Mining a Chinese hyperthermophilic metagenome

Morne Graham Du Plessis

UNIVERSITY of the
WESTERN CAPE

A thesis submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Biotechnology,

University of the Western Cape

Bellville

Supervisor: Prof. D.A. Cowan

Co-supervisor: Dr Z Arieff

November 2007

DECLARATION

I declare that *Mining a Chinese hyperthermophilic metagenome* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Morne Graham Du Plessis November 2007

UNIVERSITY of the
WESTERN CAPE

Abstract

Metagenomic sequencing of environmental samples provide direct access to genomic information of organisms within the respective environments. This sequence information represents a significant resource for the identification and subsequent characterization of potentially novel genes, or known genes with acquired novel characteristics. Within this context, the thermophilic environments are of particular interest due to its potential for deriving novel thermostable enzymes with biotechnological and industrial applications. In this work metagenomic library construction, random sequencing and sequence analysis strategies were employed to enhance identification and characterisation of potentially novel genes, from a thermophilic soil sample. High molecular weight metagenomic DNA was extracted from two Chinese hydrothermal soil samples. This was used as source material for the construction of four genomic DNA libraries. The combined libraries were estimated to contain in the order of 1.3 million genes, which provides a rich resource for gene identification. Approximately 70 kbp of sequence data was generated from one of the libraries as a resource for sequence-based analysis. Initial BLAST analysis predicted the presence of 53 ORFs/partial ORFs. The BLAST similarity scores for the investigated ORFs were sufficiently high (>40%) to infer homology with database proteins while also being indicative of novel sequence variants of these database matches. In an attempt to enhance the potential for deriving more full length ORFs a novel strategy, based on WGA technology, was employed. This resulted

UNIVERSITY of the
WESTERN CAPE

in the recovery of the near complete sequence of partial ORF5, directly from the WGA DNA of the environmental sample. While the full length ORF5 could not be recovered, the feasibility of this novel approach, for enhanced metagenomic sequence recovery was proved in principle. The implementation of multiple *in-silico* strategies resulted in the identification of two ORFs, classified as homologs of the DUF29 and Usp protein families respectively. The functional inference obtained from the integrated *in-silico* predictions was furthermore highly suggestive of a putative nucleotide binding/interaction role for both ORFs. A putative novel DNA polymerase gene (denoted TC11pol) was identified from the sequence data. Expression and characterization of the full length TC11pol did however not result in detectable polymerase activity. The implementation of a homology modeling approach proved succesfull for deriving a structural model of the polymerase that was used for: (i) deriving functional inferences of the potential activities of the polymerase and (ii) deriving a 5' exonuclease deletion mutant for functional analysis. Expression and subsequent functional characterization of the putative 5'exo$^-$ TC11pol mutant resulted in detectable polymerase and 3'-5' exonuclease activity at 37 and 45 ^{o}C, following a heat denaturation step at 55 ^{o}C for 1 hour. It was, therefore concluded that the putative 5'exo$^-$ TC11pol mutant was functionally equivalent to the Klenow fragment of *E. coli,* while exhibiting increased thermostability.

.

UNIVERSITY of the
WESTERN CAPE

Acknowledgements

I would like to express my most sincere gratitude and appreciation to the following people for their assistance and support throughout this study:

My supervisor, Professor Don Cowan for his patience and unwavering support.

My co-supervisor, Doctor Zainu Arieff.

My family and friends for their continued encouragement.

Everyone from the IMBM family for their continued encouragement and assistance.

To Chantel who stood by me and supported me through everything.

The National Research Foundation for providing the necessary funding.

UNIVERSITY of the
WESTERN CAPE

Abbreviations

DNA	deoxyribonucleic acid
µl	microlitre
3D	three dimensional
BAC	Bacterial artificial chromosome
cm	centimeter
CTAB	Cetyl Trimethyl Ammonium Bromide
dATP	deoxy-adenine 5'-triphosphate
dCTP	deoxy-cytosine 5'-triphosphate
ddH2O	deionized distilled water
dGTP	deoxy-guanine 5'-triphosphate
dNTP	deoxyribonucleotides
dTTP	deoxy-thymine 5'-triphosphate
EC	Enzyme Commission Classification
EDTA	ethylenediamine tetra-acetic acid
EGTs	environmental gene tags
et al.	*et alia* (and others)
g	gram
Gbp	gigabasepairs
GO	Gene Ontologies
h	hours
HMM	hidden markov model
IPTG	isopropyl-b-D-thiogalactopyranoside
kbp	kilobasepairs
kV	kilovolts
LB	Luri Bertani
M	molar
MDA	multiple displacement amplification

UNIVERSITY of the
WESTERN CAPE

min	minutes
ml	milliliter
NMR	nuclear magnetic resonance
OD	optical density
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PSSM	position specific scoring matrices
PVPP	Polyvinylpolypyrrolidone
rpm	revolutions per minute
SAP	Shrimp alkaline phosphatase
SARS	severe acute respiratory syndrome
SDS	sodium dodecyl sulphate
TB	Terrific broth
U	units
UV	ultraviolet
v/v	volume per volume
W	watts
w/v	weight per volume
WGA	whole genome amplification

UNIVERSITY of the
WESTERN CAPE

Table of contents

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

List of Tables

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

List of Figures

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

UNIVERSITY of the
WESTERN CAPE

Chapter 1

Literature Review

1.1 Introduction

Recent advances in molecular microbiological research have extensively focused on the improvement of DNA extraction, cloning and sequencing techniques as alternatives to the culture dependant strategies. The implementation of these strategies has subsequently led to improved access to the genomic complements of organisms as is demonstrated by the recovery of genes and enzymes of industrial and biotechnological significance (Lorenz and Eck, 2005, Lee et al., 2006) and the sequencing of complete microbial genomes. This successful implementation of these DNA extraction, cloning and sequencing techniques also formed the basis for the attempted recovery of complete metagenomes from diverse environments such as the Sargasso Sea (Venter et al., 2004), Whale falls (Tringe et al., 2005), Minnesota soil (Tringe et al., 2005) and Acid mine drainage (Tyson et al., 2004).

UNIVERSITY of the
WESTERN CAPE

Analysis of these random shotgun-sequencing approaches has subsequently contributed significantly to the understanding of both the microbial diversity, gene complement and community biochemistries of these environments. A significant proportion of this random shotgun sequences also display novel sequence features and represents a vast resource for novel gene identification (Venter et al., 2004; Tringe et al., 2005). The implementation of standard sequence alignment strategies however, has not been particularly useful in this context. *In-silico* strategies that could potentially enhance the characterization of these putative novel genes implement the use of sequence profiles (Altschul et al., 1997; Eddy, 1998) and structural modeling (David et al., 2000; Xiang, 2006) for functional inference. Alternatively, the genomic context mapping approaches (Dandekar et al., 1998; Pellegrini et al., 1999), which derive information from the conservation of gene order, could also be implemented.

Another observation that has emerged from the construction of metagenomic libraries is that the implemented molecular biology strategies do not always result in the optimal recovery, cloning and sequencing of the environmental DNA in an unbiased manner. These observations can be attributed to: (i) insufficient sampling procedures (ii) the fact that all organisms are inherently not equally amenable to the same DNA extraction methods and (iii) the fact that not all DNA sequences are equally amenable to cloning strategies. The characteristics of the environment ultimately determine the nature of the organisms that inhabit it and significantly affect the genome compositions of these organisms (Foerstner et al., 2005). The type of environment is therefore potentially also an important determinant in terms of the efficiency of DNA extraction, cloning and sequencing efficiencies.

UNIVERSITY of the
WESTERN CAPE

The thermophilic environments are representative of metagenomes for which the microbial population and possibly the DNA composition is influenced by high temperatures (>70^{o}C) (Musto et al., 2004). These thermophilic environments are of particular interest due to the potential applications of thermostable proteins in biotechnological and industrial applications (Cowan,1996; Hough and Danson, 1999; Eichler, 2001). As a result, several studies have been conducted on the discovery of novel thermostable proteins such as the xylanases (Bok et al., 1994; Paula et al. 2002), cellulases (Kengen et al., 1993; Ando et al., 2002) and DNA polymerases (Jones and Foulkes, 1989; Pantazaki et al., 2002). Studies on the feasibility of metagenomic shotgun sequencing approaches for obtaining thermostable enzymes from thermophilic environmental samples are however lacking.

The Archaea and a small number of bacteria (Zierenberg et al., 2000) dominate the thermophilic environments. Among the characteristics that are unique to the thermophiles and could potentially affect the implementation of standard metagenomic DNA extraction, cloning and sequencing strategies are: (i) The fact that these organisms have cell membranes that are distinct from their mesophilic counterparts (Herbert and Sharp, 1992) and (ii) The higher GC content in coding sequences in thermophiles as compared to mesophiles which affects the amino acid content and hence protein stability (Trivedi et al., 2006) with some exceptions (Paz et al., 2004; Hickey and Singer, 2004). Exactly to which degree, if any, these factors might influence the implementation of metagenomic sequencing strategies, remains to be investigated.

UNIVERSITY of the
WESTERN CAPE

The literature review aims to highlight the current strategies for deriving sequence information from metagenomic environments, which include DNA extraction, sequence library construction and sequencing. The literature review further reports on the implementation of these strategies in the context of the published metagenomic random shotgun sequencing projects. The second part of the review focuses on the strategies for obtaining functional inferences from a sequence based perspective.

1.2 Environmental DNA extraction

The efficient extraction of environmental DNA requires careful consideration of a number of factors such as the soil type, the initial level of biomass available for extraction, the degree of contamination of the sample as well as the diversity of the sample. The current methods are implemented by either extracting DNA directly from soil (Selenska and Klingmüller 1991; Tebbe and Vahjen 1993; Wikstrom et al. 1996) or the organisms can be extracted first and subsequently lysed (Steffan et al. 1988). The lysis methods, broadly classified as either mechanical or enzymatic, have been investigated for diverse environments to date and are summarized in Table 1.1. Both direct and indirect extraction methods are plagued, to various degrees, by limitations that include incomplete cell lysis, the absorption of DNA to soil surfaces, the co-extraction of enzymatic inhibitors from soil, as well as the loss, degradation and damage of DNA (Rochelle *et al*., 1992; More *et al*., 1994; Frostegard *et al*., 1999).

UNIVERSITY of the
WESTERN CAPE

Table 1.1 Methods used for cell lysis in direct extraction procedures

Lysis methods	*Lysis strategy*	*Reference*
Mechanical	Bead-beating	Moré et al. 1994; Berthelet et al. 1996; Purdy et al. 1996 ; Miller et al. 1999
	Sonication	Degrange and Bardin 1995
	Freeze-thawing	Lee et al. 1996
Enzymatic	Liquid nitrogen	Johnston and Aust 1994; Volossiouk et al. 1995
	Proteinase K	Wikstrom et al. 1996; Zhou et al. 1996
	Lysozyme	Porteous et al. 1994

UNIVERSITY of the
WESTERN CAPE

Direct extraction methods generally produce a higher yield of soil DNA as opposed to indirect methods, and therefore considered more representative of the total microbial population (von Witzingerode et al., 1997). Indirect methods, in turn, produce DNA of greater purity than direct methods (Steffan et al. 1988). It must however be noted that direct comparisons of multiple extraction methods within these two classes have, however, proven inconclusive for diverse soil samples due to various inconsistencies (Cullen and Hirsch 1998; Miller et al. 1999).

The most important disadvantage of the direct lysis methods relate to the co-extraction of humic acids, that typically interfere with DNA quantitation as well as the efficiency of downstream cloning and PCR applications (Tsai and Olson 1992, Zhou et al. 1996). Several divergent strategies have subsequently been implemented to remove co-extracted humic acids from DNA. (Table 1.2). As with the comparison of the extraction methods, the strategies for the removal of co-extracted humic acids from DNA also result in various inconsistencies.

Table 1.2 Methods used for removal of humic acids from extracted DNA.

Method	*Reference*
Hexadecyltrimethylammonium bromide (CTAB)	Zhou et al. 1996
Caesium chloride density gradients	Holben et al. 1988
Polyvinylpolypyrrolidone (PVPP)	Frostegård et al. 1999
Gel filtration resins	Jackson et al. 1997
Ion exchange and size-exclusion chromatography	Kuske et al. 1998; Hurt et al. 2001

UNIVERSITY of the
WESTERN CAPE

The implementation of a combination of extraction methods (Picard et al. 1992; Tebbe and Vahjen 1993; Kauffmann et al., 2004) results in optimal DNA extraction. Because of the variable complexities of different soil samples, the efficient implementation of these strategies for a specific environment is furthermore often dependant on extensive optimization.

1.3 Metagenomic libraries for sequence-based approaches

The construction of environmental DNA libraries are subject to the same techniques as the cloning of genomes of single microorganisms, i.e. fragmentation of genomic DNA, ligation into an appropriate vector system, and transformation to a host organism, usually *E. coli* (Daniel, 2005). The selection of the appropriate vector system is generally dependant on the downstream application for the specific source DNA.

Expression screening benefit from the use of large insert libraries, which enables the co-expression of all essential co-factors that may be required for expression of a gene (Wolfgang and Schmitz, 2004). The larger gene information content of the large insert libraries also improves the probability of recovering entire operons or multiple genes co-operative in related biochemical pathways (Beja et al, 2002). These advantages are also applicable to sequencing approaches assuming that the functions of these genes are identifiable from known homologous genes. These large insert libraries are typically constructed in BAC (Shizuya et al., 1992), Fosmid (Stein et al., 1996) and Cosmid (Entcheva et al., 2001) systems, which essentially differ in terms of the size of insert that it accommodates as well as the amount of input DNA required for construction (Beja, 2004; Kimura, 2006). The quantity of starting DNA for construction of the large insert libraries is, however, significantly high and often limits its use to environments with high biomass (Beja, 2004).

UNIVERSITY of the
WESTERN CAPE

The majority of metagenomic sequencing strategies implement the use of small to medium insert libraries, constructed in basic plasmid cloning vectors (Hallam et al., 2004; Tyson et al., 2004; Venter et al., 2004; Tringe et al., 2005). The advantages that smaller inserts offer are: (i) the quantity of input DNA required is significantly less than that of large insert libraries, (ii) it allows for maximal recovery of sequence information from the clone-end reads and (iii) the microorganisms can be lysed by harsh methods, which lead to extensive shearing, providing a good representation of the community.

While small-insert libraries are not useful for capturing complex pathways requiring many genes, they are a suitable resource for discovery of new metabolic functions encoded by single genes (Tyson et al., 2004; Venter et al., 2004). When larger DNA segments are required for analyses, the overlapping short sequence reads can be assembled into larger contiguous sequences (Tyson et al., 2004; Venter et al., 2004). In a study by Handelsman et al., (1998) it was estimated that more than 10^7 clones of size 5 kbp would be required to be representative of the metagenome of a typical soil sample. The level of reconstruction attainable will therefore depend on the sequence coverage the small insert library as well as the depth of sequencing undertaken.

UNIVERSITY of the
WESTERN CAPE

While the downstream application of the library influences the type of library constructed (small or large insert), it is evident that the quantity of available starting material is the absolute determinant. This especially holds true for those environments with low microbial biomass. It is often impossible to recover sufficient amounts of DNA for the construction of large insert libraries from these environments. Due to the smaller requirements of input material, it could be argued that the implementation of sequence-based approaches ultimately provides improved access to more environments in comparison to the expression based approaches.

1.4 Metagenomic shotgun sequencing projects

Metagenomic shotgun sequencing is broadly defined as the random sequencing of DNA libraries that are representative of the entire microbial population of an environment. These strategies implement the same standard sequencing technologies implemented in sequencing single microbial genomes (Tringe et al., 2005). This section of the literature review aims to focus purely on the sequencing, annotation strategies and relevant outputs generated in the context of the metagenomic sequencing projects.

UNIVERSITY of the
WESTERN CAPE

1.4.1 The Sargasso Sea

The metagenomic sequencing of the surface water microbial community of the Sargasso Sea focused on the extraction of planktonic microbes. A total of more than 1.6 Gbp of DNA sequence was generated from the pooled clone sequencing of seven independent libraries. While previous studies indicated a diversity of 200 species (Curtis et al., 2002), analysis of the Sargasso Sea sequence data was suggestive of more than 1000 species (Venter et al., 2004). The attempted assembly of the sequence data, however, resulted in the near complete genomes of only three organisms. Nearly half of the sequences could furthermore only be associated as bidirectional pairs because of the fact that they represent the end sequences of common clones (Venter et al., 2004).

The annotation of sequence data was performed through standard BLAST homology searches and in the order of 1.2 million gene sequences was identified. Investigation of the annotated genes (30% of the total) reportedly resulted in identification of numerous rhodopsin-related genes and genes involved in phosphorus uptake and metabolism (Venter et al., 2004).

The genes that could not be functionally annotated represented approximately 68% of the entire analyzed gene sequences. Approximately 38 000 of the remaining genes were assigned to the class of "unknown functions" and 794 000 as "conserved hypothetical" (Venter et al., 2004). The abundance of these uncharacterized genes therefore raises important questions: (i) Are they indicative of potentially entirely novel biochemical pathways or (ii) are they simply extremely divergent homologs of known genes?

UNIVERSITY of the
WESTERN CAPE

Despite these limitations, the study also provided a vast resource of sequence data for gene mining approaches. Studies that have been reported to date include gene mining for chitinases (LeCleir et al., 2004), proteorhodopsins (Sabehi et al., 2004), and electron transport proteins (McDonald and Vanlerberghe, 2005).

1.4.2 The acid mine biofilm

The metagenomic sequencing of a low-complexity acid mine drainage microbial biofilm involved the random shotgun sequencing of 76.2 million bp of sequence data was produced from 103 462 high quality sequence reads (Tyson et al., 2004). Given this low complexity the genomes of primarily two organisms (Leptospirillum group II and Ferroplasma type II) could be assembled to near completion. The gene prediction analysis reported the presence of 2 180 ORFs for Leptospirillum group II of which 63% had putative assigned functions and 1 931 ORFs for Ferroplasma type II of which 58% had assigned functions. Apart from these assembled genomes, the partial genomes of three additional species could furthermore also be identified (Tyson et al., 2004).

UNIVERSITY of the
WESTERN CAPE

The analysis of the genes was approached in the context of the defining the metabolic pathways as well as the roles of the organisms within the biofilm community. Riesenfeld et al, (2004), extensively reviewed significant observations from this study. A summary of the findings include: (i) genes that would enable most of the organisms to fix carbon via the reductive acetyl coenzyme A pathway were identified within each of the genome sequences, (ii) on the basis of the overrepresentation of genes with similarity to sugar and amino acid transporters the *Ferroplasma* spp. identified from this environment were proposed to prefer a heterotrophic lifestyle, (iii) The only N_2 fixation genes that were identified in the metagenome belonged to the genome of the *Leptospirillum* group III population. This was of specific significance due to the essential role of the N_2 fixation process in such nutrient limited environments and (iv) Other genes that are potentially responsible for microaerophilic survival, biofilm formation, acid tolerance, and metal resistance were observed.

1.4.3 The whale falls and Minesota soil

Metagenome shotgun sequencing analyses of complex, nutrient rich environments were first conducted on a soil sample as well as three samples from sunken whale skeletons (Tringe et al., 2005). The complexity of these environments was apparent from the lack of assembly for any genomes from these environments. The sequence data represented single sequence reads and subsequently the information content for identification of full length genes was limited. These short single sequence reads were instead annotated as Environmental Gene Tags (EGTs), which contain fragments of functional genes, considered adaptive for that specific environment (Tringe et al., 2005).

UNIVERSITY of the
WESTERN CAPE

These EGTs was subsequently used as a basis for comparison between the different metagenomes of the Minnesota soil, whale falls, Sargasso Sea and AMD. Important findings from this study included: (i) Similar environments, while geographically distinct, such as two whale skeletons 8000 miles apart on the ocean floor, have similar gene content. (ii) Environments were dominated by specific genes that reflect the biochemistry of that environment. For example, numerous homologs of cellobiose phosphorylase, an enzyme involved in the breakdown of plant material, were specific to the soil sample. The Sargasso Sea samples were dominated by sodium transport and osmoregulation proteins associated by the high sodium content of seawater and (iv) The uncharacterized genes and processes are among the most significant overrepresentations in each sample (Tringe et al., 2005).

1.5 Functional annotation by homology based transfer

The basic principle of homology transfer is based on the observation that sequences with significant degree of similarity have evolved from a common ancestor and therefore have identical or similar functions (Whisstock and Lesk, 2003). Given a sequence or structure of interest, it is therefore possible to define aspects of its function from a similar sequence or structure with, known function. The level of functional inference that can be achieved from this approach is generally considered to be directly dependant upon the degree of homology that is detectable between query and template. Therefore, sequence or structure pairs that share a high degree of homology are predicted to share related functions with higher confidence. More specifically, homology at the superfamily level relate to proteins that share the same fold and are proposed to have a common evolutionary origin, but have low levels of sequence identity Homologues at the family level have sequence identity of greater than 30-35% and certainly have a common evolutionary origin. While this holds true for most part there are, however, several aspects for consideration when inferring related functions by means of homology.

UNIVERSITY of the
WESTERN CAPE

The definition of function is a complex and therefore, not well defined term. This is because the function refers to biochemical, cellular, developmental and physiological aspects of a protein (Rost et al., 2002). It is therefore often difficult to define all of these varied aspects of function in a comprehensive manner. The three most widely used vocabularies, which attempt to achieve this, are SWISS-PROT, the Enzyme Commission Classification (EC) and the Gene Ontologies (GO) system. The SWISS-PROT classification assigns keywords that reflect aspects of function and other aspects such as post-translational modification (Boeckmann et al., 2003). With the use of the SWISS-PROT classification, however, there is often a low overlap of information between proteins with known relationships (Devos and Valencia, 2000). The use of the Enzyme Commission (EC) (Webb, 1992) allows for a more standardized classification scheme (Shah and Hunter, 1997). This system classifies proteins according to the biochemical reactions they catalyze. This is a four level hierarchy that starts with a general classification and progresses to more specific (http://www.hem.qmul.ac.uk/iubmb/enzyme/). The level at which the EC numbers correspond between homologous proteins indicate that all four classes are conserved at only high levels of sequence similarity (Devos and Valencia, 2000). Homologues with lower sequence similarities therefore do not entirely benefit from functional inference using this system. Another limitation of the EC system is the fact that additional aspects of function that extend beyond enzyme activity are not included in this classification. The Gene Ontology (GO) classification is currently the approach of choice for annotation (Ashburner et al., 2000). This

UNIVERSITY of the
WESTERN CAPE

system describes three aspects of protein function: molecular function, biological process, and cellular location. Beyond this classification, the ontologies are also represented on a graph that describes whether functions are involved in more than a single biological process, cellular compartment, or molecular activity (Ashburner et al., 2000). While this represents the most comprehensive approach toward the classification of proteins even this system does not include all aspects of function. None of the current functional classification systems are therefore complete and as a result frequently leads to incorrect or incomplete annotation by homology transfer.

Compositional biases in protein sequences can result in significant similarity scores that are otherwise biologically meaningless (Sharon et al., 2005). This is particularly observed for low complexity sequences that contain amino acids such as alanine (A), serine (S), proline (P), and glycine (G) at high frequency while cysteine (C) and tryptophan (W) are very rare (Romero *et al.*, 2001; Alba *et al.*, 2002). The repetitive nature of these specified amino acids often result in high scoring alignments that are, erroneously, considered as significant in the context of the established alignment scoring systems. As a means to prevent this, several applications that mask these sequences (Claverie *et al.*, 1993; Wootton and Federhen, 1993) have been developed. These are, however, by no means complete in identifying all low complexity sequences. Compositional biases are, however, not only restricted to low complexity sequences as demonstrated in a study by Reichsman et al, (1999). In that study alignment between Wingless/Wnt

UNIVERSITY of the
WESTERN CAPE

and secreted phospholipase A2 produced an E-value score < 10^{-3} which is considered biologically significant. Further investigation of the structural features of this alignment however demonstrated entirely divergent folds for this protein pair. It was found that the erroneous alignment was as a result of fortuitous matches of cysteines.

Multi-domain proteins also impose limitations on the ability to detect homology between sequence pairs (Hoffman, 2000). This result from the fact that proteins may share, for example, related binding domains for a specific substrate while they also contain a divergent catalytic domain (Todd et al., 2001). In this context, the homologous relationship of the conserved domains alone, could lead to erroneous functional assignment in the context of the full length protein.

UNIVERSITY of the
WESTERN CAPE

High levels of overall homology alone are often not sufficient to directly infer related function between proteins. Rost (2002) has established that even at high sequence similarity rates, enzymatic function may not necessarily be conserved. Identity scores of at least 40% were reported to be essential for establishing catalytic mechanism similarity, while 60% identity is required for substrate similarity (Tian and Scholnick, 2003).

While these observations generally represent the exceptions to the rule, they also demonstrate that functional inference by homology transfer is by no means a trivial task. On the basis of the high success rate for correctly predicting functional aspects of proteins, annotation by homology transfer still remains the most widely used approach.

Several computational approaches are currently available that enables both sequence and structural comparisons as a means to detect homology between protein pairs. These approaches for determining sequence-sequence homology, sequence-structure homology and structure-structure homology are reviewed below.

1.5.1 Homology detection by sequence similarity

Basic homology searches, which detect sequence similarity from pairwise alignments, are routinely performed using BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988) algorithms. The BLAST suite of tools is preferred for the detection of local sequence similarity while FASTA is more suited to global sequence alignments.

UNIVERSITY of the WESTERN CAPE

With these approaches, sequence homology is generally inferred when the alignment generated between a sequence of interest and a sequence in the queried database exceed a specific alignment score, S. (Hofmann, 2000). The biological significance of the alignment, in turn, is quantified by a statistical E-value. The E-value represents the number of different alignments with scores equivalent to, or better than, S that are expected to occur in the database search simply by chance (Hofmann, 2000). In this context, lower E-values are therefore considered more biologically meaningful than larger E-values.

Factors that require consideration when making functional inferences on pairwise alignment scores and E-values, relate to: (i) the compositional biases observed for certain amino acids (as previously described) and (ii) the size of the database used in the sequence comparison. The effect of database size relates to the fact that an E-value generated for a query against a small database (eg. 6000 genes) would approximately be 100 times smaller than the E-value for the same query against a larger database (eg. 600 000 genes) (Ponting et al., 2001). In the context of the biological significance assigned to E-values, this observation would lead to a gross overestimation of the significance of the query sequence in question.

UNIVERSITY of the
WESTERN CAPE

Traditional pairwise sequence alignment methods are used to assign folds to sequences with obvious evolutionary relationships to a known structure. For sequences with identities >30%, fast searching methods such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul and Gish, 1996) are capable at detecting related proteins by scoring pairwise comparisons. The BLAST programme is therefore also routinely implemented during the initial analysis of large datasets such as genome- or metagenome sequences (Venter et al., 2004).

However, when sequence identities fall below 30%, conventional pairwise sequence comparison methods fail to detect homologous relationships (Brenner *et al.*, 1998). In these instances improved homology detection can be attained through sequence profile-profile, sequence-structure and structure-structure comparisons.

1.5.2 Homology detection by sequence profiles

The problem with detecting homology beyond the accepted benchmark of 30% identity arises from the fact that the similarity between sequences becomes less obvious. The non-essential amino acids within the sequences have evolved substantially so that the only detectable sequence identities represent the active site- and binding site residues or motifs that underlie the functional aspects of the protein. Strategies that implement either position specific scoring matrices (PSSM) (Altschul et al., 1997) or derive the information from hidden markov models (HMMs) (Eddy, 1998) have been shown to succesfully detect these sequence patterns between distant homologs (Muller et al.,1999).

UNIVERSITY of the
WESTERN CAPE

PSSMs implement a weighting score for depicting the frequency with which each specific amino acid is expected to appear at a specific position in a sequence. This results in: (i) an improved estimation of the probabilities with which amino acids appear at various positions and (ii) the precision with which it assigns the boundaries of motifs (Altschul et al., 1997). HMMs in turn assign probability values to the occurrence of amino acids at specific positions within a sequence.

In this context the HMMs could be considered as equivalent to PSSM generated profiles (Butcher et al., 1996). HMMs can, however, also incorporate additional properties such as secondary structure information, surface accessibility and hydrogen bonding etc (Eddy, 1998). This information can essentially all also be incorporated into a linear description (profile) of a sequence that can be compared against a library of profiles for annotated structures.

Profiles and HMMs are routinely applied in various homology detection search algorithms such as PSI-BLAST (Altschul et al., 1997). The program initially generates gapless alignments between a query sequence and the significant hits in the database, which is then converted into a profile. The profile is then implemented to identify further homologues through several iterations of the BLAST program (Altschul et al., 1990), resulting in improved sensitivity.

UNIVERSITY of the
WESTERN CAPE

Profiles and HMMs are also used for the representation of sequences in domain and motif databases such as the PROSITE library (Hofmann et al., 1999), which implements profiles while Pfam (Bateman et al., 1997) and SMART (Schultz et al., 1998) utilize HMMs.

The profile-profile alignments fare better than the sequence-sequence or profile-sequence alignments in terms of both sensitivity (Rychlewski et al., 2000) and alignment accuracy (Jaroszewski et al., 2000). More specifically, Ohlson et al, (2004) have demonstrated that profile-profile methods performed at least 30%

better than standard sequence-profile methods both in their ability to recognize superfamily-related proteins and in the quality of the obtained alignments.

1.5.3 Functional inference by means of structure homology

The catalytic activity of a protein, associated with a specific function, is essentially classified by the positions of specific residues in an ordered 3D arrangement. As a consequence, proteins that share an identical structural arrangement of these functionally significant residues are more often than not related in terms of their functional roles (Pazos and Sternberg, 2004). Sequence representations of proteins, on the other hand, do not reflect this organization of residues in a three dimensional space. While considerable similarity might exist between protein pairs, it is therefore not a guarantee for similar 3D folds (Gan et al., 2002). Additionally, while sequences diverge as a result of evolutionary pressures, the ability to detect the relationship between homologous sequence pairs becomes lower. Structure however evolves more slowly than sequences (Silberberg, 2000) and while the sequence features have diverged beyond recognition the structural features remain intact. On the basis of these observations, homologous relationships between proteins are more accurately detected at the level of structural conservation as opposed to sequence features (Brenner et al., 1998).

UNIVERSITY of the
WESTERN CAPE

Computational programs that aim to detect structural conservation between a query structure and the structures in the PDB database include: CE (Shindyalov and Bourne, 2001), DALI/FSSP (Holm and Stander, 1993; Holm and Stander, 1997), FATCAT (Ye and Godzick, 2003), VAST (Madej et al., 1995), FAST (Zhu and Weng, 2005), Matras (Kawabata, 2003; Kawabata and Nishikawa, 2000), DaliLite (Holm and Park, 2000), and GRATH (Harrison et al., 2003). An assessment of a number of these structure based comparison methods reveal that optimal results are obtained when implementing more than one program (Novotney et al., 2004).

Entirely novel protein structures often do not have a full length, directly comparable, structural homolog in the PDB database. In these instances information on the structure-function relationship can be derived from a comparison against a database of structural domain fragments. While this often does not result in complete functional annotation, it does provide usefull clues on the presence of functional domains of the proteins.

UNIVERSITY of the
WESTERN CAPE

When the protein of interest has no determined structure, structural prediction approaches may also provide a feasible alternative toward obtaining an approximate model of the protein sequence. These modeled structures, depending on the quality of the model, are then amenable to the same structural homology strategies for inferring functional aspects from structure.

1.5.4 Structure prediction

The availability of structural information for a protein has been shown to greatly enhance the potential for deriving functional aspects of that protein. In the absence of an experimentally determined structure, a useful 3D model for the protein can sometimes be derived through either ab initio structure prediction, fold recognition (threading) or homology modeling strategies.

The Ab-initio predictions are generated from purely physics and chemistry principles. Considerable advances have been made in the application of these principles on small proteins (Simons et al., 2000; Bonneau et al 2001). However, when the protein length increases, so does the complexity for predicting the correct structure for the protein. As a result, the Ab-initio approaches are currently primarily reserved for instances where modeling by threading or homology modeling fails.

UNIVERSITY of the
WESTERN CAPE

Structure prediction via fold recognition and homology modeling require the availability of homologous templates from the structural database for model construction. The fact that the structural databases are near complete for single domain proteins means that an adequate template for homology modeling can often be identified (Zhang and Skolnick, 2004). In theory, it should therefore be possible to solve the protein structure prediction problem with either the threading or homology modeling approach.

1.5.4.1 Structure prediction by threading approaches

The threading approaches of implements comparisons of a sequence of interest against a database of known structures as a means to suggest an appropriate fold for the sequence (Bowie et al., 1991; Godzik et al., 1992; Jones et al., 1992; Levitt, 1997; Smith et al., 1997; Torda, 1997; David et al., 2000). A number of automated bioinformatics tools have been designed in an attempt to solve the threading problem (see Table 1.3).

Table 1.3 Programs implemented in threading.

Program	*Availability*
123D	http://123d.ncifcrf.gov/123D+.html
3D-PSSM	http://www.sbg.bio.ic.ac.uk/3dpssm/
FUGUE	http://www-cryst.bioc.cam.ac.uk/fugue/
GenTHREADER	http://bioinf.cs.ucl.ac.uk/psipred/
HMAP	http://trantor.bioc.columbia.edu/hmap/
LOOPP	http://cbsuapps.tc.cornell.edu/loopp.aspx
PROSPECT	http://csbl.bmb.uga.edu/downloads/#prospect
RAPTOR	http://www.bioinformaticssolutions.com/
UCLA-DOE	products/raptor/
Server	http://fold.doe-mbi.ucla.edu/

UNIVERSITY of the
WESTERN CAPE

All of the threading approaches have the same basic requirements which include: (i) A means of representing the query sequence, (ii) A library of structural models for comparison (iii) Methods for generating and scoring alignments between sequence and template (iv) Methods for selecting the best model from the library.

The query sequence is most often represented by a sequence profile which contains information from multiple sequence alignments and possibly secondary structure features. The template sequences, in turn, can either be represented as a linear model (1D) or higher order model (2D), depending on the program.

A number of algorithms have been implemented to generate sequence-structure alignments for the threading approach. Amongst these are: (Recursive) Dynamic progamming (Jones et al., 1992; Thiele et al., 1999; Westhead et al., 1995; Skolnick and Kihara, 2001; Zhang et al., 1997), Neural network (Jones et al., 1999), Divide and conquer algorithm (Xu et al., 1998; ,Xu and Xu., 2000), Branch and bound algorithm(Lathrop and Smith), Linear programming approach (Xu et al., 2003), Monte Carlo procedure (Mirny and Shakhnovich, 1998).

UNIVERSITY of the
WESTERN CAPE

Scoring the sequence-structure alignments (objective function) between the query and template are based on finding the equivalent of an optimized energy function between a query sequence and structure (Huang et al., 1996; Maiorov and Crippen 1994; Bryant and Lawrence 1993; White et al., 1994). These are used to distinguish correct structural folds from the incorrect ones and to distinguish accurate alignments against the correct structural folds from inaccurate alignments.

Subsequent to aligning and deriving an optimal score for the sequence query against a structure template, the final step involves selecting the most probable structure. This is achieved by either choosing a structure based on the alignment score or by utilizing the probability score for the model across the alignments of the sequence to the model.

The best fold-recognition algorithms are able to make up to 40 % of correct structural predictions for targets that cannot be detected by BLAST or PSI-BLAST searches run with default parameters. (Bujnicki et al., 2001). Examples of threading prediction methods that have been successfully implemented include GenThreader (Jones 1999), 3D_PSSM, (Kelly et al., 2000), FFAS (Rychlewski et al., 2000) and PROSPECTOR (Skolnick and Kihara, 2001).

Apart from the individual threading programs, several meta-servers have also emerged. These meta-servers integrate predictions from different individual threading predictors in order to generate models and have been shown to outperform the individual threading servers (Fischer et al., 2003). Examples of consensus meta-servers include 3D-Jury (Ginalski et al., 2003) and LOMETS (Wu and Zhang., 2007). The 3D-Jury server implements a set of eight servers (ORFeus, SamT02, FFAS03, mGenTHREADER, INBGU, RAPTOR, FUGUE-2, 3D-PSSM) for consensus prediction. The LOMETS server takes predictions from nine different servers, i.e. FUGUE, HHSEARCH, PROSPECT2, SAM-T02, SPARKS2, SP3, PAINT, PPA-I and PPA-II to generate a consensus.

UNIVERSITY of the WESTERN CAPE

As evidence of the accuracy of threading approaches, a large number of protein structures have been predicted prior to the solution of their experimental structures. Examples of such predictions include an obese gene (Madej et al., 1995), vitronectin (Xu et al., 2001), and a SARS protein (von Grotthuss et al., 2003). Apart from its application in protein structure predictions, the threading programs have aslo been implemented for: (i) annotating genomes (McGuffin et al., 2004), (ii) modeling protein complex structures (Lu et al., 2003), (iii) protein design (Soren-son and Head-Gordon, 1999) and experimental design and functional studies of proteins (Xu et al., 2001).

UNIVERSITY of the
WESTERN CAPE

1.5.4.2 Structure prediction by homology modeling

Homology modeling approaches are based on the construction of a structural model of a query sequence on the basis of similarity with a known structure (template) in the database (Xiang, 2006). There are currently a number of programs that have been developed for this purpose (See Table 1.4).

Table 1.4 Homology modeling programs

Programs	*Availability*	*Methods*
NEST	http://trantor.bioc.columbia.edu/programs/jackal/	Artificial evolution
COMPOSER	http://www-cryst.bioc.cam.ac.uk/	Rigid-body assembly
MODELLER	http://guitar.rockefeller.edu/modeller/modeller.html	Spatial restraints
SWI-MOD*	http://www.expasy.ch/swissmod/	Rigid-body assembly
WATIF	SWISSMODEL.html	Rigid-body assembly
SEGMOD	http://swift.cmbi.kun.nl/whatif/	Segment matching
DRAGON	Module in Look, sold to Celera in 1999	Spatial restraints
JCM	Contact Robin Munro at rmunro@nimr.mrc.ac.uk	Rigid-body assembly

*SWISS-MODELER

All of these homology modeling programs implement the same basic strategy, which include: (i) the identification and selection of a template for modeling, (ii) generating an optimal query-template alignment, (iii) the actual model construction and (iv) validating the correctness of the model (Xiang, 2006).

The different modeling programs have varied requirements with regard to the template selection and alignment approach. The fully automated programs such as SWISS-MODEL, select a template and generate alignments as part of the homology modeling process (Schwede et al., 2003). Other programs such as Modeler however require the alignment of the query and template as part of the input to generate a model (Sali and Blundell, 1993). In this instance, template selection and alignment methods for the homology modeling depend on the level of sequence similarity detectable between the query and the sequence of the structure templates in the PDB library. These methods, as discussed previously, are either: (i) pairwise sequence-sequence comparisons, (ii) sequence profile analysis or (iii) threading approaches.

UNIVERSITY of the
WESTERN CAPE

In terms of the method for model construction, four basic approaches are used: rigid body assembly (Browne et al., 1969; Blundell et al., 1987; Greer, 1990), segment matching, modeling by satisfaction of spatial constraints (Sali et al., 1990; Sali and Blundell, 1993; Sali and Overington, 1994; Fiser et al., 2000) and artificial evolution (Xiang, 2006). Rigid-body assembly methods use the core of the aligned structure as a basis to generate a close structural representation of the template. The assembly itself involves fitting the rigid structure bodies onto the framework and rebuilding the nonconserved parts (Greer, 1990). The segment-matching approach uses the aligned regions as a reference for deriving representative atomic coordinates of the structures. These coordinates are then used for the recovery of related structural segments from a database (Levit, 1992). The strategy based on satisfaction of spatial constraints derives restraints, for modeling the template, from the alignment of the structure. During the modeling these restraints are implemented to a mininmum in order to produce the best possible structure (Sali et al., 1990).

UNIVERSITY of the
WESTERN CAPE

From a comparison of a number of programs that implement the above algorithms (Wallner and Elofsson, 2005) it was deduced that: (i) none of the programs outperform the others, in all the implemented tests, (ii) the Modeller, nest, and SegMod/ENCAD generally perform better than the others, and (iii) none of the models were entirely efficient at building side chains.

The performance of several modeling programs were also investigated across a number of proteins for which the identities range from 19% to 76% (Nayeem et al., 2006). The study investigated the accuracy of the sequence alignments as well as the correctness of the constructed models. None of the programs appeared to be distinct, in terms of generating alignments, for proteins with ≥38% sequence identity. At lower sequence identities the programs Profit and Prime stood out from the rest. In terms of model construction, all programs fared equally satisfactory at sequence identity >40%. At low levels (19%) only the program Prime performed reasonably well.

Complications from the model construction primarily arise from the incorrect fold assignment for those regions between the query and template that do not have clearly comparable sequence relationships. These regions often correspond to the surface loops. One of the approaches for loop construction involves searching a database of possible folds to find a region comparable to the query loop (Chothia and Lesk, 1987). Alternatively, *ab-initio* approaches (Shenkin et al., 1987) are employed. While considerable success has been achieved with both of these approaches on particularly short loop fragments, modeling of the longer fragments remain a problem.

UNIVERSITY of the
WESTERN CAPE

Upon completion of the modeling process a well constructed model is proposed to have: (i) the correct fold, (ii) an alignment that conforms to the predicted fold, (iii) acceptable stereochemistry and (iv) acceptable spatial features. There are currently a number of Web based server programs that evaluate all these aspects of structural models (see Table 1.5). The useful application of modeled structures is dependant on the predicted accuracy of the modeling process which in turn is dependant on the degree of homology between the query-template pair. However, there are also examples in CASP in which good models are built at quite low levels of sequence identity. It is generally accepted that models based on query/template alignment of greater than about 40% have comparable accuracy to NMR structures (Sali et al., 1995).

Table 1.5 Programs for the evaluation of structural models

Feature investigated	*Programme*	*Reference*
Stereochemistry	PROCHECK	Laskowski et al., 1998
	PROCHECK-NMR	Laskowski et al., 1996
	AQUA	Laskowski et al.,1996
	SQUID	Oldfield, 1992
	WHATCHECK	Hooft et al., 1996b
Spatial features	VERIFY3D	Luthy et al., 1992
	ANOLEA	Melo and Feytmans, 1998

UNIVERSITY of the WESTERN CAPE

The applications of comparative models have been extensively reviewed by Renom et al (2003) and were reported to include : (i) designing mutants to test hypotheses about the protein's function; (ii) identifying active and binding sites; (iii) searching for, designing, and improving ligand binding strength for a given binding site; (iv) modeling substrate specificity; (v) predicting antigenic epitopes; (vi) simulating protein-protein docking; (vii) inferring function from calculated electrostatic potential around the protein; (viii) facilitating molecular replacement in X-ray structure determination; (ix) refining models based on NMR constraints; (x) testing and improving a sequence-structure alignment; (xi) confirming a remote structural relationship; and (xii) rationalizing known experimental observations.

1.6 Functional annotation from genomic context approaches

The genomic context approaches for functional inference are based on aspects of the conservation of gene order or gene occurrence in multiple organisms. These approaches infer functional relationships between proteins, as opposed to the homology transfer approaches that infer relationships from other proteins (Doerks et al., 2004). The basic principle that enables this type of analysis relates to the fact that genes are co-located on genomes to allow for co-expression in reponse to (i) a common activator or (ii) a common involvement in a specific biochemical pathway. The argument that stems from this is therefore that knowledge of the function of some of the genes at a locus could contribute to an understanding of possible roles of uncharacterized genes at the same locus

UNIVERSITY of the
WESTERN CAPE

This inferred relationship for co-occurring and conserved genes have lead to the establishment of various genomic context approaches. The first involves the analysis of the local gene neighbourhood. In this strategy, genes are evaluated for their presence in conserved, putative operons (Dandekar et al., 1998). In an alternative strategy the analyses focuses on phylogenetic patterns by investigating the presence or absence of gene clusters across genomes (Pellegrini et al., 1999). The third approach takes advantage of the occurrence of gene fusion events (Enright et al., 1999; Marcotte et al., 1999). In this approach, it was argued that the functions of two domains found separately in two proteins, are predicted to be linked if these domains are found together in a second protein (Enright et al., 1999).

Several algorithms have subsequently been developed to investigate these observations of co-evolution and chromosomal proximity as functional predictors. These include, among others, the Phydbac2 server (Enault et al., 2003), SNAP (Kolesov et al., 2001) and SynFPS (Li et al., 2007).

A study by Makarova et al, (2002) on the implementation of a genomic context approach involved the identification partially conserved neighbourhood consisting of more than 20 genes discovered in most Archaea and some bacteria. The gene composition and gene order in this neighbourhood vary greatly among species, but all versions have a conserved core of five genes that is proposed to constitute a previously undetected DNA repair system. Other studies included the implementation of genomic context approaches to: (i) annotate uncharacterized proteins from numerous organisms (Doerks et al., 2004), and (ii) to predict gene function and gene correspondence in whole genome comparisons (Li et al., 2007).

UNIVERSITY of the
WESTERN CAPE

1.7 Aims and strategy of the study

The broader aim of this work was to investigate the implementation of metagenomic library construction and sequencing-based approaches, as a basis for gene identification and functional characterization, from a novel thermophilic environment.

Specific aims were:

- To recover genomic DNA from a novel uncharacterized Chinese hydrothermal metagenomic soil sample using direct and indirect extraction methods.
    - *This soil sample investigated was representative of a low biomass, humic acid contaminated metagenomic environment.*
    - *The extraction and purification procedures investigated were aimed at generating DNA of sufficient quantity and quality for downstream molecular biology manipulation.*
- To construct small to medium insert sequence libraries from the extracted metagenomic DNA as a resource for sequencing and identification of putative ORFs.
    - *A small scale random sequencing strategy was investigated to assess the feasibility of such an approach for the constructed library.*
    - *Initial analyses of putative ORFs are routinely performed with standard sequence alignment tools. The BLAST program*

UNIVERSITY of the WESTERN CAPE

was implemented for ORF identification from the sequence data.

- To implement and optimize the whole genome amplification (WGA) strategy on DNA extracted from the thermophilic metagenomic environment.
  - *Given the limited studies on this approach, in the context of metagenomics, this was implemented to investigate the feasibility of WGA on a novel thermophilic environment.*
- To demonstrate the implementation of the environmental WGA DNA as a resource for the enhanced recovery of ORFs identified in the metagenomic library
  - *The WGA strategy is reported to generate a significant amount of non-specific spurious products. This work investigated the feasibility of recovering gene sequences from WGA DNA through the use of highly specific gene targeted primers.*
- To demonstrate the recovery of flanking sequence information for an ORF in the metagenomic library, directly from the WGA DNA of the environmental sample.
  - *This work aimed to demonstrate a novel application for WGA DNA in the context of providing improved access to the genome complement of an environment.*

UNIVERSITY of the
WESTERN CAPE

- To assign possible functional roles for metagenomic sequence-derived ORFs, by means of integrating multiple *in-silico* approaches.
    - *An understanding of the potential functions of novel/uncharacterized genes are proposed to lead to improved assessment on the (i) biotechnological potential of the gene, if any and (ii) improved experimental design.*
    - *This work aimed to investigate a novel arrangement of a pair of ORFs on a metagenomic clone, within the context of deriving potential related functions for these ORFs.*
- To demonstrate the experimental functional characterization of a putative metagenomic sequence derived ORF. This strategy was proposed to include: (i) the sequence-based characterization of the ORF, (ii) the structural classification by means of homology modeling, (iii) expression analyis and (iv) functional assay of the expression product.
    - *This strategy was implemented to demonstrate the identification and characterization of a protein with thermostable properties, with potential application in biotechnology.*

UNIVERSITY of the
WESTERN CAPE

Chapter 2

Materials and Methods

2.1 Materials

The suppliers of materials used in this study are given in Table 2.1.

Table 2.1 Suppliers of materials

Supplier	*Reagent*
Promega	JM109 (DE), pGEM®T-Easy vector system, dNTPs, LM-SIEVE Agarose
Fermentas	Restriction enzymes, T4 DNA ligase, InsT/Aclone ™ PCR product cloning kit, Klenow Fragment, pUC18
Invitrogen	pCR® T7 TOPO® TA Expression kit
Amersham	GFX PCR DNA and Gel band purification kit, GFX Micro plasmid prep kit
Qiagen	QIAEX II Gel extraction system
Sigma	Chemical reagents
Roche	Shrimp Alkaline Phosphatase
Bioline	DNA polymerase, Agarose
IDT	Oligonucleotide primers
BIO-RAD	Gene Pulser® cuvette
TaKaRa	T4 DNA polymerase
Merck	Chemical reagents
Pharmacia	Sephadex G100

UNIVERSITY of the
WESTERN CAPE

2.2 Bacterial strains and plasmids

Bacterial strains and plasmid vectors used in this study are listed in Table 2.2.

Table 2.2 Bacterial strains and plasmids used in this study

Strains/Plasmids	*Genotype or relevant characteristics*	*Source or Reference*
E. coli (DH5α)	*[supE44 ΔlacU169 (φ80 lacZMΔ15) hsdR17 relA1 gyrA96 thi-1 recA1]*	UWCculture Collection
E.coli BL21[DE3]	[(rB-mB-*ompT*F$^-$) [*lon*] *hsd*SB] with DE3 a λ prophage carrying the T7 RNA polymerase.	Novagen
pCR® T7/CTTOPO	Size 2702 bp, AP^r, T7 promoter, V5 epitope, Zeocin^r, polyhistidine region, pUC origin, TOPO® cloning site	Invitrogen
pUC18	High copy number cloning vector. Insertion into multiple cloning sites disrupts lacZα reading frame. Recombinants are identified by blue white screening. (Ap^r)	(Yanisch-Perron et al. 1985)
pUCPOL	pUC18 derivative carrying full length Polymerase gene fused to lacZ gene	This study
pGEM T-easy	Size 3015 bp, T7 promoter, SP6 promoter, Ampr, *lac* operator, *LacZ* start codon, phage f1 region, pUC M13 priming sites, 3' – T overhangs	Promega
pGEXPOL	pGEX 6P-2 derivative carrying full length Polymerase gene fused to His tag	This study
pGEXPOL (exo-)	pGEX 6P-2 derivative carrying N-terminal truncated Polymerase gene fused to His tag	This study

UNIVERSITY of the
WESTERN CAPE

2.3 Media

All media used in this study is listed below:

LB Medium (Luria-Bertani Medium)

Constituent	L^{-1}
Tryptone	10 g
Yeast extract	5 g
NaCl	10 g

The pH was adjusted to pH 7.0 with 5 N NaOH

SOB medium

Constituent	L^{-1}
Tryptone	20 g
Yeast extract	5 g
NaCl	0.5 g
KCl (250 mM)	10 ml

UNIVERSITY of the
WESTERN CAPE

The pH was adjusted to 7.0 before autoclaving. After autoclaving the broth was cooled to ~50 ^{o}C and the following filter sterilized solution (L^{-1}) added aseptically:

MgCl2 (2 M)	5 ml

SOC Medium

Constituent	L^{-1}
Tryptone	20 g
Yeast extract	5 g

NaCl	0.5 g
KCl (250 mM)	10 ml

The pH was adjusted to 7.0 before autoclaving. After autoclaving the broth was cooled to ~50°C and the following filter sterilized solutions added aseptically:

MgCl2 (2M)	5 ml
Glucose (1M)	20 ml

Super Broth

Constituent	L^{-1}
Tryptone	32 g
Yeast Extract	20 g
NaCl	5 g
NaOH (1 M)	5 ml

UNIVERSITY of the
WESTERN CAPE

2.4 DNA extraction procedures

2.4.1 Zhou method

Community DNA extractions were performed according to the modified Zhou protocol (Stach et al., 2001). Aliquots of environmental soils (5 g) were weighed out into sterile 30 ml Nalgene® centrifuge tubes followed by the addition of 6.75 ml soil extraction buffer (1% CTAB [w/v]; 100 mM Tris, pH 8.00; 100 mM NaH_2PO_4, pH 8.00; 100 mM EDTA; 1.5 M NaCl; 0.02% Protease K [w/v]). The tubes were incubated horizontally at 37 °C for 30 min with shaking. 750 µl 20% [w/v] SDS was added to each tube followed by additional 2 h incubation at 65 °C with gentle inversions every 20 min. Following incubation, the tubes were centrifuged at 3000 × g for 10 min at room temperature and the supernatant pooled into a sterile Nalgene® 30 ml centrifuge tube. An equal volume Phenol/Chloroform/Isoamyl was added and mixed gently followed by centrifugation at 16 000 × g for 10 min. Supernatants were again transferred to sterile Nalgene® 30 ml centrifuge tubes with the addition of an equal volume of chloroform. After careful mixing the tubes were centrifuged at 16 000 × g for 10 min at room temperature and supernatants recovered. Chloroform washes were repeated until the supernatants were clear. Once all washes were complete 0.6 volumes of isopropanol was added to the supernatants and DNA precipitation allowed to take place overnight at room temperature. DNA was pelleted by centrifugation at 10 000 × g for 10 min, washed with 70% ethanol, recentrifuged at 10 000 × g for 5 min, and air dried in a sterile hood. UHQ Millipore water was used to resuspend the DNA pellet and a small fraction was analysed by gel electrophoresis.

UNIVERSITY of the
WESTERN CAPE

2.4.2 Miller

DNA from all environmental soil samples were extracted using the Miller protocol (Miller et al., 1999). Between 0.5 and 1 g of soil was added to sterile 2 ml screw cap tubes containing 0.5 g sterile Quartz sand, followed by 300 μl phosphate buffer, pH 8.00, 300 μl lysis solution (0.5 M Tris-HCl, pH 8.00, 10% SDS [v/v], 100 mM NaCl) and 300 μl chloroform. The sample tubes were mixed and either shaken in a bead beater (Bio101 FastPrep FP120, Savant Instruments Inc. Holbrook, NY) at 4.5 $m.s^{-1}$ for 40 s or vortexed for 1 – 1.5 min at full speed, followed by centrifugation for 5 min at 13 000 × g. Supernatants were transferred to clean 1.5 ml Eppendorf tubes with the addition of 7 M NH_4AOc to achieve a final concentration of 2 M. Tubes were inverted several times until white flocculates appeared and centrifuged for 5 min at 13 000 × g. The supernatants were recovered and transferred to clean centrifuge tubes after which 0.6 volumes of isopropanol was added, the tubes inverted several times and incubated at room temperature for 15 min. DNA was collected at room temperature by centrifugation at 13 000 × g for 10 min, washed with 70% EtOH and air dried in a sterile hood. UHQ Millipore water was used to resuspend the air-dried DNA pellet and a small fraction was analysed by gel electrophoresis.

UNIVERSITY of the
WESTERN CAPE

2.5 DNA purification procedures

2.5.1 Sephadex G100/PVPP column purification

To remove humic acid contamination from environmental DNA samples, a modified Sephadex G100/PVPP method was employed (Stach *et al*., 2001). The end of a 1ml syringe was plugged with glass wool and packed with a 50% Sephadex G100 slurry (Pharmacia). Once the matrix had settled to approximately 400 µl, a 10% (w/v) PVPP solution was added to a final volume of 800 µl. Up to 100 µl of sample was loaded and the column was placed inside a 15 ml falcon tube and centrifuged for 3 min at 1500 rpm in an Eppendorf 5810R bench top centrifuge. The DNA was precipitated using a standard ethanol DNA precipitation method (Sambrook and Russell, 2001).

2.5.2 GFX™ column purification

Purification of DNA from either solution or agarose gels were performed using the GFX™ DNA and gel band purification kit (Amersham Biosciences) according to manufacturer's specifications.

2.5.3 Recovery of restriction digested DNA from agarose gels

Following electrophoresis and UV visualization of metagenomic DNA, the 2 – 10 kbp fraction was excised using a sterile scalpel blade. The excised gel was placed in a sterile 1.5 ml Eppendorf tube and the DNA recovered using the GFX™-gel extraction kit (Amersham Biosciences).

UNIVERSITY of the
WESTERN CAPE

2.6 DNA quantification

Environmental extracted and purified DNA was resuspended at 4 °C overnight in sterile water. Variable concentrations (5 ng, 10 ng and 20 ng) of uncut λ-DNA was electrophoresed along with the environmental DNA for 40 min at 80 V on a 1% (w/v) agarose gel. Verification of estimated DNA concentrations was performed using the Nanodrop ND-1000. The instrument was initialized using 2 μl UHQ Millipore water and thereafter blanked, at a wavelength of 260 nm, again using 2 μl UHQ Millipore water. Aliquots of 2 μl environmental resuspended DNA was used to determine concentrations of the respective samples.

2.7 Cloning procedures

2.7.1 Preparation of electro-competent cells

Electrocompetent DH5α, XL1blue and JM109 *E. coli* cells were prepared as outlined in Sambrook and Russell (2001), with slight modification. All glassware was thoroughly acid-washed with 30% H_2SO_4, rinsed and autoclaved prior to use. A single colony of the *E. coli* strain was inoculated into 30 ml of LB-broth and incubated at 37 °C with shaking until stationary phase. 10 ml of the culture was transferred to two aliquots of 500 ml of LB-broth and incubated at 30 °C until mid-logarithmic phase (OD_{600} of 0.4). The flasks were rapidly cooled in ice-water for 20 min and the cells were collected in polypropylene tubes by centrifugation at 1000 × g for 10 min in an Eppendorf 5810 R swing bucket centrifuge. The supernatant was decanted and the cells resuspended in equal volume ice-cold Millipore water. After harvesting the cells as above, the pellets were resuspended in 250 ml 10% glycerol, collected by centrifugation and the supernatant carefully decanted. The cell pellet was resuspended in 1 ml GYT medium and the cell density at OD_{600} adjusted to between 2×10^{10} to 3×10^{10} cells.ml^{-1}. The cells were aliquotted into 40 µl volumes, and stored at -80°C until required.

UNIVERSITY of the
WESTERN CAPE

2.7.2 Preparation of chemically competent cells

Chemically competent cells were prepared according to the method of Hanahan (1983) with slight modification. All glassware was thoroughly acid-washed with 30% H_2SO_4, rinsed and autoclaved prior to use. A single colony of the *E. coli* strain was inoculated into 30 ml of LB-broth and incubated at 37 °C with shaking until stationary phase. 1 ml of the culture was transferred to 100 ml of LB-broth and incubated at 30 °C until mid-logarithmic phase (OD_{600} of 0.5). The flasks were rapidly cooled in ice-water for 20 min and 60 ml of the cells were collected in polypropylene tubes by centrifugation at 1000 × g for 10 min in an Eppendorf 5810 R swing bucket centrifuge. After discarding the supernatant, the cells were resuspended in 0.5 × volume filter sterilized competency buffer (0.1 M $CaCl_2$ [w/v], 0.07 M $MnCl_2$ [w/v] and 0.04 M NaOAc [w/v], pH 5.5) and incubated at 4 °C for 30 min. Following incubation the cells were harvested by centrifugation at 1000 × g for 5 min and resuspended in 7.5 ml competency buffer. 575 µl 80% glycerol was added thoroughly mixed and the competent cells dispensed into 100 µl aliquots and stored at -80 °C until required.

UNIVERSITY of the
WESTERN CAPE

2.7.3 End treatment of DNA

2.7.3.1 5'-dephosphorylation of vector DNA

Dephosphorylation of 5' ends of GFX™-purified genomic DNA was performed using Shrimp Alkaline Phosphatase (SAP), Roche, according to manufacturer's specifications. The reaction was made up to a final volume of 50 µl by adding 12 µl SAP at 1 $U.\mu l^{-1}$ and 5 µl 10 × buffer (50 mM Tris-HCl, 5 mM $MgCl^2$, pH 8.5) to the eluted GFX™-purified DNA, followed by incubation at 37 °C for 1 h.

2.7.3.2 A-tailing of 3' termini of restriction digested metagenomic DNA

The addition of single adenosine nucleotides to the 3' ends of the partially digested metagenomic DNA was performed using a standard PCR reaction. Following restriction endonuclease digestion (Section 2.8.4) reaction mixtures were used directly in A-tailing PCR reactions at a concentration not exceeding 30% (v/v). PCR reactions (60 μl) contained 20 μl restriction digested metagenomic DNA reaction, 1 × NEB PCR buffer (20 mM Tris-HCL (pH 8.8), 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 1% [v/v] Triton X – 100), 0.1 mM dATP, and 1.5 μl Taq DNA polymerase. The PCR reaction mixtures were placed into an Applied Biosystems thermocycler Gene Amp® 2700 using the following conditions: 72°C for 30 min followed by rapid cooling to 4 °C and gel electrophoresis as described in Section 2.9.1.

UNIVERSITY of the
WESTERN CAPE

2.7.4 Ligation protocols

2.7.4.1 DNA insert ligation into vector DNA

A 20 µl reaction mixture was prepared to contain appropriate concentrations of vector and insert at a 1:3 molar ratio, reaction buffer at 5X final concentration, 10 U T4 DNA ligase and water to the final 20 ul volume. The reaction mixture was incubated overnight at 18^{o}C, followed by a heat inactivation step at 65 ^{o}C for 10 minutes.

2.7.4.2 Self circularisation ligation of linear DNA

A 20 µl reaction mixture was prepared to contain 50 ng of linearised DNA, reaction buffer at 5X final concentration, 5 U T4 DNA ligase and water to the final 20 µl volume. The reaction mixture was incubated overnight at 18 ^{o}C, followed by a heat inactivation step at 65^{o}C for 10 minutes.

2.7.5 Transformation protocols

2.7.5.1 Electroporation

An Eppendorf tube containing 40 μl of electrocompetent cells was removed from -80 °C and allowed to thaw on ice. 2 μl of ligation mix was added to the thawed cells and gently mixed. The mixture was returned to ice for ~ 1 min then pipetted into a pre-cooled 0.1 cm sterile electroporation cuvette (Bio-Rad Laboratories, Hercules, CA, USA). Electroporation was performed using the following conditions: 1.25 – 1.8 kV, 25 μF, 200 Ω. Immediately following electroporation, 950 μl TB broth, pre-warmed to 37 °C, was added to the cuvette, the cells transferred to a 15 ml Falcon tube and incubated at 37 °C for 1 h with agitation. The cells were plated in aliquots of 5 to 50 μl onto LB-agar plates supplemented with the appropriate antibiotic. Where applicable, recombinant transformants were selected by blue/white colour selection based on insertional inactivation of the lacZ gene. For this purpose, the cells were spread together with 40 μl of X-gal (2% [v/v] stock solution) and 10 μl IPTG (100 mM stock solution) over the

UNIVERSITY of the
WESTERN CAPE

surface of LB-agar plates, supplemented with the appropriate antibiotic and incubated overnight at 37 °C.

2.7.5.2 Chemical transformation

An Eppendorf tube containing 100 μl of chemically competent cells was removed from -80 °C and allowed to thaw on ice. 2 μl of ligation mix was added to the thawed cells and gently mixed. The mixture was incubated on ice for 30 min then heat-shocked at 42 °C for 90 s in a water bath. The Eppendorf tube was returned to ice for 2 min where after 900 μl of sterile LB-broth was added and the Eppendorf tube incubated at 37 °C for 1 h with agitation. The cells were plated in aliquots of 100 to 200 μl onto LB-agar plates supplemented with the appropriate antibiotic. Where applicable, recombinant transformants were selected by blue/white colour selection based on insertional inactivation of the *lac*Z gene. For this purpose, the cells were spread together with 40 μl of X-gal (2% [v/v] stock solution) and 10 μl IPTG (100 mM stock solution) over the surface of LB-agar plates, supplemented with the appropriate antibiotic and incubated overnight at 37 °C.

UNIVERSITY of the
WESTERN CAPE

2.7.5.3 Cloning in pCR TOPO TA Expression plasmid

The ligation reaction was set up as described in the pCRTOPO TA Expression kit instruction manual (Invitrogen). Dephosphorylated DNA (4 µl) was added to 1 µl (10 ng) pCRT7/CT-TOPO vector and 1µl salt solution (300 mM NaCl, 15 mM

MgCl2). The ligation mix (2 µl) was used to transform 40 µl One ShotÒ TOP10F' electrocompetent *E. coli* cells (transformation efficiency 109 CFU/1µg control plasmid DNA). Electroporation was performed using a BIO-RAD Gene Pulser. The transformation mix was added to a chilled Gene Pulser cuvette (0.1 cm electrode gap). The electroporation conditions were as follows; 1.5 Kv, 200 W and 25 µFD. After electroporation, 960 µl warm TB (1.2% w/v tryptone, 2.4% w/v yeast extract, 0.4% v/v glycerol) was added and the cells were allowed to recover by incubation at 37 °C for one hour.

2.7.5.4 Cloning in pGEMT-easy

Cloning with the pGEMT-easy vector system was performed as described in the cloning kit instruction manual (Promega).

UNIVERSITY of the
WESTERN CAPE

2.8 Screening strategies

2.8.1 Minipreps

2.8.1.1 Alkaline lysis method

Colonies were picked from the agar plates, inoculated into 5 ml of LB-broth supplemented with the appropriate antibiotic, and incubated overnight at 37 °C with agitation. Plasmid DNA was isolated from the cultures by the alkaline lysis method (Birnbiom and Doly, 1979), with the following modifications. After incubation, cells from 2 ml of each culture was collected in 2 ml Eppendorf tubes by centrifugation at 10000 × g for 1 min at room temperature. The supernatant was discarded and the bacterial pellet suspended in 400 µl of Solution 1 (50 mM glucose, 25 mM Tris·HCl, pH 8.0, 10 mM EDTA, pH 8.0). After incubation at room temperature for 10 min, 400 μl of Solution 2 (1% [w/v] SDS, 0.2 N NaOH) was added and the tubes were incubated on ice for 10 min. Following the addition of 300 μl of 7.5 M ammonium acetate (pH 7.6), the tubes were incubated on ice for 10 min, and then centrifuged at 13 000 × g for 5 min at room temperature. The plasmid DNA was precipitated from the supernatant by the addition of 650 μl isopropanol for 10 min at room temperature. The precipitated plasmid DNA was collected at room temperature by centrifugation at 12 000 × g for 10 min and the supernatant discarded before addition of 100 μl of 2 M ammonium acetate (pH 7.4). The tubes were incubated on ice for 10 min. Following ambient centrifugation at 12 000 × g for 5 min, 110 μl of isopropanol was added to the supernatant and the tubes incubated at room temperature for 10 min. Precipitated DNA was collected and the pellets washed with 70% ethanol to remove residual salts from the DNA. The DNA was air-dried and resuspended in UHQ Millipore water. Plasmid DNA was analyzed on a 1% (w/v) agarose gel as described in Section 2.9.1.

UNIVERSITY of the
WESTERN CAPE

2.8.1.2 Talent Kit

Plasmid extractions performed for subsequent nucleotide sequence analysis was performed using the Talent plasmid purification kit.

2.8.2 PCR protocols

2.8.2.1 Standard PCRs

In order to amplify target DNA, 0.2 ml thin walled tubes were used in a GeneAmp PCRsystem 2700 (Applied Biosystem) or Eppendorf Mastercycler gradient thermocycler equipped with a heated lid. Unless otherwise stated, the standard 50 µl-PCR reaction contained the following reagents: 1x PCR buffer, DNA template (20 ng plasmid or 50 ng chromosomal DNA), 0.5 µl of the upstream and downstream primers, 200 µl dNTPs mixture (dATP, dCTP, dGTP and dTTP), 1.25 U of Taq DNA polymerase. Reactions were made up to 50 µl with sterile ddH2O. PCR additives such as 1.0 % (v/v) DMSO and 200 mM betaine were added as required and the final reaction volumes were adjusted appropriately. Pfu DNA polymerase was used instead of Taq DNA polymerase to amplify PCR products with blunt ends, when necessary.

UNIVERSITY of the
WESTERN CAPE

Thermocycling conditions: 94 ^{o}C for 4 min (20-30) cycles: 94 ^{o}C for 30 sec, X ^{o}C, for 45 sec, 72 ^{o}C for 2 min and 1 cycle: 72 ^{o}C for 10 min (X) denotes relevant annealing temperature which was chosen 5 ^{o}C below the assumed primer melting temperatures calculated using the following formula (Tm= [no. of GC] x 4 + [no. of AT] x 2 ^{o}C).

The oligonucleotides used in this study are listed in Table 2.3. References with regard to which experiments the primers were implemented is included in the text.

Table 2.3 Oligonucleotides used in the study

Identifier	*Oligonucleotide sequence*
DUFiF	5’ TGCCGGATGTTGTTCAGATG 3’
DUFiR	5’ AAGTTGGCACAGCTTGGAAG 3’
DUF29exF	5’ ATAGGATCCATGGTTGTAAAAGATATA 3’
DUF29exR	5’ ATTAAGCTTGATTAAACTCTCTTTCCAT 3’
P2.96F	5’ ACTAATCTAGAGCGCCCGTTATTGGTG 3’
P2.96R	5’ GAGACAAGCTTAAGCTGCATACGTGAG 3’
PIF	5’ GGTGCAGCTTGACTTAGA 3’
PIR	5’ AAGCAATCGCGAATAACCA 3’
PIF2	5’ GCTCCTCACTGAGACGTT 3’
ArgJf	5’ ATGAATATTAAAATGGGTGTTGC 3’
ArgJr	5’ CTAAGTAGTATACTCTGCATA 3’
DNAPOLr	5’ CGGATCACGCAGTTGGTGC 3’
DPOLf	5’ AACGACGAGCTGCGAACGC 3’
PolfullF	5’ AGAGAATTCGCGCCCGTTATTGGTGTTAGT 3’
PolfullR	5’ TAGTCTAGACTCAATGGCGCAAGCAATCG 3’
USPf	5’ ATGTATAAAAATATCCTGGTTGG 3’
USPr	5’ TTAGGGAACTACCAAAACCTC 3’
T7	5’ AATACGACTCACTATAGG 3’
V5	5’ ACCGAGGAGAGGGTTAGGGAT 3’

UNIVERSITY of the
WESTERN CAPE

2.8.2.2 Colony PCR

The putative recombinants were aseptically inoculated into LB-broth containing the appropriate antibiotic (Section 2.1.2) and incubated overnight at 37 °C with agitation. Cultures were further analysed by pipetting 200 μl of each into 0.6 ml PCR tubes. The tubes were centrifuged at 13 000 × g to pellet the cells and the supernatant discarded. The cells were resuspended in 200 μl UHQ Millipore water and lysed by incubation at 98 °C for 5 min. Tubes were then centrifuged at 13 000 × g for 5 min to pellet cell debris. The DNA-containing supernatant served as template in 30 μl PCR reactions performed essentially as described in Section 2.6 (the annealing temperature was lowered to 49 °C). An aliquot of each PCR reaction was analysed by gel electrophoresis as described in Section 2.9.1.

UNIVERSITY of the
WESTERN CAPE

2.8.3 Whole genome amplification

Whole genome amplifications were performed with the Repli-G kit (Amersham biosciences), according to manufacturer's specifications. The amount of input DNA ranged between 10 ng and 20 ng as required.

2.8.4 Restriction digestions

All restriction enzyme digestions were performed in sterile Eppendorf tubes in small reaction volumes (10 – 20 μl). The reactions contained the appropriate volume of 10 × buffer supplied by the manufacturer for the specific enzyme, and 5 – 10 U of enzyme per μg of plasmid or genomic DNA. Reactions were incubated for either short periods, 0.5 – 1.5 h, or overnight in a water bath at 37 °C, unless specified otherwise. When digestions included two enzymes requiring different salt concentrations for optimal activity, the enzyme requiring a lower salt concentration was used first after which the salt concentration was adjusted and the second enzyme added. The digestion products were analyzed by gel electrophoresis in 1% or 2% (w/v) agarose gels as described in Section 2.9.1.

2.8.5 Sequencing

Sequencing of cloned insert DNA was performed using the MegaBACE 500 Automated Capillary DNA Sequencing System (Amersham Biosciences).

2.9 Electrophoresis

2.9.1 Agarose gel electrophoresis

Analysis of DNA was performed using agarose gel electrophoresis (Sambrook et al., 1982). Horizontal 0.8% – 2% (w/v) TBE agarose slab gels were cast and electrophoresed at 100 V in 0.5 × TBE buffer (40 mM Tris·HCl, 1 mM EDTA, 20 mM boric acid, pH 8.5). Where DNA was to be recovered and used in downstream applications, 0.8% – 2% (w/v) TA agarose slab gels were cast and electrophoresed at 100 V in 0.5 × TA (20 mM Tris-HCl, 10 mM glacial acetic acid, pH 8.5). To allow visualization of the DNA on a UV transilluminator, the gels were supplemented with 0.5 μg.ml^{-1} ethidium bromide. The DNA fragments were sized according to their migration in the gel as compared to that of standard DNA

UNIVERSITY of the
WESTERN CAPE

molecular markers (Lamda DNA restricted with PstI; 100 bp Hyperladder, Bioline).

2.9.2 Denaturing SDS-polyacrylamide gel electrophoresis

Denaturing sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was carried out according to the method of Laemmli (1970) using a Mighty smallTM SE 280 vertical Slab unit (Hoefer Inc, USA) with 1 mm or 1.5 mm gels containing 10-15% acrylamide (Table 2.3 and 2.4). Protein samples (8 µl) were prepared by suspending in SDS-PAGE loading buffer (2 µl) and were denatured by boiling in sealed tubes for 5 min, followed by centrifugation (16000 x g, 2 min, at room temperature). The samples were loaded onto the gel and electrophoresed at room temperature under constant voltage 60 V until the dye front migrated into the separating gel, followed electrophoresis at 120 V. Pre-stained SDS-PAGE protein molecular weights (Perfect Protein^{TM} Markers) were from Novagen and contained 15, 25, 35, 50, 75, 100 and 150 kDa sizes.

UNIVERSITY of the
WESTERN CAPE

2.9.3 Staining and Destaining of PAGE gels

Following electrophoresis, gels were stained with three Coomassie Brilliant Blue staining (CBS) solutions as follows: Gels were firstly soaked in 50 ml CSMI (10% (v/v) acetic acid; 0.125% (w/v) coomassie brilliant blue G; 25% (v/v) isopropanol) , heated for 30 sec in a microwave, followed by incubation at room temperature with shaking for 20 min. The CSMI solution was then discarded, the gels rinsed in dH2O, followed by soaking in CSMII (10% (v/v) acetic acid; 0.003% (w/v) coomassie brilliant blue G; 10% (v/v) isopropanol) for 20 min with shaking after brief heating for 30 sec in the microwave. Staining with CSMIII (10% (v/v) acetic acid; 0.003% (w/v) coomassie brilliant blue G) was performed using the same procedure. Destaining was carried out overnight by soaking in SDS-PAGE destaining solution (10% (v/v) acetic acid; 1% (v/v) glycerol).

.

2.10 Protein expression and analysis

2.10.1 Expression protocol

Super Broth media plus amplicillin was inoculated with 100 ul of a log phase culture of DH5α containing the expression construct. The culture was grown to an OD600 of 0.3 at 37 °C, and then induced with IPTG (0.5, 1.0 and 2.0 mM). Aliquots were removed after time increments of 2, 4 and 16 hours. An aliquot was stored at -20 °C for SDS-PAGE (crude extract) and an aliquot was sonicated with a Sonopulse sonicator (soluble extract) and stored at -20 °C.

UNIVERSITY of the
WESTERN CAPE

2.10.2 Purification protocol

Cells were centrifuged and resuspended in 3 ml buffer A (50mM Tris HCl, pH 7.9, 50mM dextrose, 1mM EDTA containing 4mg/ml lysozyme), and incubated 15 min at room temperature. Three ml of buffer B (10mM Tris-HCl, pH 7.9, 50mM KCl, 1mM EDTA, 0.5% Tween 20, 0.5% Nonidet P40) was added and the mixture was incubated for 60 minutes at either 50 or 60 °C, as specified. Cell debris and denatured protein were removed by centrifugation at 12 000 g for 10 minutes at 4 °C in a Sorvall RC2B centrifuge using a SS34 rotor. This lysate was mixed with an equal volume of storage buffer (50 mM Tris-HCl, pH 8.0, 100 mM NaCl, 0.1 mM EDTA, 0.5 mM DTT, 1% Triton X-100) containing 50% glycerol, followed by addition of an equal volume of storage buffer containing 75% glycerol. This mixture was then stored at -20 ^{o}C.

2.10.3 Polymerase activity assay

Standard PCR reaction mixtures were prepared as outlined in Section 2.8.2.1, with the following modifications. Fractions of the cellular lysate were diluted 1:10, 1:100 and 1:1000 and were added in 1ul aliquots, instead of commercial DNA polymerase, to test for DNA polymerase activity. Amplification was performed with the Uspf and Uspr primers with clone 2.142 as template for amplification of the Usp ORF (Section 2.8.2.1).

UNIVERSITY of the
WESTERN CAPE

2.10.4 Activity assay for the 5' exo^{-} deletion mutant polymerase

Template DNA was prepared by restriction enzyme digestion of 2 ug of puC18 plasmid DNA with *Hind*III and *Sac*I (Section 2.8.4). The linearised vector backbone was recovered by excision from a 1% agarose and purified by GFX^{TM} column purification (Section 2.5.2). The linearised pUC18 vector backbone (0.4 μg) was treated with 1 ul aliquots of the cellular expression extract, 0.5 mM of each dNTP (dATP, dCTP, dGTP and dTTP), 2 µl 10X Klenow Buffer and H_2O to a final volume of 20 μl. The same protocol was implemented for the control reaction where 5 U commercial *E .coli* Klenow fragment (Fermentas) was used, instead of the cellular expression extract. All reaction mixtures were incubated at

37, 45 and 50 ^{o}C for 30 min, followed by a heat inactivation step at 70 ^{o}C for 10 min. The end-treated pUC18 vector backbone was used as template, at 50 ng, for recircularisation (Section 2.7.4.2). Transfromation of the reaction mixture was performed as desribed in Section 2.7.5.1. The number of transformants was counted for each of the plates at the different temperatures.

2.11 Computational strategies

2.11.1 BLAST analysis

Unless otherwise stated, all BLAST analyses were performed under the default conditions set for the program (www. ncbi.nlm.nih.gov/BLAST).

2.11.2 Multiple sequence alignment

UNIVERSITY of the
WESTERN CAPE

Multiple sequence alignments were performed using the ClustalX programme on the Bioedit platform (Hall, 1999). Unless otherwise stated, alignments were run under the BLOSUM 60 algorithm.

2.11.3 Signal sequence prediction

SignalP (Hendrik et al., 1997) was used to predict the presence, location and probable cleavage site of signal sequences that might be present in the query. The sequences of amino acid queries were submitted in FASTA format and evaluated with neural networks trained on both Gram positive and Gram negative bacteria. The outputs were reported as the probability of the sequence containing a signal peptide.

2.11.4 Secondary structure prediction

The NPS server (http://npsa-pbil.ibcp.fr/cgibin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) predicts the secondary structure of a amino acid query sequence from the consensus of multiple individual secondary structure prediction programs. The amino acid sequence of the query was read as input into the NPS server. The consensus of the query sequence was derived from the outputs of DSC, MLRC and PHD. The program reported an alignment of the DSC, MLRC and PHD outputs along with the consensus generated from these programs.

UNIVERSITY of the
WESTERN CAPE

2.11.5 Prediction of trans-membrane domains

2.11.5.1 Hydrophobicity plot determination

Hydrophobicity plots were constructed from the amino acid sequences of the relevant queries. The plots were constructed in Bioedit with the algorithm of Kyte and Doolittle. The window size for calculation of the hydrophobicity plot was set at 20 amino acids.

2.11.5.2 SOSUI

The SOSUI program calculates the average hydrophobicity for an input sequence and reports a prediction on whether the protein is soluble or insoluble (Hirokawa et al., 1998). All SOSUI queries were performed in accordance with the standard procedures outlined for the program. The program reports the calculated average hydrophobicity for the input sequence as well as a prediction on whether the protein is soluble or insoluble.

2.11.6 MOTIF searches

The amino acid translations of nucleotide sequences were submitted to the MOTIF server (http://motif.genome.jp/) in FASTA format. All queries were run under the default conditions outlined for the program. The program reports the presence of domains that display sequence similarity with the query.

2.11.7 PFP prediction

UNIVERSITY of the WESTERN CAPE

The PFP program predicts molecular function, biological process and cellular component from the Gene Ontologies (Hawkins et al., 2006). The program also predicts binding sites and functionally important residues for the query sequence. The scores for the predictions are reported in order of relative probability within the respective functional categories, as opposed to global probability. The amino acid translations of nucleotide sequences were submitted to the PFP server in FASTA format. All queries were run under the default conditions set for the program.

2.11.8 Detection of distant homologues with HMMs

The amino acid translations of nucleotide sequences were submitted to the FUGUE (Shi et al., 2001), GenTHREADER (Liam and Jones, 2003), SAM T02 (Karplus et al., 2003) and LOMETS servers in FASTA format. These programs are all fully automated and subsequent to submission require no further user input. All queries were run under the default conditions set for the programs. The FUGUE, GenTHREADER and SAM T02 servers report a list of matching template PDB identifiers as well as the statistical significance of the match.

The LOMETS metaserver reports the highest matching templates from FUGUE (Shi et al., 2001), HHSEARCH (Soding, 2005), PROSPECT2 (Xu and Xu, 2000), SAM-T02 (Karplus et al., 2003), SPARKS2 (Zhou and Zhou, 2004), SP3 (Zhou and Zhou, 2005), PAINT, PPA-I and PPA-II (Wu and Zhang, 2007). The LOMETS server also reports a list of templates that is predicted as a consensus of the multiple programs.

UNIVERSITY of the
WESTERN CAPE

2.11.9 Homology modeling

2.11.9.1 Modeller

The Modeller program (Eswar et al., 2000) constructs homology models for query sequences from the atomic coordinates of a relevant template structure. The programme requires: (i) an alignment file of the query and template, (ii) the atomic coordinates of the template and (iii) a script file for the actual model construction. All alignment files were generated from ClustalW. The atomic coordinates of templates were retrieved from the PDB database (www.pdb.org). The actual model construction was run under the Automodel function in the Modeller program.

2.11.9.2 SWISS-Model

SWISS-Model automatically identifies suitable homologs for a query sequence from which a homology model is constructed (Schwede et al., 2003). Unless otherwise stated, all queries were performed under the first approach options set for the program (http://swissmodel.expasy.org//SWISS-MODEL.html).

2.11.10 Model validation

2.11.10.1 RAMPAGE

Model validation on the RAMPAGE server (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) was performed under the standard procedures outlined for the program.

UNIVERSITY of the
WESTERN CAPE

2.11.10.2 Verify3D

Model validation on Verfy3D (Eisenberg et al., 1997) was performed under the standard procedures outlined for the program.

2.11.11 Genomic context mapping

Standard BLAST searches were implemented in order to identify homologs of query ORFs. The organisms that harbor the homologous genes were identified from the genome mapping information at the NCBI (http://www.ncbi.nlm.nih.gov/). Information on genes that flank the loci of the homologous genes was extracted from genomic maps of the organisms. Functional information for the selected genes at the loci was assigned as reported from the NCBI database.

2.11.12 Mapping ORFs to Clones

The aligned regions of the partially predicted ORFs, from the BLAST analysis, were compared against the sequence information of the respective clones. The presence of the full length ORFs was evaluated on the basis of: (i) the size of the entire clone and, (ii) the presence of additionally mapped ORFs for the same clone.

UNIVERSITY of the
WESTERN CAPE

Chapter 3

Library Construction and Sequencing

3.1 Introduction

The prokaryotes are the most diverse group of organisms on earth, represent the largest biomass and are intimately involved in all aspects of making the planet liveable (Rodriguez-Valera, 2004). As a consequence microbes are also the biggest contributors of bioactive compounds and enzymes with applications in biotechnology. Such enzymes have primarily been sourced from the cultivatable microbial population, which represents roughly 1% of the entire microbial continuum (Aman et al., 1995). This inability to cultivate a greater proportion of the microbial continuum has become the rate limiting step in the identification of novel enzymes (Lorenz and Eck, 2005). Improvements in cloning and sequencing technologies now allow for the direct extraction and cloning of genomic DNA as a means to circumvent these limitations. These technologies have further been extended to the study of entire microbial populations in a discipline termed environmental metagenomics, which involves the construction of environmental libraries that are subsequently screened by either function driven or sequence driven approaches.

UNIVERSITY of the
WESTERN CAPE

The success of metagenomic sequencing is dependant on the selection of the environmental sample, the sampling strategies employed, storage of the samples, extraction techniques, selection of appropriate vector systems for cloning and the screening strategy (Rodrigez-Valera, 2004). To realize the full potential of soil microorganisms, as a resource for novel enzyme identification, the full complement of genetic information should be accessible for investigation using molecular biological strategies (Gabor *et al*., 2003). In addition, all possible steps should be taken to reduce the possibility of any degree of bias being introduced in the library construction process.

The implementation of these strategies has been demonstrated by the metagenomic sequencing of diverse environments (Table 3.1).

UNIVERSITY of the WESTERN CAPE

Table 3.1. Summary of published metagenome sequencing strategies.

Community	*Species*	*Sequence (Mbp)*	*Reference*
Acid mine biofilm	5	75	Tyson et al. (2004)
Sargasso Sea	1,800	1,600	Venter et al. (2004)
Minnesota soil	3,000	100	Tringe et al. (2005)
Whale falls	150	25	Tringe et al. (2005)
Deep-sea sediment	?	111	Hallam et al. (2004

The published environmental metagenomic sequencing projects provide insight into the gene diversity, species diversity and the biochemistry of the respective environments, but also highlight limitations in current sequence assembly technologies for environments with high diversity (Venter et al., 2004; Tringe et al., 2005 and Hallam et al., 2004). The large number of novel and uncharacterized genes identified from these studies further highlights the potential of a metagenomic approach for identification of enzymes with novel attributes (Tyson et al., 2004; Venter et al., 2004; Tringe et al., 2005 and Hallam et al., 2004).

Within this context, the thermophilic metagenomes potentially represent a significant resource for the identification of a greater number of thermostable proteins. The first step toward exploiting this resource would be to evaluate the application of the current DNA extraction, cloning and sequencing approaches on these novel metagenomes.

UNIVERSITY of the WESTERN CAPE

3.2 Aim

The specific aim of this section was to construct metagenomic libraries from the hyperthermophilic environmental samples. The objectives of the current section are to illustrate the implementation of such a strategy through:

- The extraction of metagenomic DNA from a Hyperthermophilic hydrothermal soil sample.
- The subsequent manipulation and cloning of the metagenomic gene complement as small to medium insert libraries.
- Generating sequence information from the metagenomic libraries
- Analysis of sequence information through basic sequence homology strategies (BLAST).

3.3 Results

3.3.1 Sample Selection

The environmental conditions of the samples collected at locations in Tengchong County, Yunnan Province, SW China are outlined in Table 3.2. These samples were extracted using the Zhou method. Samples TC 9 and TC 11 were selected for extraction based on their neutral pH values and high sampling temperatures, which serves as an ideal resource for identification of thermostable enzymes.

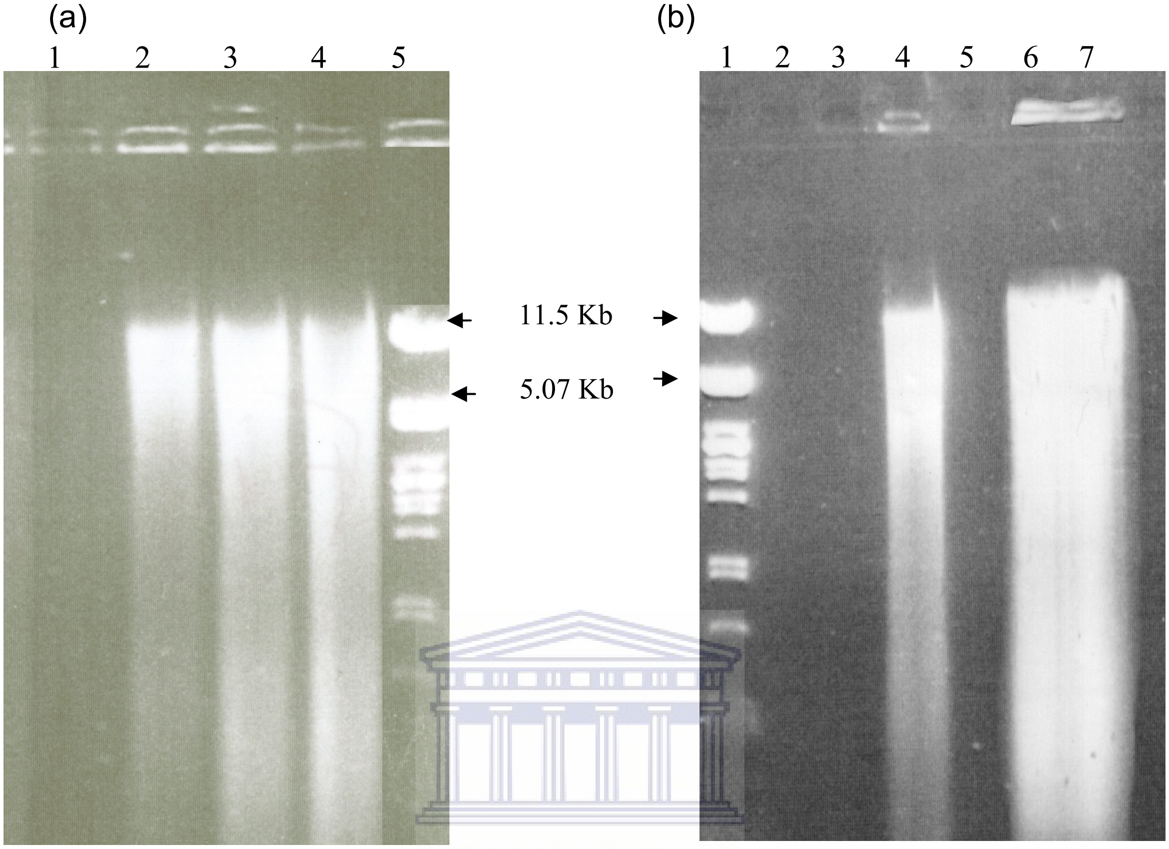
Table 3.2. Properties of Chinese hydrothermal samples

Sample	*pH*	*Temperature (^{0}C)*	*DNA yield (ug/g)*
TC1	9	79-80	<0.1
TC3	5.5	92-96	0.2
TC5	6	69.8	0.4
TC7	6	65.2	40
TC9	8	82-84	0.1
TC10	8	66	25
TC11	8	78	22
TC12	2	63.8	0
TC13	2	79-80	0

UNIVERSITY of the
WESTERN CAPE

3.3.2 Metagenomic DNA Library Construction

Metagenomic DNA from the hyperthermophilic samples TC9 and TC11 was extracted, in triplicate, using the Miller and modified Zhou methods, as outlined in sections 2.4.2 and 2.4.1. The electrophoresis of the extraction products is represented in Figures 3.1 (a) and (b). A large proportion of the extracted product was within the range of high molecular weight DNA, while significant shearing could also be observed.



UNIVERSITY of the
WESTERN CAPE

Figure 3.1. DNA extraction from environmental samples
(a) DNA extracted from sample TC9 using the Miller and Zhou methods
Lane 1, sterile water as negative control; lanes 2 and 3, TC9 DNA extracted using the Zhou method; lane 4, TC9 DNA extracted using the Miller method; lane 5, λ Pst molecular weight marker.
(b). DNA extracted from sample TC11 using the Miller and Zhou methods.
Lane 1, λ Pst molecular weight marker; lane 3, sterile water as negative control; lane 4, TC11 DNA extracted using the Miller method; lanes 6 and 7, TC11 DNA extracted using the Zhou method.

The extracted DNA was purified using a three step, modified Sephadex spin column clean-up procedure as outlined in Section 2.5.1. The DNA samples were quantitated spectrophotometrically (Section2.6) for protein contamination ($A_{260/280}$) and humic acid contamination ($A_{260/230}$) during each stage of purification (Table 3.3). Samples were stored at -80^{o}C and -20^{o}C.

Table 3.3 Quantitation of soil extracted DNA

Sample	Extraction method	Directly after extraction		After three rounds of sephadex treatment	
		$A_{260/280}$	$A_{260/230}$	$A_{260/280}$	$A_{260/230}$
TC9	Zhou	1.40	1.61	1.54	1.74
	Miller	1.43	1.68	1.52	1.72
TC11	Zhou	1.38	1.65	1.54	1.69
	Miller	1.45	1.66	1.51	1.74

The extracted, purified DNA was electrophoresed and a fraction in the region of 2 – 10 Kb was recovered (Section 2.5.3) by excision (Fig 3.2) and subsequently purified using the GFX purification kit (Section 2.5.2). Recovery of the appropriate fraction was demonstrated by electrophoresis (Fig 3.3).The purified DNA fragments were A-tailed (Section 2.7.3.2), dephosphorylated with shrimp alkaline phosphatase (Section 2.7.3.1) and cloned into the TOPO C/T vector (Section 2.7.5.3). The yields of recombinant clones in each library are shown in Table 3.4

UNIVERSITY of the
WESTERN CAPE

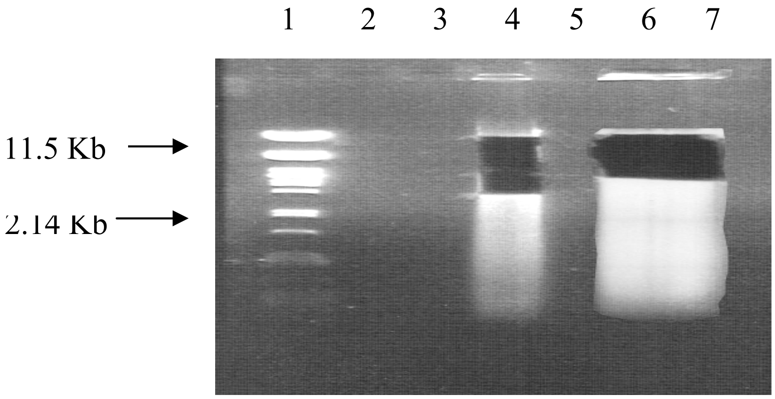


Figure 3.2 Recovery of 2 – 10 Kb DNA fraction.
Lane 1, λ Pst molecular weight marker; lane 4, DNA excised in the 2 – 10 Kb region; lanes 6 and 7, DNA excised in the 5 – 10 Kb region.

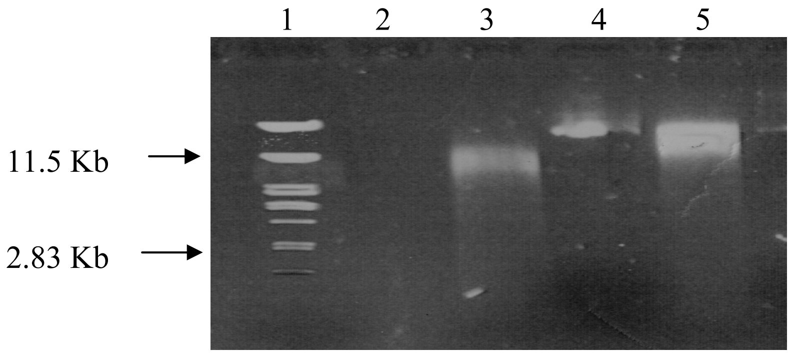


Figure 3.3 Confirming the excised DNA fraction for metagenomic cloning.
Lane 1, λ Pst molecular weight marker; lane 3, DNA excised in the 2 – 10 Kb region; lane 5, DNA excised in the 5 – 10 Kb region

Table 3.4 Library sizes for samples TC9 and TC11

Sample	DNA extraction method	Library identifier	Number of Clones
TC 9	Miller	9A	±270 000
TC 9	Zhou	9B	±290 000
TC 11	Miller	11A	±300 000
TC 11	Zhou	11B	±300 000

UNIVERSITY of the WESTERN CAPE

3.3.3 Metagenomic DNA library screening

Colony PCR (Section 2.8.2.2) and restriction analysis (Section 2.8.4) was used to determine the average insert size in the respective metagenomic library clones. Approximately 800 clones for each library were screened. The PCR amplicons or restriction fragments were electrophoresed on agarose gels and visualized against molecular weight standards, in order to get an estimation of the insert sizes (Figs 3.4 – 3.7). The average insert sizes for the screened libraries are indicated in Table 3.5.

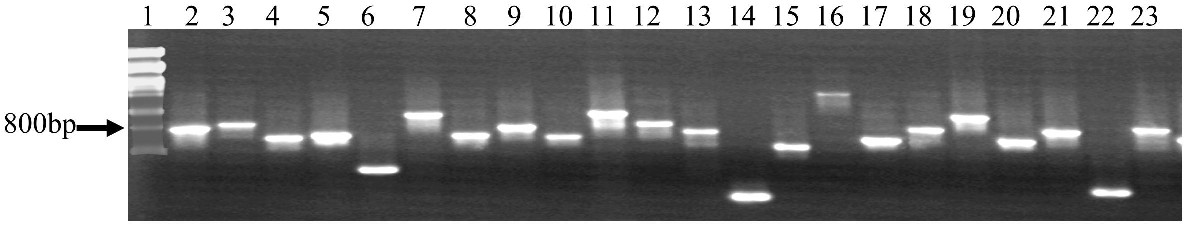


Figure 3.4 Colony PCR of selected clones from library 9A..
Lane 1, λ Pst molecular weight marker; lanes 2-23, Colony PCR amplicons from selected clones from library 9A.

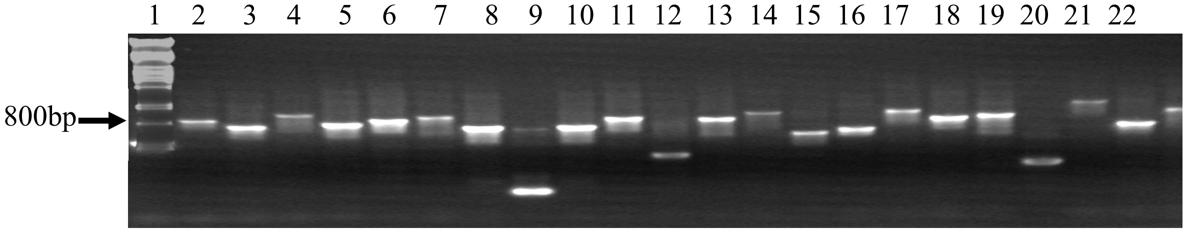
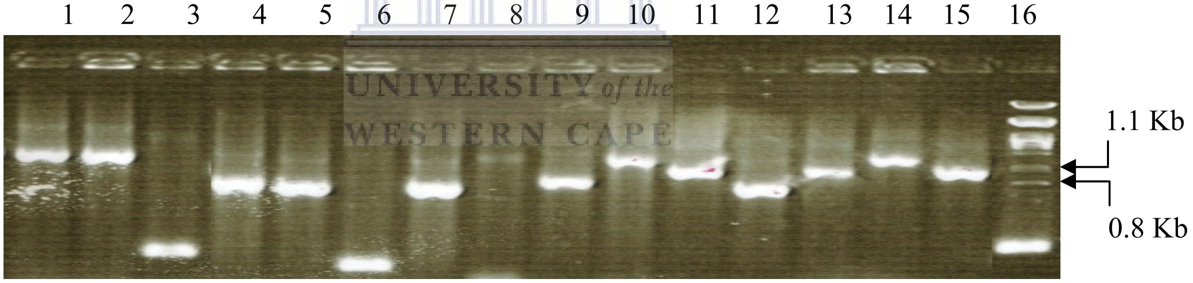


Figure 3.5 Colony PCR of selected clones from library 9B..
Lane 1, λ Pst molecular weight marker; lanes 2-23, Colony PCR amplicons from selected clones from library 9A.



UNIVERSITY of the
WESTERN CAPE

Figure 3.6 Colony PCR of selected clones from library 11A..
Lanes 1-15, Colony PCR amplicons from selected clones from library 11A; lane 16, λ Pst molecular weight marker.

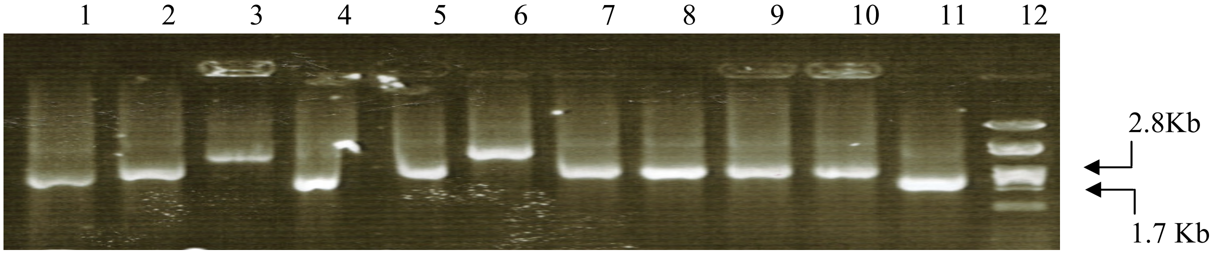


Figure 3.7 Colony PCR of selected clones from library 11B..
Lanes 1-11, Colony PCR amplicons from selected clones from library 11B; lane 12, λ Pst molecular weight marker.

Table 3.5 Average insert sizes for constructed libraries.

Library screened	*Total colonies screened*	*Avg insert size*
9A	800	800
9B	780	850
11A	840	760
11B	800	2000

3.3.4 Metagenomic DNA library sequencing

Clones containing inserts of greater than 2 Kbp were selected for 5’ and 3’ sequencing using the TOPO C/T vector specific primers (Section 2.8.2.1). A total of approximately 70 000 bp of sequence was generated from the 11B library using the MegaBACE 500 (Amersham Biosciences) sequencer (Section 2.8.5). This sequence data represented the bi-directional end-sequencing of 40 clones.

UNIVERSITY of the
WESTERN CAPE

3.3.5 Metagenomic DNA library sequence analysis

All sequences were subjected to BLASTx analysis (Section 2.11.1), using the default parameters for comparison against the NCBI microbial database, in order to establish significant matches to putative genes for further analysis (Table 3.6). A significant proportion (80%) of the sequenced clones however showed significant matches to the same putative gene. This gene was annotated as a conserved hypothetical protein from *Pelobacter carbinolicus DSM 2380 (accession gi|77543718).*

.

Table 3.6. Clone-end sequencing results from library 11 B

Clone ID	*Clone size*	*Highest BLAST match*	*e-value*	*Score*	*Clone region*	*Matching region*
2.91 T7	2.0 Kbp	conserved hypothetical protein [*Pelobacter carbinolicus DSM 2380*]	3e-27	125	92 – 553	18 – 171
2.91 V5	2.0 Kbp	Sodium/calcium exchanger region [*Moorella thermoacetica ATCC 39073*]	2e-17	92.8	526 – 323	22 – 89
2.136 T7	2.0 Kbp	Flagellar protein FliS [*Marinobacter aquaeolei VT8*]	8e-15	84.0	71 – 367	23 – 121
2.136 V5	2.0 Kbp	hook-filament junction protein 1 [*Salmonella typhimurium LT2*]	4e-11	71.6	16 – 513	406 – 552
		Flagellin, N-terminal [*Clostridium thermocellum ATCC 27405*]	2e-04	47.8	61 – 219	1 – 53
2.140 T7	2.0 Kbp	Cobaltochelatase, CobN subunit [*Chloroflexus aurantiacus J-10-fl]*	5e-71	270	27 – 476	1240 – 1389
2.140 V5	2.0 Kbp	Cobaltochelatase, CobN subunit [*Chloroflexus aurantiacus J-10-fl]*	8e-75	283	730 – 215	1216 – 1389

UNIVERSITY of the
WESTERN CAPE

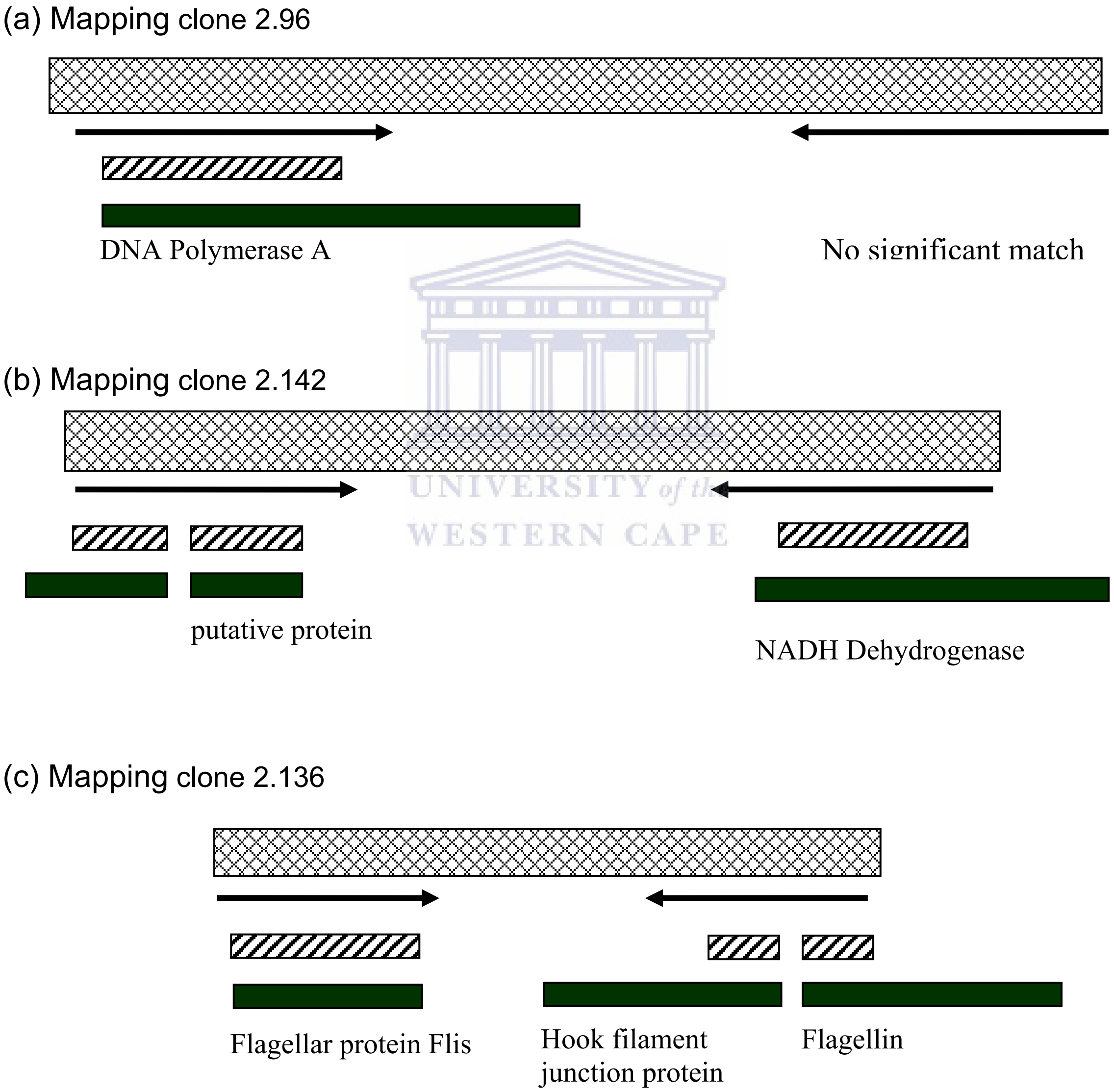
Table 3.6 (continued). Clone-end sequencing results from library 11 B

Clone ID	*Clone size*	*Highest BLAST match*	*e-value*	*Score*	*Clone region*	*Matching region*
2.139 T7	2.0 Kbp	Phosphoenolpyruvate synthase [*Trichodesmium erythraeum IMS101*]	4e-78	293	35 - 736	453 - 684
2.142 V5	2.0 Kbp	putative protein[*Hydrogenobacter thermophilus*]	1e-58	229	251 - 748	37 - 199
2.142 T7	2.0 Kbp	NADH dehydrogenase I chain G [*Aquifex aeolicus VF5]*	2e-08	62.4	45 – 473	359 - 498
2.96 T7	2.0 Kbp	No significant match				
2.96 V5	5.2 Kbp	DNA Polymerase A [*Chlorofexus aurantiacus*]	6e-72	273	281 – 664	8 - 135
2.141V5	2.0 Kbp	hypothetical protein pEA28_01 [*Erwinia amylovora*]	5e-88	66	54 - 490	2 - 436

UNIVERSITY of the WESTERN CAPE

3.3.6 Prediction of entire ORFs from sequence data

The reported gene matches from the BLASTx analysis were mapped to the respective clones (Section 2.11.12). Predictions were made for the presence of entire ORFs for the reported BLASTx matches (Figure 3.8).

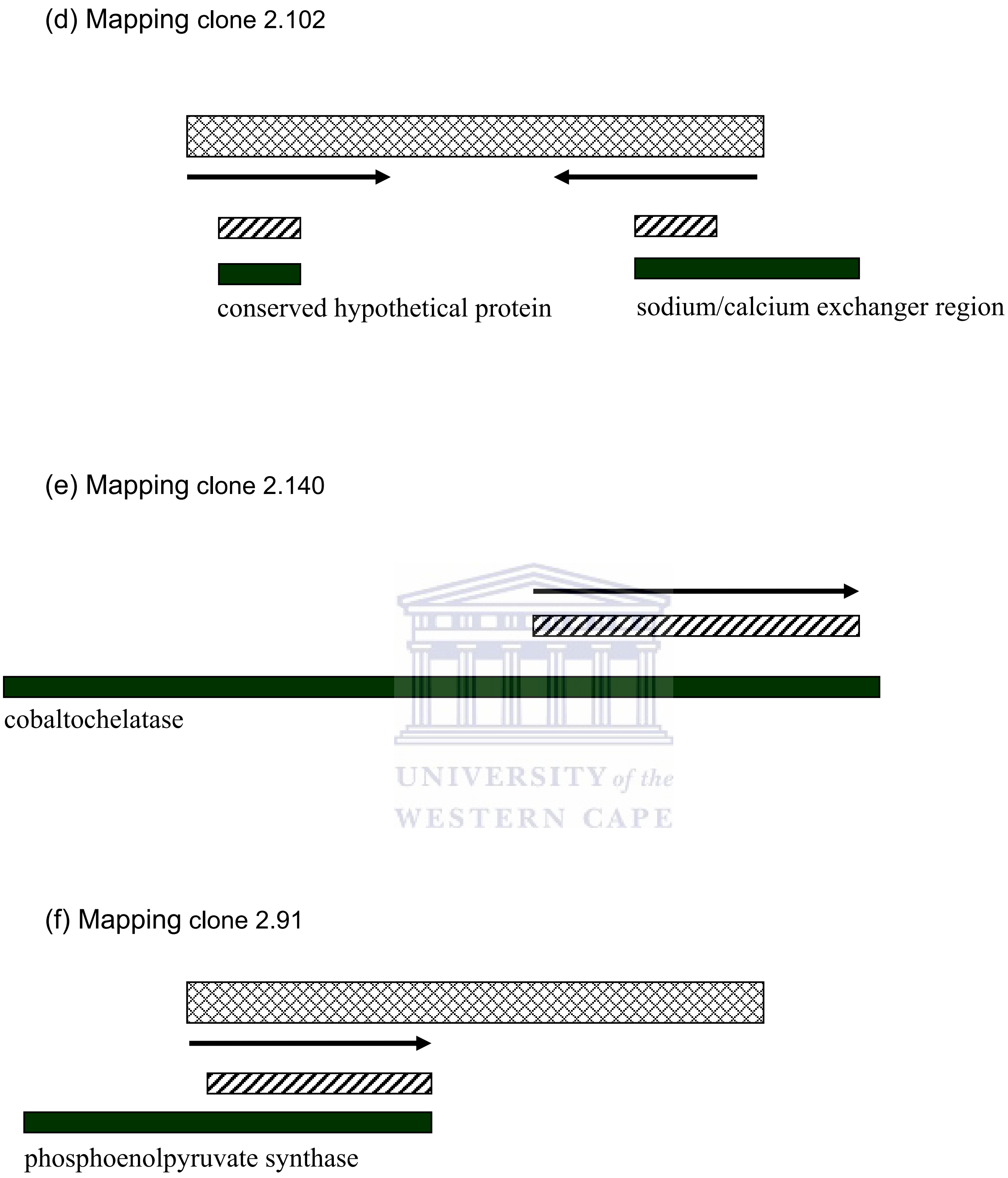

(a) Mapping clone 2.96

(b) Mapping clone 2.142

(c) Mapping clone 2.136

UNIVERSITY of the
WESTERN CAPE

Figure 3.8 Mapping the sequenced clones.
▩, relative size of the clone; →, region of the clone sequenced and orientation of sequencing; ▨, segment of the sequenced region with significant BLASTx result; ■ , predicted full length gene from highest BLASTx match.


(d) Mapping clone 2.102

(e) Mapping clone 2.140

(f) Mapping clone 2.91

conserved hypothetical protein
sodium/calcium exchanger region
cobaltochelatase
phosphoenolpyruvate synthase

UNIVERSITY of the
WESTERN CAPE

Figure 3.8 Mapping the sequenced clones. (continued)
▦, relative size of the clone; →, region of the clone sequenced and orientation of sequencing; ▨, segment of the sequenced region with significant BLASTx result; ■ , predicted full length gene from highest BLASTx match.

3.3.7 Continued sequencing of selected clones

The results of the BLASTx gene mapping for clone 2.96 indicated that sequence data was generated for 30% of the clone. The 5’ end sequence match to the DNA polymerase was predicted to span 3 Kb of the clone. The 3’ end sequence data yielded no significant BLASTx matches, prompting continued sequencing of the central region. Analyses of clone 2.142 revealed sequence coverage of 40%. The mapping data reveal that ORFs were predicted for the 5’ and 3’ regions. The unsequenced central region, spanning 2 Kb was selected for further sequencing.

A primer walking strategy was employed for the continued sequencing using the 5’ and 3’ end sequences as templates for primer design (Section 2.8.5).

UNIVERSITY of the
WESTERN CAPE

3.3.8 Prediction of gene sequences from additional sequence data

The internal sequence data for clones 2.142 and 2.96 were compared to the non-redundant Swiss-Prot protein database using the BLASTx programme (Section 2.11.1) The identified gene matches are indicated in Table 3.7.

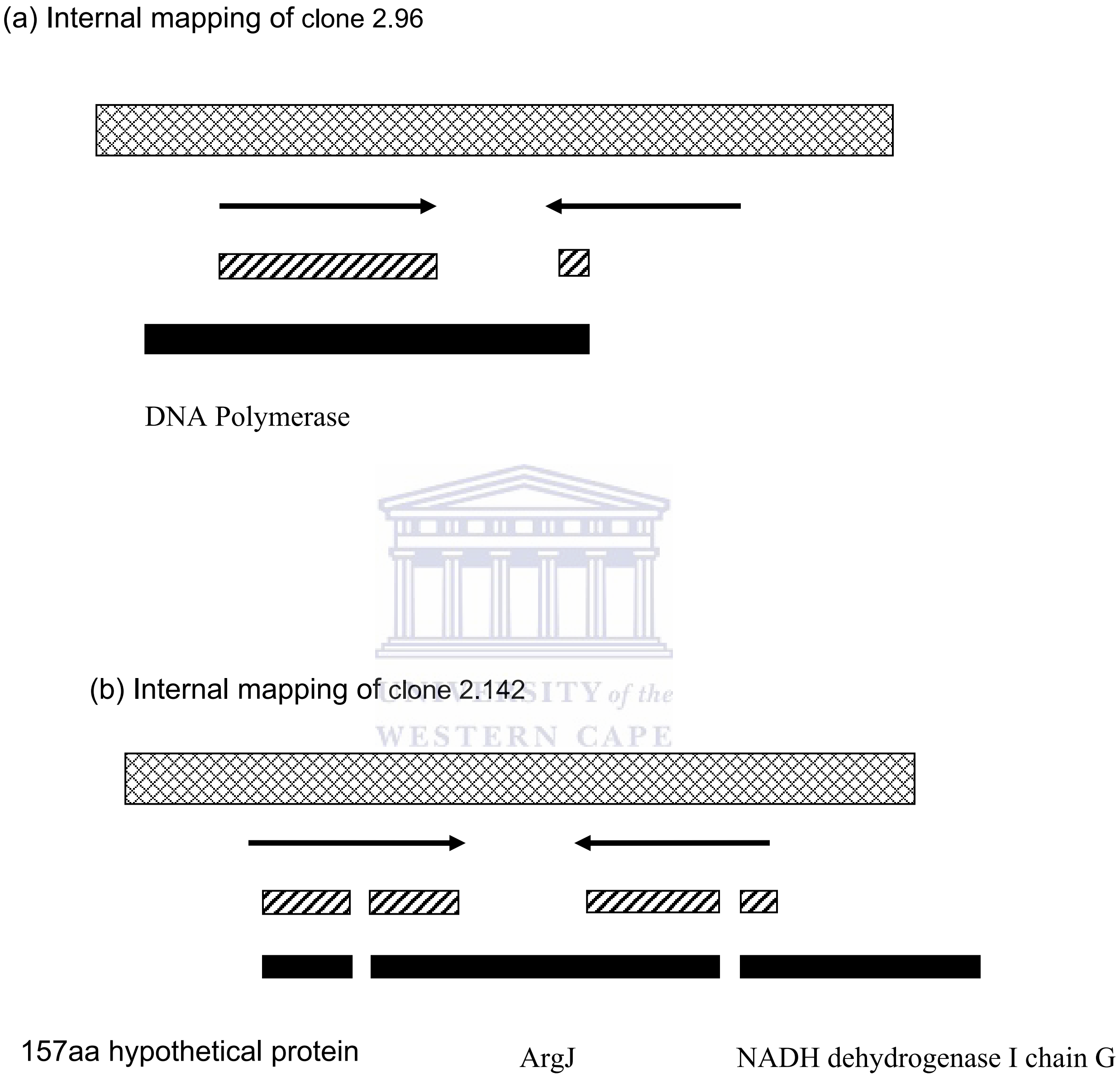
3.3.9 Mapping of the predicted genes

The matching BLASTx reported gene sequences for clones 2.142 and 2.96 were analysed by comparison against the respective clone sequences. This visualization allowed for prediction of the entire reading frames of the reported gene matches (Figure 3.9).

Table 3.7 Internal sequencing results from library 11B

Clone ID	*Clone size*	*Highest BLAST match*	*e-value*	*Score*	*Clone region*	*Matching region*
2.142 iF	2.0 Kbp	glutamate N-acetyltransferase [*Aquifex aeolicus VF5*]	2e-24	116	424-834	1-138
2.142 iR	2.0 Kbp	glutamate N-acetyltransferase [*Aquifex aeolicus VF5*]	1e-25	115	886-557	270-379
		157aa long conserved hypothetical protein [*Sulfolobus tokodaii str. 7*]	2e-15	84.7	551-84	2-156
2.96 iF	5.2 Kbp	DNA Polymerase A [*Chlorofexus aurantiacus*]	4e-72	66	445 - 3	732 - 882
2.96 iR	2.0 Kbp	No significant hit	-	-	-	-

UNIVERSITY of the
WESTERN CAPE


(a) Internal mapping of clone 2.96

(b) Internal mapping of clone 2.142

UNIVERSITY of the
WESTERN CAPE

Figure 3.9 Mapping the internal sequence of selected clones.
▩, relative size of the clone; →, region of the clone sequenced and orientation of sequencing; ▨, segment of the sequenced region with significant BLASTx result; ■ , predicted full length gene from highest BLASTx match.

3.4 Discussion

DNA extraction and library construction

The success of a metagenomic sequencing approach relies heavily on the quality of the input DNA. Factors affecting the DNA quality are the source of the sample, the sampling and storage strategies as well as the extraction methods employed (Tyson et al., 2004; Venter et al., 2004; Tringe et al., 2005 and Hallam et al., 2004). The degree of cell damage incurred during any subsequent freeze-thaw cycles of the samples depends on many parameters including the structure and content of the cell itself, the rate of freezing, the rate of thawing, the temperature of storage in the frozen state, and the medium in which the cells are suspended (Mazur, 1970; Morris *et al*., 1988). DNA is relatively stable during extended storage at temperatures varying from –70°C to 4°C (Shikama et al., 1965; Ross et al., 1990; Jerome et al., 2002). The samples used in these experiments were typically stored at - 80^{o}C and progressively thawed at – 20^{o}C o/n and – 4^{o}C o/n before extraction, reducing the possibility of significant degradation resulting from the freeze-thaw cycles.

UNIVERSITY of the
WESTERN CAPE

The Zhou and Miller extraction methods indicated that there were little differences in DNA yields between the two extraction methods for samples TC9 and TC11. These values were in accordance with results obtained from previous studies involving extraction of the same samples (personal communications). The Zhou extraction generally results in high DNA yields while ensuring a large population of intact high molecular weight DNA (Zhou and Tiedje, 1996). High molecular weight

DNA was successfully extracted using both methods, but the samples did exhibit a degree of degradation. The Zhou and Miller methods reportedly also result in reduced bias in comparison to methods that use initial extraction followed by lysis (von Wintzingerode et al., 1997). A potential drawback of the Zhou method is the high degree of humic acid contamination that remains subsequent to extraction (Kauffmann et al., 2004). Humic acid contamination affects spectrophotometric quantitation of extracted DNA as well as digestion and ligation of DNA, resulting in decreased transformation efficiencies (Tebbe and Vahjen, 1993). Most methodologies for separation of DNA from humic materials are dependent upon either differential levels of binding of humic substances and nucleic acids to a polymeric matrix or differential size fractionation (Holben et al., 1997; Kowalchuk et al., 1997; Stefan and Atlas, 1988; Stefan et al., 1988). The method by Stach et al (1997) which implements Sephadex G100/PVPP matrix columns was included in the purification procedure due to its relative simplicity and reported superior results. Spectrophotometric analysis of the sample indicated decreased $A_{260/230}$ ratio following several purification steps, which suggests that the humic acids had been removed (Yeates et al., 1998). The spectrophotmetric analysis of the sample is however limited as an indicator of humic acid removal due to the variability in spectral properties of different Humic acid preparations *(*Zipper et al., 2003). The latest extraction and purification strategies demonstrate use of a combination of previously described methods for efficiently removing humic acids from a variety of soil samples (Lakay et al., 2007).

UNIVERSITY of the
WESTERN CAPE

The use of small insert libraries allows for direct sequencing of clones in environmental metagenomic sequencing studies and eliminates the need for subcloning which is required for large insert libraries. The limitations of the small insert size can however be overcome by ensuring a good depth of sequence coverage, which allows for the formation of tighter overlapping scaffolds in the assembly process. The depth of coverage required is a direct function of the diversity of the sample (Tringe et al. 2005).

Four small to medium insert libraries were successfully constructed from the Chinese hyperthermophilic hydrothermal samples. The 9A, 9B and 11A libraries were calculated to contain in the order of 216 000 kbp, 232 000 kbp and 252 000 kbp of sequence respectively. Assuming a low degree of redundancy and an average gene size of 1kbp these libraries are estimated to contain 700 000 genes. These libraries constitute a potential rich source of novel genes/enzymes.

UNIVERSITY of the
WESTERN CAPE

The reported average insert size for libraries 9A, 9B and 11A is around of 800bp. This was significantly smaller than the expected 2 Kbp inserts, predicted on the basis of the library construction process. The inefficient removal of small DNA fragments prior to cloning typically results in the overrepresentation of these fragments in genomic libraries. Smaller DNA fragments are preferentially cloned due to higher ligation efficiency (Singh et al., 2003). The standard method for removal of smaller fragments from the DNA pool involves electrophoresis followed by excision of the required fraction (Osoegawa *et al.* 1998). Where a sufficiently large quantity of DNA starting material is available, a two-step size selection strategy can be employed in order to eliminate the potential of small insert contamination (Baba et al., 1999). However, due to low yields of DNA from Chinese hyperthermophilic samples, this two-step strategy was not a viable option. Instead, the method by Osoewaga (1998) was implemented in a single step for the recovery of the required fraction prior to cloning and could have resulted in the inefficient removal of the smaller than expected inserts.

Another possibility for the presence of small DNA fragments could be due to the persistent presence of degrading agents present in the extracted DNA. Humic acids represent a potential source of peroxy and hydroxyl radicals as well as hydrated electrons, hydrogen peroxide, singlet oxygen and superoxide. (Cooper et al., 1989). These chemical species in turn have the potential to promote redox reactions from oxidants generated by iron-mediated Fenton reactions, leading to DNA damage (Kayer and Imlay, 1996). The incomplete removal of humic acid substances, in conjunction with potential high iron content for the Chinese hydrothermal samples, could therefore have accounted for the presence of small DNA inserts in the constructed libraries.

UNIVERSITY of the
WESTERN CAPE

Library 11B contained inserts in the range of the expected 2 – 5 kbp, and based on the library size estimated to contain 600 000 genes. The relatively large inserts increase the statistical likelihood of recovering entire ORFs and this library was selected for metagenomic sequencing. The average size of a single sequence read is approximately 800 bp and sequencing from the ends of the 2 kbp clones provides sufficient sequence coverage for identification and prediction of potential ORFs.

Sequencing and analysis

The initial annotation for the Chinese Geothermal derived sequence data, for library 11B, was routinely performed using BLASTx analysis of the sequence data (Section 2.11.1). This same strategy was employed for the initial identification of genes in all the metagenomic sequencing strategies to date (Venter et al., 2004; Schmeisser et al.,2003 and Tringe et al., 2005). A total of 53 genes or partial gene sequences were identified from the 70 kbp of sequence data generated from the Chinese hyperthermophilic library 11B. This constituted 80% of the total sequence reads. The total number of identified genes is generally consistent with the calculated values from the studies of Scmeisser et al, 2003 (45 genes from 70kbp) and Venter et al., 2004 (50 genes from 70kbp), considering the volume of sequencing undertaken.

UNIVERSITY of the
WESTERN CAPE

Putative ORFs with: (i) matches to functionally annotated proteins, (ii) matches to hypothetical proteins and (iii) no significant matches in the SWISSprot database were identified based on the highest scoring E-value match. For each of the ORFs, similarity scores were sufficiently high (>40%) to infer homology with the database proteins while also being indicative of novel sequence variants of these database matches.

On the basis of the small scale sequencing of the Chinese hypertehrmophilic library, approximately 80% of the sequence data was redundant. This number constituted the fraction of sequence reads that reported the same BLAST match.

The redundancy observed in the sequence data could have resulted due to the overrepresentation of specific genes within the sequence library. Possible reasons include: (i) the fact that the overrepresented genes constituted the largest proportion of genomic material subsequent to the extraction and purification procedure, which was all that was essentially therefore cloned; (ii) certain specific gene sequences were preferentially cloned and were therefore overrepresented or (iii) while the constructed library constituted a diverse gene population, specific genes were preferentially amplified prior to plating the libraries and were therefore overrepresented. This was observed in the study by Tringe et al., (2005) on the Minnesota farm surface soil, where an amplification step was included prior to plating the library. The reported number of redundant sequence reads resulting from this, constituted 25% of the total number of reads. While a different cloning system was used in this study, the implemented amplification of the library, prior to plating, could possibly have resulted in the observed redundancy.

UNIVERSITY of the
WESTERN CAPE

Chapter 4

Whole Genome Amplification for Improved Recovery of Metagenomic Sequence information

4.1 Introduction

Metagenomic sequence libraries typically yield vast numbers of novel gene sequences as well as known genes that displays highly divergent sequence properties in comparison to the curated sequences in the databases (Venter et al., 2004). Database sequence comparisons are therefore often not sufficient to verify whether these putative ORFs represent actual novel genes or are a result of sequencing or cloning aberrations.

UNIVERSITY of the
WESTERN CAPE

Errors in metagenomic library construction typically involve cloning genes/ORFs from sources other than the intended environmental sample as well as the formation of cloning aberrations via chimera formation. An example of this was observed in one of the samples in the Sargasso Sea sequencing project. The dominating species identified in the library constructed from this sample could be sufficiently assembled but not re-discovered in an independent sample from the same site. As a result, it could not be excluded that the sample contained a certain fraction of clonally expanded, contaminating microbes (Delong, 2005).

Sequence-derived genes should therefore ideally be validated by additional molecular biology strategies. This validation procedure typically entails recovering the metagenomic sequence derived genes/ORFs from the source sample ensuring that subsequent downstream analysis is not based on data from erroneous sequences. Identification and recovery of specific genes as a means for validating the source of library derived genes is routinely performed using sequence specific PCR technology. Utilising primers generated from the 5'- and 3' ends of the library derived gene(s) to recover these genes from the environmetal source sample is considered sufficient to validate both the origin and arrangement of these novel genes.

The successful implementation of PCR on soil-extracted DNA is dependant on the quality, purity and availability of sufficient amounts of DNA as template (Tsai and Olson, 1992; Zhou et al., 1996). The presence of co-extracted inhibitory substances such as humic acids is potentially the single biggest factor affecting quality and purity of extracted DNA (Wechter et al., 2003). Strategies for improved quality and purity of soil-extracted DNA involve improved extraction and purification protocols but involve associated biases toward certain organisms and loss of DNA (Wikstrom et al. 1996; Steffan et al. 1988; von Wintzingerode et al.1997). The availability of sufficient DNA template may be dependant on the source of the sample, as some environments harbour low levels of biomass (Abulencia et al., 2006). Methods that enable the pre-amplification of the source sample have been demonstrated to improve both the template availability and DNA purity, to levels that allow for downstream molecular biology manipulation (Dean et al., 2002)

UNIVERSITY of the
WESTERN CAPE

Whole genome amplification (WGA) using the multiple displacement amplification properties of ϕ29 DNA polymerase allows for unbiased, highly efficient amplification of DNA from samples (Gonzalez et al., 2005). This strategy has been implemented for the pre-amplification of contaminated, low biomass environments and has enabled improved analysis of the microbial diversity of these environments (Gonzalez et al., 2005; Abulencia et al., 2006). This does not, however, guarantee that the WGA strategy can be equally applied to all environments.

Limitations of the WGA strategy include: (i) the significant degree of background amplification as a result of primer-primer interactions and (ii) the chimera formation that results from the multiple displacement action of ϕ29 polymerase (Lasken and Stockwell, 2007). The implication is that WGA DNA invariably contains a significant proportion of undesired products in addition to the amplified DNA. To what degree this affects subsequent downstream analysis and manipulation of the WGA DNA is still the subject of ongoing research

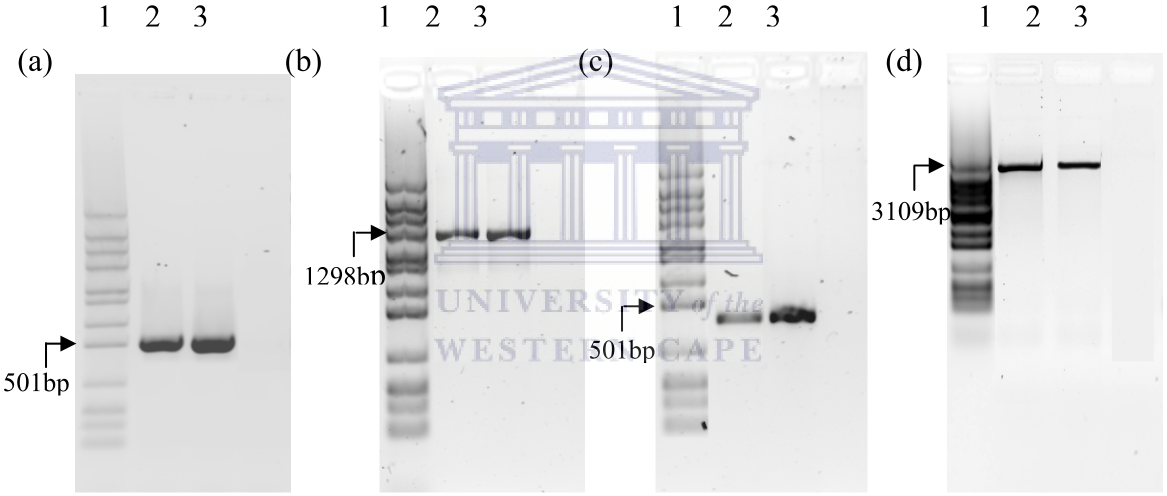
UNIVERSITY of the
WESTERN CAPE

4.2 Aims:

To investigate the implementation of a whole genome amplification strategy, of the thermophilic metagenome, as a resource for: (i) the recovery of selected metagenomic ORFs, derived from the library sequence and (ii) the development of a novel strategy for recovery of flanking sequence data for one of the library derived ORFs.

4.3 Deriving amplicons for full length ORFs from the library clones

PCR primers were designed for the full length Usp, ArgJ, DUF and Polymerase ORF sequences from the metagenomic library derived sequences. The 5’ and 3’ primers were denoted: Usp (Uspf and Uspr); ArgJ (ArgJf and ArgJr); DUF29 (DUF29exF and DUF29exR) and Polymerase (P2.96F and P2.96R). PCR conditions were optimized on the respective library clones from which the genes were derived (Figure 4.1) and the optimum reaction conditions were determined as in Section 2.8.2.1.



UNIVERSITY of the WESTERN CAPE

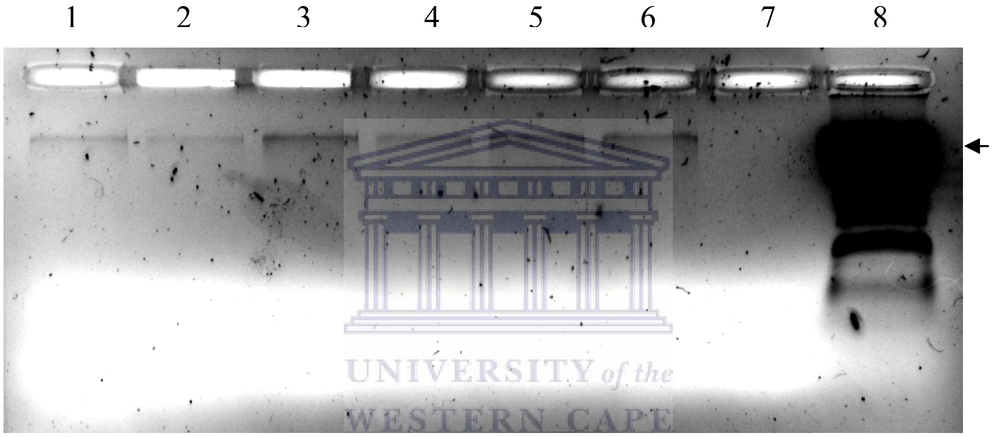
Figure 4.1 Amplification products from the library clones
(a) DUF29, (b) ArgJ, (c) Usp and (d) DNA Polymerase . For gels (a) – (d) : lane 1, Molecular weight marker; lanes 2 and 3, amplification products of respective genes.

4.4 Deriving amplicons for full length ORFs from the metagenomic DNA

The initial strategy was to use primer sets designed from library derived sequences of the Usp, ArgJ, DUF and Polymerase ORFS in order to amplify these genes directly from the TC11 source sample. The strategies and results for the DNA extraction and subsequent amplification are reported in sections 4.4.1 and 4.4.2, respectively.

4.4.1 Extraction and purification of metagenomic soil sample

Extractions were done using the modified Zhou method (Section 2.4.1) and extracted DNA was purified by three passes through PVPP columns (Section 2.5.1). High molecular weight DNA was succesfully extracted, as visualized by agarose gel electrophoresis (Figure 4.2). The extracted DNA was spectrophotometrically quantitated through all purification steps, as outlined in Table 4.1.



UNIVERSITY of the
WESTERN CAPE

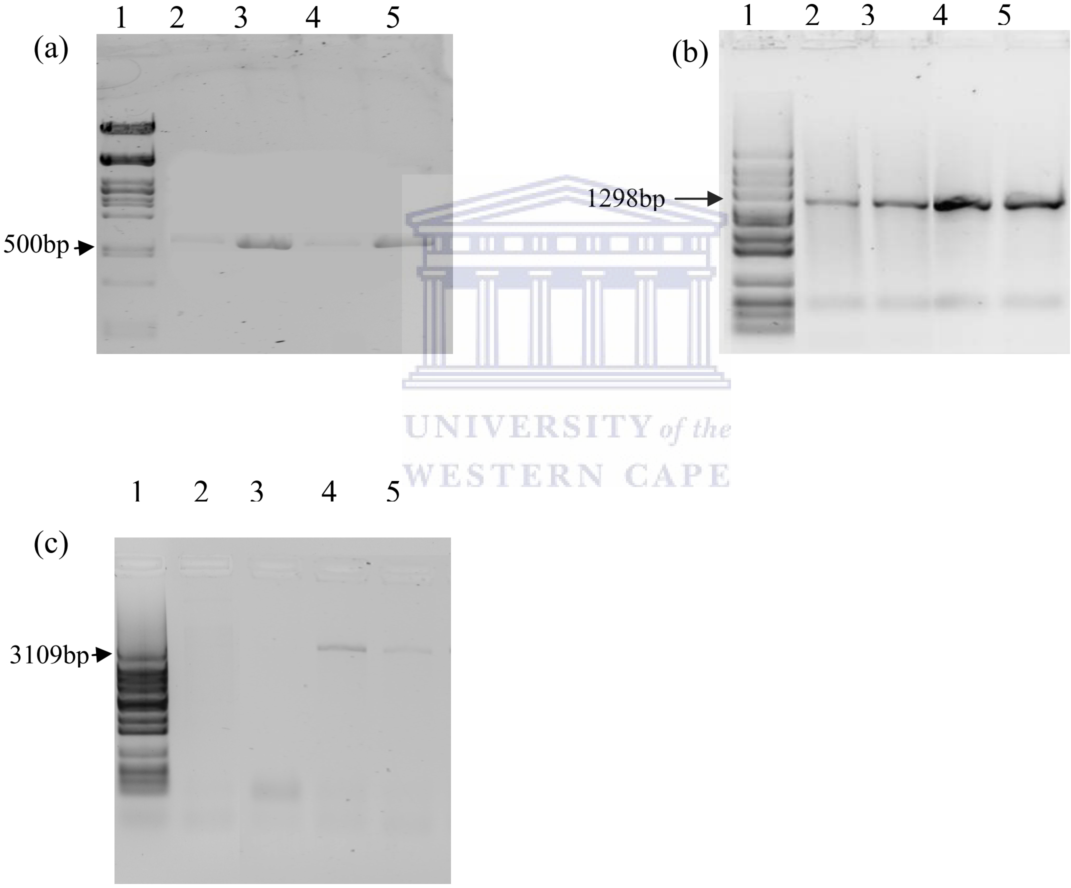
Figure 4.2. TC11 extracted DNA.
Lanes 1-6, Multiple DNA extractions from sample TC11; lane 7, negative control (H20); lane 8, λ Pst molecular weight marker.

Table 4.1. Quantitation of soil extracted DNA.

Sample	*Purification*	*Concentration(ug/g)*	*A260/280*	*A260/230*
No. 1	Crude extract	0.54	1.47	1.67
	3 Step PVPP	0.39	1.56	1.81
No.2	Crude extract	0.47	1.44	1.62
	3 Step PVPP	0.36	1.66	1.73

4.4.2 PCR amplification from the source sample

The optimized reaction conditions were used to amplify the target ORFs directly from crude extract and PVPP treated DNA of samples 1 and 2. Amplifications products were visualized by agarose gel electrophoresis (Figure 4.3) and also quantitated using the Nanodrop ND-1000 (Table 4.2)



UNIVERSITY of the WESTERN CAPE

Figure 4.3 Amplification products from environmental DNA
(a) DUF29, (b) ArgJ and (c) DNA Polymerase from soil extracted DNA template.
For gels (a) – (c); lane 1, Molecular weight marker; lane 2, PCR amplicon from sample 1 crude extract DNA template; lane 2, PCR amplicon from sample 1 PVPP treated DNA template; lane 3, PCR amplicon from sample 2 crude extract DNA template; lane 4, PCR amplicon from sample 2 PVPP treated DNA template.

Table 4.2. PCR amplicon concentrations for soil-extracted DNA template.

ORF	Sample no.	Template	Conc. ng/ul
DUF	1	Crude extract	10
		PVPP treated	25
	2	Crude extract	8
		PVPP treated	27
USP	1	Crude extract	8
		PVPP treated	10
	2	Crude extract	4
		PVPP treated	2
ArgJ	1	Crude extract	55
		PVPP treated	43
	2	Crude extract	31
		PVPP treated	27
Pol	1	Crude extract	5
		PVPP treated	6
	2	Crude extract	24
		PVPP treated	16

UNIVERSITY of the
WESTERN CAPE

4.5 Improved amplification of full length ORFs from the WGA soil DNA

Attempts to efficiently amplify the Usp, ArgJ, DUF and Polymerase ORFs directly from the source sample proved unsuccessful. The whole genome amplification of the TC11 source sample was implemented in an attempt to improve the quality and template availability of the source sample. The optimization of the whole genome amplification on sample TC11 is reported in section 4.5.1. The amplification of the Usp, ArgJ, DUF and Polymerase ORFs from the whole genome amplified template DNA is reported in section 4.5.2.

4.5.1 Whole genome amplification of DNA

The whole genome amplification was carried out using the Repli-G whole genome amplification kit (Section 2.8.3). Reaction conditions were optimized at 10ng and 20ng of crude extract and PVPP treated purifications for samples 1 and 2. Whole genome amplification products of the samples are shown in Figure 4.4. and product concentrations in Table 4.3.

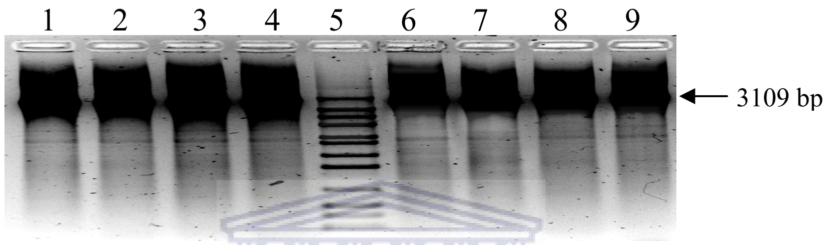


Figure 4.4. WGA yields from environmental DNA
Lane: 1, Sample 1 - crude extract at 10ng template; 2, Sample 1 – 3 Step PVPP treated sample at 10 ng template; 3, Sample 2 - crude extract at 10ng template; 4, Sample 2 – 3 Step PVPP treated sample at 10 ng template; 5, Molecular weight marker; 6, Sample 1 - crude extract at 20ng template; 7, Sample 1 – 3 Step PVPP treated sample at 20 ng template; 8, Sample 2 - crude extract at 20ng template; 9, Sample 2 – 3 Step PVPP treated sample at 20 ng template

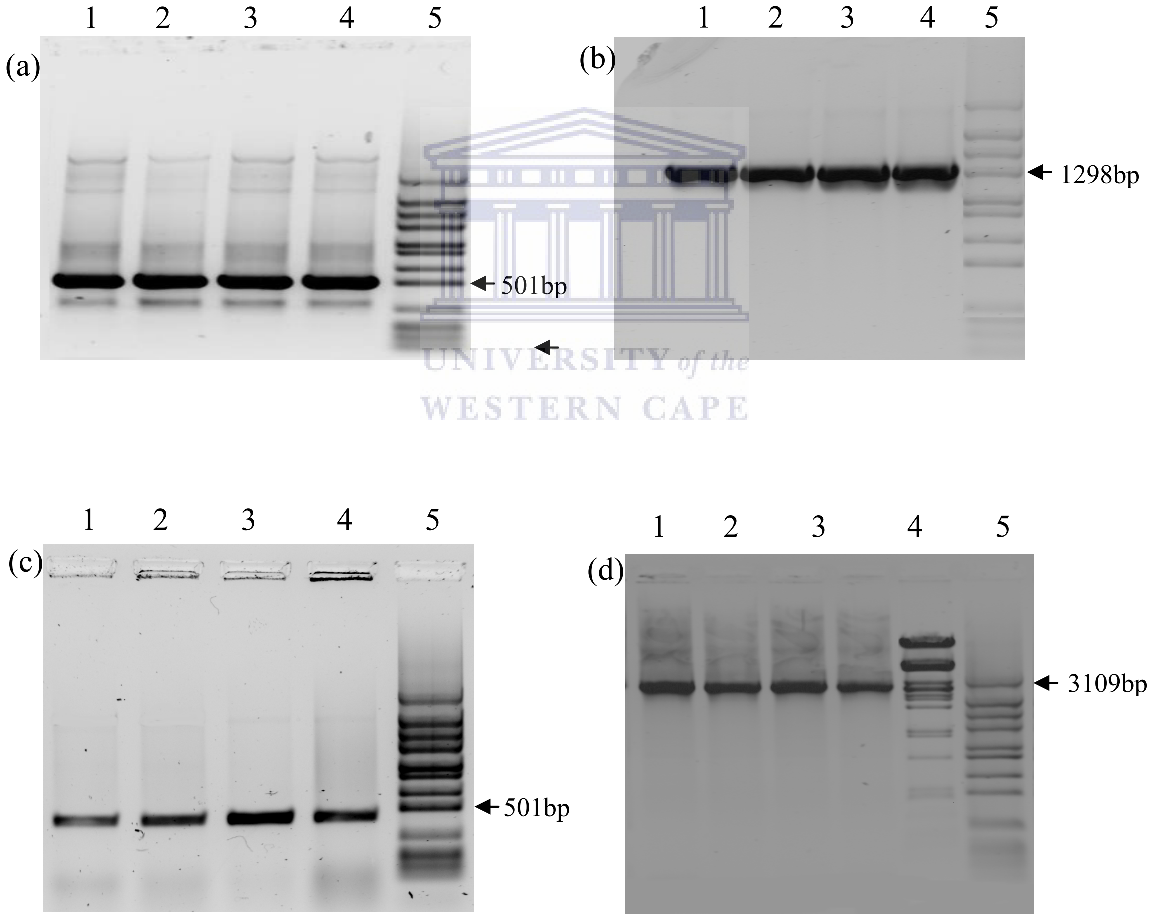
UNIVERSITY of the
WESTERN CAPE

Table 4.3 Optimised template concentrations for WGA

Sample	*Template conc.*	*Purification template*	*Conc. ng/ul*
1	10 ng	Crude extract 3 Step PVPP	2779 2412
	20 ng	Crude extract 3 Step PVPP	854 1015
2	10 ng	Crude extract 3 Step PVPP	2443 2871
	20 ng	Crude extract 3 Step PVPP	1120 1365

4.5.2 Amplicons from WGA DNA

The Repli-G whole genome amplified DNA was diluted to the 50 ng and used as template for amplification of the ORFs (Section 2.8.2.1). The succesfull amplification of the Usp, ArgJ, DUF and Polymerase ORFs was assessed by agarose gel electrophoresis (Section 2.9.1) as illustrated in Figure 4.5 and through observation of amplicon concentration (Section 2.6) as reported in Table 4.5.



UNIVERSITY of the
WESTERN CAPE

Figure 4.5 Amplification products from WGA template.
For gels (a) – (d); lane 1, PCR amplicon from sample 1 crude extract WGA template; lane 2, PCR amplicon from sample 1 PVPP treated DNA WGA template; lane 3, PCR amplicon from sample 2 crude extract WGA template; lane 4, PCR amplicon from sample 2 PVPP treated DNA WGA template; lane 5, Molecular weight marker.

Table 4.4 Relative amplicon concentrations from WGA template

ORF	Sample number	WGA template	Concentration/ul
DUF	1	Crude extract	112
		3 Step PVPP	116
	2	Crude extract	104
		3 Step PVPP	98
USP	1	Crude extract	122
		3 Step PVPP	101
	2	Crude extract	121
		3 Step PVPP	115
ArgJ	1	Crude extract	141
		3 Step PVPP	134
	2	Crude extract	143
		3 Step PVPP	133
Pol	1	Crude extract	145
		3 Step PVPP	162
	2	Crude extract	170
		3 Step PVPP	177

UNIVERSITY of the
WESTERN CAPE

4.6 Confirmation of metagenomic DNA amplified PCR amplicons

The amplicons from the library-derived and WGA samples were analysed by restriction analysis (Section 2.8.4), as reported in Figure 4.6. The same restriction digestion patterns were detectable for the putative Usp and DUF29 amplicons generated from the respective library clones as well as the WGA DNA.

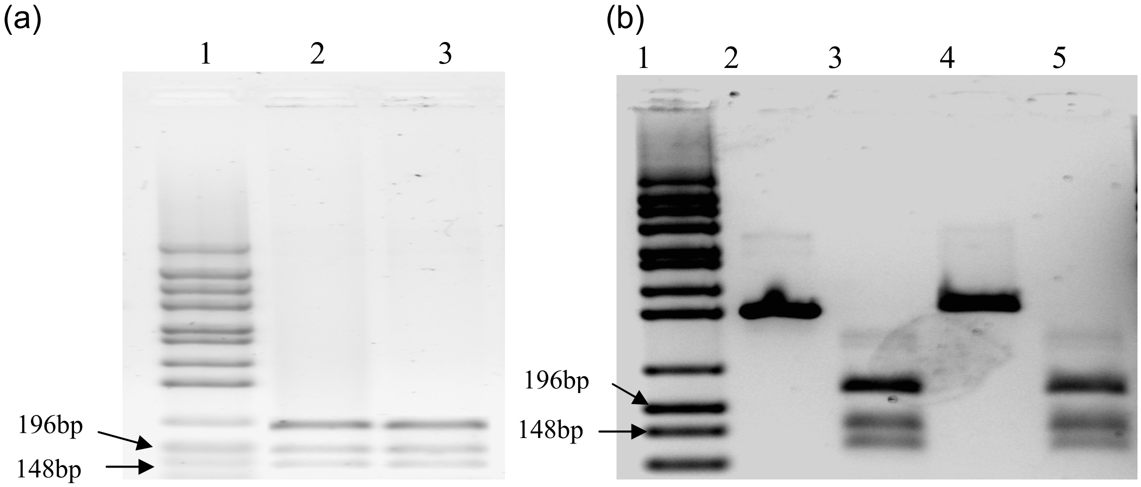


Figure 4.6 Restriction analysis patterns for library derived and WGA ORFs.
(a) Lane 1, Molecular weight marker; lane 2, Universal stress protein ORF amplified from library clone digested with Alu I; lane 3, Universal stress protein ORF amplified from WGA DNA digested with Alu I. (b) Lane 1 Molecular weight marker; lane 2, DUF29 amplicon generated with library clone as template, lane 3, amplicon generated with library clone as template digested with DraI, lane 4, DUF29 amplicon generated with WGA DNA as template; lane 5, DUF29 amplicon generated with WGA DNA as template digested with DraI.

UNIVERSITY of the
WESTERN CAPE

4.7 Utilising WGA source DNA for acquiring sequence data, flanking a library derived clone

Clone-end sequencing of clone 2.96 (Chapter 3) revealed the presence of a full length ORF for a DNA polymerase at the 5’ end as well as the N-terminal sequence of a hypothetical gene at the 3’ end of the clone. The hypothetical protein from the 3’- end sequence read (Figure 4.7) displayed significant similarity to the N-terminal portion of hypothetical protein CaggDRAFT_0969 and was denoted ORF 5-partial transcript.

```
   >gi|118047163|ref|ZP_01515804.1|hypothetical protein CaggDRAFT_0969 [Chloroflexus
aggregans DSM 9485]
 gi|117996331|gb|EAV10531.1|hypothetical protein CaggDRAFT_0969 [Chloroflexus aggregans DSM
9485]
Length=660

 Score =  175 bits (443),  Expect(2) = 7e-44
 Identities = 100/142 (70%), Positives = 116/142 (81%), Gaps = 1/142 (0%)
 Frame = -2

Query  489  MRFYFSRMQllqsaallvlvvlslsatwllaSRPWRIDAVIGGADSAIVGTGFFAKEQTP  310
            MR++FSR+QLL  A LL L+V+SLSATW+LASRPWRIDAVIGGADSA+VG+GFF KE +
Sbjct  1    MRYHFSRIQLLPLAMLLALLVISLSATWILASRPWRIDAVIGGADSALVGSGFFTKELSS  60

Query  309  TGMPFRWTSGSAVINLPPVHSAYLVTLHAYIPGDV-PAYVTIGDRAFPVVTIVDTDDGKA  133
             G PFRWTSG A+INLPPVH+ Y+VT+ AY+P DV P YV I DRAFPV TIV TD   A
Sbjct  61   DGTPFRWTSGPAIINLPPVHARYIVTMRAYVPSDVIPYYVEIKDRAFPVATIVVTDQLPA  120

Query  132  FRHYHLLWQSPPTYHWLELLLP  67
            FR YH+LWQSP TYHWL+L  P
Sbjct  121  FRRYHILWQSPVTYHWLDLFTP  142
```

Figure 4.7 BLAST analysis of region of 2.96

4.7.1 Restriction digestion based approach for deriving flanking sequences

UNIVERSITY of the
WESTERN CAPE

The WGA DNA was utilized as template in an attempt to recover additional sequence for ORF5-parial transcript. The sequence derived from clone 2.96 was investigated for restriction enzymes that do not to cleave within this sequence. The rationale for using these enzymes on the WGA was that the restriction fragments would include an extension of the sequence derived from clone 2.96. The WGA DNA (1ug) was restriction digested with HindIII and Xba I, (Figure 4.8), as outlined in Section 2.8.4.

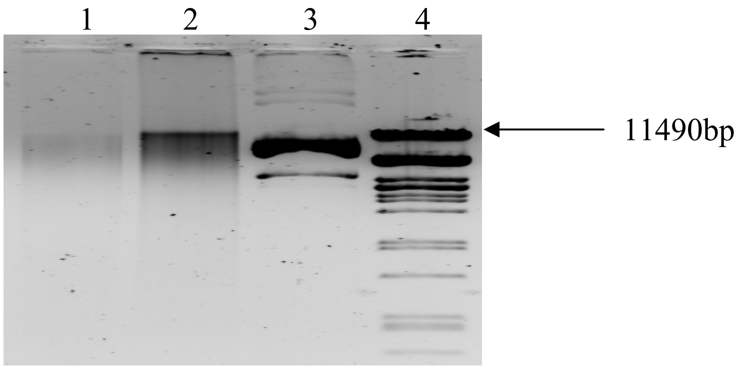


Figure 4.8 Restriction digestion of WGA with HindIII and XbaI
Lanes: 1, HindIII and XbaI double digested WGA DNA from soil sample TC11; 2, Undigested WGA DNA from soil sample TC11; 4, Molecular weight marker.

Based on the sequence information for clone 2.96 these enzymes are known not to cleave within the polymerase gene or the known flanking region. The digestion reaction was ligated with a HindIII and XbaI digested, dephosphorylated pUC18 vector (Section 2.7.4). The ligation mixture was transformed into *E.coli* and incubated for 16 hours (Section 2.7.5.1).

The *E.coli* cells were recovered by centrifugation, lysed by heat treatment and diluted in fractions of 1:10, 1:100 and 1:1000. These fractions were used as template for amplification with the DNA polymerase specific N-terminal primer (P2.96F) and the vector specific M13R primer (Section 2.8.2.1). A positive amplicon was detected with the polymerase (P2.96F) and M13 primers (Figure 4.9). The amplicon size was predicted to be approximately 4 Kb. This exceeded the 3.5 Kb size of clone 2.96 and potentially represented an extension of the clone.

UNIVERSITY of the
WESTERN CAPE

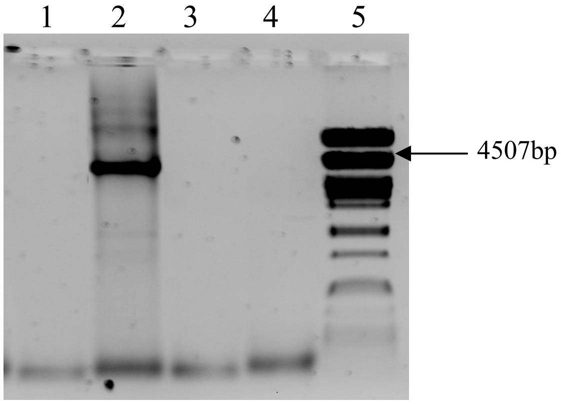


Figure 4.9 PCR recovery of restriction digestion extension product
Lane: 2, Amplification products from WGA sample with POLF and M13 primer; lane 5, λPst molecular weight marker

4.7.2 Sequencing of the extension product

The positive amplicon was sequenced (Section 2.8.5) with DNA polymerase specific N-terminal primer and the M13 primer (Figure 4.10(a)). The DNA polymerase specific N-terminal primer sequence yielded a match to the library derived DNA polymerase indicating that the correct sequence was derived using the polymerase specific primer. The M13 primer yielded a match to the central region of hypothetical protein CaggDRAFT_0969 indicating that the library derived ORF5-partial region was extended from the WGA DNA (Figure 4.10(b)).

(a) >M13 sequence product from extended sequence

```
AGGATGACTTGTCGCCAACGCAGCGGCTCATGCCATAGCAAGGCCAGCAATGCCAATGCGGCAAAACTGACCATTGGCGCA
TCAGTGCTGGCAAAAACGCCATGGGAGATCACCACCGGCGCTGCTGCCATCACTGCGCCGGCTACCCAAGCATACGATCGA
TGTATCACTCGCCATGCCATAAGCGTGGTAGCGCCGATCGTAATGAATGTAGCGATTGCGCCAGCGATTCGCCCAAACAGC
AGCGCTTGATCGCCAGCCAACTGCCCTAGCCATAGAATCAGCGGGAAACCGAACGGGTAAAAAGGATGGGCGCGGAATACG
CGCCGCCAAGAGCCGCCATGGTACAAATGCCAATAATAATCAGGCCCCTCAACGTACCAATGCCATTGCAAGGTTGCTATG
ACAGAGAACGCCAAAAGCATAACGATCAGCATGGCCAAGAGTGCGCCGCTACGCGCGCGAGCGGCGCTTTGCGCGAAGAAG
GCGCCAATGGTTAATGCCGCTGCCATGCCCGGCAACCACGAAAAAGGCAACCACGTGTAGCCGCCGATCGCAGGCGGATGC
CAGACCAGCACGTCGTACACGAGTGGCAGTAAGAGCGCAATTCCCGCCAACCCGATCAGCCTTCGCCCGCGCAATGGCCAC
AATAGACCAGCGCTTGCCAGCACCGTTAAGCCCATCGTTGCCAGCGGCAACACCGGCGCAGCCGCCAACGACGAGCTGCGA
ACGCTGACATGCCTTATTGCGATCCCAATAATCGGGTATCATCTTGAGAAGACGCTGAC
```

```
 (b)   >gi|118047163|ref|ZP_01515804.1| hypothetical protein CaggDRAFT_0969 [Chloroflexus
aggregans DSM 9485]
Length=660
 Score =  228 bits (582),  Expect = 2e-58
 Identities = 175/265 (66%), Positives = 205/265 (77%), Gaps = 0/265 (0%)
 Frame = -1
Query  831  IIGIAIRHvsvrssslaaapvlplATMGLTVLASAGLLWplrgrrliglagialllpLVY  652
            ++GIA+  + +RSS+  A PV+PL T+GLT+L  A LLWPLRG+RL+  A +AL+LP+ Y
Sbjct  161  LLGIAVSQLHIRSSNTLAVPVMPLITVGLTLLGFAHLLWPLRGKRLVWFAVVALILPVGY  220

Query  651  DVLVWHPPAIGGYTWLPFSWLPGMAAALTIgaffaqsaararsgallamlivmllAFSVI  472
            D+LVWHP     YTWLP SWLPGM AA  IG  FAQ AA +R GA  A LIV+LL  +VI
Sbjct  221  DLLVWHPLQGNDYTWLPLSWLPGMVAASVIGVAFAQRAALSRGGAWFAALIVILLMVAVI  280

Query  471  ATLQWHWYVEGPDYYWHLYHGGSWRRVFRAHPFYPFGFPLILWLGQLAGDQALLFGRiag  292
             TLQWHW VEGPDY+WHL HGGSWRRVFR+HPFYPFG PLIL++GQLAGDQALLFGRIAG
Sbjct  281  TTLQWHWLVEGPDYHWHLNHGGSWRRVFRSHPFYPFGLPLILYVGQLAGDQALLFGRIAG  340

Query  291  aiatfitigatTLMAWRVIHRSYawvagavmaaapvVISHGVFASTDAPMVSFaalalla  112
            A+ T + I A  L+ WRVI  +YAWVAG +M A+PVV+SHG  ASTDAPM   A LALLA
Sbjct  341  AVTTSVAIVAVVLLVWRVIAPAYAWVAGMIMLASPVVVSHGALASTDAPMTGLATLALLA  400

Query  111  lLWHEPLRWRQVILAGMALGFAYFF  37
            LLWHE LRW Q+ LAGM LG AY F
Sbjct  401  LLWHERLRWLQIALAGMCLGLAYLF  425
```

Figure 4.10 BLAST result for restriction digested, extension product

(a) M13 primer derived sequence, (b) BlastX result from M13 derived sequence.

UNIVERSITY of the
WESTERN CAPE

4.8 Discussion

A number of recent studies have investigated the whole genome amplification strategy by multiple displacement amplification as a pre-amplification step in the recovery of low biomass DNA (Gonzalez et al., 2005; Abulencia et al., 2006). These studies specifically focused on improved recovery of ssu rRNA genes for phylogenetic analysis as well as the improved recovery of total environmental DNA for metagenomic library construction. The success of these studies furthermore suggested that WGA could be used effectively to recover library derived sequences from the original source sample.

WGA for recovery of genes from environmental samples

The efficient recovery of specific genes from varied environments requires stringent primer design and extensive optimization of amplification conditions for the respective genes. Amplification of the ORFs for the DUF29, USP, ArgJ and DNA Polymerase genes were successfully optimised from the TC11 metagenomic library clones with the respective sequence derived primers.

UNIVERSITY of the
WESTERN CAPE

DNA extraction yields from soil are primarily dependant on the efficiency of the extraction strategy employed and the inherent biomass of the specific environment (Abulencia et al., 2006). For inherently low biomass environments, the purification procedure can result in further losses of DNA. DNA extractions from the TC 11 soil sample yielded high molecular weight DNA but low yields for all the samples. Concentrations of 0.54 and 0.47 ug/g were observed for the crude extract of samples 1 and 2 respectively. The Zhou method, which was employed in this study, typically results in lower extraction efficiencies but produces intact high molecular weight DNA (Zhou and Tiedje, 1996). This method was employed in order improve the presence of the full length ORFs for DUF29, USP, ArgJ and DNA polymerase in the source soil sample.

The corresponding $A_{260/230}$ and $A2_{60/280}$ absorbance ratios for the crude extract samples were suggestive of both humic acid and protein contamination for the crude extract DNA. A three step PVPP column purification strategy gave improved absorbance ratios for both samples 1 and 2. These values were still below the accepted values of >2.0 and >1.7 that indicate pure DNA. Samples with A260/280 ratios >1.5 have been demonstrated to be sufficient for PCR amplification (LaMontagne et al., 2001) and PVPP treated samples were therefore included.

UNIVERSITY of the
WESTERN CAPE

The use of optimized conditions with the primer combinations for the DUF29, USP, ArgJ and DNA Polymerase ORFs on both the crude extracts and PVPP treated samples 1 and 2 did not consistently yield amplicons. This was expected, due to the higher levels of inhibitory substances indicated by the spectrophotometric values (LaMontagne et al., 2001). Higher PCR efficiencies were, however, expected for the PVPP treated samples 1 and 2, due to the improved purity following PVPP treatment. It should however be noted that A260/230 ratios are not entirely accurate measures of the level of humic acid contamination (ref). Another aspect for consideration is the possibility of low template availability for the investigated ORFs that could have resulted due to the purification procedure.

The successful use of WGA DNA as template for downstream amplification has been demonstrated on various source materials such as blood and tissue culture cells (Dean et al., 2002) and environmental DNA (Gonzalez et al., 2005; Abulencia et al., 2006). Due to the variable nature of different source materials, WGA has to be specifically optimized whenever implemented on different starting materials. The optimal DNA template concentration for WGA by MDA on both crude and PVPP treated samples 1 and 2 was determined at 10ng. Higher template concentrations (20ng) appeared to have an inhibitory effect on the efficiency, with final amplication yields of approximately 50% compared to the 10ng template. The high amplicon yields for both crude and PVPP treated samples were indicative of the ability of Phi29 polymerase to function in the presence of high levels of inhibitory substances.

UNIVERSITY of the
WESTERN CAPE

Amplification efficiency of the library derived ORFs, following WGA, was approximately 10 times higher than that for the direct soil extracted template. The WGA therefore appears to have improved both the template availability and purity to appreciable levels for efficient amplification with the ORF specific primers. This is consistent with other previous studies (Gonzalez et al., 2005; Abulencia et al., 2006).

Comparison of restriction digestion products of the amplification products generated from the library clone templates and the WGA template confirmed the recovery of the library derived DUF29 and Usp ORFs. Comparison of the 5' DNA polymerase sequence from the WGA template with that of the library derived sequence confirmed the presence of the gene. WGA has been demonstrated to generate spurious amplification products as well as chimera formation even in the absence of any input template DNA (Lasken and Stockwell, 2007). While the WGA DNA generated from our source sample could potentially have contained these by-products, the use of highly specific gene-targeted primers appears to have avoided recovery of non-specific amplification products.

WGA for extended coverage of environmental library derived sequence

Library construction from low biomass environments typically would not be expected to provide complete coverage of the metagenomes, disfavoring those organisms present at low frequencies (Gonzalez et al., 2005; Abulencia et al., 2006). The consequence of this is that the entire gene complement for a specific organism would possibly not be observed within the library. The WGA strategy potentially provides improved access to the genomes of these organisms. Those regions not represented in the library could therefore be directly assessed from the environmental soil DNA.

UNIVERSITY of the
WESTERN CAPE

This study was aimed at addressing this question in the context of deriving flanking sequence data for clone 2.96, which was derived from the thermophilic, metagenomic library. The metagenomic library clone 2.96 was annotated to contain a full length polymerase gene at the 5' end and a partial ORF (denoted partial-ORF5) of 142 amino acids, homologous to hypothetical protein

CaggDRAFT_0969, at the 3' end. The full length hypothetical protein CaggDRAFT_0969 totals 550 amino acids which meant that metagenomic library derived partial-ORF5 was proposed to be devoid of an additional 308 amino acids. A restriction digestion based approach, complemented with a gene specific amplification step was implemented to recover additional sequence information for the partial transcript of clone 2.96 directly from WGA DNA of the thermophilic sample from which the clone was derived.

The rationale for the approach was based on the following arguments:

(i) The clone sequence (2.96) from the metagenomic library (further termed the library clone) represents a small fragment of a larger genome contained within the environmental DNA.

(ii) Restriction enzymes that do not cut within the library clone are however proposed to cleave at some position within the larger genome sequence, of which the library clone forms a part.

(iii) Among the large number of fragments produced, there should be specific fragments that contain the sequence of the library clone as well as additional sequence that extend to the cut sites in the genome (these fragments are further termed the environmental extension product).

(iv) Subsequent to cloning all restriction fragments, the environmental extension product could be recovered with a primer that is designed from the sequence of the library clone in conjunction with a vector specific primer.

UNIVERSITY of the
WESTERN CAPE

This approach implemented on the WGA DNA resulted in the recovery of a DNA fragment, which represents an extension of ORF5 partial, containing 425 amino acids homologous to hypothetical protein CaggDRAFT_0969. This therefore represented an extension of 849 bp (283 amino acids) of partial-ORF5. The recovery of this environmental extension product was achieved with a primer specific for the polymerase ORF from clone 2.96 and the vector specific primer.

While the entire flanking region of ORF5, homologous to hypothetical protein CaggDRAFT_0969, could not be produced, the feasibility of this novel WGA strategy was proven in principle.

UNIVERSITY of the
WESTERN CAPE

CHAPTER 5

In-Silico Functional Inference of Metagenomic Sequence Derived Open Reading Frames

5.1 Introduction

Identification and characterization of novel genes/enzymes of biotechnological importance is primarily driven by industrial requirements and for use in molecular biology manipulations (Lorenz et al., 2002). A typical strategy involves defining the industrial process, identifying potential genes/enzymes that could improve this process and subsequently either mining for these genes or traits in a relevant environment or alternatively engineering the desired traits. This rationale has the advantage of effectively eliminating the need for expensive and laborious screening strategies through an array of potentially unrelated and insignificant additional genes. The disadvantage is that the potential of identifying genes/enzymes with potential novel uses in industrial or molecular biology applications is not fully realized.

UNIVERSITY of the
WESTERN CAPE

Sequence data derived from environmental metagenome libraries constitute the biggest resource of novel and hypothetical genes (Venter et al., 2004; Tringe et al., 2005). Characterization of these novel or hypothetical genes holds the potential for identifying enzymes with entirely novel biotechnological applications or novel characteristics. The analysis of novel or hypothetical genes derived from thermophilic environments is furthermore of particular interest due to the reported thermostabilities of protein products from this environment. The time and resources that would however be required for the experimental assessment of all novel proteins would be tremendous and is the biggest limitation for such an approach.

The implementation of various bioinformatics approaches serves as a viable alternative for functional characterization of novel or hypothetical genes (Sivashankari and Shanmughavel, 2006). These *in-silico* strategies are broadly defined as either homology or non-homology based (Friedberg, 2006). The homology based strategies use comparison between either sequences or structures as a basis for functional inference. The non-homology based strategies use information on the interactions, co-location and co-regulation of proteins. (Danderkar et al., 1998).

UNIVERSITY of the
WESTERN CAPE

There are currently numerous programs that implement either one or both of these strategies for the prediction of functional aspects of genes. Each of these programs is defined by a different set of strengths and weaknesses. Improved functional inferences are therefore often obtained through integration of multiple divergent strategies.

5.2 Aim

To derive functional annotations for thermophilic metagenomic sequence-derived ORFs, through integration of *in-silico* approaches.

5.3 Results

5.3.1 A tandem arrangement of the ORFs 1 and 2 on the library clone.

BLASTx analysis (Section 2.11.1) of clone 2.142 from the metagenomic library revealed the presence of 2 ORFs, which match a putative protein from *Hydrogenobacter thermophilus* and a 157aa long conserved hypothetical protein from *Sulfolobus tokodaii str. 7* respectively (Chapter 3). Failure to detect any organisms with a similar tandem arrangement of these ORFs in the database meant that this library derived clone possibly represented a novel organization of these ORFs on the genome of an unknown organism (Figure 5.1).

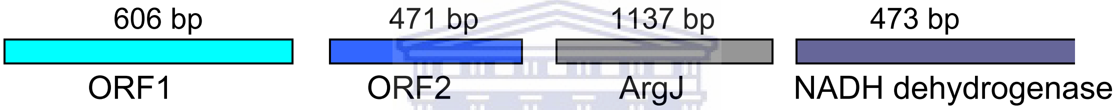
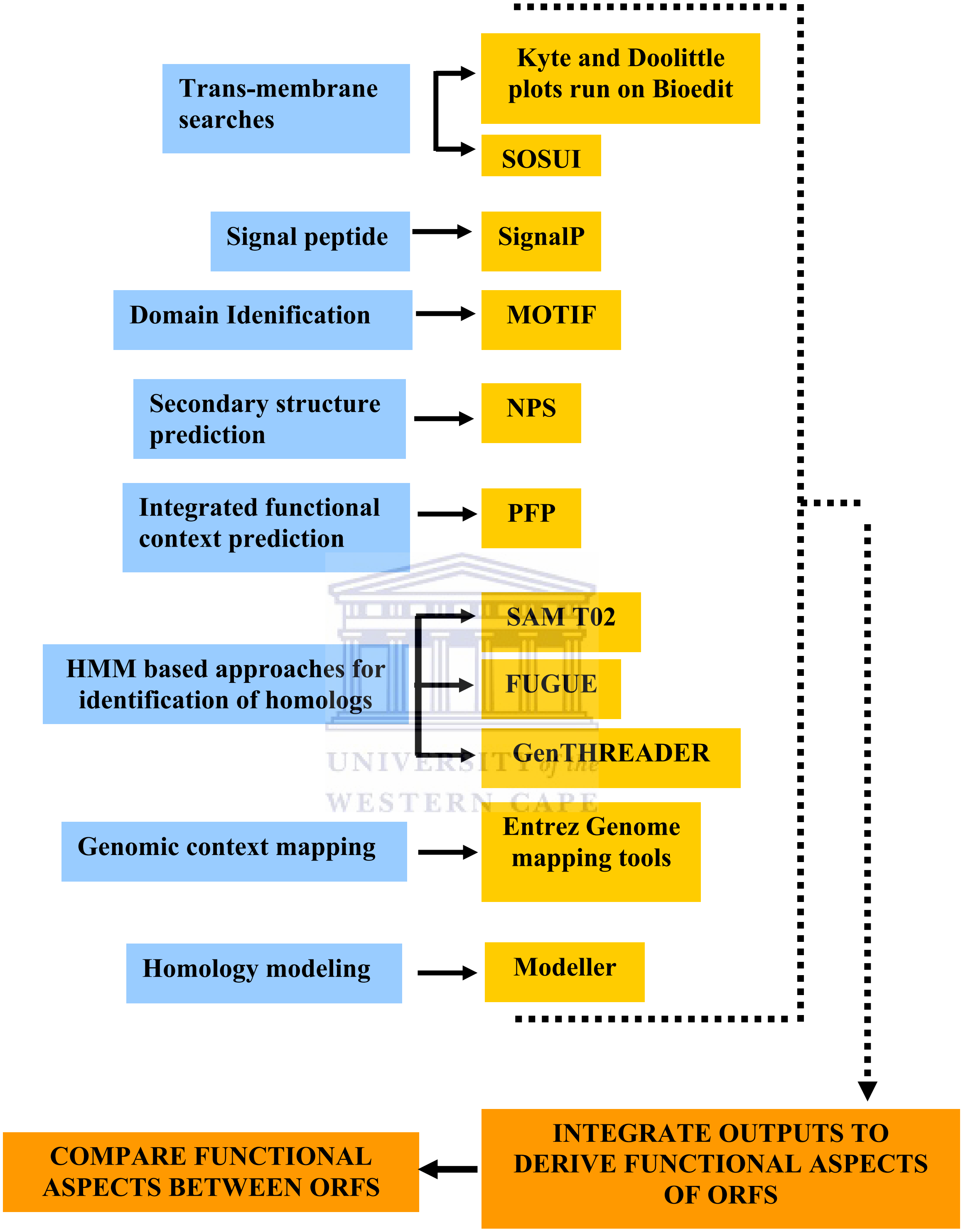


Figure 5.1 Gene organisation of clone 2.142
The locations of ORF1 and ORF2 are indicated with associated flanking genes ArgJ and NADH dehydrogenase. The sizes (in bp) of each of the ORFs are indicated above.

UNIVERSITY of the
WESTERN CAPE

5.3.2 Strategy for functional annotation of ORF1 and ORF2

Genes at an associated physical location often share common functional properties in terms of related catalysed reactions, common binding partners, and common expression patterns and/or associated roles in common biochemical pathways (Rogozin et al., 2004). A similar relationship for ORF1 and ORF2 was therefore assumed as a working hypothesis, based on the tandem arrangement of these genes on the metagenomic derived clone. Each ORF was characterized individually, after which their respective potential functional roles were compared in order to test this hypothesis. The various strategies that were implemented in recovering functional information are outlined in Figure 5.2


Trans-membrane searches
Kyte and Doolittle plots run on Bioedit
SOSUI
Signal peptide
SignalP
Domain Idenification
MOTIF
Secondary structure prediction
NPS
Integrated functional context prediction
PFP
HMM based approaches for identification of homologs
SAM T02
FUGUE
GenTHREADER
Genomic context mapping
Entrez Genome mapping tools
Homology modeling
Modeller
INTEGRATE OUTPUTS TO DERIVE FUNCTIONAL ASPECTS OF ORFS
COMPARE FUNCTIONAL ASPECTS BETWEEN ORFS

UNIVERSITY of the
WESTERN CAPE

Figure 5.2 Outline of the strategy for functional inference of ORFs 1 and 2.
The multiple divergant approaches that were implemented on ORFs 1 and 2 are shaded in blue. The programs used in these approaches are shaded in yellow. The outputs from the programs were integrated for improved prediction accuracy for the individual ORFs. The individually assigned functions were subsequently compared between the ORFs.

5.3.3 Secondary structure predictions

Predictions of the secondary structures of ORF1 and ORF2 were performed on the NPS consensus server (Section 2.11.4). The prediction for ORF 1 suggested the presence of 5 α-helices with a short central extended β-sheet between helices 2 and 3 (Figure 5.3). The prediction for ORF 2 in turn was suggestive of an α/β fold with alternating β sheets and α helices (Figure 5.3).

```
(a)                10        20        30        40        50        60        70
                    |         |         |         |         |         |         |
UNK_264340 MVVKDISKEQLKQLYNKDYPLWVEINLQLLKEKAYELVDWDNLLEEIEDMGRSDLKECISYLAVILEHMY
DSC        ??????????????????????????????????????????????????????????????????????
MLRC       cecccchhhhhhhhhcccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhchhhhhhhhhhhhhhhhhhh
PHD        ccchhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
Sec.Cons.  c?c???hhhhhhhhhccccc??hhhhhhhhhhhhhhhhhhhhhhhhhhhh?hhhhhhhhhhhhhhhhhhh
                   80        90       100       110       120       130       140
                    |         |         |         |         |         |         |
UNK_264340 KWDNFKHLAGGETAGSSWKRSIYTSRNNIEALLEIYPSLKSKLPNEVGTAWKISKARLKNWLIRNNLNLK
DSC        ??????????????????????????????????????????????????????????????????????
MLRC       hhccccccccccccccchhheeeccchhhhhhhhhhhhhhhccccccchhhhhhhhhhhhhhhccccccc
PHD        hhhhhhhhcccccccccccеeeecchhhhhhhhhhhhhhhccccchhhhhhhhhhhhhhhhhhhhhhccc
Sec.Cons.  hh??????ccccccccc???eeecc?hhhhhhhhhhhhhh?cccc???hhhhhhhhhhhhhhh????ccc
                  150       160       170       180       190
                    |         |         |         |         |
UNK_264340 DFSIPENCPYTYEQAMEREFNKIKKERKMYKNILVVMMGQMLPQRLWIELYLLQN
DSC        ???????????????????????????????????????????????????????
MLRC       cccccccccchhhhhhhhhhhhhhhhhhhhhhheehhhhhhchhhhhhhhhhhcc
PHD        cccccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhc
Sec.Cons.  cccccccc??hhhhhhhhhhhhhhhhhhhhhhh??hhhhhh?hhhhhhhhhhh?c
```

UNIVERSITY of the
WESTERN CAPE

```
(b)                10        20        30        40        50        60        70
                    |         |         |         |         |         |         |
UNK_266660 MYKNILVGYDGSDASTKALDRAISIAKLTGGKVHIVGVVKPLDFGYIDYVSPEEIDVYEKEEISKEEKLL
DSC        cccceeeecccccchhhhhhhhhhhhhhhccceeeeeccccccccceeeccccchhhhhhhhhhhhhhhh
MLRC       cceeeeeecccchhhhhhhhhhhhhhhhcccceeeeeeeccccccccccccchhhhhhhhhhhhhhhhhhh
PHD        ccceeeeeeccchhhhhhhhhhhhhhhhccceeeeeeeeeecccccccccchhhhhhhhhhhhhhhhhhhh
Sec.Cons.  ccceeeeeecccchhhhhhhhhhhhhhhhcccceeeeeeeccccccccccccchhhhhhhhhhhhhhhhhhh
                   80        90       100       110       120       130       140
                    |         |         |         |         |         |         |
UNK_266660 KKAIEKVKQENLETVYKILEGDPAXELMSYADENNIDLIVVGRKGAGMLKRILMGSTSLSLVKYANQEVL
DSC        hhhhhhhhhhhhhhheecccccccchhhhhhhhhccceeeeecccchhhhhheeccccccccchhccccceе
MLRC       hhhhhhhheeeeeeeeeeeeeeeeeeehhhhccccceeeeeeccchhhhhhhhhccchhhhhhhccccee
PHD        hhhhhhhhhccceeeeeeeeccchhhhhhhhhhccceeeeecccchhhhhhhhhchhhhhhhhhccccee
Sec.Cons.  hhhhhhhhh???eeeeeeeeccc?hhhhhhhhhccceeeeecccchhhhhhhhhccchhhhhhhccccee
```

Figure 5.3 NPS results for (a) ORFs 1 and (b) ORF 2

The secondary structure predictions from the DSC, MLRC and PHD servers are annotated accordingly. The consensus from these servers is designated as Sec.Cons. Predictions of the secondary states of residues are colour coded as: coiled regions, yellow; helices, black and extended sheets, red.

5.3.4 Motif searches

The domain queries for ORF1 and ORF2 were performed with the MOTIF program (Section 2.11.6) The best scoring matches were against DUF29 (Domain of unknown function 29) and Usp (Universal stress protein) respectively (Table 5.1).

Table 5.1 Highest scoring domain matches from PFAM

Query	*Aligned region*	*Raw score*	*E-value*	*Domain match*
ORF 1	9 -109	85.4	4.7e-24	DUF29
ORF 2	15 -156	80.1	3.2e-23	Usp

5.3.5 Trans-membrane domain search

UNIVERSITY of the WESTERN CAPE

Extended hydrophobic regions in protein sequences are generally considered as indicative of potential transmembrane regions. The amino acid sequences of ORF 1 and 2 were investigated for potential transmembrane regions using the method of Kyte and Doolittle, run on the BioEdit program (Section 2.11.5.1). In order to confirm the predictions from the Kyte and Doolittle plots the potential presence of trans-membrane domains were also investigated with the SOSUI (Hirokawa et al., 1998) program (Section 2.11.5.2).

The Kyte and Doolittle suggested that no significant transmembrane spanning regions could be detected for ORF 1 (Figure 5.4 (a)). The prediction for ORF 2 indicated the presence of an extended stretch of hydrophobic residues between amino acids 25-45 (Figure 5.4 (b)). The peak hydrophobicity reported across this region (0.8) however falls below the suggested value (1.6) for a transmembrane region. The two ORFs could therefore be predicted to be soluble. The results from the SOSUI prediction reflected the same observation reporting that both ORFs are predicted to be soluble (Table 5.2).

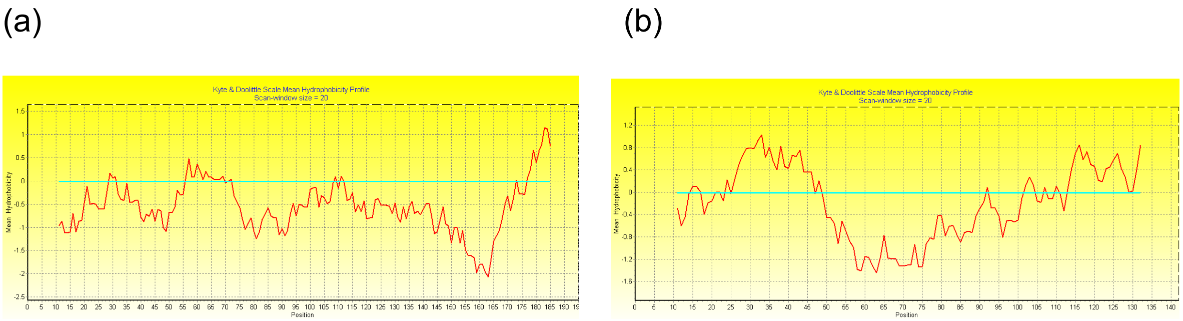


Figure 5.4 Kyte and Doolittle hydrophobicity plots
(a) ORF1 and (b) ORF 2.

Table 5.2 SOSUI results for ORFs 1 and 2

ORF ID	*AVG HYDROPHOBICITY*	*SOLUBLE/INSOLUBLE*
TC11.1	-0.501538	SOLUBLE
TC11.2	-0.092908	SOLUBLE

UNIVERSITY of the
WESTERN CAPE

5.3.6 Signal sequence prediction

The presence of extended hydrophic regions in proteins are not only suggestive of transmembrane regions but could also be suggestive of putative signal sequences. SignalP (Henrik et al., 1997) was used to predict the presence of signal sequences that might be present in the ORFs (Section 2.11.3). The results suggested that no signal sequences could be detected for either ORF 1 or 2 from datasets trained on both Gram positive (Tables 5.3 and 5.4) and Gram negative bacteria (Tables 5.5 and 5.6).

Table 5.3 SignalP-NN results trained on Gram positive bacteria

Query	*Measure*	*Position*	*Value*	*Cutoff*	*Signal peptide?*
DUF29	max. C	35	0.106	0.52	No
	max. Y	35	0.094	0.32	No
	max. S	28	0.430	0.97	No
	mean S	1-34	0.068	0.51	No
USP	max. C	27	0.162	0.52	No
	max. Y	27	0.134	0.32	No
	max. S	24	0.730	0.97	No
	mean S	1-26	0.125	0.51	No

Table 5.4 SignalP-HMM results trained on Gram positive bacteria

Query	*Prediction*	*Signal peptide Probability*	*Max cleavage site probability*
DUF29	Non-secretory protein	0.000	0.000
USP	Non-secretory protein	0.000	0.000

UNIVERSITY of the
WESTERN CAPE

Table 5.5 SignalP-NN results trained on Gram negative bacteria

Query	*Measure*	*Position*	*Value*	*Cutoff*	*Signal peptide?*
DUF29	max. C	35	0.027	0.52	No
	max. Y	35	0.026	0.33	No
	max. S	24	0.037	0.92	No
	mean S	1-34	0.019	0.49	No
USP	max. C	19	0.046	0.52	No
	max. Y	36	0.029	0.33	No
	max. S	24	0.109	0.92	No
	mean S	1-35	0.030	0.49	No

Table 5.6 SignalP-HMM results trained on Gram negative bacteria

Query	*Prediction*	*Signal peptide Probability*	*Max cleavage site probability*
DUF29	*Non-secretory protein*	*0.000*	*0.000*
USP	*Non-secretory protein*	*0.000*	*0.000*

5.3.7 Hidden markov model approaches for the identification of homologs

Identification of distant homologues requires the use of algorithms that implement position specific scoring matrices. Both ORFs were queried for structure-function homologues (Section 2.11.8) using the SAM T2K (Table 5.7), GenTHREADER (Table 5.8) and FUGUE (Table 5.9) programs.

None of the programs predicted any significant scoring homologs for ORF 1. All three programs however predicted significantly scoring templates for ORF 2. Several of the significantly scoring templates could also be observed across all three of the programs and therefore improved the accuracy of the prediction.

UNIVERSITY of the
WESTERN CAPE

Table 5.7 SAM T2K results for ORFs 1 and 2

ORF ID	*PDB sequence ID*	*Length**	*E-value*
ORF 1	1gv3A	248	4.0822e+00
	1cp9A	205	8.4036e+00
	2awpA	198	8.6025e+00
	1xreA	202	1.1520e+01
	1ja0A	620	1.2055e+01
	1j9zA	622	1.2098e+01
	2cn5A	329	1.6037e+01
	1uerA	192	1.6594e+01
	1y67A	229	1.7090e+01
ORF 2	1mjhB	162	1.0921e-20
	1tq8A	163	3.3568e-20
	1mjhA	162	6.6028e-20
	2dumA	170	1.3980e-18
	1wjgA	137	2.1305e-18
	2gm3A	175	4.3091e-17
	1jmvA	141	1.2464e-16
	2pfsA	150	1.2685e-11
	1q77A	138	3.0807e-03

*Length of template sequence

Table 5.8 GenTHREADER results for ORFs 1 and 2

Orf id	*PDB Alignment*	*P-value*	*Aln Len*	*Conf.*
ORF 1	1tv7A0	0.068	157	LOW
	1sedA0	0.073	69	LOW
	1ybvA0	0.133	188	GUESS
	1gw5A0	0.139	160	GUESS
	1bdb00	0.153	188	GUESS
	1ka2A0	0.199	156	GUESS
	1e0cA0	0.199	184	GUESS
ORF 2	1wjgA0	6e-11	135	CERT
	1mjhA0	6e-11	137	CERT
	1jmvA0	8e-11	135	CERT
	2gm3A0	1e-10	141	CERT
	1sbzA0	1e-10	120	CERT
	1qnf00	4e-10	127	CERT
	1g5qA0	0.003	104	MEDIUM

Table 5.9 FUGUE results for ORFs 1 and 2

UNIVERSITY of the
WESTERN CAPE

Orf id	*PDB Alignment*	*Aln length*	*Z-score*	*Conf.*
ORF 1	1I17	107	3.00	GUESS
	2AHJ	211	2.75	GUESS
	1JB0	31	2.24	GUESS
	1CNS	243	2.20	GUESS
	2CVH	220	2.16	GUESS
	GLYCO	244	2.09	GUESS
	1SPF	35	2.05	GUESS
ORF 2	Usp	147	29.02	CERTAIN
	hsd1mjha	143	28.24	CERTAIN
	hs1jmva	140	24.29	CERTAIN
	hs2gm3a	150	23.99	CERTAIN
	hs2gm3a	150	23.99	CERTAIN
	hs1q77a	137	6.75	CERTAIN

5.3.8 Functional context prediction

The ORF1 and 2 amino acid sequences were analysed with the PFP program (Section 2.11.7) that predicts molecular function, biological process and cellular component from the Gene Ontologies (Hawkins et al., 2006). The scores for the predictions are reported in order of relative probability within the respective functional categories, as opposed to global probability. The results for ORF1 and 2 are reported in Tables 5.10 and 5.11 respectively. Multiple high scoring matches were reported under the categories of biological process and molecular function for both ORFs. These were generally considered as confident predictions for the ORFs. The matches under the cellular component category were not considered significant for use in the functional inferences of either ORF.

UNIVERSITY of the
WESTERN CAPE

Table 5.10 PFP results for ORF1

GO CATEGORIES	GO ID*	SCORE	DEFINITION
Biological Process	GO.0009168	123.15	purine ribonucleoside monophosphate biosynthesis
	GO.0009117	106.07	nucleotide metabolism
	GO.0006196	101.45	AMP catabolism
	GO.0019735	9.41	antimicrobial humoral response (sensu Vertebrata)
	GO.0007242	9.31	intracellular signaling cascade
	GO.0007165	8.35	signal transduction
	GO.0007275	8.15	development
Molecular Function	GO.0003876	314.01	AMP deaminase activity
	GO.0016787	137.07	hydrolase activity
	GO.0019239	128.91	deaminase activity
	GO.0004784	8.33	superoxide dismutase activity
	GO.0005198	7.19	structural molecule activity
	GO.0004435	6.21	phosphoinositide phospholipase C activity
	GO.0004780	5.82	sulfate adenylyltransferase (ADP) activity

*

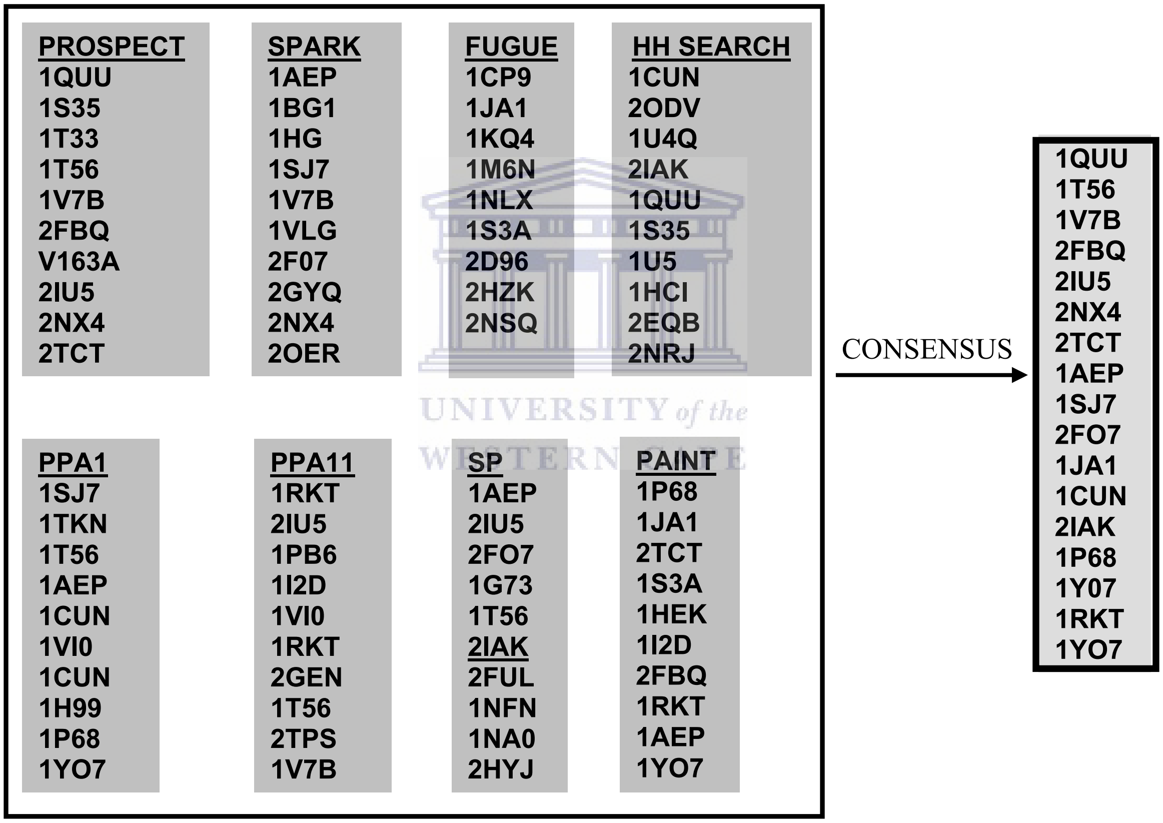
Table 5.11 PFP results for ORF 2

GO CATEGORIES	***GO ID***	***SCORE***	***DEFINITION***
Biological Process	GO.0006950	1798.61	response to stress
	GO.0006885	81.99	regulation of pH
	GO.0009415	81.54	response to water
	GO.0006470	59.74	protein amino acid dephosphorylation
	GO.0007166	43.00	cell surface receptor linked signal transduction
	GO.0007600	33.43	sensory perception
	GO.0007165	33.23	signal transduction
	GO.0006986	32.08	response to unfolded protein
	GO.0006457	32.03	protein folding
	GO.0006468	26.38	protein amino acid phosphorylation
Molecular Function	GO.0008201	81.96	heparin binding
	GO.0005524	70.71	ATP binding
	GO.0005554	62.68	molecular_function unknown
	GO.0016301	56.21	kinase activity
	GO.0016740	51.79	transferase activity
	GO.0004674	40.65	protein serine/threonine kinase activity
	GO.0016757	39.48	transferase activity, transferring glycosyl groups
	GO.0016308	39.42	-phosphatidylinositol-4-phosphate 5-kinase
	GO.0004679	38.20	activity
	GO.0000155	35.01	AMP-activated protein kinase activity
			two-component sensor molecule activity

UNIVERSITY of the
WESTERN CAPE

5.3.9 Consensus based homology searches

The lack of significant sequence-structure homologues for ORF1 from the SAM T2K, FUGUE and GenTHREADER programs prompted additional investigation through the implementation of a consensus approach. This represented a score-independent approach where homologues were investigated by virtue of multiple occurrence across different threading programs (Section 2.11.8). The LOMETS server (Wu and Zhang, 2007) takes predictions from nine different servers that represent a diverse set of state-of-the-art threading algorithms, i.e. FUGUE (Shi et al., 2001), HHSEARCH (Soding, 2005), PROSPECT2 (Xu and Xu, 2000), SAM-T02 (Karplus et al., 2003), SPARKS2 (Zhou and Zhou, 2004), SP3 (Zhou and Zhou, 2005), PAINT, PPA-I and PPA-II (Wu and Zhang, 2007). A total of 17 potential homologs were identified across more than one server and are listed, according to PDB code, in Figure 5.5.



PROSPECT
1QUU
1S35
1T33
1T56
1V7B
2FBQ
V163A
2IU5
2NX4
2TCT

SPARK
1AEP
1BG1
1HG
1SJ7
1V7B
1VLG
2F07
2GYQ
2NX4
2OER

FUGUE
1CP9
1JA1
1KQ4
1M6N
1NLX
1S3A
2D96
2HZK
2NSQ

HH SEARCH
1CUN
2ODV
1U4Q
2IAK
1QUU
1S35
1U5
1HCI
2EQB
2NRJ

PPA1
1SJ7
1TKN
1T56
1AEP
1CUN
1VI0
1CUN
1H99
1P68
1YO7

PPA11
1RKT
2IU5
1PB6
1I2D
1VI0
1RKT
2GEN
1T56
2TPS
1V7B

SP
1AEP
2IU5
2FO7
1G73
1T56
2IAK
2FUL
1NFN
1NA0
2HYJ

PAINT
1P68
1JA1
2TCT
1S3A
1HEK
1I2D
2FBQ
1RKT
1AEP
1YO7

CONSENSUS

1QUU
1T56
1V7B
2FBQ
2IU5
2NX4
2TCT
1AEP
1SJ7
2FO7
1JA1
1CUN
2IAK
1P68
1Y07
1RKT
1YO7

UNIVERSITY of the
WESTERN CAPE

Figure 5.5. Consensus HMM results for ORF 1
All reported templates from the LOMETS server is displayed under their respective search algorithms in the box to the left. The templates that are reported across more than one algorithm are dispayed in the box on the right.

5.3.10 Secondary structure comparisons for ORF1

The sequence of ORF1 was implemented in a secondary structure based homology detection program (Section 2.11.13) developed by Kim and Xie (2006). The analysis was performed with the full-length sequence of ORF1 (Table 5.12) as well as the N-terminal region of ORF1 spanning the first 72 amino acids in a Helix-Coil-Helix-Coil-Helix conformation (Table 5.13). This region was specifically investigated to establish a potential conservation of the helix-coil-helix fold observed in the transcription factors that was reported as templates across multiple programs in LOMETS.

Table 5.12 Secondary structure homologs for full length ORF1

PDB ID	*SCOP CLASS*	*DESCRIPTION*	*SCORE*	*Z-SCORE*
1qkm	a.123.1.1	HUMAN OESTROGEN RECEPTOR BETA LIGAND-BINDING DOMAIN IN COMPLEX WITH PARTIAL AGON...	71.9	3.37
1axi ch A	a.26.1.1	STRUCTURAL PLASTICITY AT THE HGH: HGHBP INTERFACE	70.5	3.24
1a52 ch A	-	ESTROGEN RECEPTOR ALPHA LIGAND-BINDING DOMAIN	70.5	3.24
1ere ch A	-	HUMAN ESTROGEN RECEPTOR LIGAND-BINDING DOMAIN IN COMPLEX WITH 17 BETA-ESTRADIOL...	70.2	3.21
1nde	-	ESTROGEN RECEPTOR BETA WITH SELECTIVE TRIAZINE MODULATOR	70.1	3.21
1uom	-	THE STRUCTURE OF ESTROGEN RECEPTOR	69.9	3.18
1g50 ch A	-	CRYSTAL STRUCTURE OF A WILD TYPE HER ALPHA LBD AT 2.9 ANGSTROM RESOLUTION	69.9	3.18
1m47	a.26.1.2	CRYSTAL STRUCTURE OF HUMAN INTERLEUKIN-2	69.7	3.16

UNIVERSITY of the
WESTERN CAPE

Table 5.13 Secondary structure matches for N-terminal region of ORF1

PDB ID	*SCOP CLASS*	*DESCRIPTION*	*SCORE*	*Z-SCORE*
1f4m ch A	-	P3(2) CRYSTAL STRUCTURE OF ALA2 ILE2-6, A VERSION OF ROP	82.7	4.72
1df4	-	INTERACTIONS BETWEEN HIV-1 GP41 CORE AND DETERGENTS	82.1	4.66
1rop	-	TRANSCRIPTION REGULATION	79.5	4.45
1h2s ch B	f.13.1.1	MOLECULAR BASIS OF TRANSMENBRANE SIGNALLING BY SENSORY RHODOPSIN II-TRANSDUCER	78.5	4.37
1jpx ch A	-	MUTATION THAT DESTABILIZE THE GP41 CORE	77.6	4.29
1nkd	a.30.1.1	ATOMIC RESOLUTION (1.07 ANGSTROMS) STRUCTURE OF THE ROP MUTANT	77.6	4.29
1joy ch A	a.30.2.1	SOLUTION STRUCTURE OF THE HOMODIMERIC	77.4	4.28
1b6q	a.30.1.1	DOMAIN OF ENVZ FROM ESCHERICHIA COLI	76.8	4.23
1gmg ch A	-	ALANINE 31 PROLINE MUTANT OF ROP PROTEIN	76.8	4.23
1rpo	-	ALANINE 31 PROLINE MUTANT OF ROP PROTEIN, MONOCLINIC FORM	76.3	4.19
1ec5 ch A	-	TRANSCRIPTION REGULATION	76.3	4.18

UNIVERSITY of the
WESTERN CAPE

Analysis of the full length ORF demonstrates multiple matches to the secondary structure of estrogen receptor. The actual alignments, however, shows several gaps in the alignment. The N-terminal region had matches to proteins involved in transcriptional regulation. These regions displayed fewer gaps in the alignments but did not show concordance between the lengths of the respective α-helices and β-sheets (Figure 5.6)

```
(a) 1rop TRANSCRIPTION REGULATION
Z-Score: 4.45
Experimental residue alignment:
16 - 66     length 66
1 - 55     length 56
E-HHHHHHHHHHHHHHHHHHHHHHHHHHHHC--HHHHHHHHHHHHHHHHHHHHH---
CCHHHHHHHHHHHHHHHHHHHHHHHHHH—CCCHHHHHHHHHHHHHHHHHHHHHHHH
```

```
(b) 1nkd ATOMIC RESOLUTION (1.07 ANGSTROMS) STRUCTURE OF THE ROP MUTANT
Z-Score: 4.29
Experimental residue alignment:
16 - 66     length 66
1 - 58     length 59
E-HHHHHHHHHHHHHHHHHHHHHHHHHHHHC----HHHHHHHHHHHHHHHHHHHHH----
CCHHHHHHHHHHHHHHHHHHHHHHHHHH--CCCCCHHHHHHHHHHHHHHHHHHHHHHHHH
```

```
(c)1rpo TRANSCRIPTION REGULATION
Z-Score: 4.19
Experimental residue alignment:
16 - 66     length 66
1 - 58     length 61
E-HHHHHHHHHHHHHHHHHHHHHHHHHHHHC----HHHHHHHHHHHHHHHHHHHHH----
CCHHHHHHHHHHHHHHHHHHHHHHHHHH--CCCCCHHHHHHHHHHHHHHHHHHHHHHHHH
```

Figure 5.6 Secondary structure alignments for ORF1
The actual alignments of the secondary structure features are shown between ORF1 and transcriptional regulators: (a) 1rop, (b) 1nkd and (c) 1rpo. For all alignments: the secondary structure conformation (C-coil, E-extended sheet and H-helix) of the ORF1 N-terminal query is show at the top and the corresponding secondary structure of the respective templates are shown below.

UNIVERSITY of the
WESTERN CAPE

5.3.11 Homology modeling

Homology models were constructed for ORFs, if a specific homolog was reported by FUGUE, GenTHREADER and SAM T02, and was scored as a significant match. No such homologs were identified for ORF 1. The PDB template MJH was reported as a homolog for ORF 2 from the FUGUE, GenTHREADER and SAM T02 results. All three programs reported this template within the range of significant matches. MJH was therefore selected as template for homology modeling of ORF 2 with the Modeller program (Section 2.11.9.1). The model for ORF 2 is reported in Figure 5.7 along with the published structure of MJH.

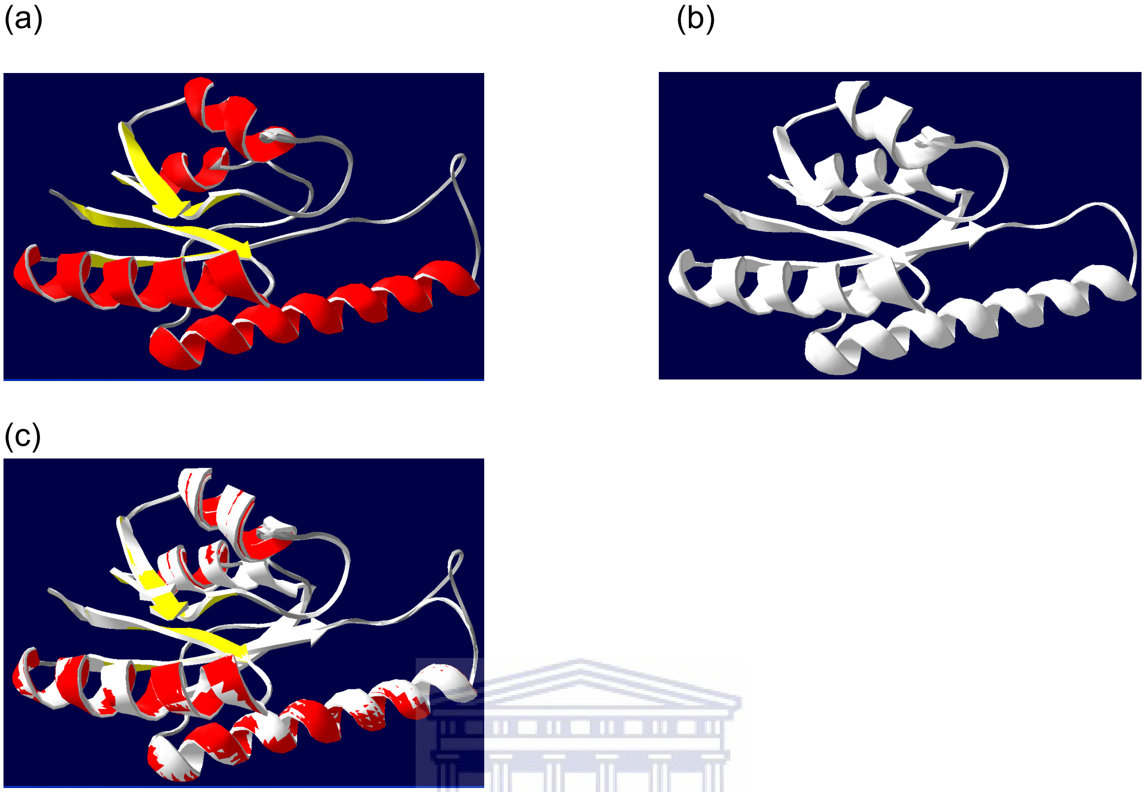


Figure 5.7 Cartoon representation of the ORF 2 – UspA homology model.
(a) The predicted fold for ORF 2 based on the UspA template. (b) The UspA structure. (c) Overlay of the ORF 4 structure prediction with UspA.

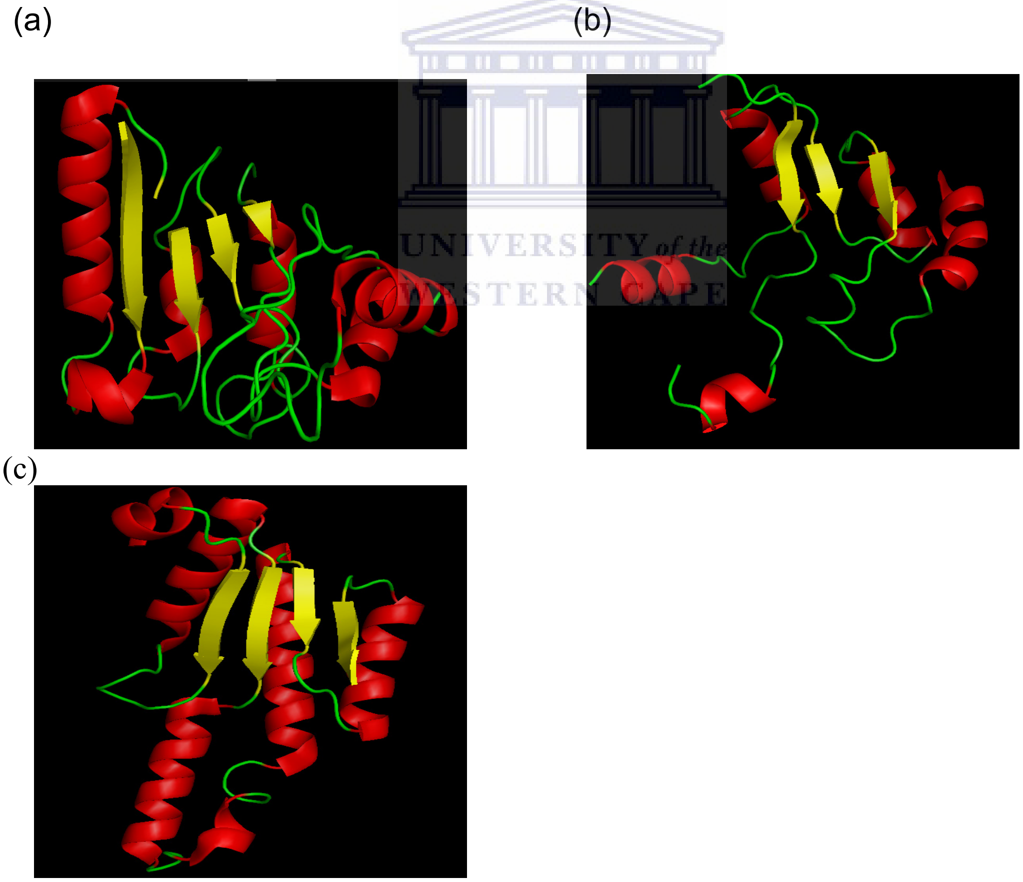
UNIVERSITY of the
WESTERN CAPE

The ORF 2 model displayed a conservation of the α/β fold of the MJH template and corresponds to the predicted secondary structure of ORF 2. The MJH template corresponds to the UspA proteins with known ATP binding capacity.

5.3.12 Identifying Usp homologues with known functions

The published expression data on the Usp protein family suggests a functional relationship between the UspA, UspC and UspD protein members (Gustavsson et al., 2002). The exact function is however unknown. It was therefore argued that information on functional homologs of these proteins might aid in the annotation process. The sequences of ORF 2 (homolog of UspA), UspC and UspD were therefore queried against the FUGUE (Section 2.11.8) server in order to retrieve potential homologs. The highest scoring templates with known function were investigated for possible conserved homologous domains and binding sites.

The highest scoring functionally annotated templates for ORF2 (UspA), UspC and UspD were dihydropyriminidase, dihydroorotase and glycolate oxidase respectively. The respective regions of these homologs that align with ORF2 (UspA), UspC and UspD were observed as a α/β fold (Figure 5.8).



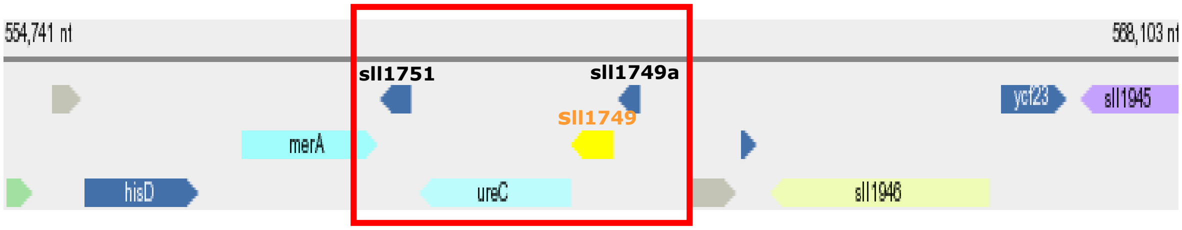
UNIVERSITY of the WESTERN CAPE

Figure 5.8 Structural features of the Usp homologs
(a) OGJ region aligning with ORF 2 (UspA), N-terminal domain, (b) GOX region aligning with UspC, (c) 1GKP – region aligning with UspD and (d) 1O4V- region aligning with ORF 2 (UspA), C-terminal domain.

5.3.13 Genomic context mapping

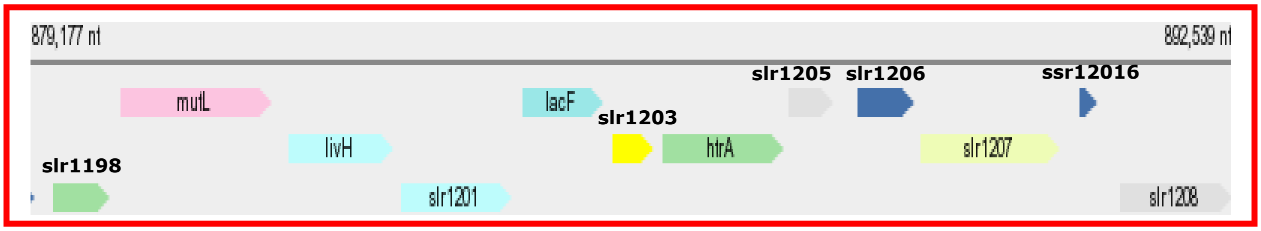
The implementation of multiple divergent strategies for functional annotation of hypothetical/novel genes has obvious advantages over the use of one strategy. Based on the argument that genes within a close physical proximity on a genome potentially have related functional properties (von Mering et al., 2003) a comparative mapping approach was implemented for the DUF29 gene (Section 2.11.11). The genes flanking the multiple DUF29 loci for Synechococcus spp. were investigated to determine possible relationships of the DUF29-like genes. All DUF29 flanking genes, of the same transcriptional orientation were identified (Figure 5.9) and functionally annotated (Table 5.14). The functionally annotated genes flanking the DUF29 loci for Synechococcus were compared in terms of possible catalysed reactions, potential relationships in biochemical pathways and common interacting protein partners. The same strategy was also implemented for additional genomes of organisms with the highest scoring homologs of DUF29 domains (Figure 5.10) as well as organisms with UspA domains (Figure 5.11 and Table 5.15).

UNIVERSITY of the
WESTERN CAPE

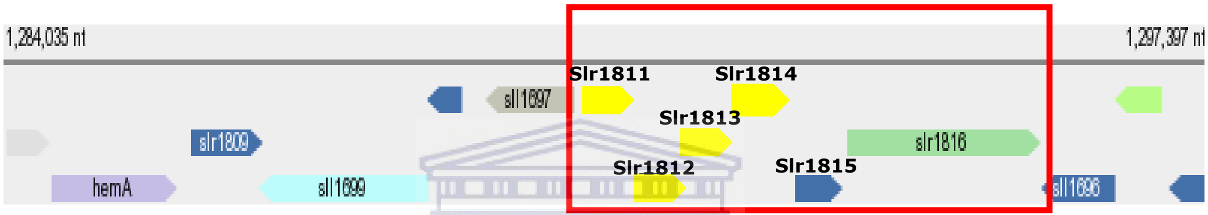


Arrangement of genes flanking locus Sll1749
sll1751, hypothetical prot sll1751; ureC, urease alpha subunit; sll1749, hypothetical prot sll1749; sll1749a, hypothetical prot sll1749a

Figure 5.9 The DUF29 loci of *Synechococcus elongates*.
Genes at the DU29 locus, within the same transcriptional orientation are boxed in red

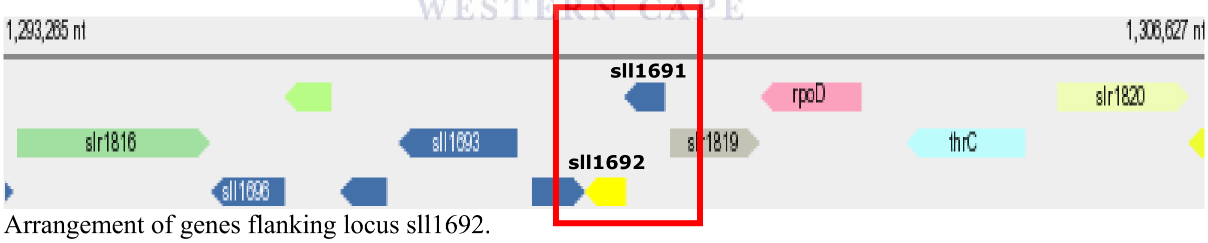


Arrangement of genes flanking locus slr1203
slr1198, rehydrin; mutL, DNA mismatch repair protein; livH, high-affinity branched-chain amino acid transport; slr1201, hypothetical protein slr1201; lacF, lactose transport system permease protein; slr1203, hypothetical protein slr1203; htrA, serine protease; slr1205, ferredoxin component; slr1206, hypothetical protein slr1206; slr1207, hypothetical protein slr1207; ssr12016, hypothetical protein ssr12016;

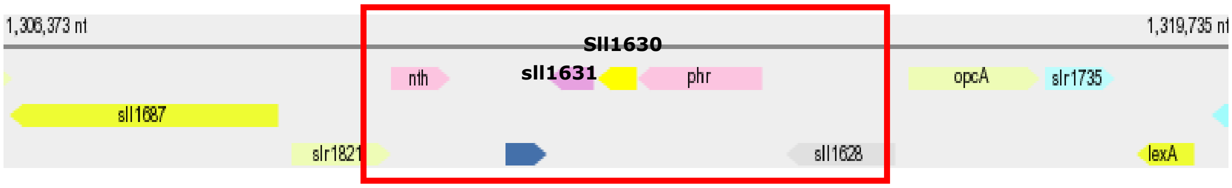


Arrangement of genes flanking loci slr1811, slr1812, slr1813 and slr1814
slr1811, hypothetical protein slr1811; slr1812, hypothetical protein slr1812; slr1813, hypothetical protein slr1813; slr1814, hypothetical protein slr1814; slr1815, hypothetical protein slr1815;
slr1816, hypothetical protein slr1816

UNIVERSITY of the
WESTERN CAPE

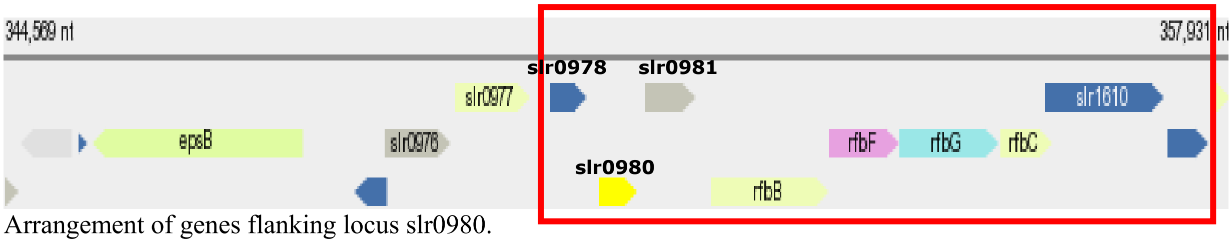


Arrangement of genes flanking locus sll1692.
sll1692, hypothetical protein sll1692; sll1691, hypothetical protein sll1691

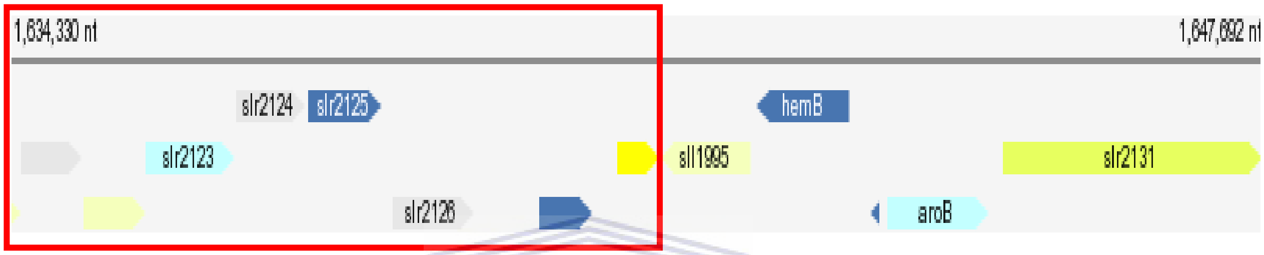


Arrangement of genes flanking locus sll1630.
sll1631, hypothetical protein sll1631; sll1630, hypothetical protein sll1630; phr, DNA photolyase; sll1628, hypothetical protein sll1628; opcA, OxPPCycle

Figure 5.9 (Continued) The DUF29 loci of *Synechococcus elongates*.
Genes at the DU29 locus, within the same transcriptional orientation are boxed in red

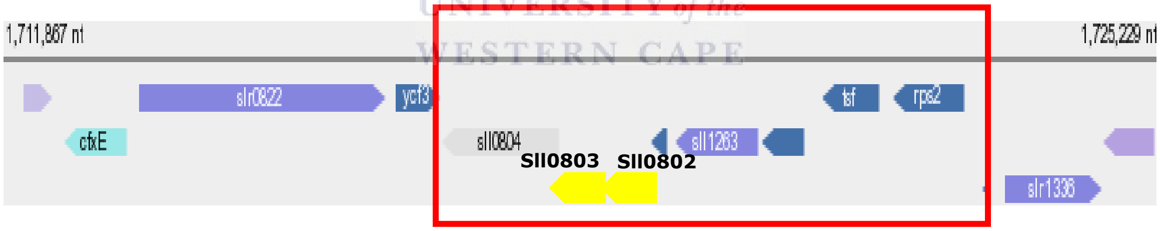


Arrangement of genes flanking locus slr0980.
slr0976, hypothetical protein slr0976; slr0977, ABC transporter; slr0978, hypothetical protein slr0978; slr0980, hypothetical protein slr0980; slr0981, hypothetical protein slr0981 ; rfbB, ABC transporter; rfbF, alpha-D-glucose-1-phosphate cytidylyltransferase ; rfbG, CDP-glucose-4,6-dehydratase; rfbC, dTDP-6-deoxy-L-mannose-dehydrogenase; slr1610, hypothetical protein slr1610

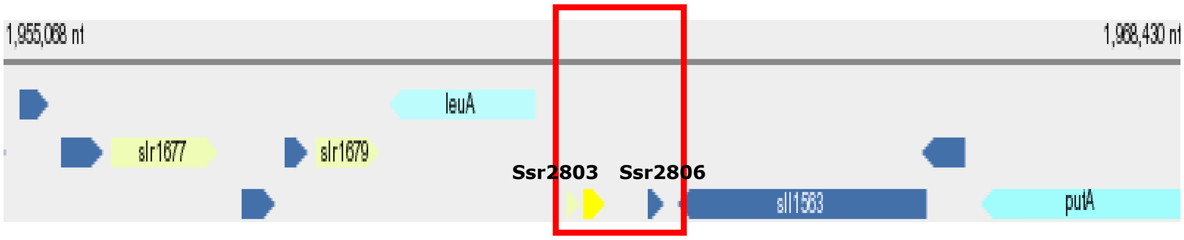


Arrangement of genes flanking locus slr2128.
slr2123, isomer specific 2-hydroxyacid dehydrogenase; slr2124, short-chain alcohol dehydrogenase family; slr2125, hypothetical protein slr2125; slr2126, hypothetical protein slr2126; ???; slr2128, hypothetical protein slr2128.

UNIVERSITY of the
WESTERN CAPE

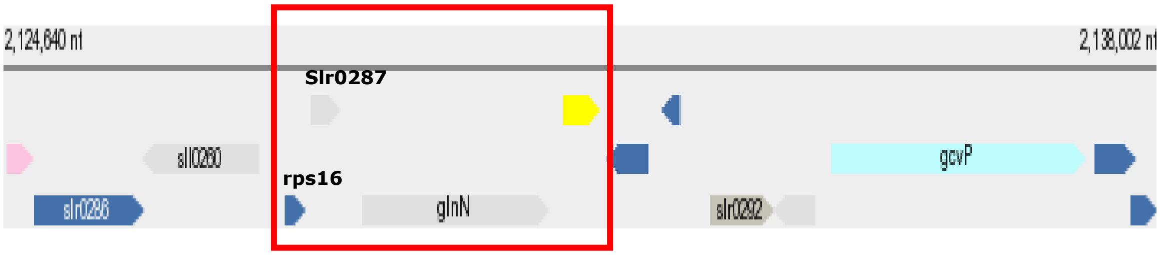


Arrangement of genes flanking loci sll0803 and sll0802.
sll0804, hypothetical protein sll0804; sll0803, hypothetical protein sll0803; sll0802, hypothetical protein sll0802; ssl1498, hypothetical protein ssl1498; sll1263, hypothetical protein sll1263; sll1262, hypothetical protein sll1262; sll1261, elongation factor Ts; sll1260, 30S ribosomal protein S2;

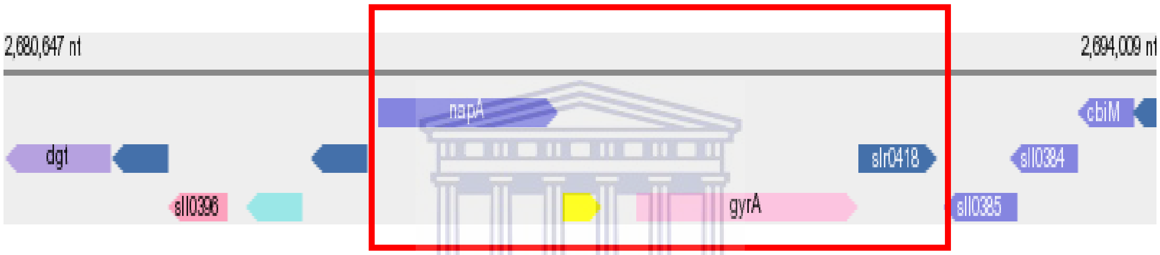


Arrangement of genes flanking locus ssr2803.
ssr2803, hypothetical protein ssr2803; ssr2806, hypothetical protein ssr2806.

Figure 5.9 (Continued) The DUF29 loci of *Synechococcus elongates*
Genes at the DU29 locus, within the same transcriptional orientation are boxed in red

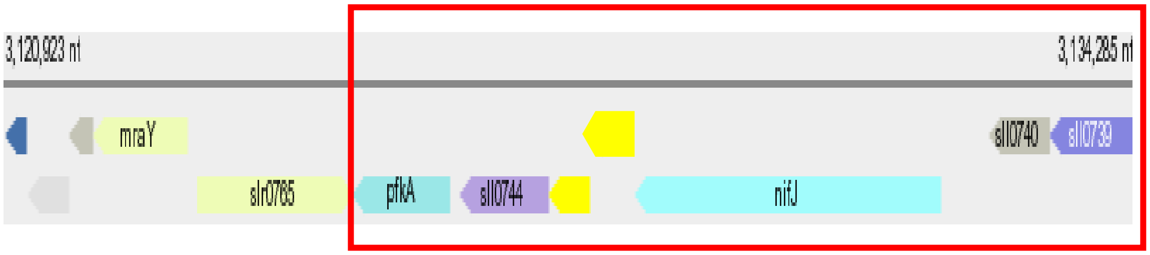


Arrangement of genes flanking locus slr029.
rps16, 30S ribosomal protein S16; slr0287, hypothetical protein slr0287; glnN, glutamate--ammonia ligase; slr0291, hypothetical protein slr0291.



Arrangement of genes flanking locus slr0416.
slr0415, Na(+)/H(+) antiporter; slr0416, hypothetical protein slr0416; gyrA, DNA gyrase subunit A; slr0418, hypothetical protein slr0418.

UNIVERSITY of the
WESTERN CAPE



Arrangement of genes flanking loci sll0743 and sll0742
pfkA, phosphofructokinase; sll0744, dihydroorotate dehydrogenase; sll0743, hypothetical protein sll0743; sll0742, hypothetical protein sll0742; nifJ, pyruvate oxidoreductase; sll0740, hypothetical protein sll0740; sll0739, ABC transport

Figure 5.9 (Continued) The DUF29 loci of *Synechococcus elongates.*
Genes at the DU29 locus, within the same transcriptional orientation are boxed in red

Table 5.14 Annotated functions of genes at the DUF29 loci.

DUF locus	*Flanking genes*	*Function*
sll1749	urease alpha subunit	An enzyme that catalyzes the hydrolysis of urea into carbon dioxide and ammonia.
slr1203	rehydrin	Peroxiredoxin
	DNA mismatch repair protein	Contributes to the overall fidelity of DNA replication by targeting mispaired bases that arise through replication errors during homologous recombination and as a result of DNA damage. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex
	high-affinity branched-chain amino acid transport	Responsible for the high affinity transport of branched-chain amino acids
	lactose transport system permease protein	Lactose transport
	serine protease	The serine proteases are a family of enzymes that cut certain peptide bonds in other proteins
	ferredoxin component	Electron carrier proteins with an iron-sulphur cofactor that act in a wide variety of metabolic reactions
slr1811, slr1812, slr1813 and slr1814		Hypothetical proteins
sll1692		Hypothetical protein

UNIVERSITY of the WESTERN CAPE

Table 5.14 (continued) Annotated functions of genes at the DUF29 loci.

DUF locus	*Flanking genes*	*Functional features reported from the NCBI*
sll1630	DNA photolyase	Binds to and repairs cyclobutane pyrimidine dimers induced by UV radiation.
	OxPPCycle	-
slr0980	ABC transporter	Transport protein
	alpha-D-glucose-1-phosphate cytidylyltransferase	Starch and sucrose metabolism, Nucleotide sugars metabolism
	CDP-glucose-4,6-dehydratase	Intramolecular oxidation-reduction that involves an internal hydrogen transfer from C-4 of the substrate to C-6 of the resulting product.
	dTDP-6-deoxy-L-mannose-dehydrogenase	Oxidoreductases. Acting on the CH-OH group of donors
slr2128	isomer specific 2-hydroxyacid dehydrogenase	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
	short-chain alcohol dehydrogenase family	Large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. The first member of this family to be characterized was Drosophila alcohol dehydrogenase
sll0803 and sll0802	elongation factor Ts	Incoming amino acid monomers enter the ribosomal A site in the form of aminoacyl-tRNAs complexed with elongation factor Tu (EF-Tu) and GTP.
	30S ribosomal protein S2	Ribosomes are the particles that catalyze mRNA-directed protein synthesis in all organisms. The codons of the mRNA are exposed on the ribosome to allow tRNA binding.

UNIVERSITY of the WESTERN CAPE

Table 5.14 (continued) Annotated functions of genes at the DUF29 loci.

DUF locus	*Flanking genes*	*Functional features reported from the NCBI*
ssr2803		
slr029	*30S ribosomal protein S16*	*Ribosomes are the particles that catalyze mRNA-directed protein synthesis in all organisms. The codons of the mRNA are exposed on the ribosome to allow tRNA binding.*
	glutamate--ammonia ligase	*Catalyses the conversion of ATP, l-glutamate, and nh3 to ADP, orthophosphate, and l-glutamine*
slr0416	*Na(+)/H(+) antiporter*	*Transport*
	DNA gyrase subunit A	*gyrA subunit of DNA gyrase could participate in the repair of certain types of DNA damage, such as cross-links, in a mode independent of SOS-regulated excision repair and post-replication repair.*
sll0743 and sll0742	*ABC transporter*	*Transport protein*
	dihydroorotate dehydrogenase	*Catalyzes the fourth step in the de novo biosynthesis of pyrimidine, the conversion of dihydroorotate into orotate.*
	phosphofructokinase	*Catalyses the phosphorylation of fructose-6-phosphate to fructose-1,6- bisphosphate, a key regulatory step in the glycolytic pathway*

UNIVERSITY of the
WESTERN CAPE

Cyanothece sp. CCY0110

CY0110_26502	hypothetical protein (Permeases of the drug/metabolite transporter)
CY0110_26507	hypothetical protein
CY0110_26512	hypothetical protein
CY0110_26517	**hypothetical protein (DUF29)**
CY0110_26522	hypothetical protein (COG1943 Transposase)
CY0110_26527	hypothetical protein
CY0110_26532	ribonucleotide reductase subunit alpha
CY0110_26537	hypothetical protein (COG0553 Superfamily II DNA/RNA helicases)
CY0110_26542	hypothetical protein (COG1196 Chromosome segregation ATPases)
CY0110_26547	hypothetical protein (COG5637 Predicted integral membrane protein)
CY0110_26552	zeta-carotene desaturase

Rhodospirillum rubrum ATCC 11170

Rru_A3565	Rhomboid-like protein
Rru_A3566	HemY-like
Rru_A3567	Uroporphyrinogen III synthase HEM4
Rru_A3568	Porphobilinogen deaminase
Rru_A3569	O-sialoglycoprotein endopeptidase
Rru_A3570	**Protein of unknown function DUF29**
Rru_A3571	Glycerol-3-phosphate dehydrogenase (NAD(P)+)
Rru_A3572	YCII-related
Rru_A3573	Protein of unknown function DUF589
Rru_A3574	Rieske (2Fe-2S) region
Rru_A3575	Acetate--CoA ligase
Rru_A3576	UDP-glucuronate 5'-epimerase

UNIVERSITY of the
WESTERN CAPE

Hydrogenobacter thermophilus

korA	2-oxoglutarate ferredoxin oxidoreductase alpha subunit
korB	2-oxoglutarate ferredoxin oxidoreductase beta subunit
Orf3	putative protein
Orf4	**putative protein**
leuS	leucine-tRNA ligase

Figure 5.10 Arrangement of genes at the DUF29 loci of related organisms.
The different organisms are assigned in bold letters at the top of the list of genes at the locus. Genes are positioned in the order in which they appear at the respective loci and the DUF29 gene is underlined.

Nitrosococcus oceani ATCC 19707] (locus 1)

Noc_1889	Enoyl-CoA hydratase/isomerase
Noc_1890	Propionyl-CoA carboxylase
Noc_1891	Acetyl-CoA C-acetyltransferase
Noc_1892	hypothetical protein
Noc_1893	Methylmalonyl-CoA mutase-like
Noc_1894	**protein of unknown function DUF29**
Noc_1895	hypothetical protein
Noc_1896	hypothetical protein
Noc_1897	hypothetical protein

Nitrosococcus oceani ATCC 19707 (locus 2)

Noc_0763	hypothetical protein
Noc_0764	hypothetical protein
Noc_0765	hypothetical protein
Noc_0766	**Protein of unknown function DUF29**
Noc_0767	PpiC-type peptidyl-prolyl cis-trans isomerase
Noc_0768	Protein of unknown function DUF1568
Noc_0769	Gamma-glutamyltransferase
Noc_0770	Phosphoenolpyruvate carboxylase
Noc_0771	Glycogen/starch synthase, ADP-glucose type
Noc_0772	Glucose-6-phosphate 1-dehydrogenase

UNIVERSITY of the
WESTERN CAPE

Nostoc punctiforme PCC 73102

Npun02007017	COG2124: Cytochrome P450
Npun02007018	COG0546: Predicted phosphatases
Npun02007019	hypothetical protein
Npun02007021	hypothetical protein
Npun02007022	**Chromosome segregation ATPases (DUF29)**
Npun02007023	hypothetical protein
Npun02007024	COG2319: FOG: WD40 repeat
Npun02007025	COG2268: Uncharacterized conserved in bacteria
Npun02007026	COG0355: F0F1-type ATP synthase, epsilon subunit
Npun02007028	COG0055: F0F1-type ATP synthase, beta subunit

Figure 5.10 (continued) Arrangement of genes at the DUF29 loci of related organisms. The different organisms are assigned in bold letters at the top of the list of genes at the locus. Genes are positioned in the order in which they appear at the respective loci and the DUF29 gene is underlined.

Photorhabdus luminescens subsp. laumondii TTO1 (locus 1)	
Plu4114	gldA
Plu4115	glycerol dehydrogenase
Plu4117	hypothetical protein
Plu4118	**Similar to Unknown protein**
Plu4119	glpT
Plu4119	glycerol-3-phosphate transporter
plu4120	glycerophosphodiester phosphodiesterase
Plu4121	lysyl-tRNA synthetase

Photorhabdus luminescens subsp. laumondii TTO1 (2) (locus 2)	
Plu3927	2',3'-cyclic nucleotide 2'-phosphodiesterase
Plu3928	hypothetical protein (Some similarities with lysine-tRNA ligase LysS)
Plu3929	recombination associated protein
Plu3930	hypothetical protein (Similarities with Huntington interacting protein)
Plu3931	hypothetical protein
Plu3932	hypothetical protein
Plu3933	**Similar to Unknown protein**
Plu3934	hypothetical protein
Plu3935	hypothetical protein
Plu3936	hypothetical protein

UNIVERSITY of the
WESTERN CAPE

Photorhabdus luminescens subsp. laumondii TTO1 (3) (locus 3)	
Plu2336	hypothetical protein (Some similarities with acyl-CoA oxidase)
Plu2337	hypothetical protein
Plu2338	hypothetical protein
Plu2339	**Similar to hypothetical protein**
Plu2340	hypothetical protein (Highly similar to tyrosine-specific transport system)
Plu2341	hypothetical protein (Some similarities with decarboxylase)
Plu2342	hypothetical protein (Similar to probable oxidoreductase)
Plu2343	hypothetical protein (Similarities with Unknown protein of Photorhabdus)
Plu2344	3-ketoacyl-(acyl-carrier-protein) reductase
Plu2344	hypothetical protein (Some similarities with probable oxidoreductase)

Figure 5.10 Arrangement of genes at the DUF29 loci of related organisms.
The different organisms are assigned in bold letters at the top of the list of genes at the locus. Genes are positioned in the order in which they appear at the respective loci and the DUF29 gene is underlined.

Table 5.15 Functional annotations of genes at Usp loci.

Organism	*Gene*	*Function*
Sulfolobus tokodaii str. 7	hypothetical tldD protein	Predicted Zn-dependent proteases
	hypothetical protein ST1523	Predicted Zn-dependent proteases
	hypothetical protein ST1525	acetyl-CoA carboxylase
	hypothetical protein ST1526	Translation, ribosomal structure and biogenesis
	hypothetical NADH-ubiquinone oxidoreductase subunit K	NADH:ubiquinone oxidoreductase
	hypothetical protein ST1528	
Haloarcula marismortui ATCC 43049	hypothetical protein pNG7138	Predicted flavin-nucleotide-binding protein
	hypothetical protein pNG7140	Uncharacterized conserved protein
	hypothetical protein pNG7141	Predicted flavin-nucleotide-binding protein
	cation-transporting ATPase	Involved in Na+/K+, H+/K+, Ca++ and Mg++ transport.
Carboxydothermus hydrogenoformans Z-2901	amidohydrolase family protein	Metal dependent hydrolase superfamily that includes adenine deaminase, dihydroorotase and N-acetylglucosamine-6-phosphate deacetylases,
	nucleotidyltransferase domain protein	Catalysis of the transfer of a nucleotidyl group to a reactant
	hypothetical protein CHY_0081	-
	hypothetical protein CHY_0084	-

UNIVERSITY of the
WESTERN CAPE

Table 5.15 (continued) Functional annotations of genes at Usp loci.

Organism	*Gene*	*Function*
Methanosarcina acetivorans C2A	carbon-monoxide dehydrogenase, catalytic subunit	found in acetogenic and methanogenic organisms and is responsible for the synthesis and breakdown of acetyl-CoA, respectively.
	carbon-monoxide dehydrogenase, Fe-S subunit	Energy production and conversion
	hypothetical protein MA3285	Function unknown
Methanosarcina mazei Go1	DNA topoisomerase I	DNA-binding, ATP-binding
	hypothetical protein MM_3075	Predicted Rossmann fold nucleotide-binding protein
	mRNA 3'-end processing factor	Translation, ribosomal structure and biogenesis
	hypothetical protein MM_3079	-
Pyrobaculum aerophilum str. IM2	conserved protein with 2 CBS domains	-
	transport protein part 2	-
	hypothetical protein PAE2387	-
	hypothetical protein PAE2388	-

UNIVERSITY of the
WESTERN CAPE

5.4 Discussion

Functional inferences for ORF 1

ORF 1 was predicted to contain a DUF29 domain from the MOTIF database. This domain constitutes a region of 120 amino acids from the proteins in which they occur (Bateman et al., 2002). This observation of conserved multiple matches to conserved protein domains is often a good indication of homology (Thompson et al., 1994), which was inferred between ORF1 and the DUF29 proteins. The DUF29 domain is highly represented in the Cyanobacteria, as reported in the PFAM database (Bateman et al., 2002). It was further established that this domain is specifically over-represented in *Synechococcus elongates,* where the domain occurs at 16 loci in the genome (Bateman et al., 2002). Genes that occur at multiple loci within genomes often tend to perform regulatory functions. Literature searches on the DUF29 domain however revealed that no function has been assigned to this protein family to date.

UNIVERSITY of the
WESTERN CAPE

On the basis of the Secondary structure predictions, SignalP (Henrik et al., 1997), Hydrophobicity plots (Thompson et al., 1994) and SOSUI outputs (Hirokawa et al., 1998) ORF1 was predicted to encode a protein with an all α-helix conformation lacking a signal peptide or transmembrane spanning regions and therefore could be regarded as a soluble protein.

The search for functionally annotated homologs of ORF1 yielded matches to the PDB structures of 1gv3A, 1tv7A0 and 1i17 from the SAM T02, GenTHREADER and FUGUE programs respectively. These highest scoring matches for the individual programs (Tables 5.7 to 5.9) were well below the values considered for significant matches (Karplus et al., 2003; Liam and Jones, 2003; Shi et al., 2001). None of these HMM matches could therefore be confidently assigned as homologs for ORF1. It is therefore possible that ORF1 represents an entirely novel fold with no comparable sequence-structure match in the database. The likelihood of this is however small given that the PDB library is presumed to be complete for single domain protein structures at low to moderate resolution (Zang and Skolnick, 2004).

The more plausible argument is that a homolog for ORF 1 could still be present among the low scoring hits, but remains undetected due to the limitations of HMM tools. In this regard, proteins with shared structural relationships has been shown to generate similar E-value scores as false positive matches (Petry and Honig, 2005), thereby complicating the process of selecting the correct one. As a result, low scoring E-values are therefore often not an absolute indicator of false positive or incorrect alignments.

UNIVERSITY of the
WESTERN CAPE

The search for potential templates was therefore extended toward a score-independent approach, by attempting to identify homology matches that are consistently reported across multiple programs. The LOMETS meta-server (Wu and Zhang, 2007) reported matches to three transcription factors (PDB identities: 2IU5, 1T56 1V7B) across three or more programs (Figure 5.5).

Consensus approaches are reported to improve the accuracy of prediction (Wu and Zhang, 2007), therefore the transcription factors potentially represented distant homologs of ORF1. The common thread among the structures of 2IU5, 1T56 and 1V7B (Christen et al., 2006; Dover et al., 2004; Itou et al., 2005) relates to their N-terminal, helix-loop-helix (HLH), DNA binding domain (Nelson, 1995). This domain consist of a three-helix bundle, with the second and third helices comprising the HTH motif and the third helix bound in the major groove of DNA (Huffman and Brennan, 2002). The turn consists of three to four amino acids in which a glycine is usually found in the second position, and the two helices make an angle of 120° (Huffman and Brennan, 2002). The classical secondary structrure features, of the HTH motif, could however not confidently be detected within the sequence of ORF1.

UNIVERSITY of the
WESTERN CAPE

The predicted secondary structure of the N-terminal region of ORF1 was therefore also directly queried against a database of secondary structures for experimentally determined proteins (Kim and Xie,. 2006). Several structures of transcription factors, such as 1rpo, 1 rop and 1nkd (PDB codes) were also reported as templates. These transcription factors were different from those reported for the LOMETS meta-server. The aligned regions between query (ORF1) and template (1rpo, 1 rop and 1nkd) generally corresponded to the N-terminal HTH motif of these transcription factors. In ORF1, however, the turn motif is proposed to be a single glutamine residue as opposed to four residues, of which the second residue is a glutamine for 1rpo, 1rop and 1nkd. This observed variation is however not unique, as several HTH DNA binding proteins have: (i) extra α-helices and β-strands, (ii) differences in the length and angle of the turn in the HTH motif and (iii) rearrangements of the three helix bundle (Nelson 1995). In this context it is therefore still probable that ORF1 represents a HTH motif protein.

The multiple occurrence of the DUF29 domain in *Synechococcus elongatus* made it particularly suitable for implementing strategies that focus on the conservation of gene organization at specific loci (Overbeek et al., 1999; Huynen et al., 2000). Investigation of the DUF29 loci revealed the presence of multiple genes involved in the biosynthesis of nucleotides (alpha-D-glucose-1-phosphate cytidylyltransferase, dihydroorotate dehydrogenase) and genes involved in DNA damage repair systems (DNA mismatch repair protein, DNA photolyase, DNA gyrase subunit A). The remaining genes potentially have a common relationship with regard to interactions with proteins classified as carbon sugars/derivatives (elongation factor Ts, 30S ribosomal protein S2, glutamate--ammonia ligase and phosphofructokinase). Generally the conservation of genes within the same physical location in different genomes increases the possibility of related function or interaction between these proteins (Dandekar et al., 1998). This relationship could relate to co-expression in response to: (i) a common activator, (ii) substrates within a related biochemical pathway or (iii) catalysis of related reactions (von Mering et al., 2003). In this context, the DUF29 genes of *Synechococcus elongatus* are therefore potentially associated with nucleotide

UNIVERSITY of the
WESTERN CAPE

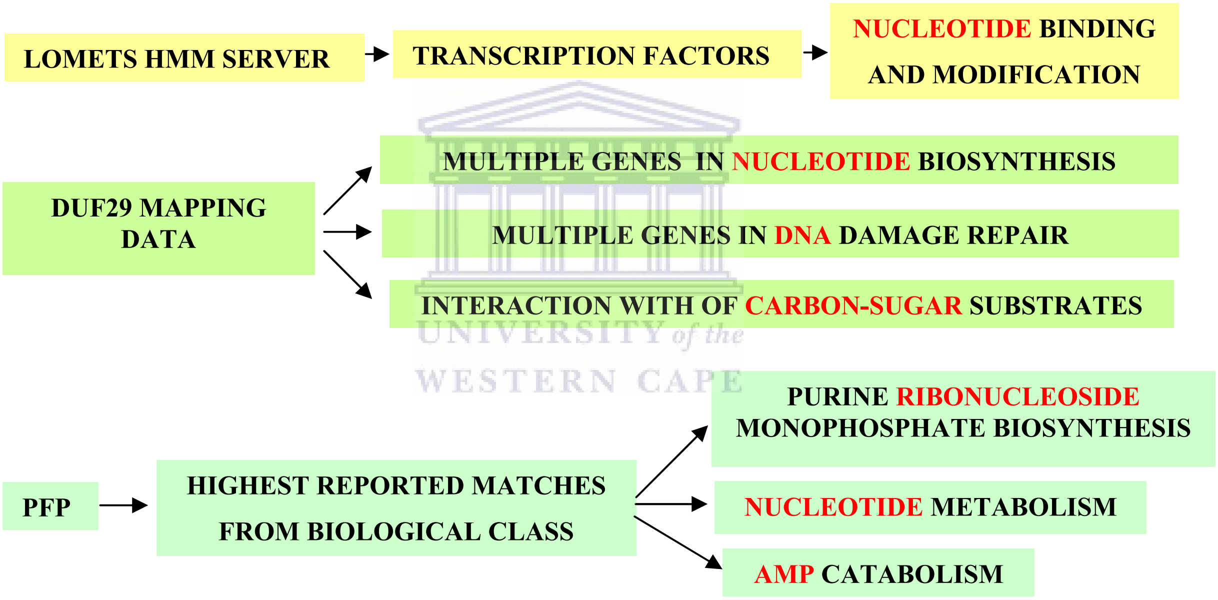
biosynthesis or DNA repair. In broader terms this relationship could potentially be defined as DNA binding/interacting/modifying function.

The genomic context mapping was further also extended to additional organisms with DUF29 domains that share significant alignment with ORF1. The organisms investigated were *Cyanothece sp. CCY0110 (1 locus)*, *Rhodospirillum rubrum ATCC 11170 (1 locus), Hydrogenobacter thermophilus (1 locus)*, *Nostoc punctiforme PCC 73102 (1 locus)*, *Nitrosococcus oceani ATCC 19707 (2 loci)* and *Photorhabdus luminescens subsp. laumondii TTO1 (3 loci)*. No clear relationship could however be deduced among these flanking genes due to the significant number of unannotated ORFs at the DUF 29 loci (Figure 5.10). Investigation of the substrate molecules, that the genes at these loci are predicted to interact with, indicates that they are all carbon sugars or derivatives. This potentially serves as a basis for their physical co-location at the DUF29 loci and therefore of ORF 1.

UNIVERSITY of the
WESTERN CAPE

The PFP analysis of ORF1 listed the highest predicted Biological Processes under the keywords: “purine ribonucleoside monophosphate biosynthesis”, “nucleotide metabolism” and “AMP catabolism” (Hawkins et al., 2006). All of three of these ontologies were predicted with scores in the order of ten times higher than the next best scores. Given that the PFP prediction scores represent relative probabilities, these three highest scoring ontologies should be considered as significant matches (Hawkins et al., 2006). The contribution from this prediction is therefore suggestive of a role for ORF1 in nucleotide metabolism.

Following integration of the inferred results from the HMM searches, genomic context mapping and PFP predictions, all suggested a common theme of a potential nucleotide binding/interacting/modifying function for ORF1 (summarized in Figure 5.12).


LOMETS HMM SERVER
TRANSCRIPTION FACTORS
NUCLEOTIDE BINDING AND MODIFICATION
DUF29 MAPPING DATA
MULTIPLE GENES IN NUCLEOTIDE BIOSYNTHESIS
MULTIPLE GENES IN DNA DAMAGE REPAIR
INTERACTION WITH OF CARBON-SUGAR SUBSTRATES
PFP
HIGHEST REPORTED MATCHES FROM BIOLOGICAL CLASS
PURINE RIBONUCLEOSIDE MONOPHOSPHATE BIOSYNTHESIS
NUCLEOTIDE METABOLISM
AMP CATABOLISM

UNIVERSITY of the
WESTERN CAPE

Figure 5.12 Summary of ORF1 function prediction

Functional inferences for ORF2

Secondary structure predictions of ORF2 indicated an α/β fold for the protein. Visualisation of the Kyte and Doolittle hydrophobicity plot reveal the presence of possible hydrophobic regions at the N-terminus as well as the C-terminus. The hydrophobicity values across these regions are however observed at levels lower than expected for transmembrane domains. Predictions from the SOSUI program furthermore also suggested that ORF2 is potentially a soluble protein.

From the HMM search of ORF2, the protein structure of 1mjh was consistently reported as a significant match across the FUGUE, SAM T02 and GenThreader programs. The probability score reported from all three programs were >99%. The consistently high probabilty scores served as a clear indicator of a homologous relationship between ORF1 and template (Yost et al., 2003). The 1mjh protein is classified as a member of class I of the Usp protein family of which ORF 2 is therefore potentially also a member.

UNIVERSITY of the
WESTERN CAPE

The superfamily of Usp proteins in *E.coli* are comprised by the familes of *uspA, uspC, uspD, uspE, uspF* and *uspG*. Further classification places *uspA, uspC* and *uspD* in class I, *uspF* and *uspG* in class II and the two usp domains of *uspE* are separated and placed into classes III and IV (Nachin et al., 2005).

To date the atomic structures of two *UspA* domains have been resolved, namely 1mjh from *Methanococcus jannaschii* and 1mjv from *H.influenzae*. The solved USPA domain structures (1mjh and 1jmv) show that they fold into an alpha/beta conformation featuring a slightly twisted planar surface of several parallel beta strands, with alpha helices adjacent on either side of this surface (Zarembinski et al., 1998; Sousa and McKay, 2001). This same conformational fold was observed from the homology model of ORF 2 (Figure 5.7). The 1mjh structure binds ATP through interactions with residues D13, V41, G127, G130, G140, S141, V142 and T143 (Zarembinski et al., 1998). Alignment of ORF 2 with 1mjh reveals the conservation of all these residues in ORF 2, beside T142 and S143. The ATP binding function of 1mjh could therefore not directly be inferred for ORF 2. The close structural relationship between ORF 2 and MJH does however suggest a potential related function for ORF 2.

The exact functions of the Usp proteins remain elusive but several experimental studies have been done to address this. Gustavsson et al., 2000, have deduced a role for the Usps in oxidative stress resistance, adhesion, motility and defence against DNA damage, based on the analysis of expression studies. In another study this specific group of proteins was demonstrated to be up-regulated in response to a variety of stresses including carbon, nitrogen, phosphate, sulphate and amino acid starvation (Kvint et al, 2003). It was also demonstrated that some stressors actually represses the expression of the *uspA* proteins, of which the important ones are extreme temperatures and tetracycline (Kvint et al, 2003).

UNIVERSITY of the
WESTERN CAPE

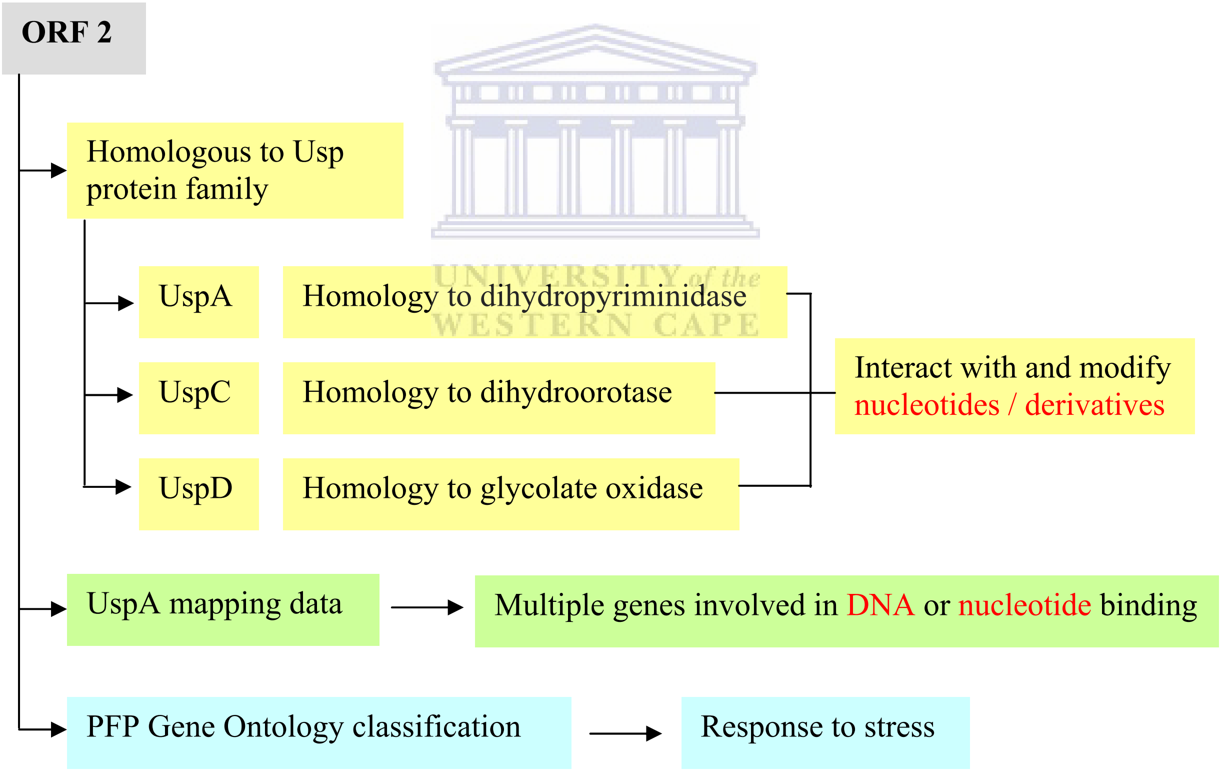
Increasing studies on expression patterns of UspA, UspC and UspD proteins have proposed distinct but related functions for these class I Usp proteins and that they co-operate in the same biochemical pathway (Gustavsson et al., 2002). When the exact function of proteins are not known, functional information can often be derived from closely related members within the same protein superfamily. This formed the argument for identifying and characterizing homologs of the UspA (ORF2), UspC and UspD proteins to improve understanding of this protein family.

The highest scoring homologs of ORF 2 (UspA), UspC and UspD, with known functions, were dihydropyriminidase, dihydroorotase and glycolate oxidase. The dihydropyriminidase and dihydroorotase homologs are both classified as amidohydrolases (Altenbuchner et al., 2001) that display a common TIM barrel fold (Holm and Sander, 1997; May et al., 1998). This TIM barrel fold reflects the nucleotide binding capacity of these proteins that is essential for their roles in pyrimidine biosynthesis (Lohkamp et al., 2006). Glycolate oxidase is involved in glyoxylate and dicarboxylate metabolism from glycolate and also has a propensity to bind the ucleotide derivative FAD (Lau and Armbrust, 2006). The alignments of the UspA, UspC and UspD queries however do not extend across the entire sequences of the homologs. As a result, the reported functions of these homologs are therefore not directly comparable to the corresponding ORF2 (UspA), UspC and UspD queries.

UNIVERSITY of the
WESTERN CAPE

Genomic context mapping of the USP loci reported the presence of flanking genes that catalyse diverse reactions. The nature of the co-location of these genes could therefore not be ascribed to a common reaction mechanism. There were however multiple protein encoding genes that interact with either nucleotides or nucleotide deriviatives which include an amidohydrolase family protein, nucleotidyltransferase domain protein and DNA topoisomerase I (Table 5.15). This suggested that the co-location of the genes at the Usp loci could be due to a common nucleotide/ nucleotide derivative binding or interacting relationship.

The integration of these HMM searches; structural modeling and PFP predictions are all suggestive of a potential role for ORF2 in interacting with or modifying nucleotides as a substrate (Figure 5.13).


ORF 2
Homologous to Usp protein family
UspA
Homology to dihydropyriminidase
UspC
Homology to dihydroorotase
UspD
Homology to glycolate oxidase
Interact with and modify nucleotides / derivatives
UspA mapping data
Multiple genes involved in DNA or nucleotide binding
PFP Gene Ontology classification
Response to stress

UNIVERSITY of the
WESTERN CAPE

Figure 5.13 Summary of ORF 2 function prediction

A potential relationship between ORFs 1 and 2 - testing the hypothesis

The functional inferences for ORFs 1 and 2 could only be attained at the level of a general classification as nucleotide binding/modifying. Essentially the hypothesis for related functions of ORF1 and ORF2, on the basis of their co-location on the library derived clone, holds true at this broad level of functional assignment. While the presence of known nucleotide binding motifs or sequence features could not be derived in this study, these proteins could likely represent entirely novel folds or binding domains. A similar observation is apparent from the sequences of several NADP dependent enzymes that do not include the NADP consensus sequence, like cytochrome P450 reductase (Vogel and Lumper, 1986). Proteins that have associated functions furthermore may share a related fold but contain functionally equivalent residues at non-homologous positions (Hasson et al., 1998). This invariably complicates the analysis for a relationship between structurally related proteins that shows extensive sequence divergence. This observation potentially explains the reported structural relationship of ORF1 specifically with multiple transcription factors while the sequences varied significantly.

More accurate predictions of the functional characteristics of the ORFs could potentially only be attained when: (i) improved homology identification strategies become available, (ii) closer related homologs with known functions for the ORFs become available.

UNIVERSITY of the WESTERN CAPE

Chapter 6

Characterisation of a Metagenomic Sequence Derived DNA Polymerase

6.1 Introduction

The polymerase chain reaction has led to rapid advances in molecular biology research due to its ability to amplify DNA. Central to this strategy is the DNA polymerase which catalyzes the primer dependant elongation of the DNA strands. Initial applications saw the implementation of mesophilic DNA polymerases however with limited efficiency (Saki et al., 1985). The mesophilic polymerases were plagued by the low nucleotide incorporation rates at high temp and also the constant requirement to replace the polymerase at high frequency.

UNIVERSITY of the WESTERN CAPE

These temperature limitations were addressed via implementation of thermostable polymerases such as BstI (Mead et al., 1991), Deep Vent (Cline et al.,1996), Pfu (Lundberg et al., 1991), Pwo (Frey and Suppman 1995), Taq (Jones and Foulkes, 1989), Tfl (Perler et al., 1996), Tth pol (Myers and Gelfand, 1991) and Vent pol (Perler et al., 1996). While this lead to significantly improved application of the PCR technologies, the lack of fidelity continues to remain a problem.

Current advancement of DNA polymerases are aimed at improving the fidelity of the PCR, development of strategies for more economical use of the enzymes and ensuring rapid multiplication from fewer strands Research on polymerases geared at addressing these aspects are resolved through discovery of novel polymerases, the mutation of the currently known polymerase active sites (Patel and Loeb, 1999) and by domain swapping between known polymerases (Villbrandt et al., 2000). Deriving structural and functional information on more polymerases would therefore: (i) lead to improved understanding of the sequence and based features that are involved in functional activity and (ii) allow for rational design of mutagenesis experiments to enhance functional activity.

6.2 Aims:

The characterization of a putative DNA polymerase gene, derived from a thermophilic, metagenomic sequence library. The objectives were: (i) to classify the putative polymerase, through homology searches and multiple sequence alignments with related polymerases, (ii) to demonstrate the conservation of the functional domains of the putative DNA polymerase, through the construction of homology models, and (iii) to demonstrate thermostable activity for the expressed, recombinant DNA polymerase.

UNIVERSITY of the
WESTERN CAPE

6.3 Results:

6.3.1 Sequencing of a metagenomic library derived DNA polymerase

The full length ORF sequence for the DNA polymerase was generated through bi-directional primer walking sequencing of clone 2.96 (Section 2.8.5). The resultant nucleotide sequence for clone 2.96 corresponded to an ORF of 2796 bp. Translation of the ORF represented a protein of 932 amino acids, denoted TC11pol.

6.3.2 Identification of homologous DNA polymerases

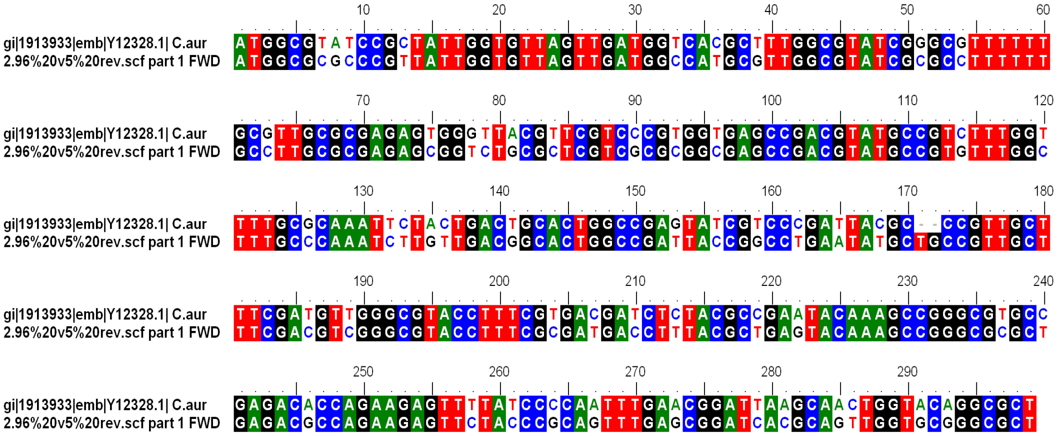
The full length TC11pol amino acid sequence was used to retrieve sequence homologues using the BLASTp programme (Section 2.11.1) and the highest scoring matches corresponded to the polymerase protein family (Table 6.1). The top scoring matches were against the polymerases of *C. aurantiacus* and *C. aggregans* (Table 6.1).

The nucleotide sequence of TC11pol was also aligned (Section 2.11.2) against the corresponding sequences of *C. aurantiacus* and *C. aggregans.* Similarity scores of 84% and 72% were reported for these organisms' polymerases respectively. The alignment of the N-terminal TC11pol nucleotide sequence, generated with the V5 primer, against the corresponding *C. aurantiacus* nucleotide sequence is shown in Figure 6.1. The result indicates a variation in regard to the codon usage between the polymerases of TC11pol and that of *C. aurantiacus.*

UNIVERSITY of the
WESTERN CAPE

Table 6.1 List of highest scoring polymerase matches.

Protein accession	*Organism*	*E-value*	*Identity*
ZP_00768405.1	Chloroflexus aurantiacus J-10-fl	9e-174	95%
ZP_01515805.1	Chloroflexus aggregans DSM 9485	4e-149	86%
YP_001277723.1	Roseiflexus sp. RS-1	1e-104	63%
ZP_01529401.1	Roseiflexus castenholzii DSM 13941	2e-103	62%



```
                                      10        20        30        40        50        60
gi|1913933|emb|Y12328.1| C.aur        ATGGCGTATCCGCTATTGGTGTTAGTTGATGGTCACGCTTTGGCGTATCGGGCGTTTTTT
2.96%20v5%20rev.scf part 1 FWD        ATGGCGCGCCCGTTATTGGTGTTAGTTGATGGCCATGCGTTGGCGTATCGCGCCTTTTTT

                                      70        80        90       100       110       120
gi|1913933|emb|Y12328.1| C.aur        GCGTTGCGCGAGAGTGGGTTACGTTCGTCCCGTGGTGAGCCGACGTATGCCGTCTTTGGT
2.96%20v5%20rev.scf part 1 FWD        GCCTTGCGCGAGAGCGGTCTGCGCTCGTCGCGCGGCGAGCCGACGTATGCCGTGTTTGGC

                                     130       140       150       160       170       180
gi|1913933|emb|Y12328.1| C.aur        TTTGCGCAAATTCTACTGACTGCACTGGCCGAGTATCGTCCCGATTACGC--CCGTTGCT
2.96%20v5%20rev.scf part 1 FWD        TTTGCCCAAATCTTGTTGACGGCACTGGCCGATTACCGGCCTGAATATGCTGCCGTTGCT

                                     190       200       210       220       230       240
gi|1913933|emb|Y12328.1| C.aur        TTCGATGTTGGGCGTACCTTTCGTGACGATCTCTACGCCGAATACAAAGCCGGGCGTGCC
2.96%20v5%20rev.scf part 1 FWD        TTCGACGTCGGGCGTACCTTTCGCGATGACCTTTACGCTGAGTACAAAGCCGGGCGCGCT

                                     250       260       270       280       290
gi|1913933|emb|Y12328.1| C.aur        GAGACACCAGAAGAGTTTTATCCCCAATTTGAACGGATTAAGCAACTGGTACAGGCGCT
2.96%20v5%20rev.scf part 1 FWD        GAGACGCCAGAAGAGTTCTACCCGCAGTTTGAGCGGATCACGCAGTTGGTGCGGGCGCT
```

Figure 6.1 Alignment of N-terminal nucleotide sequences of polymerases
C. aurantiacus is denoted (gi 1913933) and TC11pol is denoted (1FWD). Identical residues are colour coded while non-identical residues are uncoloured.

UNIVERSITY of the
WESTERN CAPE

6.3.3 Construction of TC11pol primers for expression cloning

The full length TC11pol ORF sequence was used to generate primer sequences which would enable cloning of the gene for expression analysis. The N-terminal PolF primer was engineered to contain a HindIII restriction site and the C-terminal PolR primer a XbaI restriction site. TC11pol amplicons were generated under optimized PCR reaction conditions (Section 2.8.2.1). The polymerase gene was cloned into the pUC18 and pGEX 6P-2 vectors to yield pUC-TC11pol (Section 6.3.3.1) and pGEXTC11pol (Section 6.3.3.2) expression constructs respectively.

6.3.3.1 pUC-TC11pol vector construction

TC11pol amplicons were cloned into the pGEM T-easy TA cloning vector (Section 2.7.5.4) and restriction digested with HindIII and XbaI (Section 2.8.4). The resultant polymerase fragment was subsequently ligated into HindIII and

XbaI restriction digested pUC18 vector, yielding the pUC-TC11pol vector (Section 2.7.4). Construction of the polpUC18 expression vector resulted in a fusion of the partial polymerase ORF with the N-terminal portion of the β-galctosidase gene (Figure 6.2). The strategy ensured directional, in-frame cloning of the polymerase to enable efficient expression.

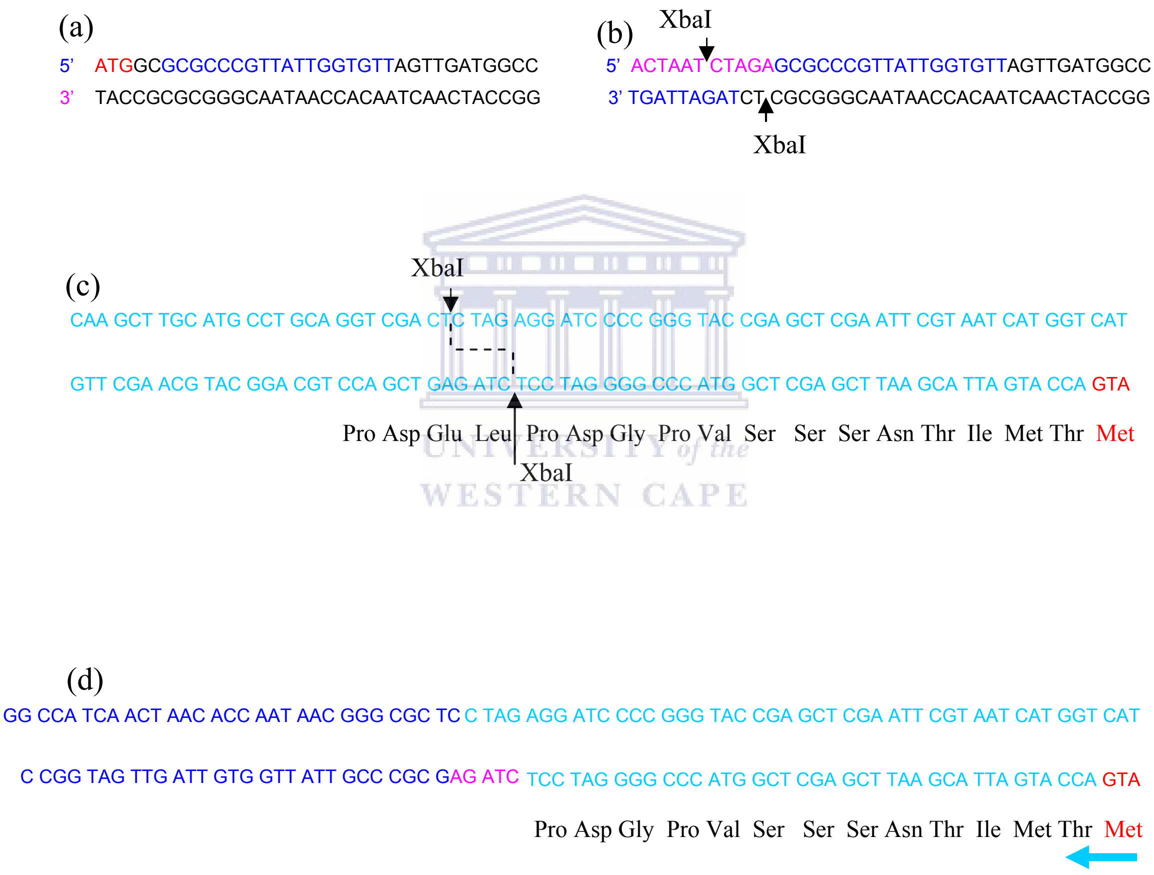
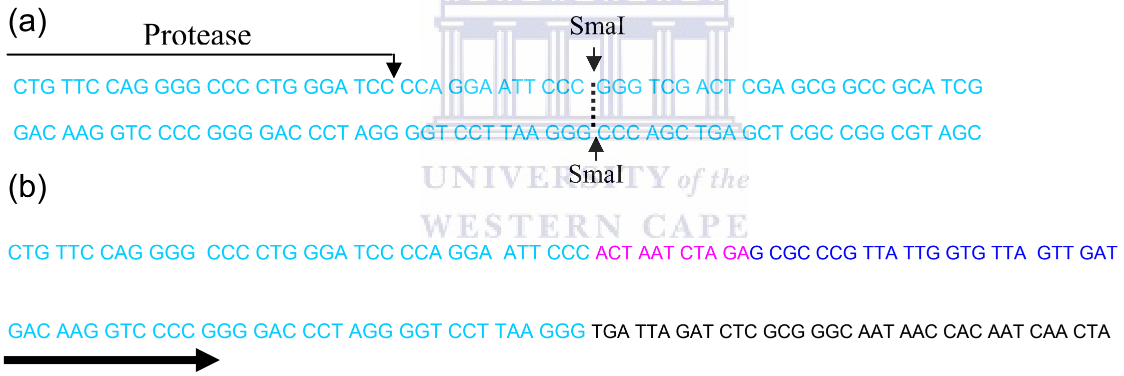

(a)
5' ATGGCGCGCCCGTTATTGGTGTTAGTTGATGGCC
3' TACCGCGCGGGCAATAACCACAATCAACTACCGG
(b)
XbaI
5' ACTAATCTAGAGCGCCCGTTATTGGTGTTAGTTGATGGCC
3' TGATTAGATCTCGCGGGCAATAACCACAATCAACTACCGG
XbaI
(c)
XbaI
CAA GCT TGC ATG CCT GCA GGT CGA CTC TAG AGG ATC CCC GGG TAC CGA GCT CGA ATT CGT AAT CAT GGT CAT
GTT CGA ACG TAC GGA CGT CCA GCT GAG ATC TCC TAG GGG CCC ATG GCT CGA GCT TAA GCA TTA GTA CCA GTA
Pro Asp Glu Leu Pro Asp Gly Pro Val Ser Ser Ser Asn Thr Ile Met Thr Met
XbaI
(d)
GG CCA TCA ACT AAC ACC AAT AAC GGG CGC TC C TAG AGG ATC CCC GGG TAC CGA GCT CGA ATT CGT AAT CAT GGT CAT
C CGG TAG TTG ATT GTG GTT ATT GCC CGC GAG ATC TCC TAG GGG CCC ATG GCT CGA GCT TAA GCA TTA GTA CCA GTA
Pro Asp Gly Pro Val Ser Ser Ser Asn Thr Ile Met Thr Met
UNIVERSITY of the
WESTERN CAPE

Figure 6.2 pUC-TC11pol construct
(a) The native 5' terminus of the TC11pol showing the start codon (red) and the region that constitutes the polF primer without the engineered Xba site(blue). (b) The 5' amplicon sequence for TC11pol showing the complete polF primer (underlined), the engineered Xba site (pink), the remainder of the primer is blue. The introduced XbaI cut sites are indicated by arrows. (c) The multiple cloning site of pUC18 demonstrating the XbaI cut sites. (d) The ligation of the 5' XbaI region of the HindIII and XbaI digested TC11pol, to yield pUC-TC11pol. The vector ATG codon is coloured red; the pUC18 vector sequence included in the expression fragment of TC11pol is coloured light-blue; the remaining engineered XbaI site is coloured pink and the 5' region of the TC11pol is coloured blue.

6.3.3.2 pGEX-TC11pol vector construction

Blunt-ended amplicons were generated for TC11pol using the polF and polR primers and Deep Vent DNA polymerase (Section 2.8.2.1). The polymerase amplicons were blunt-end ligated into the SmaI site of the pGEX 6P-2 vector (Section 2.7.4). Polymerase inserts of the correct orientation were identified by PCR amplification with the pGEX2 6P-2 vector specific 5' -end primer (pGEX 5') and the polymerase specific polR primer (Section 2.8.2.1). This enabled recovery of TC11pol in a directional, in-frame manner to enable efficient expression of the gene (Figure 6.3)


(a)
Protease
SmaI
CTG TTC CAG GGG CCC CTG GGA TCC CCA GGA ATT CCC GGG TCG ACT CGA GCG GCC GCA TCG
GAC AAG GTC CCC GGG GAC CCT AGG GGT CCT TAA GGG CCC AGC TGA GCT CGC CGG CGT AGC
SmaI
(b)
CTG TTC CAG GGG CCC CTG GGA TCC CCA GGA ATT CCC ACT AAT CTA GAG CGC CCG TTA TTG GTG TTA GTT GAT
GAC AAG GTC CCC GGG GAC CCT AGG GGT CCT TAA GGG TGA TTA GAT CTC GCG GGC AAT AAC CAC AAT CAA CTA

UNIVERSITY of the WESTERN CAPE

Figure 6.3 pGEX-TC11pol construct
(a) SmaI cloning site of the native pGEX 6P-2 vector indicated with arrows. (b) Orientation of the 5' primer region of the TC11pol (underlined) with the engineered site (pink) and the polymerase binding region (blue) within the pGEX 6P-2 vector (light-blue).

6.3.4 Expression conditions for the polymerase gene

The polpGEX 6P-2 and polpUC18 constructs were transformed into the *E. coli* BL21 (DE3) expression strain (Section 2.7.5.2). Expression was assessed at varied temperatures (25 - 37°C), induction times (3 hours – overnight) and IPTG concentrations (0.05 - 2.0 mM) as outlined in Section 2.10.1. Positive expression of the polymerase protein was investigated through sodium dodecyl sulphate-polyacrylamide gel electrophoresis (Section 2.9.2) of soluble and insoluble fractions of the host cell lysate. The presence of a protein band corresponding to the expected 95 kDa for TC11pol could not be positively confirmed for the investigated expression conditions.

6.3.5 Purification protocol

Proteins that exhibit thermostability above 37 ^{o}C can be readily separated and purified from the *E.coli* native proteins through a heat denaturation step at 60 ^{o}C for 1 hour (Pluthero, 1993). Our library derived TC11pol polymerase was presumed to exhibit thermostability because of the source sample derived from a high temperature environment as well as the sequence similarity with known thermostable polymerases. The SDS-PAGE (Section 2.9.2) analysis of the heat denatured cell lysate (Section 2.10.2) could however not confirm the presence of a protein band in the expected 95KDa range of TC11pol.

UNIVERSITY of the
WESTERN CAPE

6.3.6 Activity assay

In the absence of any detectable level of expressed protein by SDS-PAGE, an activity-based strategy was implemented. DNA polymerase activity was assayed by implementation of the heat denatured cell lysate in the polymerase chain reaction (Section 2.10.3). The metagenomic derived clone 2.142 was used as template for amplification of the Usp ORF, with primers Uspf and Uspr.

The presence of amplicons was assessed by agarose gel electrophoresis (Section 2.9.1). The lack of any observeable amplicons was indicative that the TC11pol could not be successfully expressed under the implemented conditions.

6.3.7 Homology modeling of the TC11pol and Chloroflexus aurantiacus polymerases

The lack of observable expression or activity of the TC11pol polymerase impeded further experimental analysis of the functional properties of the gene. The availability of crystal structure data of related polymerases did however allow for construction of structural models for both the TC11pol and highly similar Chloroflexus aurantiacus polymerases via homology modeling. The experimental layout for the homology modeling approach is outlined in Figure 6.4 and the resultant outputs are presented in the relevant sections.

UNIVERSITY of the
WESTERN CAPE

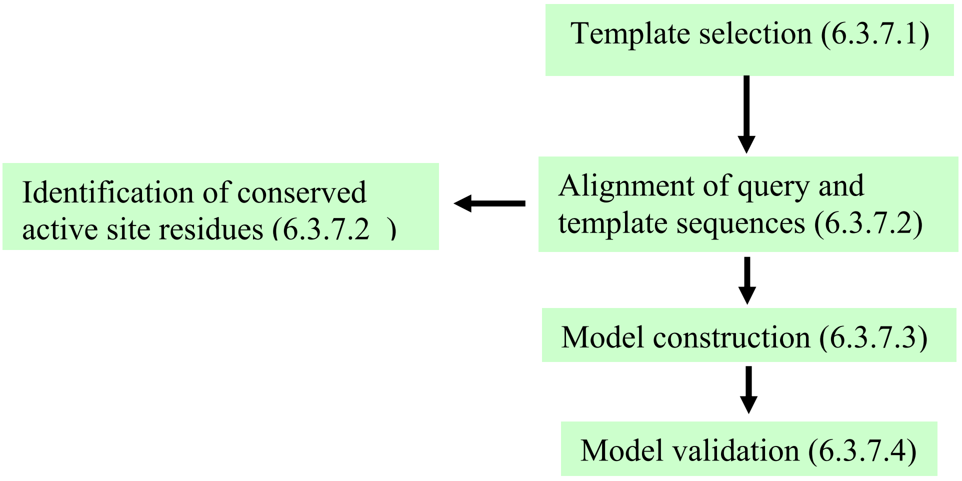

Template selection (6.3.7.1)
Identification of conserved active site residues (6.3.7.2)
Alignment of query and template sequences (6.3.7.2)
Model construction (6.3.7.3)
Model validation (6.3.7.4)

Figure 6.4 Outline of the structural modeling strategy.

6.3.7.1 Template selection

The TC11pol 3’-5’ exonuclease, 5’-3’ exonuclease- and polymerase domains were compared against the Swissprot database using the BLASTp tool in order to recover the closest related structural homologues for the respective domains (Section 2.11.1). The closest structural domains were defined as the highest scoring BLAST matches with an experimentally derived structure available in the PDB database (www.pdb.org). The TC11pol 5’-3’ exonuclease- and polymerase domains displayed the highest degree of similarity with the corresponding domains of *T. aquaticus* while the 3’-5’ exonuclease domain matched that of *E. coli* (Table 6.2). These were in accordance with the previously reported results of the 3’-5’ exonuclease, 5’-3’ exonuclease- and polymerase domains of *C. aurantiacus* (Tvermeyer et al., 1998) which has 96% identity to TC11pol. The respective domains of C. *aggregans* were also evaluated against the same templates based on the high sequence conservation (86%) with TC11pol (Table 6.2).

UNIVERSITY of the
WESTERN CAPE

Table 6.2 List of highest scoring domain matches

Template domain	*Query*	*% Identity with template*
5’-3’ exonuclease (T.aquaticus)	*C. aurantiacus*	*42.7*
	C. aggregans	*42.5*
	TC11pol	*42.4*
3’-5’ exonuclease (E.coli)	*C. aurantiacus*	*43.1*
	C. aggregans	*41.75*
	TC11pol	*42.5*
polymerase (T.aquaticus)	*C. aurantiacus*	*58*
	C. aggregans	*57.4*
	TC11pol	*58*

6.3.7.2 Multiple sequence alignments and identification of conserved active site residues

The sequence features required for functional activity of the 3’-5’ exonuclease, 5’-3’ exonuclease- and polymerase domains of the DNA polymerases are outlined in Table 6.3. The multiple sequence alignment of *C. aurantiacus, C. aggregans and* TC11pol against the respective *T. aquaticus* and *E. coli* domains are shown in Figures 6.5 to 6.7 (Section 2.11.2). This multiple sequence alignment was investigated to observe whether all active site residues of the 3’-5’ exonuclease, 5’-3’ exonuclease- and polymerase domains were conserved.

The sequence alignments indicated that all the defined sequence features required for 3’-5’ exonuclease, 5’-3’ exonuclease- and polymerase activities were present in the polymerases of TC11pol, *C .aurantiacus* and *C. aggregans* (Figures 6.5 to 6.7).

UNIVERSITY of the
WESTERN CAPE

Table 6.3 Conserved motif features

Domain	*Motifs*	*Specific residues/motif descriptors*
3’-5’ exonuclease	Exo I-III	D355, D424, D501 and E357
5’-3’ exonuclease	A-F	14 invariant amino acids 9/14 amino acids are acidic D13, Y77, G185, G195 mutations cause reduced activity
polymerase	1, 2a, 2b, 3-5	motifs 3 and 5 are conserved in all classes of polymerases motif 4 is conserved in all DNA-dependent polymerases

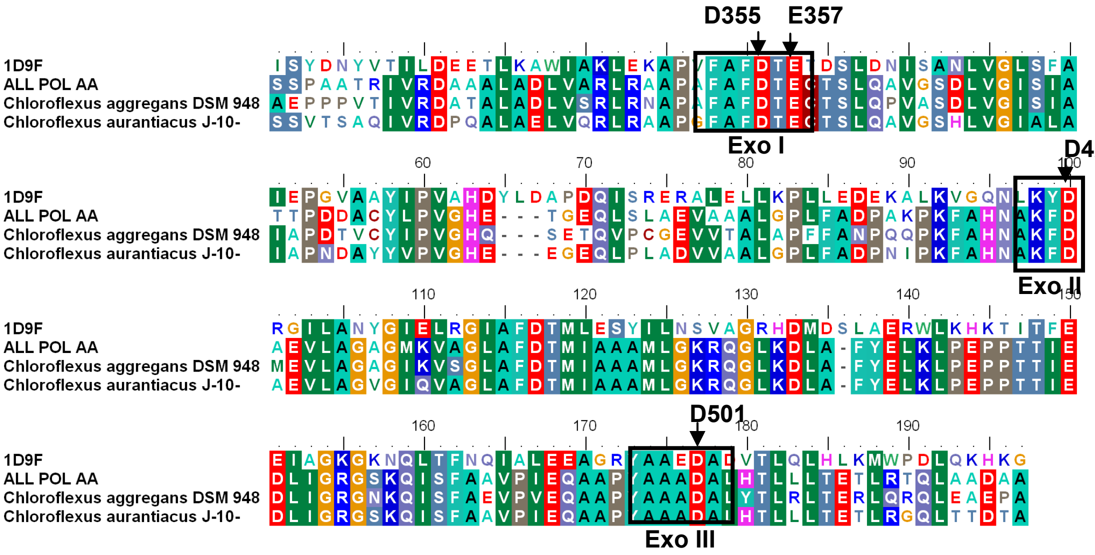

D355
E357
D424
D501
Exo I
Exo II
Exo III
1D9F
ALL POL AA
Chloroflexus aggregans DSM 948
Chloroflexus aurantiacus J-10-

Figure 6.5 Multiple sequence alignments of 3’-5’ exonuclease domains.

Sequences are denoted*:* *E. coli* (1D8Y), *C. aurantiacus* (Cau), *C. aggregans* (Cag) and TC11pol (All POL AA). Motifs ExoI-III are boxed and labeled accordingly. Residues D355, D424, D501 and E357are indicated with arrows.

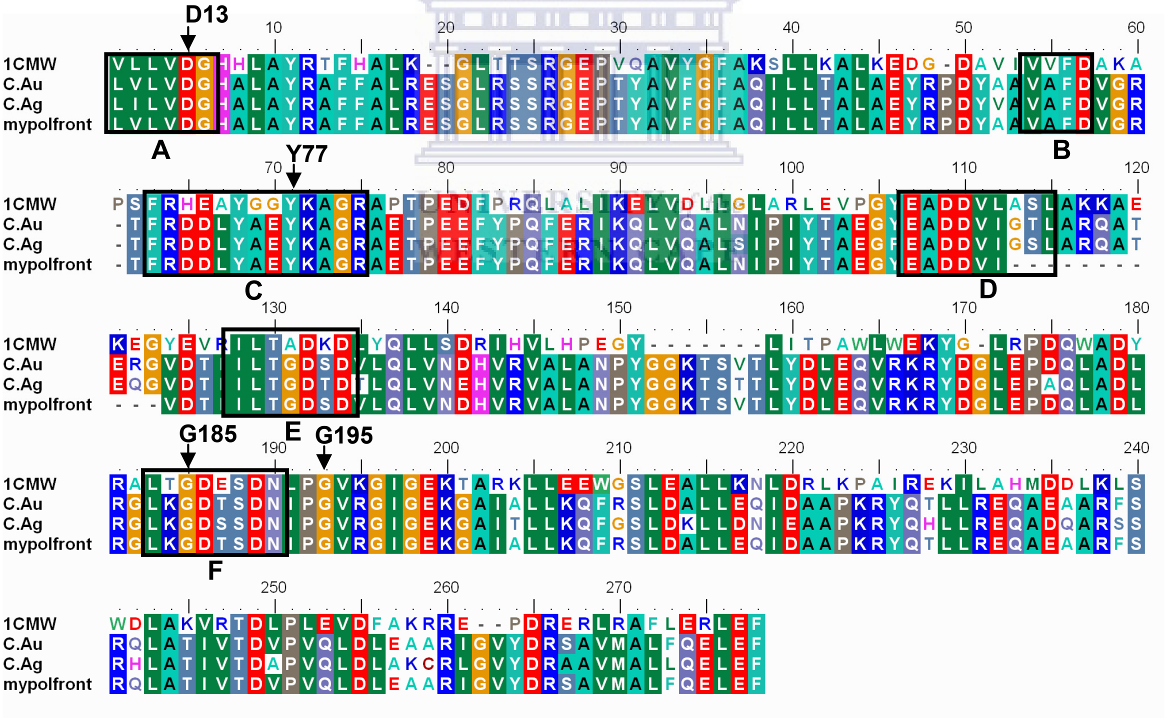


Figure 6.6 Multiple sequence alignments of the 5’-3’ exonuclease domains.

Sequences are denoted: *T .aquaticus* (1CMW), *C. aurantiacus* (Cau), *C. aggregans* (Cag) and TC11pol (mypolfront). Motifs A-F are boxed and labeled accordingly. Residues D13, Y77, G185 and G195 are indicated with arrows.

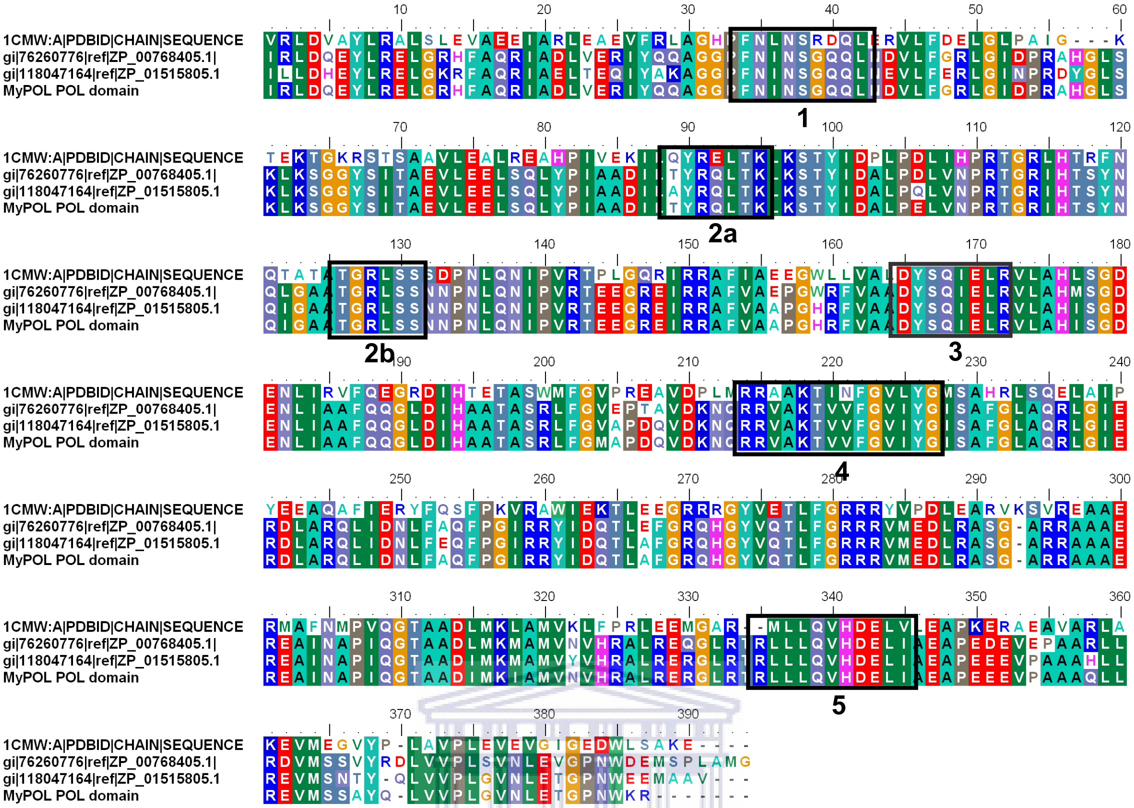


Figure 6.7 Multiple sequence alignments of the polymerase domains.
Sequences are denoted: *T. aquaticus* (1CMW), *C. aurantiacus* (gi 7626077), *C. aggregans* (gi 118047164) and TC11pol (MyPOL). Motifs 1-5 are boxed and labeled accordingly.

UNIVERSITY of the
WESTERN CAPE

6.3.7.3 Construction of structural models for TC11pol and C. *aurantiacus* polymerase

Models were constructed for the respective 5'-3' exonuclease-, 3'-5' exonuclease- and polymerase domains of TC11pol and C. *aurantiacus* polymerase. All models were constructed with the SWISS-Model program on the Expasy server and visualisation of the topology and structural features of the domains were done in PyMOL.

The 3'-5' exonuclease domain

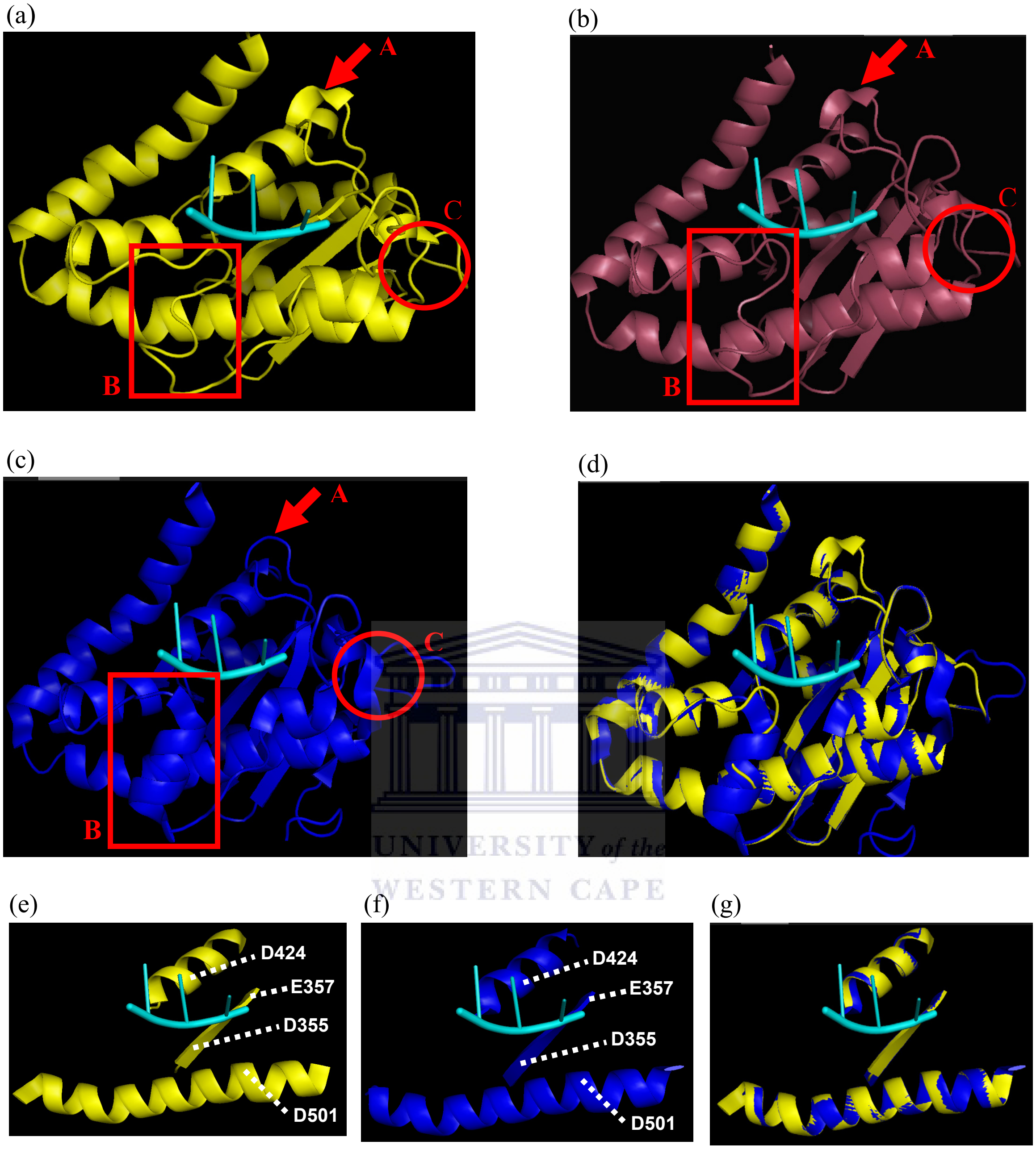
The models of the 3'-5' exonuclease domains of TC11pol and C. *aurantiacus* polymerase display an overall comaparable topology with the 3'-5' exonuclease domain of *E. coli* (Figure 6.8). The observed variations refer to regions that were modeled as helices but are observed as loops or β-sheets in the template structure and vice versa. The structural variations that were observed are outlined in Table 6.4 and correspondingly indicated in Figure 6.8.

Table 6.4 Structural variations in TC11Pol models

Variation	*TC11pol*	*C. aurantiacus*	*E.coli*
A	Helix	Helix	Loop
B	Loop	Loop	Helix
C	Loop	Loop	Sheet

UNIVERSITY of the
WESTERN CAPE

Investigation of the structural arrangement around the active site residues D355, D424, D501 and E357 reveal that these regions are conserved for both the modeled regions of TC11pol and the experimentally derived *E. coli* structure (Figure 6.8 e-g). These regions are observed as two a helices and beta sheet in both modeled polymerases and the *E. coli* template. The location of the active sites residues are also conserved within these folds.



UNIVERSITY of the
WESTERN CAPE

Figure 6.8. Structural models of the 3'-5' exonuclease domains.
(a) TC11pol; (b) *C. aurantiacus*; (c) the crystal structure of the *E. coli* 3'-5' exonuclease domain and (d) the superimposed structures of the TC11pol and E.coli 3'-5' exonuclease domains. The structural variations observed between the structures are denoted A, B and C. The structural components that harbor the 3'-5' exonuclease active site residues (D355, D424, D501 and E357) are shown for (e) TC11pol and (f) *E. coli*. The 3'-5' exonuclease active site components for TC11pol and *E. coli* are superimposed in (g).

The 5’-3’ exonuclease and polymerase domains

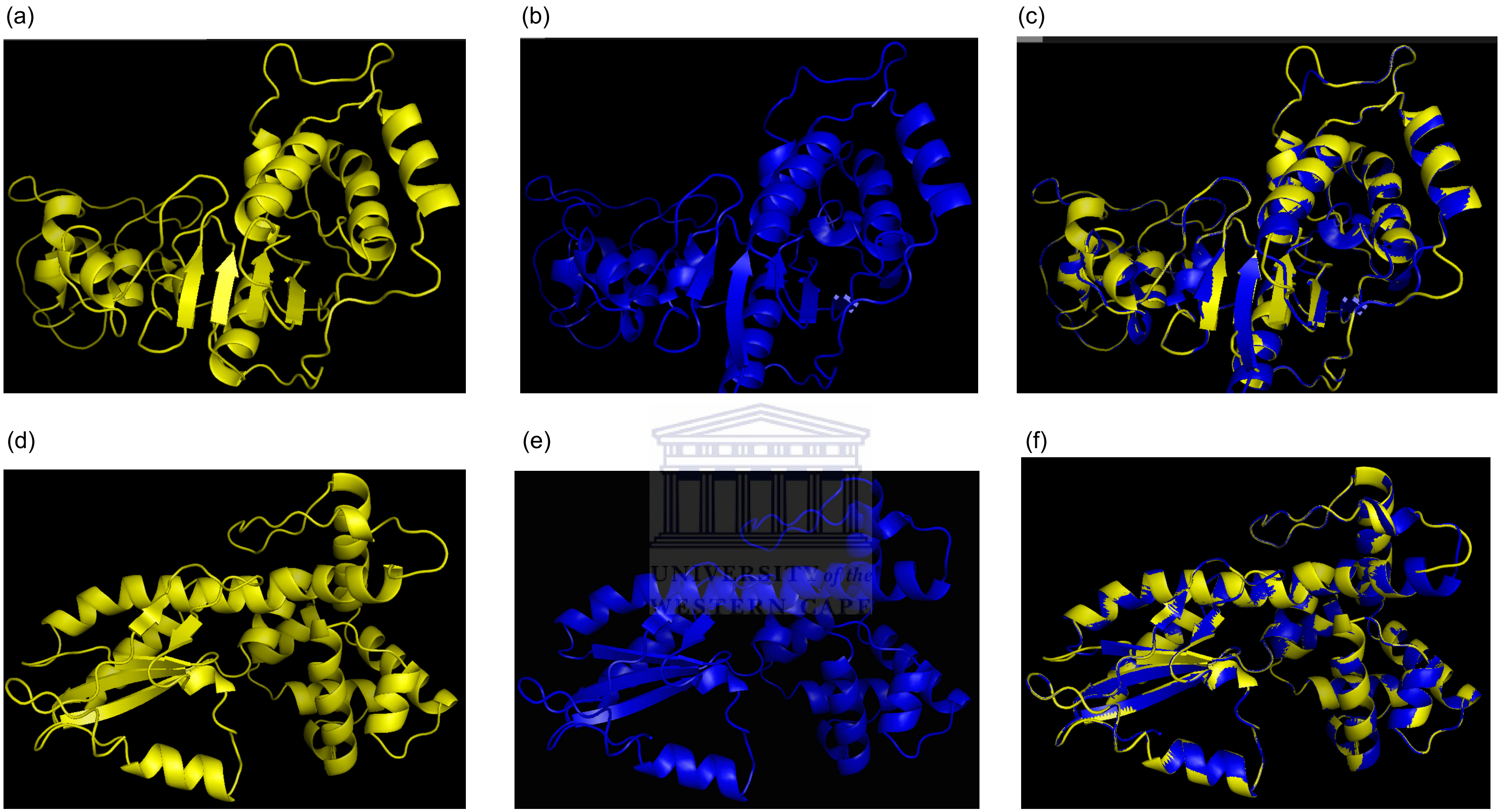
The topologies of the 5’-3’ exonuclease- and polymerase domains of both the TC11pol and C. *aurantiacus* models were comparable to the corresponding domains of *T. aquaticus* (Figure 6.9). All α-helix, β-sheet and loop assigned regions were therefore observed in similar arrangements across the above-mentioned structures. The models for these structures did display variations in the lengths of assigned β-sheets and α-helices but no gross variations such as the complete removal or addition of structural features were observed.

The structural conservation of all active site residues was also investigated for the 5’-3’ exonuclease- and polymerase domains of TC11pol-, *C. aurantiacus*- and *T. aquaticus* polymerase. Where residues of the active sites were observed in similar structural folds in TC11pol, *C. aurantiacus* and *T. aquaticus* they were considered structurally conserved, otherwise they were considered non-conserved (Table 6.5). The analysis demonstrated the complete conservation of the active site residues and motifs in related structural folds for TC11pol, *C. aurantiacus* and *T. aquaticus*.

Table 6.5 Conservation of active site residues

Domain	*Catalytic site residues/motifs*	*Polymerase query*	*Structural features*
5’-3’ exo	D355,D424,D501and E357	C. *aurantiacus*	All conserved
		TC11pol	All conserved
polymerase	Motifs 3, 4 and 5	C. *aurantiacus*	All conserved
		TC11pol	All conserved

UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 6.9 Topology the 5’-3’ exonuclease- and polymerase domains
(a) the modeled 5’-3’ exonuclease domain of TC11pol, (b) the experimentally derived structure of the 5’-3’ exonuclease domain of *T. aquaticus* and in (c) these domains are superimposed. The cartoon models of: (d) the modeled polymerase domain of TC11pol, (e) the experimentally derived structure of the polymerase domain of *T. aquaticus* and in (e) these domains are superimposed.

6.3.7.4 Model validation

The quality of the models was assessed on the outputs from ramachandran plots (RAMPAGE) as well as the fitness of the residues within the modeled environment (VERIFY3D). The RAMPAGE output (Table 6.6) demonstrates that all the modeled structures display good overall stereochemistry. In all the models >90% of residues were identified within favoured and allowed regions. The Verify3D graph for the 3'-5' exonuclease domain model of TC11pol is shown in Figure 6.10. The graph shows that all the residues throughout the structure are modeled within favorable environments (all positive scores). When the Verify3D graph displays negative values across a stretch of residues it generally implies that the model requires additional investigation for incorrect aligned and modeled regions. The Verify 3D graphs for all the modeled structures reflected positive scores which are suggestive of good model quality.

UNIVERSITY of the
WESTERN CAPE

Table 6.6 RAMPAGE evaluation of homology models

Region	*Model*	*RAMPAGE evaluation (% of residues)*		
		Favoured region	*Allowed region*	*outlier region*
3'-5' Domain	Lib Pol	92.7	4.9	2.4
	C. auriantiacus	91.5	6.8	1.7
5'-3' Domain	Lib Pol	74.5	17.9	7.7
	C. auriantiacus	72.1	18.8	9.9
POL Domain	Lib Pol	92.1	7.6	0.4
	C. auriantiacus	91.3	8.2	0.5

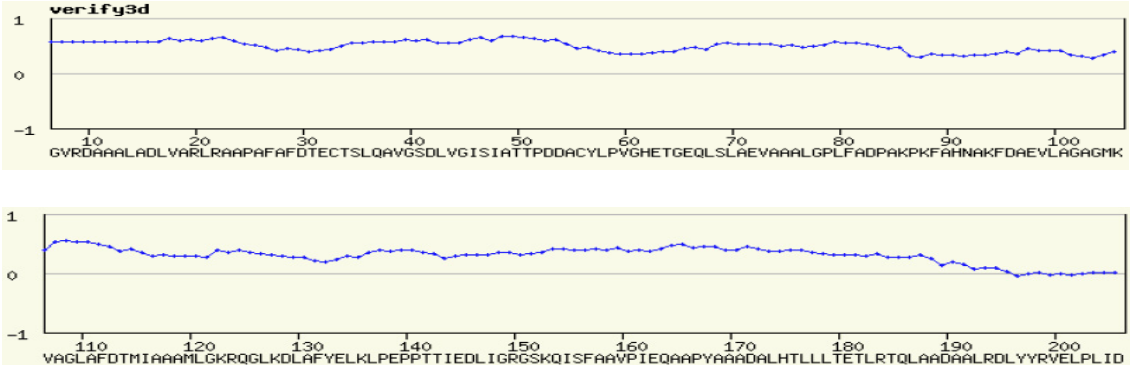


Figure 6.10 Verify3D evaluation of the 3’ domain model of TC11pol

6.3.8 Deriving a mutant polymerase from the structure models

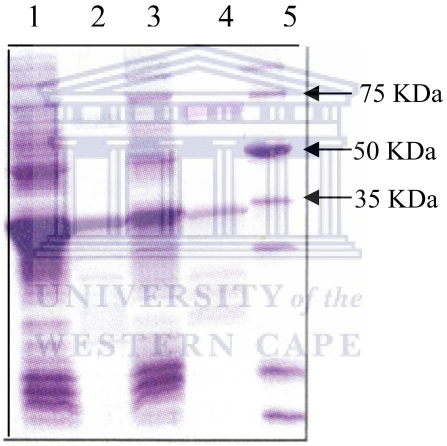
The lack of expression demonstrated for the entire ORF of the TC11pol prompted investigation of the expression of a shortened Klenow-like fragment for TC11pol. The modeling data identified region 270-930 for TC11 pol to be representative of the domains for 3’-5’ exonuclease and polymerase domain, which is subsequently refered to as 5’ exo$^-$ TC11pol.

6.3.9 Generation of the 5’ exo$^-$ TC11pol construct

Blunt ended PCR products for 5’ exo$^-$ TC11pol was generated with Deep Vent Polymerase and primer set PIF and 2.96R (Section 2.8.2.1). The amplicon was digested with HindIII which cleaves the engineered HindIII cut site of the 2.96R primer region (Section 2.8.4). The pUC18 vector was cleaved with HindIII and Ecl136II which enabled directional in-frame ligation of 5’ exo$^-$ TC11pol (Section 2.8.4). The construct was sequenced from the 5’ (M13F) and 3’ (M13R) ends to ensure that the insert was cloned in the correct orientation (Section 2.8.5).

6.3.10 Expression and purification of the 5'-3' exo$^-$ deriviative of TC11pol

Expression and purification was attained as previously described in Sections 2.10.1 and 2.10.2 respectively. A protein product of 74 KDa was predicted for 5' exo$^-$ TC11pol. SDS-PAGE following heat treatment at 60^{o}C however resulted in no observable protein bands. The heat denaturation temperature was subsequently lowered to 55^{o}C and SDS-PAGE analysis revealed the presence of multiple protein bands, which included a band of the expected 74 KDa size in the cellular lysate (Figure 6.11).

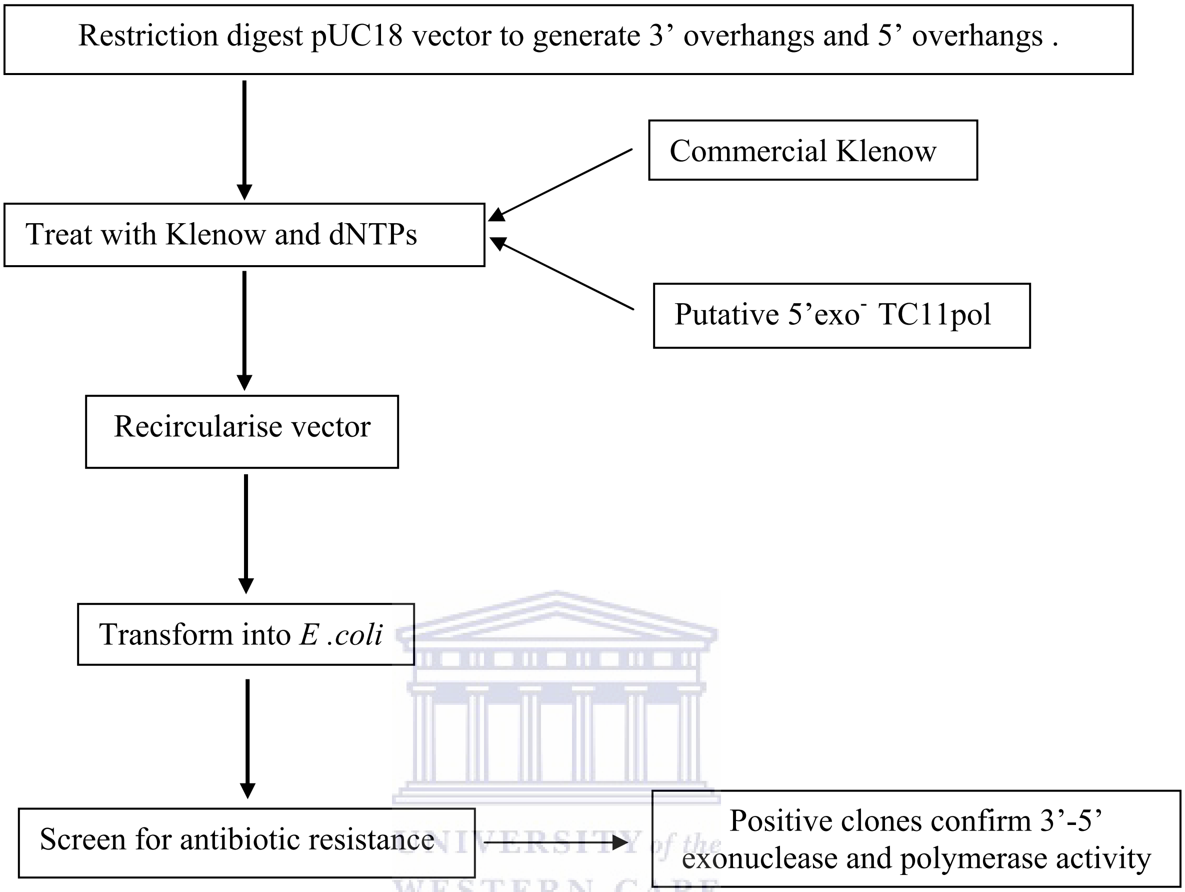


UNIVERSITY of the
WESTERN CAPE

Figure 6.11 SDS PAGE gel of heat denatured TC11pol expression extract
Lane: 1, Uninduced E.coli BL21(DE3) containing 5'-3' TC11pol deletion mutant; 2, Heat treated uninduced E.coli BL21(DE3) containing 5'-3' TC11pol deletion mutant; 3, Induced E.coli BL21(DE3) containing 5'-3' TC11pol deletion mutant; 4, Heat treated induced E.coli BL21(DE3) containing 5'-3' TC11pol deletion mutant.

6.3.11 Deriving evidence for potential 3'-5' exonuclease and polymerase activity of 5' exo$^-$ TC11pol

The heat denatured cellular extract of the *E. coli* BL21(DE3) host strain, proposed to contain 5' exo$^-$ TC11pol, was subsequently investigated for 3'-5' exonuclease and polymerase activity. The strategy for the assay is outlined in Figure 6.12 and described in Section 2.10.4.


Restriction digest pUC18 vector to generate 3’ overhangs and 5’ overhangs .
Commercial Klenow
Treat with Klenow and dNTPs
Putative 5’exo- TC11pol
Recircularise vector
Transform into *E .coli*
Screen for antibiotic resistance
Positive clones confirm 3’-5’ exonuclease and polymerase activity

UNIVERSITY of the
WESTERN CAPE

Figure 6.12 Strategy for 3’ exonuclease and polymerase activity assay.

The strategy was based on the demonstrated 3’-5’ exonuclease and polymerase activity of the *E.coli* Klenow fragment. The same activities were predicted for 5’ exo$^-$ TC11pol based on the structural similarity between 5’ exo$^-$ TC11pol and the family of type I polymerases. The cellular fractions and controls that were investigated for 3’-5’ exonuclease and polymerase activity are listed in Table 6.7. All fractions that were subjected to heat treatment were incubated at 55^{o}C for 1 hour. Activity was assayed at 37, 45 and 50 ^{o}C. The results for the experiments, performed in triplicate, are reported in Table 6.8.

Table 6.7 Fractions investigated

Fraction investigated	Description
Untreated Klenow fragment (Commercial)	-
Heat treated Klenow fragment (Commercial)	-
Heat treated E.coli (BL21), uninduced, (-5'exo$^-$ TC11pol)	Uniduced host cell that lacks the 5'exo$^-$ TC11pol construct
Heat treated E.coli (BL21), induced, (-5'exo$^-$ TC11pol)	Induced host cell that lacks the 5'exo$^-$ TC11pol construct
Heat treated E.coli (BL21), uninduced, (+5'exo$^-$ TC11pol)	Uninduced host cell that contains the 5'exo$^-$ TC11pol construct
Heat treated E.coli (BL21), induced, (+5'exo$^-$ TC11pol)	Induced host cell that contains the 5'exo$^-$ TC11pol construct

Table 6.8. 3'-5' exonuclease and polymerase activity assay

UNIVERSITY of the
WESTERN CAPE

Fraction assayed	Temp Of Assay	Colonies counted per experiment		
		Ex 1	Ex 2	Ex 3
Untreated Klenow fragment (Commercial)	37	>500	>500	>500
Heat treated Klenow fragment (Commercial)	37	8	7	3
Heat treated E.coli (BL21), uninduced, (-5'exo$^-$ TC11pol)	37	4	6	4
Heat treated E.coli (BL21), induced, (-5'exo$^-$ TC11pol)	37	4	3	8
Heat treated E.coli (BL21), uninduced, (+5'exo$^-$ TC11pol)	37	5	5	3
Heat treated E.coli (BL21), induced, (+5'exo$^-$ TC11pol)	37	63	52	33
	45	57	31	26
	50	3	3	4

The large number of positive transformants for the untreated Klenow fragment were indicative of successful blunt-end treatment via 3'-5' exonuclease and polymerase activity. Heat treatment of the Klenow fragment resulted in loss of activity as indicated by the low number of positive transformants. A lack of activity was also observed for (i) heat treated extract, from induced and uninduced *E. coli* (BL21) without the 5’exo^{-} pol construct, and (ii) the heat treated extract, from induced *E. coli* (BL21) without the 5’exo^{-} pol construct. The heat-treated cellular extract, from induced *E. coli* (BL21) with the 5’exo^{-} pol construct extract, yielded a significant number of colonies at reaction temperatures of 37 ^{o}C and 45 ^{o}C. Activity was, however, significantly reduced at a reaction temperature of 50 ^{o}C.

6.4 Discussion

Due to its applications in the Polymerase chain reaction the DNA polymerase type I (POL I) family represents the most studied of the five families that constitute the polymerases. The Pol I family contains DNA polymerases that exhibit 5’-3’ exonuclease, 3’-5’ exonuclease and polymerase activity as well as those polymerases that lack the central 3’-5’ exonuclease domain. Sequence comparison of the metagenomic library derived DNA polymerase (TC11pol) reveal a striking similarity of this protein sequence with the polymerases of *Chloroflexus aurantiacus* (95%) and *Chloroflexus aggregans* (86%). Based on the amino acid sequence alignments, the TC11pol was putatively classified as a member of the DNA polymerase type I family that exhibit 5’-3’ exonuclease, 3’-5’ exonuclease and polymerase activity (Steitz, 1999).

UNIVERSITY of the
WESTERN CAPE

TC11pol polymerase displays nucleotide codon preferences that are distinct from *C. aurantiacus* and *C. aggregans* and, while speculative, potentially represents a polymerase from a novel organism. Due consideration is, however, also required for the possibility that the identified polymerase represents a variable strain of *C. aurantiacus*. In this event, the observed variability in terms of codon preferences, could have resulted from the different geographical distribution of this organism (Foerstner et al., 2005).

Distinct codon usage patterns are reportedly because of: (i) differential %GC content at the third base position (Alvarez et al., 1994); (ii) translational limitations due to the levels of available tRNAs for specific codons (Looyd and Sharp, 1992); and (iii) variations in the required expression levels for the specific protein in different organisms (Pouwels and Leunissen, 1994). The variable codon usage observed for the TC11pol in comparison to *C.aurantiacus* is therefore not only indicative of a different microbial origin for the gene but implies potential varied expression patterns as well.

UNIVERSITY of the
WESTERN CAPE

The strategies employed for expression of the full length TC11pol polymerase did not result in any observable expression by SDS-PAGE or detection of activity via a polymerase chain reaction assay. The variable codon usage displayed for the TC11pol could potentially explain the lack of expression of this polymerase in an *E.coli* expression system, for the reasons outlined above. Other issues for consideration, in view of the lack of expression, relate to toxicity of the expressed protein as well as use of incompatible promoters. Toxicity could however be discarded as an issue due to the proven expression of the polymerase of *C.aurantiacus* which is a close homolog of the TC11pol (Tvermyer et al.,1999). The question of promoter compatibility in turn was addressed in this study by placing the TC11pol under control of both the *tac* and *lac* promoters.

Structural models can also be used as a basis for identifying the function of individual proteins, much in the way this is accomplished with experimentally determined structures (Petrey and Honig, 2005). Aspects that require consideration during functional inferences on the basis of homology models are: (i) that there should be sufficient sequence similarity between the protein in question and the protein on which the model is constructed (ii) that the sequence based features are correctly aligned to generate a correct model and (iii) that the structure-function relationship is observable in a comparable fashion between the modeled protein and actual experimentally determined template structure (Wallner and Elofsson, 2005).

UNIVERSITY of the
WESTERN CAPE

The sequence features and structural models constructed for the TC11pol and C.aurantiacus polymerases were subsequently evaluated because of the above-mentioned criteria. The level of sequence similarity required between the novel protein and homologous structure model construction with sufficient accuracy is currently estimated at >40% (Wallner and Elofsson, 2005). Research by Sali et al, 1995, also suggests that models produced from query sequences with >40%

alignment against the template has accuracies comparable with NMR structures. Most often, the functional domains of comparable proteins attain these levels of similarity while the interspersed regions display higher divergence resulting in alignment scores that fall below the 40% benchmark. Accurate models are then often derived by focusing on only the functional and therefore structurally conserved regions for constructing the models. This strategy was implemented for the 3'-5' exonuclease domains of TC11pol and *C. aurantiacus* which displayed >40% similarity against that of *E. coli* and their 5'-3' exonuclease and polymerase domains against that of *T.aquaticus*.

The active site resdues of the 3'-5' exonuclease, 5'-4' exonuclease and polymerase domains could be observed at conserved locations for TC11pol, *C .aurantiacus, C. aggregans* and all corresponding template domains in the multiple sequence alignments. The structural features of the active site components for all modeled domains and corresponding templates were entirely conserved. The structural variations that were however observed between the TC11pol and *C. aurantiacus* polymerases in comparison to the templates did not result in an altered topology for any of the models.

UNIVERSITY of the
WESTERN CAPE

Validation of the structural models is achieved by evaluating whether the models adhere to standard steric and geometric criteria (Laskowski et al., 1993; Vriend, 1990). The Verify3D (Eisenberg et al., 1997) and RAMPAGE (ref) predictions of model quality for all domains of both TC11 and *C. aurantiacus* indicated that all models conformed to these criteria. The models were considered to be representative of the native structures of the library derived TC11 polymerase as well as the functionally characterized *C. aurantiacus* homolog.

All the sequence and structure derived evidence therefore confirms the close relationship between the TC11pol and *C. aurantiacus* polymerases. The 3'-5' exonuclease, 5'-3' exonuclease and polymerase actvities for the *C. aurantiacus* polymerase were also experimentally demonstrated in a study by Tvermyer et al.,1999. It is therefore highly suggestive that the TC11pol polymerase also posses comparable 3'-5' exonuclease, 5'-3' exonuclease and polymerase actvities.

UNIVERSITY of the
WESTERN CAPE

Apart from the demonstrated role in assigning functions for proteins, homology models also act as a resource for rational design of mutagenesis experiments. In the context of the DNA polymerases, structural models were utilized to derive truncated *T. aquaticus* polymerases mutants that demonstrated enhanced thermostability and polymerase activity (Villbrandt et al., 1997). Villbrandt et al, (2000) also implemented structural models to substitute the non-functional 3'-5' exonuclease domain of *T. aquaticus* with the functional equivalent of *E.coli*. The resulting chimera displayed characteristics from both polymerases at an intermediate temperature (Villbrandt et al., 2000).

The available structure models of the TC11pol polymerase were therefore similarly implemented to aid in the experimental design of a 5’ exonuclease deficient TC11pol protein (5’ exo^{-} TC11pol). In line with the observations from the structural models of TC11pol and the *C. aurantiacus* polymerases, the region spanning amino acids 290 to 942 was proposed as suitable to retain 3’-5’ exonuclease and polymerase activities. This 5’ exo^{-} TC11pol was proposed to be the structurally equivalent of the *E. coli* Klenow fragment (Ollis et al., 1992) and the *T. aquaticus* Klentaq (Korolev et al., 1995) fragment. The Klenow fragment displays both polymerase and 3’ exonuclease activities while Klentaq lacks 3’ exonuclease activity.

Expression analysis of the 5’ exo^{-} TC11pol mutant polymerase did not result in the observation of an enhanced purified protein product of the expected 74 kDa size. The protein could potentially have been expressed at a low level, which was why it could not be readily observed on a SDS-PAGE gel. Heat treatment of the cellular extracts (60 ^{o}C for 1 hour), proposed to denature the native *E.coli* (BL21) host proteins did not yield any products on SDS-PAGE either. This meant that: (i) the recombinant TC11POL was either not expressed or expression levels were too low to be detected by SDS-PAGE analysis or (ii) the recombinant protein did not display thermostability at the selected temperature and therefore was also denatured.

UNIVERSITY of the
WESTERN CAPE

The cell lysate was therefore heat denatured at a lower temperature of 55 ^{o}C for 1 hour. The SDS-PAGE analysis indicated a 74 kDa band that potentially represents the TC11POL mutant but also additional multiple protein bands in cellular fractions from both the uninduced and induced *E.coli* (BL21) expression strain. The number of protein bands was, however, significantly fewer than observed for the cellular fractions from the untreated *E.coli* (BL21) expression strain. This was expected due to the thermal liability of the native E.coli proteins which, with exceptions, are denatured above 37 ^{o}C. The observed bands could potentially have included native *E. coli* (BL21) proteins such as chaperones, proteases, transcription regulators, and other proteins that are activated upon heat shock (Zhao et al., 2003). Alternatively the bands represented multiple truncated forms of the TC11POL mutant, as a consequence of the variable codon usage observed in the protein.

UNIVERSITY of the
WESTERN CAPE

The cellular fraction that was heat denatured at 55 ^{o}C, was implemented in the polymerase and 3' exonuclease activity assay outlined in Section 2.10.4. The fact that the activity assay did not measure activity directly meant that it was not quantitative. Instead, the objective was to demonstrate that the putative 5'exo^{-} TC11pol possesses thermostability as well as activity beyond its mesophilic counterparts.

Based on the structural modeling information the putative 5'exo^{-} TC11pol was proposed to either demonstrate the activities associated with either the Klenow fragment of *E. coli* or the Klentaq fragment of *T. aquaticus.* The Klenow fragment displays both polymerase and 3' exonuclease activities while Klentaq lacks the Analysis of the results (Table 6.8) indicated detectable polymerase and 3' exonuclease activity for the putative 5'exo$^-$ TC11pol. In terms of functionality 5'exo$^-$ TC11pol therefore resembled the Klenow fragment of *E. coli* (Ollis et al., 1992) rather than the *T.aquaticus* Klentaq (Korolev et al., 1995) fragment. The observed polymerase and 3' exonuclease activity of the 5'exo$^-$ TC11pol, furthermore, displayed increased thermostability and activity in comparison to the Klenow fragment of *E. coli*. This was deduced from the lack of activity (at 37^{o}C) for the Klenow fragment, following heat deanturation of 55^{o}C while the putative 5'exo$^-$ TC11pol displayed activities at 37 and 45^{o}C.

UNIVERSITY of the
WESTERN CAPE

Chapter 7

Concluding Remarks

7.1 Summary of findings

Chapter 3

The successful implementation of standard extraction, cloning and sequencing procedures for a novel hyperthermophilic environment was demonstrated in this section of the work. Important findings from this work included:

- The quantity and quality of DNA extracted from the metagenomic sample was typical of a low biomass, humic acid contaminated environment.
- While degraded, sufficient high molecular weight DNA could be recovered for metagenomic library construction. This was achieved using both direct and indirect DNA extraction strategies.
- This DNA was successfully implemented in the construction of 4 small insert libraries proposed to collectively contain approximately of 1.3 million genes.
- A total of 70 Kb of sequence data was generated using a random shotgun sequencing approach.
- Initial sequence analysis revealed 53 gene sequences or partial gene transcripts from three broad categories: (i) genes with known functional homologues, (ii) genes with homologues to hypothetical ORFs and (iii) genes with no significant matches in the current protein database.

UNIVERSITY of the
WESTERN CAPE

Chapter 4

This section of the study demonstrated the succesfull implementation of the WGA strategy on the thermophilic environmental DNA. It was further demonstrated that the WGA DNA, as a resource, has the potential to enhance sequence-based analysis of low biomass thermophilic metagenomic environments.

In the first implementation, the WGA DNA was used as a resource for the PCR recovery of four metagenomic, library-derived, putative ORFs. This strategy validated: (i) the fact that these ORFs were of thermophilic origin, and not as a result of possible contamination and (ii) that the sequence data (for the ORFs) were from the metagenomic library.

Important observations from this section included:

- The environmental extracted DNA, while being of sufficient quality and quantity for direct cloning, was not sufficient for amplification of the investigated library derived ORFs (Usp, DUF29, ArgJ and Pol).
- The use of the WGA thermophilic environmental DNA resulted in the efficient recovery of these library derived ORFs (Usp, DUF29, ArgJ and Pol).
  - The whole genome amplification of the environmental DNA improved the availability and quality of the template to levels amenable by PCR amplification.

UNIVERSITY of the WESTERN CAPE

- The use of highly specific primers eliminated the potential recovery of non-specific whole genome amplification products, which are typically produced during multiple displacement amplification reactions.

In the second implementation, the WGA DNA was used as a resource for the recovery of sequence information, of a flanking region, of an ORF in the library.

- The sequence information of the metagenomic library derived ORF was used as a template for deriving a sequence specific primer for recovery of the flanking region.
- A significant proportion of the partial ORF could be recovered from the restriction digestion products of the metagenomic WGA sequence.
- This strategy represented a novel implementation of the WGA technique within a metagenomic sequencing context

UNIVERSITY of the
WESTERN CAPE

Chapter 5

In this chapter in-silico functional annotation strategies were implemented to derive functional information for two ORFs (denoted ORF1 and ORF2), identified from metagenomic sequence data.

The important findings included:

- The ORFs appeared in a novel tandem arrangement on one of the metagenomic derived clones.

- The ORFs were identified to have significant sequence homology to the DUF29 (proteins with no known function) and Usp family of proteins (proteins shown to be upregulated in response to various environmental stresses).
    - The close structural relationship between ORF2 and the Usp proteins was further demonstrated through homology modeling and comparison of the structures.
- The functional inferences from various homology prediction strategies for ORF1 could only be achieved with low confidence.
    - The consensus homology and genomic context predictions for ORF1 were suggestive of a potential role as transcriptional regulator.
- Integration of the various homology detection approaches with the genomic context mapping led to improved confidence in the predictions, but could still only classify the ORFs within the general class of putative nucleotide/DNA binding proteins.
- Within this broad classification, the hypothesis of related functions for these ORFs seems highly plausible.

UNIVERSITY of the
WESTERN CAPE

Chapter 6

Chapter 6 focused on the sequencing, homology modeling and expression analysis of a metagenomic sequence derived DNA polymerase (denoted TC11pol).

- On the basis of sequence homology searches and multiple sequence alignments the putative DNA polymerase could be classified within the type I family of polymerases.
- Attempts to express the full length protein, in two expression constructs, proved unsuccessful. This highlighted and confirmed the reported limitations of E.coli expression systems for proteins with altered codon frequencies.
- This study then further reported on the implementation of structure homology modeling of the TC11pol protein using the 3’-5’ exonuclease domain of *E.coli* and the 5’-3’ exonuclease- and polymerase domains of *T.aquaticus* as templates.
    - Analysis of the TC11pol structure model suggested that all the required sequence and structure features, essential for polymerase activity could be observed.

UNIVERSITY of the
WESTERN CAPE

- The structural model was also implemented in deriving a 5’-3’ exo^{-} deletion mutant proposed to be sufficient for 3’-5’ exonuclease and polymerase activity.
- From the expression of the 5’-3’ exo^{-} deletion mutant a putative recombinant protein of the expected 74 kDa molecular weight was observed.
- The implementation of a functional assay was indicative of low level 3’ exonuclease- and polymerase activities for the 5’-3’ exo^{-} deletion mutant.

- The observed activity indicated that the the 5'-3' exo^{-} deletion mutant was functionally equivalent to the *E.coli* Klenow fragment.
- The observed activity, subsequent to an extended heat denturation strategy, was furthermore indicative of thermostability of the deletion mutant in comparison to the mesophilic Klenow fragment.

7.2 Concluding remarks

The recent successes of the metagenome sequencing of several environmental samples have highlighted the potential for implementing this strategy on more diverse environments. These novel metagenomic environments, which will contribute further to the vast number of novel genes, is however also proposed to introduce new challenges regarding the recovery, cloning, sequencing and sequence analysis of the DNA complement. The thermophilic environments are particularly interesting in this regard, as it introduces elevated temperature as a factor, which could influence genome composition and has been shown to influence the composition of the cell wall. These features, in turn, could affect DNA extraction and cloning. Despite these potential limitations, the thermophilic environments are, however, also a potential rich source of novel thermostable proteins.

In this work metagenomic library construction, random sequencing and sequence analysis strategies were employed to enhance identification and characterisation of potentially novel genes, from a thermophilic soil sample. The extraction of metagenomic DNA from this environment yielded significant quanities of high molecular weight DNA that was suitable for further downstream analysis. This DNA was subsequently used for the construction of four metagenomic sequence libraries. From the successful implementation of these strategies, it was clearly apparent that the thermophilic environments were readily accessible through the implementation of standard DNA and extraction procedures. What was however apparent is that the DNA was contaminated with humic acid and the yields were not specifically high. These factors typically influence the downstream analysis of the DNA and result in incomplete library construction. One of the strategies implemented in this study focused on enhancing access to the metagenomes of low biomass environments. This strategy was demonstrated to allow for recovery of flanking sequence information form genes in the library, directly from the environment. Using integrated *in-silico* approaches putative nucleotide binding functions were assigned for two ORFs that occur in a tandem arrangement. The feasibility of implementing integrated but divergent *in-silico* strategies for enhanced functional inferences was demonstrated. The implementation of multiple sequence alignments and homology modeling strategies furthermore resulted in the identification of a putative DNA polymerase gene. The homology model was implemented to derive a putative deletion mutant of the polymerase. From the expression and functional characterization of the mutant polymerase it could be deduced that the putative protein demonstrated themostability.

UNIVERSITY of the
WESTERN CAPE

Appendix

Table of partial ORFs identified in this study

Clone ID	GI number of highest scoring BLAST result	Blast score	Identified functional domains
2.102 T7	gi|77543718	3e-27	No significant match
2.102 V5	gi|83572583	2e-17	No significant match
2.104 T7	gi|77543718	2e-27	No significant match
2.104 V5	gi|88947379	3e-10	No significant match
2.106 T7	gi|77543718	2e-27	No significant match
2.106 V5	gi|20515244	4e-23	No significant match
2.108 T7	gi|77543718	5e-15	No significant match
2.108 V5	gi|88947379	3e-24	No significant match
2.110 T7	gi|77543718	1e-28	No significant match
2.136 V5	gi|16419700	4e-11	No significant match
	gi|67875266	2e-04	No significant match
2.137 T7	gi|77543718	2e-27	No significant match
2.143 T7	gi|77543718	3e-06	No significant match
2.141 T7	gi|45861240	1e-53	No significant match
2.140 V5	gi|76260864	5e-71	No significant match
2.140 T7	gi|76260864	8e-75	No significant match
2.139 T7	gi|71673543	4e-78	"TIM barrel domain"
2.138 V5	gi|88947379	1e-25	No significant match

UNIVERSITY of the
WESTERN CAPE

Table of full length ORFs identified in the study

Clone ID	GI number of highest scoring BLAST result	ORF start and stop codon positions	Blast score	Identified functional domains
2.96	gi|1913934	31 – 3031	9e-174	5’-3’ exonuclease domain 3’-5’ exonuclease domain DNA polymerase domain
2.142	gi|12583694	23 – 366	4.7e-24	DUF29
	gi|13456284	372 – 882	3.2e-23	Usp
	gi|43555674	892 – 934	4.3e-43	Glutamate N-acetyltransferase
2.136	gi|77952610	55 - 443	8e-15	No significant match

UNIVERSITY of the
WESTERN CAPE

Congress contributions

National:

Du Plessis MG, Muyanga S, Cowan DA (2006) Metagenome sequencing and analysis of a Chinese geothermal library-A pilot study. 14th Biennial Congress of the South African Society for Microbiology SASM, Pretoria, April 2006.

UNIVERSITY of the
WESTERN CAPE

Reference List

Alba MM, Laskowski RA, and Hancock JM (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics. 18:672–678.

Altenbuchner J, Siemann-Herzberg M, and Syldatk C (2001) Hydantoinases and related enzymes as biocatalysts for the synthesis of unnatural chiral amino acids. Current Opinion in Biotechnology. 12:559–563.

Altschul SF, and Gish W (1996) Local alignment statistics. Methods in Enzymology. 266:460–480.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology. 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25: 3389–3402.

UNIVERSITY of the
WESTERN CAPE

Alvarez F, Robello C, and Vignali M (1994) Evolution of codon usage and base contents in kinetoplastid protozoan. Molecular Biology and Evolution. 11:790–802.

Amann RI, Ludwig W, and Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiology Reviews. 59:143–169.

Ando S, Ishida H, Kosugi Y, and Ishikawa K (2002) Hyperthermostable endoglucanase from *Pyrococcus horikoshi*. Applied and Environmental Microbiology. 68:430–433.

Ashburner M, Ball CA, Blake CA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 25:25-29.

Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, Xu Y, Lai X, Huang L, Dong X, Ma Y, Ling L, Tan H, Chen R, Wang J, Yu J, and Yang H. (2002). A complete sequence of the *T. tengcongensis* genome. Genome Research. 12:689-700.

UNIVERSITY of the
WESTERN CAPE

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, and Sonnhammer EL (2002). The Pfam protein families database. Nucleic Acids Research. 30:276-280.

Beja O, (2004) To BAC or not to BAC: marine ecogenomics. Current Opininion in Biotechnology. 15: 187–190.

Berthelet M, Whyte LG, and Greer CW (1996) Rapid, direct extraction of DNA from soils for PCR analysis using polyvinylpolypyrrolidone spin columns. FEMS Microbiology Letters. 138:17–22.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research. 31:365–370.

Bok J, Goers S, and Eveleigh D (1994) Cellulase and xylanase systems of *Thermotoga neapolitana*. ACS Symposium Series. 566:54–65.

Brenner SE, Chothia C, and Hubbard T (1998) Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. Proceedings of the National Academy of Sciences of the United States of America. 95:6073–6078.

UNIVERSITY of the
WESTERN CAPE

Bujnicki JM, Elofsson A, Fischer D, and Rychlewski L (2001) LiveBench-1: Continuous benchmarking of protein structure prediction servers. Protein Science. 10:352-361.

Chothia C, and Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. Journal of Molecular Biology. 196:901-917.

Christen S, Srinivas A, Bähler P, Zeller A, Pridmore D, Bieniossek C, Baumann U, and Erni H (2006) Regulation of the DHA operon of *Lactococcus lactis*: A deviation from the rule followed by the TetR family of transcription regulators. 281:23129-23137

Claverie JM, and States DJ (1993) Information enhancement methods for large scale sequence analysis. Computational Chemistry. 17:191–201.

UNIVERSITY of the
WESTERN CAPE

Cline J, Braman JC, and Hogrefe HH (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research. 24:3546-3551.

Cowan D (1996) Industrial enzyme technology. Trends in Biotechnology. 14:177–178.

Cullen DW, and Hirsch PR (1998) Simple and rapid method for direct extraction of microbial DNA from soil for PCR. Soil Biology and Biochemistry. 30:983–993.

Curtis TP, Sloan WT, and Scannell JW (2002) Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences of the United States of America. 99:10494–10499.

Dandekar T, Snel B, Huynen M, and Bork P (1998) Conservation of gene order: A fingerprint of proteins that physically interact. Trends in Biochemical Sciences. 23:324–328.

Daniel R (2005). The metagenomics of soil. Nature Reviews Microbiology. 3:470-478.

de la Torre JR, Christianson LM, B´ej`a O, Suzuki MT, Karl DM, Heidelberg J, and DeLong EF. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proceedings of the National Academy of Sciences of the United States of America. 100:12830–12835.

Dean FB, Hosono S, Fang SL, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M, and Lasken RS (2002) Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America. 99:5261– 5266.

UNIVERSITY of the
WESTERN CAPE

Degrange V, and Bardin R (1995) Detection and counting of Nitrobacter populations in soil by PCR. Applied and Environmental Microbiology. 61:2093–2098.

Demirijan D, Moris-Varas F, and Cassidy C (2001) Enzymes from extremophiles. Current Opinion in Chemical Biology. 5:144–151.

Devos D, and Valencia A (2000) Practical limits of function prediction. Proteins. 41:98-107.

Doerks T, von Mering C, and Bork P (2004) Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes. Nucleic Acids Research. 32:6321–6326.

UNIVERSITY of the
WESTERN CAPE

Dong G, Vieille C, Savachenko A, Zeikus G (1997) Cloning, sequencing, and expression of the gene encoding extracellular α-amylase from *Pyrococcus furiosus* and biochemical characterization of the recombinant enzyme. Applied and Environmental Microbiology. 63:3569–3576.

Dover LG, Corsino PE, Daniels IR, Cocklin SL, Tatituri V, Besra GS, and Futterer K (2004) Crystal structure of the TetR/CamR family repressor *Mycobacterium tuberculosis* EthR implicated in ethionamide resistance. Journal of Molecular Biology. 340:1095–1105.

Eddy SR (1998) Profile hidden Markov models. Bioinformatics. 14:755-763.

Eichler J (2001) Biotechnological uses of archaeal extremozymes. Biotechnology advances. 19:61–278.

Eisenberg D, Luthy R, and Bowie JU (1997) Verify3D: assessment of protein models with three dimensional profiles. Methods in Enzymology. 277:396–404.

Enault F, Suhre K, Poirot O , Abergel C, and Claverie MJ. (2003) Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. Nucleic Acids Research. 31:3720-3722.

Enright AJ, Iliopoulos I, Kyrpides NC, and Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature. 402:86–90.

Entcheva P, Liebl W, Johann A, Hartsch T, and Streit WR (2001): Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. Applied and Environmental Microbiology. 67:89-99.

Fischer D (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins. 51:434–441.

UNIVERSITY of the
WESTERN CAPE

Friedberg I (2006) Automated Protein Function Prediction: The Genomic Challenge. Briefings in Bioinformatics.

Foerstner KU, von Mering C, Hooper SD, and Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Reports. 6:1208-1213.

Frey B, and Suppman B (1995) Demonstration of the expand PCR system's greater fidelty and higher yields with a *lac*I-based fidelity assay. Biochemica. 2:34–35.

Frostegard A, Courtois S, Ramisse V, Clerc S, Bernillon D, Le Gall F, Jeannin P, Nesme X, and Simonet P (1999) Quantification of bias related to the extraction of DNA directly from soils. Applied and Environmental Microbiology. 65:5409–5420.

Gabor EM, de Vries EJ, and Janssen DB (2003). Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. FEMS Microbiology Ecology. 44:153–63.

Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, Yan S, Nicoletta A, Vafai J, Sun D, Wang L, Noah JE, Pasquali S, and Schlick T (2002) Analysis of protein sequence/structure similarity relationships. Biophysical Journal. 83:2781–2791.

UNIVERSITY of the
WESTERN CAPE

Ginalski K, Elofsson A, Fischer D, and Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 19:1015–1018.

Gonzalez JM, Portillo MC, and Saiz-Jimenez C (2005) Multiple displacement amplification as a pre-polymerase chain reaction (pre-PCR) to process difficult to amplify samples and low copy number sequences from natural environments. Environmental Microbiology. 7:1024–1028.

Greer J (1990) Comparative modeling methods:Application to the family of the mammalian serine proteases. Proteins. 7:317–334.

Gustavsson N, Diez A, and Nystrom T (2002) The universal stress protein paralogues of *Escherichia coli* are co-ordinately regulated and co-operate in the defence against DNA damage. Molecular Microbiology. 43:107–117.

Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson RM, and DeLong EF (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. Science. 305:1457-62.

Handelsman J, Rondon MR, Brady SF, Clardy J, and Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemical Biology. 5:245–249.

UNIVERSITY of the
WESTERN CAPE

Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, and Orengo C (2003). Recognizing the fold of a protein structure. Bioinformatics. 19:1748-1759.

Hasson MS, Schlichting I, Moulai J, Taylor K, Barrett W, and Kenyon GL, Babbitt PC, Gerlt JA, Petsko GA, and Ringe D (1998) Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. Proceedings of the National Academy of Sciences of the United States of America. 95:10396-10401

Hawkins T, Stanislav L, and Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Science. 15:1550–1556.

Henrik N. Engelbrecht J. Brunak S, and von Heijne D (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering. 10:1–6.

Herbert R, and Sharp R (1992) Molecular Biology and Biotechnology of Extremophiles. Chapman and Hall, NY.

Hickey A, and Singer GAC (2004) Genomic and proteomic adaptations to growth at high temperature. Genome Biology. 5:117-121.

UNIVERSITY of the
WESTERN CAPE

Hirokawa T, Seah BC, and Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics. 14:378–379.

Hofmann K (2000) Sensitive protein comparisons with profiles and hidden Markov models. Briefings Bioinformatics. 1:167-178.

Hofmann K, Bucher P, Falquet L, and Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Research. 27:215-219.

Holben WE, (1997) Isolation and purification of bacterial community DNA from environmental samples. 431–444. In Hurst CJ, Knudsen GR, McInerney MJ, Stetzenbach LD, and Walter MV (ed.). Manual of environmental microbiology. ASM Press, Washington, D.C.

Holben WE, Jansson JK, Chelm BK, and Tiedje JM (1988) DNA probe method for the detection of specific microorganisms in the soil bacterial community. Applied and Environmental Microbiology. 54:703–711.

Holm L, and Park J (2000) DaliLite workbench for protein structure comparison. Bioinformatics. 16:566-567.

Holm L, and Sander C (1993). Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology. 233:123-138.

UNIVERSITY of the
WESTERN CAPE

Holm L, and Sander C (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urea. Proteins. 28:72–82.

Holm L, and Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Research. 25:231-234.

Hooft RW, Sander C, and Vriend G (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins. 26:363-376.

Hough D, and Danson M (1999) Extremozymes. Current Opininion Chemical Biology. 3:39–46.

Huffman JL, and Brennan RG (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. Current Opinion in Structural Biology. 12:98–106.

Hurt RA (2001) Simultaneous recovery of RNA and DNA from soils and sediments. Applied and Environmental Microbiology. 67:4495-4503.

Huynen M, Snel B, LatheW, and Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Research. 10:1204–1210.

UNIVERSITY of the
WESTERN CAPE

Itou H, Okada U, Suzuki H, Yao M, Wachi M, Watanabe N, and Tanaka I (2005) The CGL2612 protein from *Corynebacterium glutamicum* is a drug resistance-related transcriptional repressor. Structural and functional analysis of a newly identified transcription factor from genomic DNA analysis. The Journal of Biological Chemistry. 280:38711–38719.

Jackson CR, Harper JP, Willoughby D, Roden EE, and Churchill PF (1997) A simple, efficient method for the separation of humic substances and DNA from environmental samples. Applied and Environmental Microbiology. 63:4993-4995.

Jaroszewski L, Rychlewski L, and Godzik A (2000). Improving the quality of twilight-zone alignments. Protein Science. 9:1487-1496.

UNIVERSITY of the
WESTERN CAPE

Jerome KR, Huang ML, Wald A, Selke S, and Corey L. (2002) Quantitative stability of DNA after extended storage of clinical specimens as determined by realtime PCR. Journal of Clinical Microbiology. 40:2609–2611.

Johnston CG, and Aust SD (1994) Detection of *Phanerochaete chrysosporium* in soil by PCR and restriction enzyme analysis. Applied and Environmental Microbiology. 60:2350–2354.

Jones DT (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology. 287:797-815.

Jones M, and Foulkes N (1989) Reverse transcription of mRNA by *Thermus aquaticus* DNA polymerase. Nucleic Acids Research. 17:8387– 8388.

Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, and Hughey R (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins. 53: 491–496.

Kauffmann IM, Schmitt J, and Schmid RD (2004) DNA isolation from soil samples for cloning in different hosts. Applied Microbiology and Biotechnology. 64:665–670.

Kawabata T (2003). MATRAS: A program for protein 3D structure comparison. Nucleic Acids Research. 31:3367-3369.

Kawabata T, and Nishikawa K (2000) Protein structure comparison using the markov transition model of evolution. Proteins. 41:108-122.

Kayer K, and Imlay JA (1996) Superoxide accelerates DNA damage by elevating free-iron levels. Proceedings of the National Academy of Sciences of the United States of America. 93:13635-13640.

UNIVERSITY of the
WESTERN CAPE

Kelley LA, MacCallum RM, and Sternberg MJ (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. Journal of Molecular Biology. 299:499-520.

Kengen S, Luesink E, Stams A, and Zehnder A (1993) Purification and characterization of an extremely thermostable b-glucosidase from the hyperthermophilic archaeon *Pyrococccus furiosus*. European Journal of Biochemistry. 213:305–312.

Kim NK, and Xie J (2006) Protein multiple alignment incorporating primary and secondary structure information. Journal of Computational Biology. 13:1615–1629.

Kim UJ, Shizuya H, de Jong PJ, Birren B, and Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Research. 20:1083-1085.

Kimura N (2006) Metagenomics: Access to unculturable microbes in the environment. Microbes Environment. 21:201-215.

Kirk O, Borchert TV, and Fuglsang CC (2002) Industrial enzyme applications. Current Opinion in Biotechnology. 13:345–351.

UNIVERSITY of the
WESTERN CAPE

Kolker E, Makarova KS, Shabalina S, Picone AF, Purvine S, Holzman T, Cherny T, Armbruster D, Munson RS, Kolesov G, Frishman D, and Galperin MY (2004) Identification and functional analysis of ‘hypothetical’ genes expressed in *Haemophilus influenzae*. 32:2353-2361.

Korolev S, Nayal M, Barnes WM, DiCera E, and Waksman G (1995) Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-A° resolution: structural basis for thermostability. Proceedings of the National Academy of Sciences of the United States of America. 92:9264–9268.

Kuske CR, Banton KL, Adorada DL, Stark PC, Hill KK, and Jackson PJ (1998) Small-Scale DNA sample preparation method for field PCR detection of microbial cells and spores in soil. Applied and Environmental Microbiology. 64:2463-2472.

UNIVERSITY of the
WESTERN CAPE

Kvint K, Nachin L, Diez A, and Nystrom T (2003) The bacterial universal stress protein: function and regulation. Current Opinion Microbiology. 6:140–145.

LaMontagne MG, Michel FC, Jr. Holden PA, and Reddy CA (2002) Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. Journal of Microbiological Methods. 49:255-64.

Laskowski RA, MacArthur MW, Moss DS, and Thornton JM (1993). PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 26:283–291.

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, and Thornton JM (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. Journal of Biomolecular NMR. 8: 477-486.

LeCleir GR, Buchan A, and Hollibaugh JT (2004) Chitinase gene sequences retrieved from diverse aquatic habitats reveal environment specific distributions. Applied and Environmental Microbiology. 70:6977-6983.

Lee M, Lee C, Oh T, Song JK, and Yoon J (2006) Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: Evidence for a new family of bacterial lipases. Applied and Environmental Microbiology. 72:7406–7409.

Lee SY, Bollinger J, Bezdicek D, and Ogram A (1996) Estimation of the abundance of an uncultured soil bacterial strain by a competitive quantitative PCR method. Applied and Environmental Microbiology. 62:3787–3793.

Levitt M (1992) Accurate modeling of protein conformation by automatic segment matching. Journal of Molecular Biology. 226: 507-33.

UNIVERSITY of the
WESTERN CAPE

Li J, Halgamuge SK, Kells CI, and Tang S (2007) Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. BMC Bioinformatics. 8:S6.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR,Loiacono KA, Lynch BA, Macneil IA, MinorC, TiongCL, Gilman M, Osburne MS, Clardy J, Handelsman J, and Goodman RM (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and Environmental Microbiology. 66:2541-2547.

Lloyd AT, and Sharp PM (1992). Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Research. 20:5289-5295.

Lorenz P, and Eck J (2005). Metagenomics and industrial applications. Nature Reviews Microbiology. 3:510–516.

Lorenz P, Liebeton K, Niehaus F, and Eck J (2002) Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Current Opinion in Biotechnology. 13:572–77.

UNIVERSITY of the
WESTERN CAPE

Lundberg K, Shoemaker D, Adams M, Short J, Sorge J, and Marthur E (1991) High-fidelty amplification using a thermostable polymerase isolated from *Pyrococcus furiosus*. Gene. 108:1–6.

Luthy R, Bowie JU, and Eisenberg D (1992). Assessment of protein models with three-dimensional profiles. Nature. 356:83-85.

Madej T, Gibrat JF, and Bryant SH (1995) Threading a database of protein cores. Proteins. 23:356-369.

Makarova KS, Aravind L, Grishin NV, Rogozin IB, and Koonin EV (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Research. 30:482–496.

Marcotte EM, Pellegrini M, Rice DW, Yeates TO, and Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science. 285:751–753.

May O, Habenicht A, Mattes R, Syldatk C and Siemann M (1998) Molecular evolution of hydantoinases. Biological Chemistry. 379: 743 – 747.

McDonald AE, and Vanlerberghe GC (2005) Alternative oxidase and plastoquinol terminal oxidase in marine prokaryotes of the Sargasso Sea. Gene. 349:15-24.

UNIVERSITY of the
WESTERN CAPE

McGuffin LJ, and Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics. 19:874–881.

Mead D, McClary J, Luckey J, Kostichka A, Witney F, and Smith L (1991) Bst DNA polymerase permits rapid sequence analysis from nanogram amounts of template. Biotechniques. 11:76–78.

Melo F, and Feytmans E (1998). Assessing protein structures with a non-local atomic interaction energy. Journal of Molecular Biology. 277:1141-1152.

Miller DN, Bryant JE, Madsen EL, and Ghiorse WC 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Applied and Environmental Microbiology. 65:4715–4724.

UNIVERSITY of the
WESTERN CAPE

Moré MI, Herrick JB, Silva MC, Ghiorse WC, and Madsen EL (1994) Quantitative cell lysis of indigenous microorganisms and rapid extraction of microbial DNA from sediment. Applied and Environmental Microbiology. 60:1572–1580.

Muller A, MacCallum RM, and Sternberg MJ (1999) Benchmarking PSI-BLAST in genome annotation. Journal of Molecular Biology. 293:1257-1271.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, and Bernardi G (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Letters. 573:73–77.

Myers T, and Gelfand D (1991) Reverse transcription and DNA amplification by a *Thermus thermophilus* DNA polymerase. Biochemistry. 30:7661–7665.

Nachin L, Nannmark U, and Nystrom T (2005) Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility. Journal of Bacteriology. 187:6265–6272.

Nayeem A, Sitkoff D, and Krystek S Jr. (2006) A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. Protein Science. 15:808-824

Nelson HCM (1995) Structure and function of DNA-binding proteins. Current Opinion in Genetics and Development. 5:180-189.

Novotny M, Madsen D, and Kleywegt GJ (2004). Evaluation of protein fold comparison servers. Proteins. 54:260-270.

Ohlson T, Wallner B, and Elofsson A (2004) Profile-profile methods provide improved fold recognition: a study of different profile-profile alignment methods. Proteins. 57:188–197.

UNIVERSITY of the
WESTERN CAPE

Oldfield TJ (1992) SQUID: A program for the analysis and display of data from crystallography and molecular dynamics. Journal of Molecular Graphics.10: 247-252.

Ollis DL, Brick P, Hamlin R, Xuong NG, and Steitz TA (1992) Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. Nature. 313:762–766.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, and Maltsev N (1999) The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America. 96:2896-2901.

Pantazaki A, Prista A, and Kyriakidis D (2002) Biotechnologically relevant enzymes from *Thermus thermophilus*. Applied Microbiology and Biotechnology. 58:1–12.

Paz A, Mester D, Baca I, Nevo E, et al. (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. Proceedings of the National Academy of Sciences of the United States of America. 101:2951-2956.

UNIVERSITY of the
WESTERN CAPE

Pazos F, and Sternberg MJ (2004) Automated prediction of protein function and detection of functional sites from structure. Proceedings of the National Academy of Sciences of the United States of America. 101:14754-14759.

Pearson WR, and Lipman DJ (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. 85:2444–2448.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America. 96:4285-4288.

Perler F, Kumar S, and Kong H (1996) Thermostable DNA polymerases. Protein Chemistry. 48:377-435.

Petrey D, and Honig B (2005) Protein structure prediction: Inroads to biology. Molecular Cell. 20: 811–819.

Petrey D, and Honig B (2000) Free energy determinants of tertiary structure and the evaluation of protein models. Protein Science. 9:2181–2191.

UNIVERSITY of the
WESTERN CAPE

Picard C, Ponsonnet C, Paget E, Nesme X, and Simonet P (1992) Detection and enumeration of bacteria in soil by direct DNA extraction and polymerase chain reaction. Applied and Environmental Microbiology. 58:2717–2722.

Pluthero FG (1993) Rapid purification of high-activity Taq DNA polymerase. Nucleic Acids Research. 21:4850–4851.

Ponting CP (2001) Issues in predicting protein function from sequence. Briefings in Bioinformatics. 2:19-29.

Porteous LA, Armstrong JL, Seidler RJ, and Watrud LS (1994) An effective method to extract DNA from environmental samples for polymerase chain reaction amplification and DNA fingerprint analysis. Current Microbiology. 29:301–307.

Pouwels PH, and Leunissen JA (1994) Divergence in codon usage of *Lactobacillus* species. Nucleic Acids Research. 22:929–936.

Premal H, Patel, and Loeb LA (2000) DNA polymerase active site is highly mutable: Evolutionary consequences. Proceedings of the National Academy of Sciences of the United States of America. 97:5095–5100.

UNIVERSITY of the
WESTERN CAPE

Purdy KJ, Embley TM, Takii S, and Nedwell DB (1996) Rapid extraction of DNA and rRNA from sediments by a novel hydroxyapatite spin-column method. Applied and Environmental Microbiololgy. 62: 3905–3907.

Reichsman F, Moore HM, and Cumberledge S (1999) Sequence homology between Wingless/Wnt-1 and a lipid-binding domain in secreted phospholipase A2'. Current Biology. 9:353-355.

Riesenfeld CS, Schloss PD, and Handelsman J (2004) Metagenomics: Genomic analysis of microbial communities. Annual Review Genetics. 38:525–552.

Rochelle PA, Fry JC, Parkes RJ, and Weightman AJ (1992) DNA extraction for 16S rRNA gene analysis to determine genetic diversity in deep sediment communities. FEMS Microbiology Letters. 79:59-65.

Rodriguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiology Letters. 231:153-158.

Rogozin IB, Makarova KS, Wolf YI, and Koonin EV (2004) Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Briefings in Bioinformatics. 5:131–149.

UNIVERSITY of the
WESTERN CAPE

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, and Dunker AK (2001) Sequence complexity of disordered protein. Proteins. 42:38–48.

Ross K.S, Haites NE, and Kelly KF (1990) Repeated freezing and thawing of peripheral blood and DNA in suspension – Effects on DNA yield and integrity. Journal of Medical Genetics. 27:569–570.

Rost B (2002) Enzyme function less conserved than anticipated. Journal of Molecular Biology. 318:595-608.

Rychlewski L, Jaroszewski L, Li W, and Godzik A (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science. 9:232-241.

Sabehi G, Beja O, Suzuki MT, Preston CM, and DeLong EF (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. Environmental Microbiology. 6:903–910.

Saiki RK, Scharf SJ, Faloona F, Mullis KB, Horn GT, Erlich HA and Arnheim N (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 230:1350–1354.

UNIVERSITY of the
WESTERN CAPE

Sali A, and Blundell TL (1993) Comparative modelling by statisfaction of spatial restraints. Journal of Molecular Biology. 234:779–815.

Sali A, Potterton L, Yuan F van Vlijmen H, and Karplus M (1995) Evaluation of comparative protein modeling by modeller. Proteins. 23:318–326.

Saunders NF, Thomas T, Curmi PM, Mattick JS, et al. (2003). Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Research. 13:1580-1588.

Schmeisser C, Stockigt C, Raasch C, Wingender J, Timmis KN, Wenderoth DF, Flemming HC, Liesegang H, Schmitz RA, Jaeger KE, and Streit WR (2003) Metagenome survey of biofilms in drinking-water networks. Applied and Environmental Microbiology. 69:7298-7309.

Schultz J, Milpetz F, Bork P, and Ponting CP (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. Proceedings of the National Academy of Sciences of the United States of America. 95:5857-5864.

Schwede T, Kopp J, Guex N, and Peitsch MC (2004) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Research. 31:3381–3385.

UNIVERSITY of the
WESTERN CAPE

Selenska S, and Klingmüller W (1991) DNA recovery and direct detection of tn5 sequences from soil. Letters Applied Microbiology.13: 21–24.

Shah I, and Hunter L (1997) Predicting enzyme function from sequence: a systematic appraisal. Proc Int Conf Intell Syst Mol Biol. 5:276-283.

Sharon I, Birkland A, Chang K, El-Yaniv R, and Yonah G (2005) Correcting BLAST e-Values for Low-Complexity Segments. Journal of Computational Biology. 12:978-1001.

Shearstone JR, and Baneyx F (1999) Biochemical Characterization of the Small Heat Shock Protein IbpB from *Escherichia coli*. The Journal of Biological Chemistry. 274:9937–9945.

Shenkin PS, Yarmush DL, Fine RM, Wang HJ, and Levinthal C. (1987). Predicting antibody hypervariable loop conformation Ensembles of random conformations for ringlike structures. Biopolymers. 26:2053-2085.

Shi J, Blundell TL, and Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environmentspecific substitution tables and structure-dependent gap penalties. Journal of Molecular Biology. 310:243–257.

UNIVERSITY of the
WESTERN CAPE

Shikama K. (1965) Effect of freezing and thawing on stability of double helix of DNA. Nature. 207:529–530.

Shindyalov IN, and Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Research. 29:228-229.

Silberberg MS (2000) Chemistry: The molecular nature of matter and change, 2nd ed. McGraw-Hill, Boston, MA.

Sivashankari S, and Shanmughavel P (2006) Functional annotation of hypothetical proteins – A review. Bioinformation. 1: 335 -338.

Skolnick J, and Kihara D (2001) Defrosting the frozen approximation: PROSPECTOR: A new approach to threading. Proteins. 42:319–331.

Smith TF, and Waterman MS (1981) Overlapping genes and information theory. Journal of Theoritical Biology. 91:379-380.

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21:951–960.

UNIVERSITY of the
WESTERN CAPE

Steffan RJ, Goksoyr J, Bej AK, and Atlas R.M (1988) Recovery of DNA from soils and sediments. Applied and Environmental Microbiology. 54:2908–2915.

Stein JL, Marsh TL, Wu KY, Shizuya H and DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology. 178:591-599.

Steitz TA (1999) DNA Polymerases: Structural diversity and common mechanisms. The Journal of Biological Chemistry. 274:17395–17398.

Streit WR, and Schmitz RA (2004) Metagenomics – the key to the uncultured microbes. Current Opinion in Microbiology. 7:492–498.

Tebbe CC, and Vahjen W (1993) Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. Applied and Environmental Microbiology. 59:2657–2665.

Thompson JD, Higgins DG, and Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 22:4673-4680.

UNIVERSITY of the
WESTERN CAPE

Tian W, and Skolnick J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? Journal of Molecular Biology. 333:863-882.

Todd AE, Orengo CA, and Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. Journal of Molecular Biology. 307:1113-1143.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, and Rubin EM (2005) Comparative metagenomics of microbial communities. Science. 308:554–557.

Trivedi S, Gehlot HS, and Rao SR (2006) Protein thermostability in Archaea and Eubacteria. Genet. Mol. Res. 5 (4): 816-827.

Tsai YL, and Olson BH (1992) Rapid method for separation of bacterial DNA from humic substances in sediments for polymerase chain reaction. Applied and Environmental Microbiology. 58:2292–2295.

Tvermyr M, Kristiansen BE, and Kristensen T (1998) Cloning, sequence analysis and expression in *E. coli* of the DNA polymerase I gene from *Chloroflexus aurantiacus,* a green nonsulfur eubacterium. Genetic Analysis: Biomolecular Engineering. 14:75–83.

UNIVERSITY of the
WESTERN CAPE

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, and Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 428:37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rush D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, SevyS, Knap H, Lomas MW, Nealson K, White O, Peterson JD, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, and Smith HO (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304:66-74.

Villbrandt B, Sagner G, and Schomburg D (1997) Investigations on the thermostability and function of truncated *Thermus aquaticus* DNA polymerase fragments. Protein Engineering. 10:1281-1288.

Villbrandt B, Sobek H, Frey B, and Schomburg D (2000) Domain exchange: chimeras of *Thermus aquaticus* DNA polymerase, *Escherechia coli* DNA polymerase I and *Thermotoga neapolitana* DNA polymerase. Protein Engineering. 13:645-654.

Vogel F, and Lumper L**.** (1986) Complete structure of the hydrophilic domain in the porcine NADPH-cytochrome P-450 reductase. Biochemical Journal. 236:871–878.

UNIVERSITY of the
WESTERN CAPE

Volossiouk T, Robb EJ, and Nazar RN (1995) Direct DNA extraction for PCR-mediated assays of soil organisms. Applied and Environmental Microbiology. 61:3972–3976.

von Grotthuss M, Wyrwicz LS, and Rychlewski L (2003) mRNA cap-1 methyl-transferase in the SARS genome. Cell. 113:701–702.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, and Snel B (2003) STRING: A database of predicted functional associations between proteins. Nucleic Acids Research. 31:258–261.

Von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, and Bork P (2003) Genome evolution reveals biochemical networks and functional modules. Proceedings of the National Academy of Sciences of the United States of America. 100:15428–15433.

Vriend G, (1990) WHAT IF: a molecular modeling and drug design program. Journal of Molecular. Graphics. 8: 52–56.

Wallner B, and Elofsson A (2005) All are not equal: A benchmark of different homology modeling programs. Protein Science. 14:1315–1327.

UNIVERSITY of the
WESTERN CAPE

Webb EC, (1992) Enzyme Nomenclature 1992. Recommendations of the Nomenclature committee of the International Union of Biochemistry and Molecular Biology, Academic Press.

Wechter P, Williamson J, Robertson A, and Kluepfel D (2003) A rapid, cost-effective procedure for the extraction of microbial DNA from soil. World Journal of Microbiology & Biotechnology. 19:85–91.

Whisstock JC, and Lesk AM (2003) Prediction of protein function from protein sequence and structure. Quarterly Reviews of Biophysics. 36: 307-340

Wierenga RK (2001) The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Letters. 492:193-198

Wikstrom P, Wiklund A, Andersson AC, and Forsman M (1996) DNA recovery and PCR quantification of catechol 2,3-dioxygenase genes from different soil types. Journal of Biotechnology. 52:107–120.

Wintzingerode F von, Gobel UB, and Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews. 21:213–229.

Wootton JC, and Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Computational Chemistry. 17:149–163.

UNIVERSITY of the
WESTERN CAPE

Wu S, and Zhang Y (2007) LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Research. 35:3375–3382.

Xu D, Baburaj K, Peterson CB, and Xu Y (2001) Model for the three-dimensionalstructure of vitronectin: Predictions for the multi-domain protein from threading and docking. Proteins. 44:312–320.

Xu Y, and Xu D (2000) Protein threading using PROSPECT: design and evaluation. Proteins. 40:343–354.

Ye Y, and Godzik A (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics. 19:246-255.

Yeates C, Gillings MR, Davison AD, Altavilla N, and Veal DA (1998) Methods for microbial DNA extraction from soil for PCR amplification. http://www.biologicalprocedures.com/. Biological Procedures Online 1: 40-47.

Yost C, Hauser L, Larimer F, Thompson D, Beliaev A, Zhou J, Xu Y, and Xu D (2003) A Computational Study of *Shewanella oneidensis* MR-1: Structural Prediction and Functional Inference of Hypothetical Proteins. A Journal of Integrative Biology. 7:177-191.

UNIVERSITY of the
WESTERN CAPE

Zareminski TI, Hung L, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, and Kim SH (1988) Structure-based assignment of the biochemical function of hypothetical protein: A test case of structural genomics. Proceedings of the National Academy of Sciences of the United States of America. 95:5189–15193.

Zhang Y, and Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57:702–710.

Zhang Y, and Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America. 101:7594-7599

Zhao Y, Liu D, Kaluarachchi WD, Bellamy HD, White MA, and Fox RO (2003) The crystal structure of Escherichia coli heat shock protein YedU reveals three potential catalytic active sites. Protein Science. 12:2303-2311.

Zhou H, and Zhou Y (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins. 55:1005–1013.

Zhou H, and Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 58:321–328.

UNIVERSITY of the
WESTERN CAPE

Zhou J, Bruns MA, and Tiedje JM (1996) DNA recovery from soils of diverse composition. Applied and Environmental Microbiology. 62:316–322.

Zhu J, and Weng Z (2005) FAST: a novel protein structure alignment algorithm. Proteins. 58:618-627.

Zierenberg RA, Adams MWW, and Arp AJ (2000) Life in extreme environments: Hydrothermal vents. Proceedings of the National Academy of Sciences of the United States of America. 97:12961-12962.

UNIVERSITY of the
WESTERN CAPE