three fragments of 219, 1,118 and 1,803 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The red marker indicates the unmodified vector shown in the results.

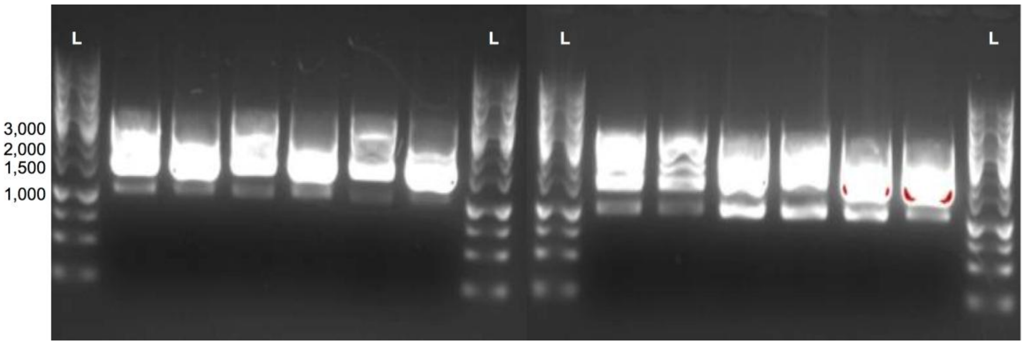


Figure E.8: The gel results for all twelve clones of pUC57-Glx2 after digestion with BglII and BglI in the second verification. Modified plasmids are expected to contain two fragments of 1,118 and 2,016 bp while unmodified are expected to contain three fragments of 219, 1,118 and 1,803 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp.

Appendix F. Construction of expression vector

Ars1

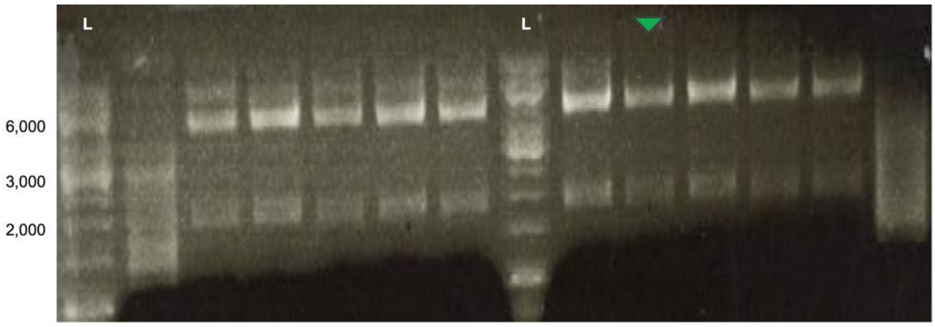


Figure F.1: The gel results for all twelve clones of pET21-Ars1 after digestion with BsaI and HindIII. Correct constructs (red markers) are expected to contain two fragments of 1,736 and 4,111 bp while re-ligated pOD1 are expected to contain one fragment of 5,631 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green marker indicates the vector with correct construct shown in the results.

Ars1-His

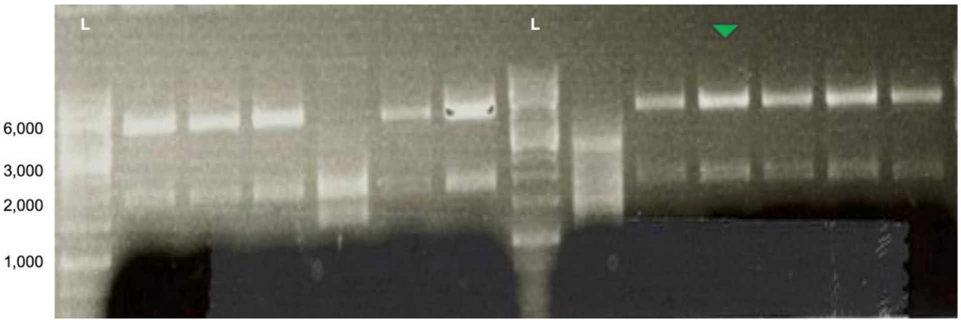


Figure F.2: The gel results for all twelve clones of pET21-Ars1-His after digestion with BsaI and HindIII. Correct constructs (red markers) are expected to contain to fragments of 1,730 and 4,111 bp while re-ligated pOD1 are expected to contain one fragment of 5,631 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green marker indicates the vector with correct construct shown in the results.

Ars2

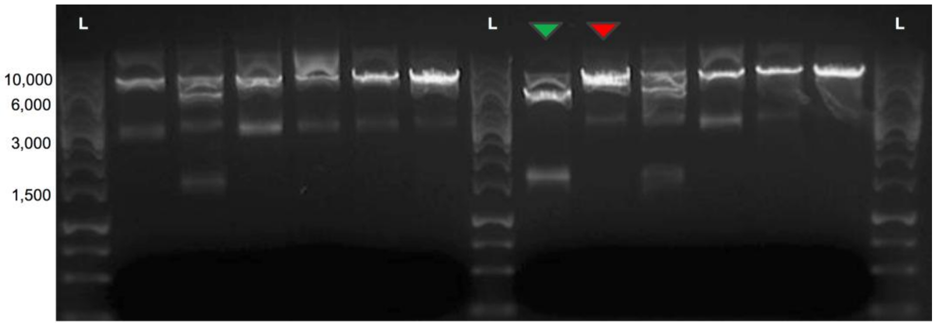


Figure F.3: The gel results for all twelve clones of pET21-Ars2 after digestion with HincII. Correct constructs are expected to contain two fragments of 1,493 and 4,339 bp while re-ligated pOD1 are expected to contain one fragment of 5,631 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) shown in the results.

Ars2-His

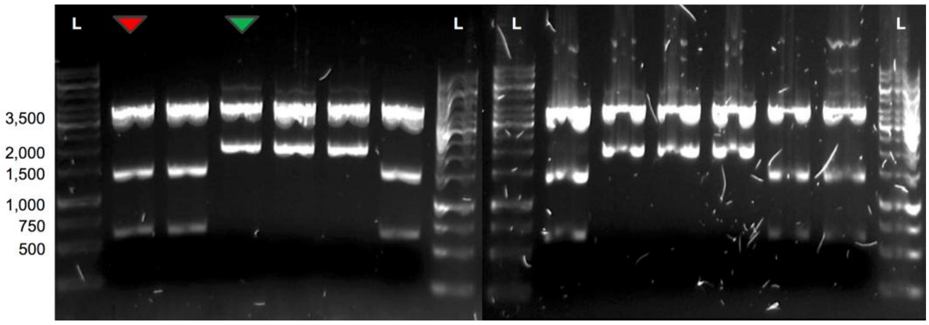


Figure F.4: The gel results for all twelve clones of pET21-Ars2-His after digestion with ApoI. Correct constructs are expected to show a digestion pattern of three fragments with 700, 1,451 and 3,664 bp while re-ligated pOD1 are expected to show two fragments with 1,956 and 3,664 bp. Digested pOD1 (P) is showed as a negative control. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) shown in the results.

Ars3

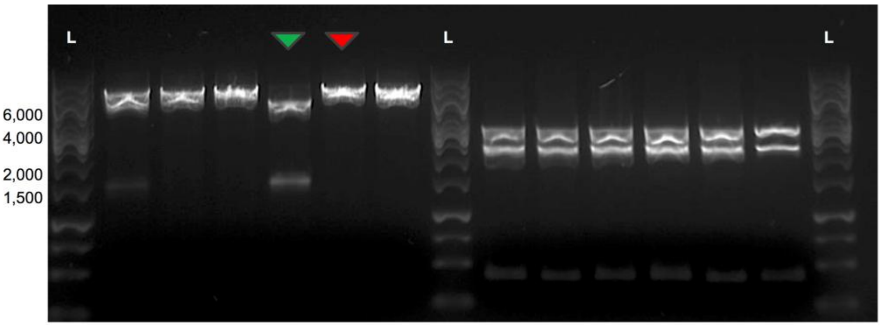


Figure F.5: The gel results for all twelve clones of pET21-Ars3 after digestion with EcoRV to the left and BsaAI to the right. Correct constructs are expected to contain two fragments for EcoRV of 1,452 and 4,389 bp while re-ligated pOD1 are expected to show one fragment with 5,631 bp for digestion with EcoRV. For BsaAI, the correct construct should have two fragments of 2,284 and 3,515 bp while three fragments of 471, 2,284and 2,876 bp were expected for pOD1. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) shown in the results.

Glx1

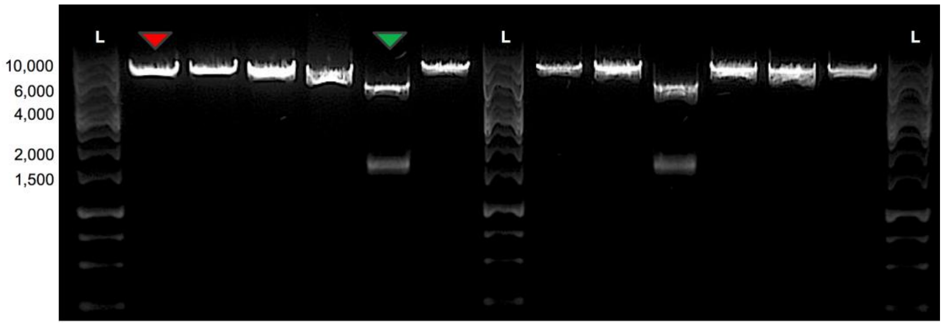


Figure F.6: The gel results for all twelve clones of pET21-Glx1 after digestion with EcoRV. Correct constructs are expected to contain two fragments of 1,493 and 4,339 bp while re-ligated pOD1 are expected to contain one fragment of 5,631 bp. The specified values for the DNA ladder (L – O'GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) shown in the results.

Glx1-His

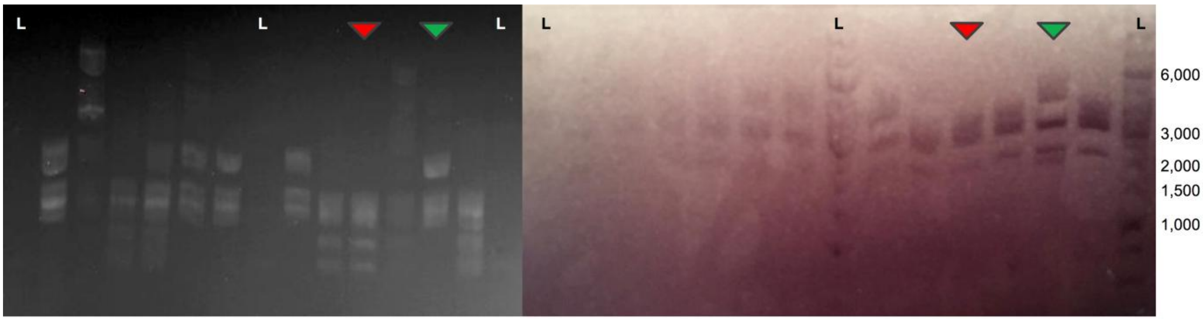


Figure F.7: The results for all twelve clones of pET21-Glx1-His after digestion with RsaI (to the left) and EarI (to the right). For the RsaI, correct constructs are expected to give three fragments of 1,567, 1,760 and 2,472 bp while re-ligated pOD1 are expected to supply four fragments with 1,027, 1,277, 1,567 and 1,760 bp.

For the EarI, correct constructs are expected to give three fragments with 1,628, 1,804 and 2,367 bp while re-ligated pOD1 are expected to supply four fragments with 621, 839, 1,804 and 2,367 bp.

The specified values for the DNA ladder (L – O’GeneRuler 1kb) are given in bp. Green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) that were tested again using RsaI and displayed in the results.

Glx2

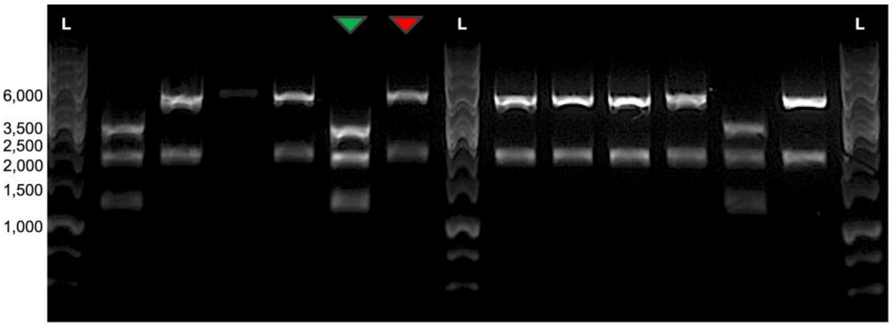


Figure F.8: The gel results for all twelve clones with pET21-Glx2 after digestion with XmnI. Correct constructs are expected to provide three fragments of 1,271, 1,934 and 2,573 bp while re-ligated pOD1 are expected to give two fragments with 1,934 and 3,697 bp. The specified values for the DNA ladder (L – O’GeneRuler 1kb) are given in bp. The green and red markers indicate the vectors (of a correct construct and a re-ligated pOD1 respectively) shown in the results.