Isolation of a strong constitutive promoter from *Euglena gracilis*

Bishal Khatiwada

Department of Chemistry and Biomolecular Sciences,
Macquarie University, Sydney, Australia

Supervisors

Prof. Helena Nevalainen

Dr. Angela Sun

MACQUARIE
University
SYDNEY·AUSTRALIA

A thesis submitted in fulfilment of the requirements for the degree of
Masters of Research
October 2016

“Imagination is the beginning of creation. You imagine what you desire, you will what you imagine and at last you create what you will.”

-George Bernard Shaw

“Let’s imagine the world without malnutrition and starvation. This research is dedicated to all those who are in hope to get proper nutrition. How about we try to solve the malnutrition by creating a better strain of Euglena that can produce required amount of nutrients.”

Table of Contents

ABSTRACT

Euglena gracilis is a unicellular protist currently used to produce nutraceuticals, cosmeceuticals and biofuel. Full exploitation of the organisms requires the development of genetic engineering tools. Chloroplast transformation using the non-native promoter (CaMV 35S) has been reported, but there is no published information about the use of native promoters for recombinant gene expression in *Euglena*.

The aim of this research was to isolate a strong constitutive promoter with a view to establish an expression vector for *E. gracilis*. Proteins produced during dark cultivation were separated using 2D SDS-PAGE and identified by mass spectrometry. One of the highly expressed proteins was glyceraldehyde 3-phosphate dehydrogenase (GAPDH). The *gapC* gene encoding GAPDH was identified using BLAST. The 5' UTR region of the gene was amplified by PCR using gDNA of *E. gracilis* as a template and inserted into a promoterless *E. coli* and yeast plasmids containing a reporter gene (*β-gal* or *eGFP*). Promoter activity was verified by successful expression of EGFP in yeast. Sequence analysis of the putative *Euglena* promoter revealed the presence of eukaryotic promoter elements like CAAT box, GC Box and myb-like transcriptional activator. Final confirmation of the promoter activity will be carried out in *Euglena* after a suitable transformation method has been established.

STATEMENT OF DECLARATION

This thesis is a presentation of my original research work carried out as part of the Master of Research program at Macquarie University and has not been submitted for a higher degree to any other university or institution.

I also certify that this thesis is an original piece of research conducted by me between January 2016 and October 2016 and it contains no material previously published or written. The contributions made by others are properly indicated with due reference to the literature and complete acknowledgement of any instruction or assistance.

Bishal Khatiwada

October 2016

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude and respect to Prof. Helena Nevalainen for giving me the opportunity to undertake the Master of Research program under her principal supervision. Without her endless support, kindness, and guidance this project would not have been possible.

I am highly indebted to my associate supervisor Dr. Angela Sun for her continuous support and assistance throughout the year. Her immense knowledge and supervision have made this project to achieve a fruitful outcome. Beside my supervisors, I would like to thank Dr Liisa Kautto for supplying me material with lab experiments and suggestions. Special thanks to Ms Mafruha Hasan for her guidance with proteins extraction and mass spectrometry works. I would also like to thank Dr Aneesh Pillai for his guidance in lab experiments and suggestions. Big thanks to my friend Mr Wisam Babar for his support throughout this project.

I would like to extend my gratitude to Dr Heinrich Kroukamp for his recommendations and materials for yeast transformation. Also, I would like to thank Mr Alexander Gissibl and Mr Dominik Kopp for their guidance. I am thankful to Zhiping Han, Anna Gryshyna, Hannah Hunt, Jodie Abrahams, Sameera Iqbal and Abdul Shathili for their company and advice in and outside the laboratory.

I wish to express my heart full gratitude to all my family members and my well-wishers for their kindness and support that always pushed me forward in my hard times. Massive thanks to my parents who always give me a positive vibe to pursue higher education and keep on telling me education is the only path to success. A very special thanks to my wife Ruby for her unconditional love and support who went through all my pains and suffering and at the same time stand strong enough to keep me on track.

Finally, I would like to thank all of them who, knowingly or unknowingly, helped me to complete my project.

Abbreviations

1D	One-dimensional gel electrophoresis
2D	Two-dimensional gel electrophoresis
ACN	Acetonitrile
bp	Base pairs
BLAST	Basic local alignment search tool
BPB	Bromo phenol blue
CDS	Coding DNA sequence
CHAPS	3 [(3cholamidopropyl) dimethylammnio]-1 propanesulfonate
dil	Dilution
DTT	Dithiothreitol
eGFP	Enhanced green fluorescent protein
ESI	Electrospray ionisation
EST	Expressed sequence tag
fig	Figure
g	RCF or G-force
gapC	Glyceraldehyde-3-phosphate dehydrogenase gene
gapC	Glyceraldehyde-3-phosphate dehydrogenase promoter
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase protein
IAA	Iodoacetamide
IEF	Isoelectric focusing
IPG	Immobilised pH gradient
kDa	Kilo Dalton
LC	Liquid chromatography
MCS	Multiple cloning sites
mg	Milligrams
min	Minutes
mM	Millimolar
MS	Mass spectrometry
MW	Molecular weight
MS/MS	Tandem mass spectrometry
NEB	New England Biolabs
nm	Nanometer
PCR	Polymerase chain reaction
PGK1	Phosphoglycerate kinase promoter
pI	Isoelectric point
sec	Second
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TFBS	Transcription factor binding sites
ura	Uracil
UTR	Untranslated region

V	Voltage
v/v	Volume per volume
w/v	Weight per volume

Chapter 1: Introduction

1.1 Overview of *Euglena gracilis*

Euglena gracilis is a unicellular protist that shares features characteristic for both plants and animals. Like plants, *E. gracilis* contains chloroplasts allowing it to obtain energy by autotrophy using sunlight, but it can also grow heterotrophically like animals utilising organic compounds as food (Ogbonna *et al.*, 2002). Based on the morphological and molecular evidence, *E. gracilis* belongs to kinetoplastids, a group of flagellated protozoans of which the molecular structure resembles that of the human pathogenic protozoa *Trypanosoma* and *Leishmania* (Henze *et al.*, 1995). Studies into the evolution of *Euglena* suggest a series of endosymbiotic events between various organisms such as eukaryotic green algae (Martin *et al.*, 1992), alpha-proteobacteria and protozoa which make organism complex to characterise (Yasuhira and Simpson, 1997).

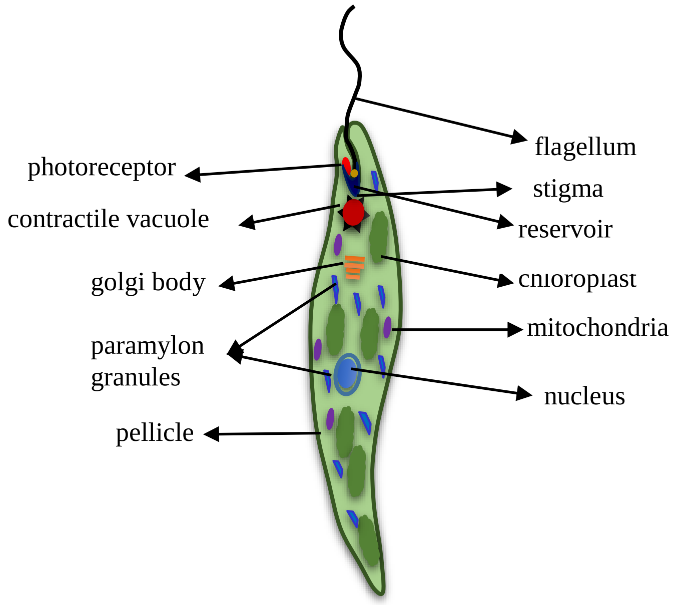


Fig 1-1: Diagrammatic representation of *E. gracilis*

E. gracilis possesses a flagellum for locomotion and an eyespot for photoreception (Fig 1-1). Though being a unicellular organism, *Euglena* possesses cell organelles similar to eukaryotes such as mitochondria, chloroplasts, vacuoles and Golgi (Kempner and Miller, 1968). These free-swimming organisms are mostly found in freshwater streams, lakes and ponds. They reproduce asexually by the binary longitudinal fission method during free-swimming flagellated form while multiple cell division is recorded during encysted condition (Tannreuther, 1923). *E. gracilis* has several features of interest in the industry. It can produce a wide variety of compounds such as vitamins, essential amino acids and fatty acids. Vitamins produced by *E. gracilis* include B (except B12), C, D, E and K (Krajčovič *et al.*, 2015). *E. gracilis* can store carbohydrates as a linear polysaccharide paramylon which is packed in granules. Paramylon is a β-1,3-glucan that contributes up to 85% of the dry weight of the organism (O'Neill *et al.*, 2015). Multiple health benefits have made paramylon a commercial product. Paramylon shows an immune stimulatory properties (Kondo *et al.*, 1992), has an anti-HIV effect (Koizumi *et al.*, 1993), cholesterol lowering properties (Šantek *et al.*, 2009) and anti-tumor activity (Watanabe *et al.*, 2013). It is used as a dietary supplement due to the similar features to the dietary fibres (Barsanti *et al.*, 2011). *Euglena* has been used as a tool to monitor ecotoxicological risk assessment by evaluating water quality by a change in growth condition, pigmentation, motility and DNA damage of the organism (Tahedl and Häder, 2001). *E. gracilis* can grow under high concentration of carbon dioxide, and it also possesses the capacity to remediate polluted water as this organism has been found to be growing in heavy metal contaminated water and dirty river water in mining zones (Krajčovič *et al.*, 2015). Polluted water can be used as a substrate for *E. gracilis* to grow and produce valuable lipids. Lipids, in turn, can be refined to generate biofuels (Krajčovič *et al.*, 2015). *Euglena* based products such as food tablets, beverages, cookies, noodles and cosmetics are currently being produced at the industrial level (http://www.euglena.jp). A considerable amount of research is currently directed towards the use of *Euglena* in biofuel production.

Klebs (1893) laid the foundation of *Euglena* research. After this, studies were undertaken in different strains of *Euglena*. The most studied strain of *E. gracilis* is the wild type Klebs strain Z isolated from fresh water by E.G. Pringsheim in 1950 (ATCC® 12894™). There are also variants of the *E. gracilis*: *Euglena gracilis* var. *bacillaris, Euglena gracilis* var. *zumsteini, Euglena gracilis* var. *fuscopunctata* amongst others. The current study was conducted using *Euglena gracilis* var. *saccharophila* which develops chloroplasts at a slower rate upon exposure to light and appear yellowish or white when grown in the dark, which will reduce the amount chloroplast-related proteins. This will support the approach for isolating a constitutive promoter; that would not be dependent on light. This strain can also metabolise glucose at a faster rate and thus shows better growth compared to the wild-type (Pringsheim, 1955), which will make it easy to culture.

1.2 Ongoing research and -omics tools developed for *Euglena gracilis*

Thus far no complete genome sequence or complete transcriptome data are available for *E. gracilis*. However, a complete chloroplast DNA sequence of the wild-type *E. gracilis* (Z strain) was reported already in 1993 (Hallick *et al.*, 1993). The size of *E. gracilis* genome is approximately two billion base pairs (2 Gbp) (O'Neill *et al.*, 2015). It contains a large number of repetitive sequences (O'Neill *et al.*, 2015). Also, it has a modified nucleotide base J (glycosylation and hydroxylation of thymidine base) that impedes genome sequencing because it will not allow the RNA polymerase to read throughout the DNA strand silencing the modified region (Borst and Sabatini, 2008, O'Neill *et al.*, 2015).

1.2.1 Chloroplast genome

The *E. gracilis* chloroplast genome consists of 143,170 base pairs, accommodating 96 genes. About 38% of the sequence features introns. Further on, there are genes encoding 16S, 5S, and 23S rRNAs. Genes encoding the synthesis of ribosomal proteins and photosynthesis-related polypeptides were discovered (Hallick *et al.*, 1993). The availability of a complete chloroplast genome sequence has opened the door for several investigations such as the origin of chloroplasts, how the photosynthetic apparatus has developed in a eukaryotic organism and the pattern of chloroplast gene evolution (Hallick *et al.*, 1993).

1.2.2 Expressed Sequence Tag (EST) Database

The length of expressed sequence tags (ESTs) ranges from 200 to 800 base pairs. As they represent an expressed portion of a genome, they are valuable for determining highly expressed gene (Parkinson and Blaxter, 2009). The Taxonomically Broad EST Database (TBestDB) features ESTs from more than 50 organisms. The majority of these ESTs have been created by the Protist EST Program, a coordinated effort among six Canadian research groups (O'Brien *et al.*, 2007). To date, nearly thirty thousand ESTs of *E. gracilis* have been deposited in the TBestDB database (O'Brien *et al.*, 2007). The deposited ESTs can be used to assess gene expression levels.

Since ESTs indicate a transcribed region of the genome, the information can be used to identify the most highly expressed genes in *E. gracilis*. Following the identification of highly expressed genes, further analysis can be carried out, for instance, to isolate promoters driving these highly expressed genes.

1.2.3 *Euglena* transcriptome

The first complete transcriptome of *E. gracilis* var. *saccharophila* was produced by O'Neill *et al.* (2015). Their study revealed that the transcriptome of *E. gracilis* contains genes homologous to prokaryotes, protozoa, animals, plants and fungi, which highlights the complex nature of the *E. gracilis* genome. The analysis exposed genes responsible for encoding proteins involved in the core metabolic and biochemical pathways for the synthesis of carbohydrates, lipids, amino acids and vitamins, and also genes for the synthesis of polyketides and non-ribosomal peptides. O'Neill *et al.* (2015) compared the transcriptomic data between samples grown in the light and dark. They showed that the number of proteins expressed in the cells cultivated in the light (22,814) was less than in the cells grown in the dark (26,738). This suggests that the expression of some proteins in *Euglena gracilis* var. *saccharophila* may be inhibited or down-regulated in the presence of light.

In 2016, Yoshida *et al.* (2016) published transcriptome dataset for *E. gracilis* (Z strain-wild type). The transcriptome was produced by growing the organism under aerobic and anaerobic conditions to understand the metabolic pathways for wax ester metabolism. RNA-Seq analysis was carried out to annotate the gene transcripts and to provide sequence information. Data analysis showed that 26,479 genes were considered to be potentially expressed of which 2080 transcripts were differentially expressed during anaerobic growth. Their analysis revealed that most of the differentially expressed genes were involved in photosynthesis, carbohydrate metabolism, oxidative phosphorylation, and nucleic acid metabolism.

Transcriptomic study of *E. gracilis* conducted by O'Neill *et al.* (2015) and Yoshida *et al.* (2016) pave the way for understanding the biochemical and metabolic pathways of this organism and engineering these to produce high amounts of valuable products at a lower cost. Data generated from these transcriptomes can be used to determine genes that are highly expressed in different cultivation conditions as a lead for the isolation of strong gene promoters active under desired conditions.

1.2.4 Central metabolites of *E. gracilis*

Comparative analysis and profiling of *E. gracilis* metabolites was originally published by Matsuda *et al.* (2011). Metabolic profiling was carried out with *E. gracilis* grown in aerated cultures in both light and dark conditions and anaerobic cultures in the dark (Fig 1-2). A change in the metabolic profile was observed during the shift from aerobic to anaerobic growth. The substantial reduction was noted in the amount of some glycolytic intermediates and amino acids while switching from aerobic to anaerobic growth resulted in a significant increase in wax ester production. This study suggests that metabolic pathways can be controlled for increase production of paramylon or wax ester by varying the level of the nitrogen supplement, type of the carbon source and the amount of aeration.

For example, paramylon production was highly favoured during aerobic growth in the dark, which suggests that genes responsible for the formation of paramylon were repressed in the presence of light (Fig 1-2). More proteins and vitamins were produced during aerobic growth in the light. Anaerobic growth in the dark was favourable only for wax ester formation. These results highlighted the fact that light was the major factor inducing degradation of paramylon which is one of the valuable products *E. gracilis* makes.

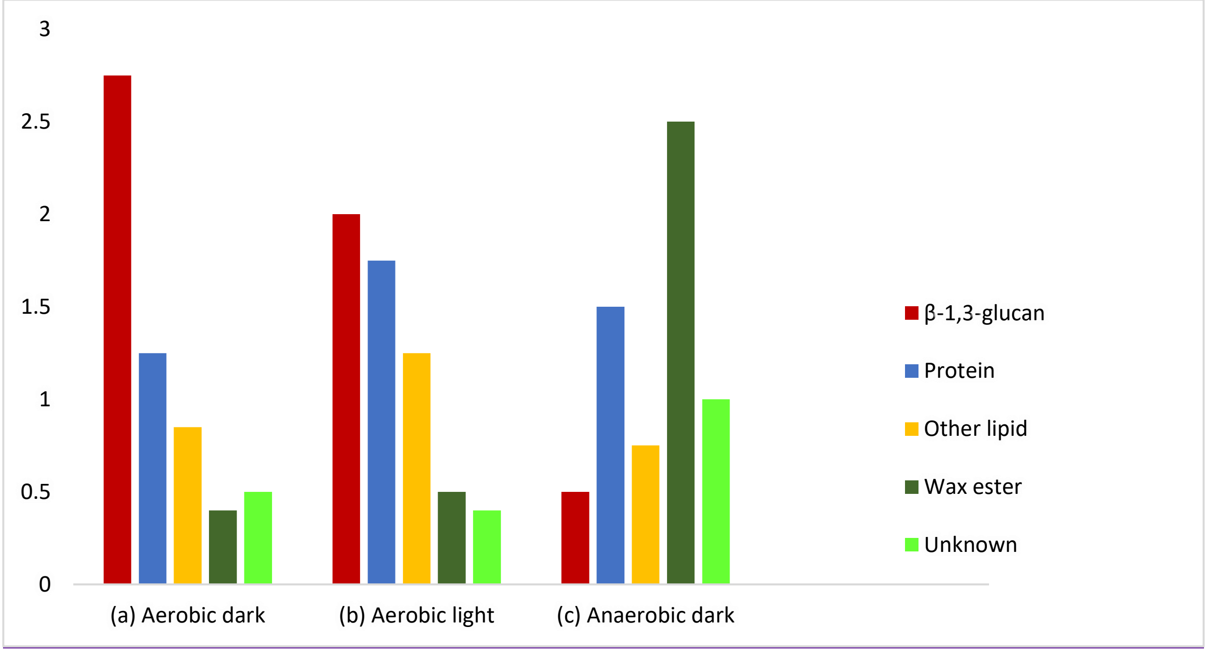


Fig 1-2: Cellular composition of *E. gracilis* growing in three types of culture conditions (a) aerobic dark, (b) aerobic light and (C) anaerobic dark (Matsuda *et al.*, 2011)

A study conducted by Barsanti *et al.* (2001) highlighted the effect of various carbon sources such as glucose, pyruvate, succinate, acetate, glutamate and ethanol on the production of paramylon. They used two different culture conditions (light and dark), and two strains (photosynthetic wild type and a non-photosynthetic mutant). Their findings suggested that the mutant strain lacking photosynthesis was less affected by light during paramylon degradation and shows slow degradation of paramylon during dark cultivation. The results agreed with the study of Matsuda *et al.* (2016) on that light triggers paramylon degradation. Thus, both studies provide valuable information for attempts to improve paramylon production in *Euglena*.

1.2.5 Chloroplast transformation of *E. gracilis* using non-native promoters

To date, there are no reports of utilisation of a native 'genomic' promoter to drive recombinant gene expression in *E. gracilis,* and all published transformation experiments have targeted the chloroplast. Chloroplast transformation was carried out by Doetsch *et al.* (2001) to insert *addA* transformation cassette (provide resistance to spectinomycin and streptomycin) using biolistic bombardment technique. An independently expressing *psbK* operon (*psbK-ycfl2-psaM-trnR*) was also inserted into the transformation vector. The *E. gracilis psbA* chloroplast gene promoter was used to express the construct in the chloroplast. The *psbK* operon also includes two group III introns and group III twintrons. The expression construct was intended to replace the gene via homologous recombination or autonomous expression of non-integrated, unrelated genes. The results revealed that the *psbK* operon was transcribed correctly, and the resulting pre-mRNA was spliced correctly. This study successfully analysed the intron splicing mechanism of *E. gracilis* chloroplast genes.

Another chloroplast transformation of *E. gracilis* was carried out by Ogawa *et al.* (2015). Cyanobacterial fructose-1,6/sedoheptulose-1,7-bisphosphatase *(FBP/SBPase*) was introduced successfully into *E. gracilis* chloroplasts by biolistic bombardment to enhance its photosynthetic ability. Successful transformation resulted in an increased level of biomass and wax ester production indicating enhanced photosynthetic ability (Ogawa *et al.*, 2015). Here a heterologous 35S promoter from the cauliflower mosaic virus (CaMV) was used to drive the expression of the *FBP/SBPase* gene. A native *E. gracilis* promoter capable of driving expression in a non-chloroplast environment would be more compatible and advantageous for increasing the level of production of compounds of industrial interest such as paramylon, vitamins, lipids and amino acids in *Euglena.*

1.3 Constructing a gene expression system for *E. gracilis*

To fully exploit *Euglena* as an industrial producer, there is a need to develop molecular tools that allow flexible genetic manipulation of the organism. There required components feature: the strain to be transformed, a method for introducing exogenous DNA into the cells, and importantly, an expression vector consisting of a promoter, multiple cloning sites(s), a selectable marker, and a transcription terminator.

An appropriate selection marker is needed to select the transformed colonies of *E. gracilis*. Previously, Doetsch *et al.* (2001) used genes encoding spectinomycin and streptomycin as an antibiotic selection marker for *Euglena* chloroplast transformation whereas Ogawa *et al.* (2015) used neomycin phosphotransferase II (*NPT II*) gene. Hygromycin resistance gene (*hph*) has been used as a selection marker for microalgae like *Chlorella* (Chow and Tung, 1999) and *Chlamydomonas* (Kumar *et al.*, 2004) for nuclear transformation. As *E. gracilis* share the features similar to microalgae, hygromycin may be a suitable selectable marker for nuclear genetic transformation. Transcription terminator sequences can be sourced from the 3’ UTR regions of identified *E. gracilis* genes.

After constructing an expression cassette, there should be a suitable transformation system to allow the introduction of the gene of interest into *E. gracilis*. Transformation approaches such as glass bead agitation, biolistic bombardment and electroporation have been used to transfer foreign DNA into microalgae (Hallmann, 2007). From these, glass bead agitation has emerged as the preferred technique (Kindle *et al.*, 1991, Economou *et al.*, 2014). Chloroplast transformation of *E. gracilis* has been successfully carried out using biolistic bombardment (Ogawa *et al.*, 2015, Doetsch *et al.*, 2001). Past reports on chloroplast transformation (Gong *et al.*, 2011, Boynton *et al.*, 1988, Sanford, 1990) have featured biolistic bombardment as the preferred method to penetrate the rigid outer cellular membrane to bring in exogenous DNA. The biolistic bombardment has been found successful in both chloroplast and nuclear transformation especially for plants, algae, yeast and fungi (Armaleo *et al.*, 1990, Boynton *et al.*, 1988, Daniell, 1997). *Agrobacterium*-mediated transformation has also been found successful in microalgae (Kumar *et al.*, 2004). Transformation using *Agrobacterium* can probably be applied to *E. gracilis* as well.

1.4 Gene promoters

Identification of a suitable promoter is a key step towards assembling the components of the expression vector. Promoter not only provides a signal for gene transcription but also promotes the expression of the gene. The promoter contains multiple cis-acting elements that function as precise binding locations for proteins that have a role in the regulation of the transcription. Different types of promoters are used for genetic manipulation and could be categorised into following types: (1) Constitutive promoters - these promoters are always turned on and are not affected by environmental factors; (2) Tissue-specific or stage-specific promoters- these promoters are only expressed during certain stages of development and in certain tissues; (3) Inducible promoters- these promoters are regulated by chemical or physical environmental factors, and (4) Synthetic promoters- these promoters are created using the defined regulatory elements such as the core promoter and the synthetic motifs for transgene expression (Hernandez-Garcia and Finer, 2014).

In the context of this research, a strong constitutive promoter would work well. Identification and isolation of new promoter will pave a way to build an expression vector system for *Euglena*, which is required to make *E. gracilis* an industrially workable organism. An expression vector system will allow improvement and modification of genes and metabolic pathways to produce vitamins, amino acids, paramylon and other products of interest. Further, the gene promoter should be of non-chloroplast origin so it would not be light-dependent (Sexton *et al.*, 1990). Culturing in the dark may be the best approach to help minimise interference of the proteins involved in photosynthesis. A strong constitutive promoter independent of light would suit our applied goal to increase paramylon production, as in previous research, Matsuda *et al.* (2011) have shown that paramylon production is favoured in the absence of light.

1.4.1 Promoter elements

A core promoter is characterised by the presence of conserved sequence(s) upstream of the transcription initiation site. In prokaryotes, the most frequently occurring elements are TATAAT at -10 and TTGACA at -35. The TATAAT box is also known as Pribnow box (Pribnow, 1975). The promoters in eukaryotes are more complicated and diverse than those in prokaryotes (Bucher, 1990).

Different sequence motifs of eukaryotic promoters are: (1) TATA box- binding site for general transcription factors (Smale and Kadonaga, 2003); (2) INR box- binding site for transcription Factor II D (Smale and Baltimore, 1989); (3) BRE- binding site for transcription factor II B (Lagrange *et al.*, 1998); (4) CCAAT box- binding site for general transcription factors (Mantovani, 1998); (5) GC box- functions as an enhancer sequence and stabilizes the transcription pre-initiation complex.

1.5 Promoter isolation approaches

The promoter of a gene is located upstream of the transcription initiation site. There are several options for isolation of gene promoters. Genome walking PCR is a commonly used approach to scan the chromosomal DNA for gene promoters and terminators walking upstream or downstream of a gene (Rishi *et al.*, 2004, Zhang and Gurr, 2000, Guo and Xiong, 2006). While genome sequencing can also identify gene promoters, it may prove complicated depending on the nature of the genome and costs involved. Towards this end, there is no available genome sequence for *E. gracilis*. In the absence of a predetermined gene target, one can start with proteomics to identify highly expressed protein(s) functional in particular growth conditions and work backwards to the corresponding DNA sequence.

A suitable method to separate and identify (highly) expressed proteins is to analyse a protein sample on one- dimensional (1D) gel electrophoresis (Weber and Osborn, 1969) or 2-dimensional (2D) gel electrophoresis (Wang *et al.*, 2003). After locating highly expressed proteins in the gel, the spots can be analysed using mass spectrometry to obtain protein identity and the amino acid sequence (Aebersold and Mann, 2003). The amino acid sequence can be back-translated into DNA sequence which enables the design of primers for the isolation of the corresponding gene and its promoter and terminator sequences by chromosome walking PCR. This method has been used successfully for the isolation of the *Trichoderma reesei hex1* gene and its regulatory sequences (Curach *et al.*, 2004). Alternatively, if there are existing transcriptomic data (e.g. ESTs), these can be used to identify highly expressed genes and thereby the proteins they encode. For example, information from the TBestDB database can be matched with the 2D gel protein identifications to validate the findings. Several PCR methods have been proposed to determine the unidentified region flanking a known DNA sequence of the genome such as a promoter and terminator plus DNA insertion or deletion sites. Following methods of genome walking are suitable for isolating both the gene promoter and transcription terminator.

1.5.1 Inverse PCR for genome walking

This genome walking technique was initially proposed by Ochman *et al.* (1988). The method allows the isolation of both upstream and downstream regions of a known DNA sequence (gene). This technique was independently further developed by two groups, Triglia *et al.* (1988) and Silver and Keerikatte (1989).

The first step in this method is digestion of the genomic DNA (Fig 1-3) with a restriction endonuclease which will produce fragments with sticky ends and that are small enough for PCR amplification (usually <1 kb). The digested products are then allowed to self-ligate to make a circular structure. Following the ligation, PCR will be carried out using a pair of primers complementary to the known region of the genome. Primers should be placed outwardly facing the known region (Fig 1-3).

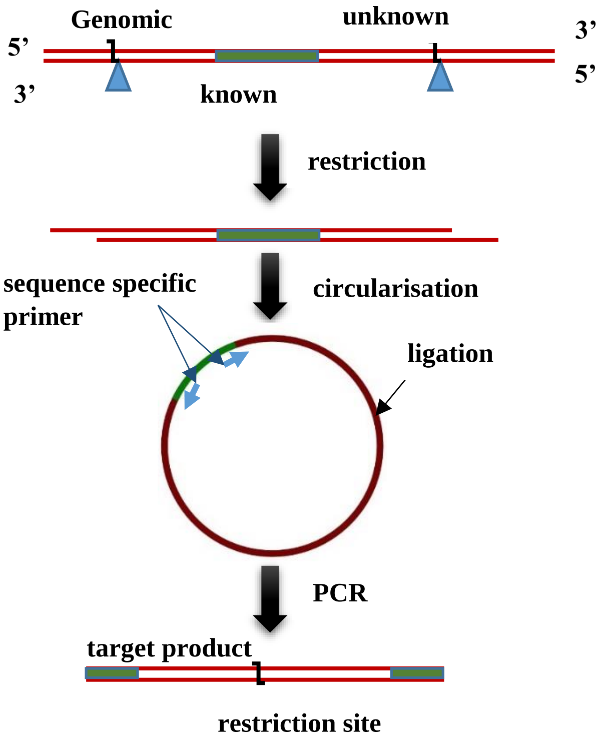


Fig 1-3: Procedure for inverse PCR. Restriction digestion sites are presented by triangle arrows. DNA fragments become circular after ligation. Performing PCR using sequence specific primers facing outward (blue arrows) allows amplification of both the upstream and downstream regions. Adapted from Ochman *et al.* (1988) and Tonooka and Fujishima (2009).

1.5.2 Cassette PCR

Genome walking using cassette PCR involves restriction digestion of the genomic DNA and then ligation of adapters (Rosenthal and Jones, 1990). PCR is performed using sequence specific and adapter specific primers (Fig 1-4). This approach was developed independently by various groups, but the leading principle is the same with some modifications to increase its efficiency. Mueller and Wold (1989) termed this method as “ligation-mediated PCR” whereas Rosenthal and Jones (1990) called it cassette PCR.

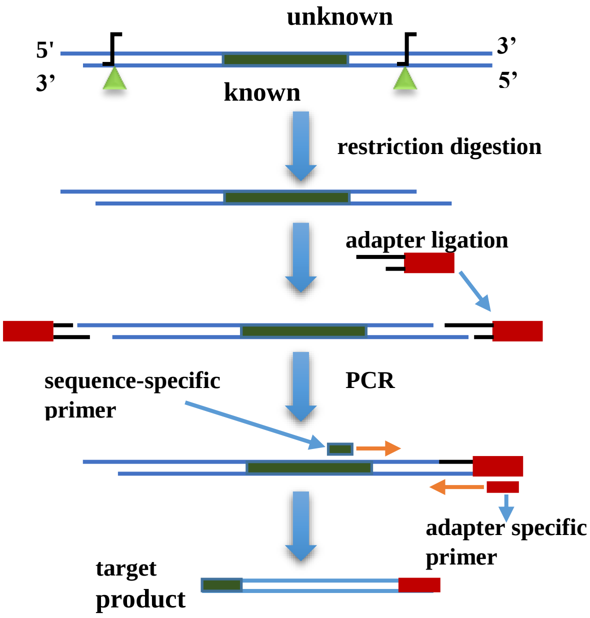


Fig 1-4: Procedure of cassette PCR. Restriction digestion sites are shown by the green triangle arrows. Ligation of the adapter is demonstrated by the small red rectangular box. PCR is performed using a sequence specific primer facing outwards and an adapter-specific primer facing inwards which allow amplification of either upstream or downstream regions that depend on the orientation of primers. Adapted from Tonooka and Fujishima (2009) and Rosenthal and Jones (1990).

The techniques discussed above have their pros and cons. For example, genome walking is valuable if the genome sequence of the organism is unknown. In the absence of an annotated genome sequence, the proteomic approach provides an alternative option for isolating a gene promoter functional under defined conditions. The availability of transcriptomic data offers an avenue to pursue identification of a promoter of a highly expressed gene, which then can be sourced from the organism of interest.

1.6 Research objectives

Euglena gracilis is a promising candidate for research and development because of its capability to produce industrially valuable compounds such as paramylon, vitamins, amino acids and lipids. To boost the level of production of these compounds, there is a need to identify strong gene promoters that can drive the expression of genes associated with the compounds of interest. Hence, the overall objective of this study was to isolate a strong constitutive promoter from the *Euglena gracilis* nuclear genome. To accomplish this, the following objectives were set:

- To identify highly expressed *E. gracilis* proteins using a proteomic approach:
    - Separation and visualisation of proteins using 1D and 2D SDS-PAGE.
    - Identification of highly expressed proteins produced under cultivation in the dark using nano-LC-ESI-MS/MS mass spectrometry.
    - Working backwards from the amino acid sequence of a chosen highly expressed protein to identify the gene corresponding to the protein.
- To isolate the promoter from the gene identified above.
- To express a reporter gene in *E. coli* or *S. cerevisiae* under the newly isolated *E. gracilis* promoter to validate its functionality.

The work contributes to a future goal for and assembling an expression vector for *E. gracilis* for genetic engineering of the organism.

Chapter 2: Materials and methods

2.1 Culture medium and cultivation

In this study, *E. coli* and yeast were cultivated on both solid and liquid media. For *E. gracilis* only liquid broth (algal) medium was required. All ingredients for the preparation of the culture media were purchased from Sigma-Aldrich, Australia.

2.1.1 *E. gracilis* culture medium and cultivation

Culture medium for *E. gracilis* was prepared using a modified recipe by Rodríguez-Zavala *et al.* (2010) based on the original medium described by Hutner *et al.* (1956).The base medium contained (per litre): 0.2 g $CaCO_3$, 0.4 g $(NH_4)_2HPO_4$, 0.2 g $(KH_2)PO_4$ and 0.5 g $MgSO_4$. Two ml of trace mineral stock A (composed of 2.2 g $ZnSO_4·7H_2O$, 2 g $MnSO_4·4H_2O$, 0.5 g $Na_2MoO_4·2H_2O$, 0.04 g $CoCl_2·6H_2O$) and one ml of trace mineral stock B (consisting of 0.078 g $CuSO_4·5H_2O$, 0.057 g H_3BO_3) were added to the base medium. As a source of nitrogen, 15 g (1.5%) of yeast extract was added to the medium. The pH of the mixture was adjusted to 3.5 before sterilisation at 121˚C for 20 min. After the medium had cooled down, 20% glucose (40 g glucose dissolved in 200 ml ddH_2O; autoclaved separately at 121˚C for 20 min), 1% of vitamin B1 stock (10 mg thiamine, 50 mg $FeCl_3$ dissolved in 10 ml ddH_2O; filter sterilised) and 0.1% of vitamin B12 (10 mg vitamin B12 dissolved to 2 ml ddH_2O; filter sterilised) stock were added to the medium.

One ml of *E. gracilis* stock (about $2×10^7$ cells) was used to inoculate 150 ml of the culture medium. The cultures were grown in the absence of light at 24 ± 1˚C in an incubator with continuous shaking (150 rpm) for 72 h for the culture reach a logarithmic growth phase, before starting the protein and genomic DNA extraction. Cell density was measured using NanoDrop 2000 (Thermo Scientific, USA) at OD600. Cultivations were conducted in the dark to minimise the expression of chloroplast-related proteins.

2.1.2 Bacterial culture medium and cultivation

Luria-Bertani (LB) broth and agar plates were prepared to culture bacterial cells. To make LB broth, 5 g tryptone, 2.5 g yeast extract and 5 g NaCl were added to 500 ml of ddH_2O (Luria and Burrous, 1957). To prepare 100 ml of super optimal broth (SOB), 0.5 g yeast extract, 0.058 g NaCl, 0.02 g KCl, 2 g tryptone, 0.2 g $MgCl_2·6H_2O$, 0.25 g and $MgSO_4$ were added to 100 ml of ddH_2O (Hanahan, 1983). Two ml of glucose stock (20% (w/v) was added to the pre-sterilised SOB solution to make super optimal broth (SOC) for catabolite repression studies. To prepare the LB-agar plates, 7.5 g of agar was added to 500 ml of LB medium. Sterilisation of media was performed by autoclaving at 121 ºC for 20 min. Fifty ml of culture liquid broth was grown in 250 ml Erlenmeyer flask continuously shaking at 200 rpm. All cultures were incubated at 37 ºC.

For the blue-white screening of *E. coli* transformants, X-Gal plates were prepared by mixing 2.5 ml of X-Gal stock containing 20 mg/ml of X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) dissolved in dimethylsulfoxide (DMSO), 2.5 ml of IPTG stock containing 238 mg of IPTG in 10 ml of ddH_2O and 250 µl of the antibiotic kanamycin (50 µg/ml). The prepared mixture was added to 500 ml of pre-autoclaved (121 ºC for 20 min) agar medium (1.5% w/v).

2.1.3 Yeast culture medium and cultivation

Two types of media were prepared for yeast cultivation and selection of transformants following the protocols by Xiao (2006). For the yeast culture and propagation of yeast plasmids, yeast extract-peptone-dextrose (YPD) broth and YPD agar medium were prepared. To make YPD broth, 10 g yeast extract, 20 g peptone and 20 g dextrose were added to the one litre of ddH_2O. To make YPD agar medium, 15 g of agar was added to the one litre of YPD broth. For the selection of yeast transformants, yeast drop-out agar medium without uracil was made by adding 20 g of bacto agar, 1.92 g of yeast synthetic drop-out medium and 6.8 g of yeast nitrogen base without amino acids in one litre of ddH_2O. The medium was autoclaved at 121 ºC for 20 min, after which sterile glucose was added to the final concentration of 2% (w/v). All cultures were incubated at 30ºC. Liquid culture cultivation was carried out in a shaking incubator at 250 rpm.

2.2 Strains and plasmids

The *Euglena gracilis* var. *saccharophila* B752 strain was obtained from UTEX Culture Collection of Algae (UTEX Number: 752). The chloroplasts of this strain do not develop if grown in the absence of light but does so at a slower rate upon exposure to the light.

The *E. coli* DH5α strain (Hanahan, 1985) was available in house. *E. coli* DH5α was used for both plasmid propagation and as a transformation host. The promoterless pSF-PromMCS-BetaGal (OG372) plasmid was purchased from Oxford Genetics. This plasmid also contains the *beta-gal* reporter gene which allows checking for the functionality of the newly identified promoter.

The *Saccharomyces cerevisiae* mutant strain BY4742 (MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0), an auxotrophic strain (uracil deficient) and the plasmid pRS426 were kindly provided by Dr Henerich Kroukamp, SynBio Yeast 2.0 group, Department of Chemistry and Biomolecular Sciences, Macquarie University. The plasmid contains the phosphoglycerate kinase (*PGK1*) promoter, enhanced green fluorescent protein (*eGFP*) reporter gene and the uracil (*URA3*) marker gene.

2.3 Antibiotics used for the selection of transformants

In order to select recombinant *E. coli* colonies transformed with the pSF-PromMCS-BetaGal (OG372) plasmid, kanamycin was used on LB agar plates at the concentration of 50 µg/ml. Ampicillin at the concentration of 100 µg/mL was used to select recombinant *E. coli* transformed with pRS426.

2.4 Extraction of total intracellular proteins of *E. gracilis*

Protein extraction was carried out according to procedures mentioned by Baumer *et al.*, 2001. An aliquot of 5 ml *E. gracilis* cells (3×10^7 number of cell per ml) was harvested and centrifuged at 1500 g for 5 min. The supernatant was discarded, and the pellet washed twice with ice cold ddH_2O. The washed pellet was resuspended in a lysis buffer containing 50 mM Tris-HCl, pH 7.4, 250 mM sucrose, 3 mM EDTA, 0.04% (v/v) β-mercaptoethanol and 100 µl Sigma Protease Inhibitor Cocktail (Product Number P 8465). All procedures were carried at 4˚C on the ice. The sample mix was sonicated for five times at 25% of maximum amplitude (10 s sonication and 30 s rest) using a Branson Digital Sonifier® 450 with a probe tip diameter of 3 mm. The sonicated sample was centrifuged at 1500 g for 5 min and supernatant transferred to a fresh tube.

Protein precipitation was performed using a mixture containing ddH_2O/methanol/chloroform at a 3:4:1 ratio. The mixture was added to the supernatant and vortexed vigorously followed by centrifugation at 10,000 g for 2 min to separate the aqueous and organic phases. The upper phase (ddH_2O and methanol) was discarded without disturbing the protein pellet. Ice cold methanol (4 x to the volume of supernatant) was added to the remaining solution followed by vigorous vortexing. The sample was then centrifuged at 10,000g for 2 min. The supernatant was discarded and the pellet air-dried for 10 min. The protein pellet was resolubilised in 200 µl of 5% (w/v) sodium dodecyl sulphate (SDS). Protein concentration was measured by the bicinchoninic acid assay (BCA) (Smith *et al.*, 1985) as per the manufacturer's protocol (Thermo Scientific Pierce, Australia).

2.5 Visualisation of highly expressed proteins

The extracted proteins were visualised using both one-dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis (1D SDS-PAGE) and the two-dimensional sodium dodecyl sulphate polyacrylamide gel electrophoresis (2D SDS-PAGE).

The 1D SDS-PAGE was carried out using NuPAGE™ Novex™ 10% Bis-Tris protein gels (Thermo Scientific, Australia) under denaturing conditions (in the presence of SDS and DTT). Protein samples for gel loading were prepared by mixing 2.5 µl of protein, 2.5µl NuPAGE® LDS loading buffer, 1 µl 1 M DTT and 4 µl ddH_2O. The mixture was heated at 70 °C for 10 min in a heating block in an Eppendorf tube. Novex® Sharp Pre-stained Protein Standard (Invitrogen, Australia) was used as the molecular weight standard, and the gel was run at 200 V for 50 min in a buffer containing 3-(N-Morpholino) propane sulfonic acid (MOPS). The gel was stained with Coomassie Brilliant Blue stain (Bio-Rad, Australia) solution overnight and then destained with 1% acetic acid (v/v) solution overnight.

Protein samples were desalted using Zeba™ Spin Desalting columns (Thermo Fisher Scientific) to perform the 2D SDS-PAGE analysis. Protein fractionation in the first dimension was carried out using immobilised pH gradient (IPG) strips of 11 cm, pH (3-10) (ReadyStrip™ Bio-Rad). A total of 35 µl of the protein sample (approximately 500 µg protein) was mixed with 1.1 µl of IPG buffer ampholytes, 4.4 µl of 1M DTT, 2.5 µl of bromophenol blue (BPB) stock (1% w/v) and 177 µl of sample rehydration buffer [7 M urea, 2 M thiourea, and 4% (w/v) CHAPS in 40 ml ddH_2O]. The sample solution was mixed thoroughly and centrifuged at 10,000 g for 5 min. From the supernatant fraction, 250 µl was transferred into another tube for rehydration of IPG strip. An IPG strip facing the gel side down was placed onto the rehydration tray containing 250 µl of the rehydration solution. Two ml of mineral oil was used to cover the strip to prevent it drying out. Active rehydration of IPG strips was performed supplying a low voltage of 50 v for 10 h. After completion of IPG rehydration, the strip was ready to undergo IEF focusing. The proteins were focused at a maximum of 8000 V for 35,000 volt-hours (V-h) at RT. After IEF completion, the strip was removed from the mineral oil and equilibrated for 10 min on a Bio-Rad re-equilibration tray. Initially, 10 ml of the re-equilibration solution 1 (50 M TRIS, 6 M urea, 30% (v/v) glycerol, 2% (w/v) SDS, 2% (w/v) DTT and 400 µL of 1% (w/v) BPB stock added in 200 ml ddH_2O) was used to re-equilibrate the strip. Next the solution 1 was decanted and the re-equilibration solution 2 that contained iodoacetamide 4.5% (w/v) instead of DTT.

For separation in the second dimension, the IPG strip was pushed down the well of a Criterion™ TGX™ (Tris-Glycine eXtended) precast gel until it was flat against the bottom of the well. Agarose sealing solution (0.5% w/v) was pipetted on top of the strip to seal it into place. The Novex® Sharp Pre-stained marker was used as the protein size marker. The reservoir on top of the gel was filled with 1 x running buffer (Tris 5% (w/v), Glycine 2% (w/v) and SDS 1.5% (w/v). Initial electrophoresis was performed at a low voltage (~60 volts) until the bromophenol blue dye front reached the gel. The running voltage was then increased to 160 volts for 1 h. The gel was then stained according to colloidal Coomassie staining protocol by (Candiano *et al.*, 2004).

2.6 Mass spectrometry analysis of proteins

Mass spectrometry was employed in this study to identify the proteins based on their peptide masses determined by the mass to charge ratios of their ions. In this study, Liquid chromatography coupled with electrospray ionisation-tandem mass spectrometry (LC-ESI-MS/MS) was performed using the Thermo Scientific™ Q Exactive™ mass spectrometer housed at APAF Ltd - Australian Proteome Analysis Facility, Macquarie University. The following steps were carried out for the peptide mass fingerprinting (PMF). The procedure to run mass spectrometer was assisted by Ms Mafruha Hasan.

2.6.1 In-gel tryptic digestion of proteins and peptide extraction

In-gel digestion and peptide extraction were carried out according to the method described by Rosenfeld *et al.* (1992). From the 2D gel, 15 large spots were excised manually with a scalpel. Excised gel pieces were cut into several pieces and put into separate fresh Eppendorf tubes

De-staining of Gel Pieces

Chopped gel pieces were destained by washing briefly with 200µl NH_4HCO_3 followed by addition of 200 µl 50% Acetonitrile/50% 100 mM NH_4HCO_3 into each tube. The tube was vortexed and incubated for ten min. The liquid was removed, and gel pieces were rewashed twice to make sure the gel pieces were completely de-stained. Following detaining, 50µl of 100% acetonitrile was added to dehydrate the gel pieces and the mixture was vortexed followed and incubated at room temperature for 5 min incubation. Acetonitrile was then removed, and the gel pieces were air-dried in the fume hood for 10 min.

Reduction and Alkylation

Gel pieces were covered with 50 µl 10 mM DTT in 50mM NH_4HCO_3**.** The reduction was done for 20 min in 37°C incubator. DTT was removed, and 50 µl of 55mM Iodoacetamide in 50 mM NH_4HCO_3 was added followed by incubation for 25 min in the dark at room temperature (RT). Iodoacetamide was discarded followed by washing of gel pieces with 100 µl of NH_4HCO_3 for 5 min with vortexing, to which 100 µl of 100% acetonitrile was added. The liquid was removed after 5 min incubation, and the gel pieces were washed with 50 µl 100 mM NH_4HCO_3 for 10 min, then twice with 200µl 50% acetonitrile / 50% 100 mM NH_4HCO_3 for 10min. Gel pieces were dehydrated with 100 µl of 100% acetonitrile for 5 min. The liquid was removed, and gel pieces were dried.

In-gel tryptic digestion

Trypsin solution was prepared as follows: 20 µl of the resuspension buffer provided by Promega was added to the one vial of Promega Sequencing Grade Trypsin, containing 20 µg of trypsin to make a 1 µg/µl solution. 1580 µl of 50mM NH_4HCO_3 was added to the solution to the enzyme suspension to make a 12.5 ng/µl solution. Freshly prepared trypsin solution (30 µl) was added to the tube containing gel pieces and allowed for 30 min gel rehydration at 4 °C (on ice). Overnight digestion was performed at 37 °C. On the following day, the supernatant was transferred into clean 0.5ml Eppendorf tubes. To the gel pieces, 50 µl of 50% acetonitrile / 2% formic acid was added and incubated for 25 min. The supernatant was removed and combined with the initial digest solution supernatant. Samples were vacuum dried in a speed vac to reduce the volume to 10 µl.

Peptide extraction and cleanup

A ziptip packed with C18 resin (Merck Millipore) was used for desalting and concentrating samples before MS. To equilibrate column, 10 µl of 90% (v/v) acetonitrile/0.1 % trifluoroacetic acid (TFA) was added to column consecutively for five times. Aliquots of 10 µl of each sample were withdrawn and dispensed into the column consecutively for ten times to ensure that the peptides have been fully bound to the resin. Bound peptides were washed with 0.1 % TFA for five times to remove unwanted salts and contaminants. Twenty µl of 70% ACN/0.1% TFA was used twice to elute peptides. Samples were vacuum dried and reconstituted in 20 ml of 2% FA and centrifuged at 15,000 *g* for 15 min. The top 10ml of the peptides was used for mass spectrometry (nanoLC ESI MS/MS).

2.6.2 Liquid chromatography-electrospray ionisation- tandem mass spectrometry (nanoLC ESI MS/MS)

The nanoLC ESI MS/MS analysis was performed as described by Thompson *et al.* (2015). The peptide mixtures were analysed by nanoLC ESI MS/MS (Thermo Scientific™ Q Exactive™). Easy Nano-LC 1000 (Thermo Scientific) was used to separate each fraction over a 60 min gradient. Ten ml of sample was injected onto a peptide trap column (3.5 cm x 100 mm) packed in-house with Halo C18 (Advanced Materials Tech) with a particle size of 2.7 µm and desalted with 20 ml of Buffer A (0.1% FA). The peptide trap was then switched online with an in-house packed reversed-phase column (10 cm x 75 mm) with Halo C18. Elution of peptides was performed using a linear solvent gradient as follows: 1 to 50% of Buffer B [99.9% (v/v) ACN, 0.1% (v/v) FA] for 50 min, 50 to 85% of Buffer B for 2 min, holding at 85% for 8 min with a flow rate of 300 ml/min across the gradient. The column eluate was directed into a nanospray ionisation source of the mass spectrometer. A 2.0 kV electrospray voltage was applied via a liquid junction upstream of the C18 column. Spectra were scanned over the range 350 to 2000 amu. Xcalibur software (version 2.06) automatically recognises peaks with dynamic exclusion and tandem MS of top ten precursor ions at 30% normalisation.

2.6.3 Database searching for protein identification

Data were analysed using the Global Proteome Machine (GPM) software (GPM Fury, version 3.0) (http://www.thegpm.org/) against the X! Tandem algorithm. The data files were first converted from raw to mzxml using the readW.exe converter. All data obtained from the 15 samples were merged into one directory, and GPM-XE was used to run a single directory search. Tandem mass spectra were searched against a *Euglena* sequence database compiled with annotated *Euglena* protein and EST (http://tbestdb.bcm.umontreal.ca/searches/organism) sequences from NCBI, JIC (http://jicbio.nbi.ac.uk/euglena/) and the TbestDB databases along with common human and trypsin peptide contaminants. The following parameters were used for peptide analysis: fragment mass error 0.4 Da; cleavage site trypsin and a tolerance up to two missed tryptic cleavages. Reverse sequence searches were performed to calculate false discovery rates (FDR); potential modifications include methionine oxidation and cysteine carbamidomethylation. Protein names and IDs were obtained based on sequence homology using the BLAST tool in UniProt (http://www.uniprot.org/blast/).

2.7 Isolation of high molecular weight DNA from *E. gracilis*

DNA extraction was based on the method described by (Mederic *et al.*, 1987). Thirty ml of *E. gracilis* cell culture grown for five days was collected (approx. 3×10^8 of cells) by centrifugation at 2,500 g for 5 min. Step one involved weakening of the cell pellicles, for which the cells were suspended in 30 ml of glycerol preservation buffer (l0 mM Tris-HCl- pH 7.4, 6 mM $MgC1_2$, 0.13 mM $MnC1_2$, 5 mM $Na_2S_20_5$, and 70% v/v glycerol) pre-chilled at -20 ˚C. The cell suspension was transferred into 50 ml falcon tubes and pelleted by centrifugation at 2,500 g for 5 min. Cells were stored at -20 ˚C for 24 hours.

Step two involved chromatin decondensation for which the cell suspension was kept at -20 ˚C in the preservation buffer. Cells were first diluted with two volumes of chromatin decondensation buffer (1 mM Tris-HC1 (pH8), 10 mM EDTA, 0.5 mM DTT, and 12.5% glycerol v/v) and centrifuged at 4 ˚C at 2500 g for 5 min. Chromatin decondensation was repeated with the addition of 10 ml of decondensation buffer with continuous shaking at 0 ˚C for 30 min. After incubation, the cell suspension was centrifuged at 2,500 g for 5 min. The third step involved cell lysis and nucleoprotein dissociation. The previously decondensed cells were resuspended in 30 ml of a dissociation buffer made up of 2% (w/v) sodium N-lauroyl-sarcosinate (SLS), 0.1 M EDTA, and 0.5 mM DTT dissolved in 200 ml saline sodium citrate (SSC) buffer that contained 0.15 M NaCl and 0.015 M sodium citrate. Fifty µg/ml of proteinase K was added to each tube and allowed to achieve dissociation for two hours at 0 ˚C. The supernatant was collected in 50 ml falcon tubes after centrifugation at 8,000 g for 10 min.

Precipitation of DNA was carried out as described by Greco *et al.* (2014). To each Falcon tube containing 15 ml of the supernatant collected above, 1/9 sample volume of 100% ethanol and 1/4 sample volume (v) of 3 M potassium acetate (pH 4.8) were added. The solution was mixed gently by inverting and one volume of chloroform: isoamyl alcohol (24:1, v/v) was added and vortexed for 1 min. The tubes were kept at 4 ˚C with continuous shaking for 30 min and then centrifuged at 8,000 g for 20 min at 4 ˚C. Six millilitres of the aqueous layer were collected to which 0.3V of 100% ethanol was added and the tube vortexed vigorously for 1 min. After this, the tubes were incubated at 4 ˚C with gentle shaking for 20 min and then centrifuged at 8,000 g for 20 min at 4 ˚C. The aqueous phase was collected, and 0.1 V of 3M sodium acetate (pH5.2) and 0.8 V of isopropanol were added and mixed gently by inverting the tube. The tubes were stored at -80 ˚C for an hour and centrifuged at 8,000 g for 30 min at 4 ˚C. The supernatant was discarded and the pellet was washed twice with 1 ml of 75% ethanol followed by centrifugation at 8,000 g for 20 min at 4 ˚C. The supernatant was discarded, and the DNA pellet was dried in a fume hood for 10 min followed by resolubilisation with nuclease free ddH$_2$O.

DNA concentration was determined using the NanoDrop 2000 (Thermo Scientific, USA) and agarose gel electrophoresis 0.9% (w/v) agarose gel at 100 V for 60 min.

2.8 Isolation of the 5'untranslated region (UTR)- promoter region by polymerase chain reaction (PCR)

PCR was used to isolate the 5' untranslated region (UTR), 1015 bp in length, of the highly expressed gene glyceraldehyde-3-phosphate dehydrogenase (*gapC*) based on the *gapC* sequence available on the European Bioinformatics Institute (EMBL-EBI) database deposited by Henze *et al.* (1995). (http://www.uniprot.org/uniprot/Q43311, sequence ID L39772.1) Primers were designed using SnapGene 3.2.1 (http://www.snapgene.com)

Primers ecoligapcFWD and ecoligapcREV (see below) were designed to isolate the 5'UTR region of *gapC* gene and incorporate restriction sites *SpeI* (highlighted in red) and *XmaI* (highlighted in blue) onto the 5'UTR region for insertion into the *E. coli* plasmid (OG372, Appendix 3).

ecoligapcFWD: 5' CGACCAACTAGTAAGCTTGGGAGAAACTATCC 3'
ecoligapcREV: 5' GCCCTACCCGGGTAGATATATCTGGAGTTGGG 3'

Primers scerevisiaegapcFWD and scerevisiaegapcREV (see below) were designed to isolate the 5'UTR region of *gapC* gene and incorporate restriction sites *XhoI* (highlighted in yellow) and *PacI* (highlighted in green) for insertion into the *S. cerevisiae plasmid* (pRS426, Appendix 4)

scerevisiaegapcFWD: 5' CCCCCCCTCGAGAAGCTTGGGAGAAACTATCC 3'
scerevisiaegapcREV: 5' AACCATTTAATTAACGATATATCTGGAGTTGGGTG 3'

PCR amplification was carried out on the Eppendorf MasterCycler EP S thermal cycler. All primers used in this study were purchased from Macrogen, Inc. (South Korea). AmpliTaq Gold® DNA Polymerase from Thermo Fisher Scientific Inc. was used in all PCR reactions in this study. PCR reaction mixture was prepared as follows: 1 µl each of 10 pmol forward and reverse primers, 5 µl of AmpliTaq Gold® reaction buffer, 1 µl of 10 mM dNTPs, 1 µl of MgCl$_2$, 2.5 µl GC enhancer and 1 µl template DNA (approx. 10 ng of DNA). The reaction mixture was then made up to 50 µl using sterile ddH_2O.

Table 2-1: PCR thermocycling parameters

Step	Temp	Time
Initial denaturation	96°C	5 mins
35 cycles	96°C 60°C 72°C	1 min 30 secs 2 minute
Final extension	42°C	5 minutes
Hold	4°C	

PCR products were purified using a Qiagen QIAquick PCR purification kit following the protocol by the manufacturer. DNA concentration was determined using the NanoDrop 200 (Thermo Scientific, USA). To confirm isolation of the 5'UTR of the *gapC* gene, the PCR product were sent for sequencing to Macrogen, Korea. The primers used for sequencing of the amplified PCR products are given below.

Primer used for sequencing the 5'UTR of the *gapC* gene amplified by PCR of the *E. coli* plasmid construct: 5' AAGCTTGGGAGAAACTATCC 3'

Primer used for sequencing the 5'UTR of *gapC* gene isolated by PCR of the *S. cerevisiae* plasmid construct: 5' CGATATATCTGGAGTTGGGTG 3'

2.9 *E. coli* expression vector construction

A promoter-less plasmid pSF-PromMCS-BetaGal (Appendix 1; section 2.2) was used to test the potential promoter function of the amplified 5'UTR of *gapC* in *E. coli* and contains multiple cloning sites (MCS) located upstream of the beta-galactosidase (*β-gal*) reporter gene without the promoter, which allows inserting any newly identified DNA for the assessment of its ability to drive *β-gal* expression. The 5'UTR of the *gapC* gene was cloned in using the *SpeI* and *XmaI* restriction sites. (Section 2.11; Fig 2-1).

2.10 *S. cerevisiae* expression vector

The plasmid pRS426 was used to create an expression construct for *S. cerevisiae* (Appendix 2; section 2.2). It is a yeast episomal plasmid (YEp) that yields high copy numbers (approx. 25 copies/cell) (Sikorski and Hieter, 1989). The MCS contains restriction sites *PacI* and *XhoI* for the cloning of the amplified 5'UTR of the *gapC* gene lifted from the *E. gracilis* genome incorporating similar restriction sites. The multiple cloning sites allow swapping of the resident promoter (*PGK1*) with the newly isolated, potential *E. gracilis gapC* promoter (Section 2.11; Fig 2-2).

2.11 Restriction digestion and ligation

To clone the 5’UTR region of the *gapC* gene into the plasmid vector OG372 (*E. coli*) (Appendix 1) and the plasmid pRS426 (*S. cerevisiae*) (Appendix 2), both vectors and the DNA amplified from the *E. gracilis* genome were subjected to restriction digestion with *SpeI* and *XmaI* for OG372, and *PacI* and *XhoI* for pRS426. All restriction enzymes used in this study were purchased from New England Biolabs (NEB) and used at a concentration of 10 units/µl to generate “sticky ends” to enable cloning. Restriction digestion of the plasmids and the inserts were performed in separate 1.5 ml Eppendorf tubes containing 40 µl of DNA (approx. 5 µg), 5 µl of each restriction enzyme, 15 µl of 10 x CutSmart buffer (NEB) and 85 µl of ddH_2O to make the volume of 150 µl. All reagents in the tube were mixed gently by pipetting, and the tube was incubated at 37 ˚C for 4 h. After completion of the digestion, the plasmid and insert DNA were analysed by electrophoresis on a 1% (w/v) agarose gel at 100 V for 60 min. The digested DNA inserts and plasmid OG372 (*E. coli*) were purified using QIAquick PCR Purification Kit while the digested plasmid pRS426 (*S. cerevisiae*), was gel purified using a QIAquick Gel Extraction Kit. The yeast *PGK1* promoter sequence was removed from plasmid pRS426 during the restriction digestion with the *PacI* and *XhoI*. The procedures for gel extraction and PCR purification were carried out according to the manufacturer’s protocol. After the purification and extraction, the DNA concentration was determined using the NanoDrop 2000 (Thermo Scientific, USA).

The purified insert and vector DNA were ligated in the ratio of 1:1, 1:3 and 3:1 (insert: vector) for *E. coli* and *S. cerevisiae* respectively. The insert: vector molar ratio was determined by the free online tool- NEBioCaculator provided by New England Biolabs. For the 1:1 ratio, the following ligation reaction was set up in a sterile 1.5 ml Eppendorf tube: 1 µl insert (30 ng of DNA), 1 µl of plasmid (150 ng of DNA), 2.5 µl T4 DNA ligase (NEB) 10× buffer, 1 µl T4 DNA ligase (NEB) and 4.5 µl ddH2O to make 10 µl reaction volume. The tubes were gently mixed by pipetting and incubated at 4 ˚C overnight. On the following day, the ligated products were analysed on a 0.1% (w/v) agarose gel at 100V for 60 min.

The plasmid vector carrying the features above was designed *in silico* using the software SnapGene® ver. 3.2.2 (from GSL Biotech; available at www.snapgene.com) in such a way that the 5’UTR promoter region of *gapC* gene (insert) is in frame with the reporter gene. To ensure the insert is in frame two extra bp (TA) was incorporated into 5' end of the 5'UTR region with the ecoligapcREV primer and one extra bp (C) was added in scerevisiaegapcREV (see below):

XmaI

ecoligapcREV: 5’ GCCCTACCCGGG**TA**GATATATCTGGAGTTGGG 3’

pacI

scerevisiaegapcREV: 5’ AACCATTTAATTAA**C**GATATATCTGGAGTTGGGTG 3’

Refer to Appendix 3 and 4 for complete 5'UTR sequence with incorporated restriction sites. The illustration figure for the translation frame along with the insert and vector is shown in Appendix 5 and 6.

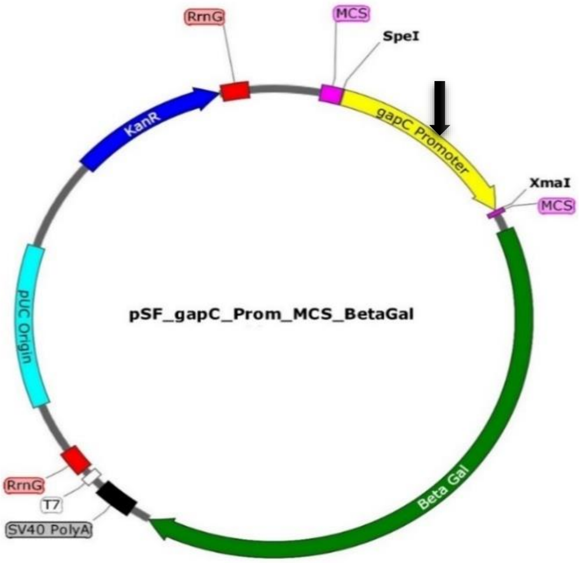


Fig (2-1): pSF-PromMCS-*gapC-BetaGal* (OG372) plasmid with the *gapC* promoter (black arrow) upstream of reporter *β-gal* gene. (Expression vector *E. coli*), Schematic generated using SnapGene® ver. 3.2.2

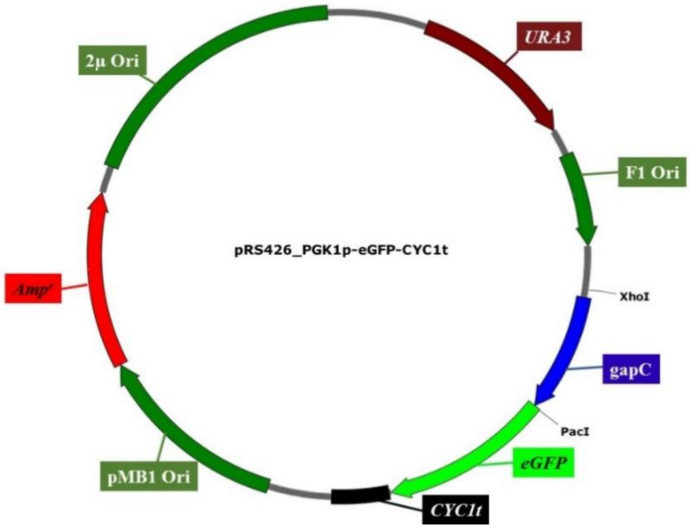


Fig (2-2): pRS426-PGAP -eGFP plasmid with the *gapC* promoter (blue box) upstream of reporter *eGFP* gene. (Expression vector *S. cerevisiae*) Schematic generated using SnapGene® ver. 3.2.2

2.12 Transformation of *E. coli*

All transformations were carried out using the *E. coli* DH5α strain and the protocols by Hanahan (1983), and Inoue *et al.* (1990) were used.

2.12.1 Preparation of competent cells

To prepare competent cells, colonies of DH5α cells freshly streaked on Luria-Bertani (LB) agar plates were inoculated into 10 ml of LB broth and grown overnight at 37˚C in a shaking incubator at 150 rpm. Next day, 40 µl of the culture was used to inoculate 40 ml SOB medium and the cultures were grown at 37˚C in a shaking incubator at 200 rpm until the OD 600 (optical density at 600 nm) was between 0.5-0.4. Optical density was measured on NanoDrop 2000 (Thermo Scientific, USA). After the culture had reached the desirable OD, it was placed on ice for 10 min. All remaining procedures were carried out at 4 ˚C. The culture was centrifuged at 2,500 g for 10 min and the pellet resuspended in 12.8 ml of ice-cold transformation buffer (TB) which contains 10 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), 12mM $CaCl_2$, 55 mM $MnCl_2$, and 250 mM KCl_2 (pH 6.7). The mixture was incubated for 10 min on ice and centrifuged at 2,500 g for 10 min. The pellet was resuspended in 3.2 ml of ice-cold TB, 224 µl (7% v/v) dimethyl sulfoxide (DMSO) was added, and the mixture was incubated on the ice for 10 min. Aliquots of 50 µl were placed in sterile 1.5 ml Eppendorf tubes and immediately frozen in dry ice/ethanol followed by storage at -80˚C.

2.12.2 Introduction of DNA into competent cells

The vector pSF-PromMCS-*gapC-BetaGal* (Fig 2-1) containing the 5'UTR of the *E. gracilis gapC* gene was transformed into DH5α cells using the heat shock method (Inoue *et al.* (1990).

The DH5α competent cells stored at -80˚C were thawed on ice for 30min. The X-Gal plates with antibiotic kanamycin (50 µg/ml) were pre-warmed at 37 ˚C incubator. Two µl (approximately 100 ng of DNA) of ligation mix (OG372 plasmid and the 5'UTR *gapC* inserts of *E. coli*) were mixed with 25 µl of competent cells in a microcentrifuge tube and mixed gently by flicking the bottom of the tube. The mixture was kept on ice for 30 min after which the tubes were heat shocked by placing the bottom 2/3 of the tube into a 42 ˚C water bath for 45 seconds. Immediately after this, tubes were kept on ice for 2 min. An aliquot of 300 µl of pre-warmed SOC medium (section 2.1.2) without an antibiotic was added into the tube, and the culture was grown on a 37 ˚C shaking incubator (300 rpm) for two hours.

Different controls were set to assure the quality of transformation: uncut vector alone, cut vector alone and cut vector + ligase were also transformed into competent cells by a similar procedure as described above. Dilutions of the transformants (50 µl) were plated in pre-warmed X-Gal plates containing kanamycin. The plates were incubated at 37 ˚C overnight. Next day, the plates were observed for colony growth and colonies were marked by a number. The marked colonies were again streaked onto fresh LB agar plates with an antibiotic kanamycin for secondary selection.

2.12.3 Plasmid extraction from recombinant *E. coli* cells

QIAprep Spin Miniprep Kit from Qiagen was used for plasmid extraction using the protocol provided by the manufacturer. The DNA concentration was determined by NanoDrop 2000 (Thermo Scientific, USA) and the plasmid DNA extracted from recombinant *E. coli* cells was visualised on a 1% (w/v) agarose gel run at 100V for 60 min.

2.13 Transformation of *S. cerevisiae*

Lithium acetate (LiAc)/single-stranded (ss) carrier DNA/ polyethylene glycol (PEG) method previously described by Gietz and Schiestl (2007) was used to carry out the transformation of *S. cerevisiae.*

Colonies of *S. cerevisiae* BY4742 (section 2.2) cells freshly streaked onto YPD agar plates (section 2.1.3) were inoculated into 5 ml YPD broth medium and grown overnight at 30 ºC in a shaking incubator at 200 rpm. Following day, 200 µl of the overnight grown culture was inoculated into 20 ml of YPD broth in a 50 ml falcon tube and the culture was grown at 30 ºC on a shaking incubator at 200 rpm until the OD 600 reached 0.5 to 0.6. After the culture had reached the required OD, cells were harvested by centrifugation at 2,000 rpm for 10 min at RT. Supernatant was discarded, and 10 ml of 0.1 M LiAc was added to the pellet and by gentle mixing. The mixture centrifuged at 2000 rpm for 10 min at RT, and the pellet was resuspended in 600 µl of 0.1 M LiAc. One ml of the sample carrier DNA (Salmon sperm DNA; Sigma-Aldrich) was denatured by boiling at 95˚C for 10 min on a heating block and then chilled immediately on ice. ssDNA was vortexed, 10 µl of it was added to pellet previously resuspended in 0.1 M LiAc vortexing lightly. To the mixture, 2 µl of ligation mix (pRS426 plasmid and *gapC* promoter insert) was added and the tube was again vortexed lightly. The ligation mixture was incubated for 30 min at RT. In the meantime, the transformation mixture was prepared as shown in the Table 2-2 below.

Table (2-2): Transformation mixture for *S. cerevisiae.*

Reagent	Final Concentration	Volume/transformation	Vol. for master mix (8 x)
50% PEG	30%	600 µl	4800 µl
1M LiOAC	0.1 M	90 µl	720 µl
DMSO	10%	100 µl	800 µl
dH2O	N/A	98 µl	784 µl
		888 µl//transformation	7104 µl total mix

After completion of incubation, the cell suspension with ligation mix and was again vortexed slightly. An aliquot of 888 µl of the transformation mixture was added on top of the ligation mixture, and the tube was mixed gently. The mixture was incubated at RT for 30 min followed by incubation at 42 ˚C for 14 min. After a heat shock treatment, cells were spun down at 8,000 rpm for 2 min. Supernatant was discarded by aspiration with a pipette, and the pellet was resuspended in 1 ml 5 mM calcium chloride ($CaCl_2$) and incubated at RT for 10 min exactly.

From the transformation mixture, a 250 µl was plated onto an appropriate selective medium: yeast drop-out agar medium without uracil and antibiotic ampicillin (section 2.1.3). Different controls were set to assure the quality of transformation: uncut vector alone, cut vector alone and cut vector + ligase were also transformed into competent cells by a similar procedure as described above. The plates were incubated for two days in 30 ˚C.

2.13.1 Plasmid extraction from recombinant *S. cerevisiae* cells

Plasmid extraction was performed using the QIAprep Spin Miniprep Kit; the protocol was modified by Singh and Weil (2002) for plasmid extraction. A fresh transformant colony was taken from the master plate and streaked onto the -ura plates. One loop full of colonies was inoculated into 10 ml of -ura broth. Cells were grown overnight at 30 ˚C on a shaking incubator at 150 rpm. Cells were harvested by centrifugation at 2,500g for 5 min at 4˚C. To the pellet, 300 µl of Buffer P1 (QIAprep Spin Miniprep Kit) was added, and the mixture was vortexed for 2 min. To the cell suspension, 100 µl lyticase (100U, Sigma Aldrich-USA), 200 µl of 1.2 M sorbitol and 200 µl 0.1 M NaPO4 (pH 7.4) were added and the mixture incubated at 37 ˚C for one hour (mixed every 15 min) to digest the cell wall. An aliquot of 600 µl of buffer P2 from the Qiagen kit was added and the tube mixed by gently by inverting several times. The cell suspension was incubated for 10 min at RT. After this, 900µl buffer N3 was added to the lysate, and the tube mixed several times by gentle inversion. The tube was incubated on ice for 30 min followed by centrifugation at 15,000 g for 10 min at 4 ˚C. The supernatant was collected and applied to QIAprep Spin columns which were spun for a minute at 13,000 rpm. The columns were washed with 500 µl of buffer PB, followed by 750 µl of buffer PE and spun for one minute at 13,000rpm. The columns were spun for an additional one minute at 13,000 rpm. Finally, the bound plasmid DNA on the columns was eluted by adding 50 µl elution buffer (Buffer EB). Plasmid DNA concentration was determined using NanoDrop 2000 (Thermo Scientific, USA). The quality of the extracted plasmid DNA was further analysed by electrophoresis on a 1% (w/v) agarose gel at 100 V for 60 min.

2.14 Characterisation of 5'UTR promoter region of *gapC* gene from *E. gracilis*

The 5’UTR promoter region of glyceraldehyde-3-phosphate dehydrogenase (*gapC*) based on the *gapC* sequence available on the European Bioinformatics Institute (EMBL-EBI) database deposited by Henze *et al.* (1995) (http://www.uniprot.org/uniprot/Q43311, sequence ID L39772.1), was used for further characterisation using MatInspector; Genomatix Software Suite (Cartharius *et al.*, 2005). MatInspector identifies transcription factor binding sites (TFBS) from a DNA sequence utilising an extensive library of weight matrices. The obtained possible TFBS were further selected by highest matrix similarity. A 100% similarity with the matrix gets a score of 1.00 (the sequence position resembles the highest conserved nucleotide at that position in the matrix). The matrix similarity above 80% is considered as a good match. The TFBS search parameter in the Genomatix was confined to algae, fungi and plants only.

Chapter 3: Results

3.1 Visualisation of highly expressed proteins of *E. gracilis* after separation on 1D SDS-PAGE and 2D SDS-PAGE

In the absence of annotated genome data, a proteomics-based approach was used to identify strongly expressed proteins for which the promoter sequence can be isolated by PCR approach from the gene encoding the highly expressed protein (Section 3.5). Proteins extracted from the wild-type *E. gracilis* Z and *E. gracilis* var. *saccharophila* (B752) were separated and visualised on 1D SDS-PAGE whereas proteins from *E. gracilis* var. *saccharophila* were also analysed on 2D SDS-PAGE. After separation, proteins were stained with Coomassie Brilliant Blue R-250 for visualisation of strong protein bands (highly expressed proteins).

3.1.1 1D SDS-PAGE analysis

Proteins separated on 1D SDS-PAGE revealed a similar expression pattern in both strains with thick bands indicating highly expressed proteins of a molecular weight of approximately 150 kDa (Fig 3-1; white box). Other highly expressed proteins of about 40 kDa, 60 kDa and 200 kDa were also be observed on the gel for both strains. There were also some differences, for example, a strong protein band of a molecular mass of 80 kDa (Fig 3-1; red oval) was only present in the B752 strain. This result could be further analysed to identify differentially expressed proteins.

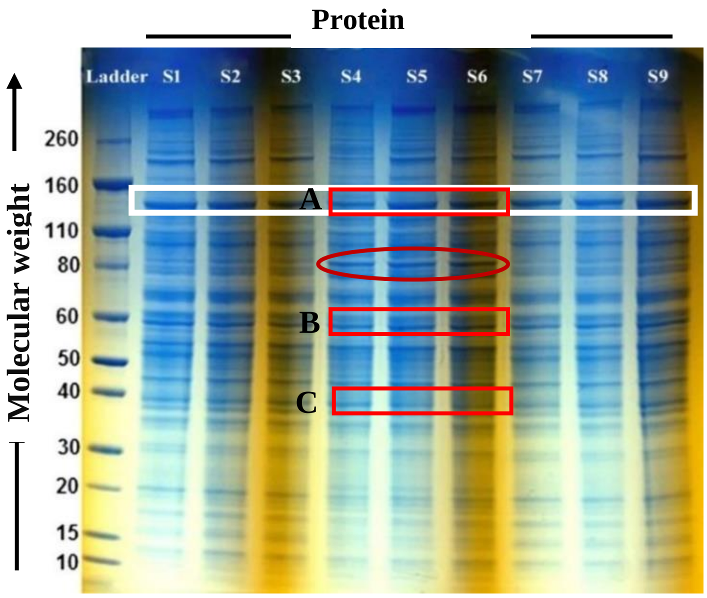


Fig (3-1): 1D SDS-PAGE analysis of the intracellular proteins extracted from *E. gracilis* on NuPAGE® Novex® Bis-Tris Mini Gel, stained with Coomassie brilliant blue. An equal amount of proteins (25 µg) was loaded in each lane. Lane 1 (Ladder): molecular weight marker (Novex® Sharp Pre-stained marker); S4, 5 and 6 indicate proteins extracted from *E. gracilis* var *Saccharophila* (B-752) and S1, 2, 3, 7, 8, 9 are the proteins extracted from *E. gracilis Z* strain. The white rectangular box marks a highly expressed protein across the two strains. Proteins extracted from the B752 strain (Fig 3-1; samples A, B and C; surrounded by a red rectangular box) were used for mass spectrometric analysis. The red oval represents a protein found in the B752 strain but not in the Z strain.

Highly expressed proteins extracted from the B752 strain (Fig 3-1; S4, S5 and S6; A, B and C) were used for mass spectrometry analysis for protein identification. Since the B752 strain does not develop large amounts of chloroplasts when grown in the dark, the majority of the proteins extracted will be non-chloroplast related.

3.1.2 2D SDS-PAGE analysis

Proteins extracted from the *E. gracilis* var. *saccharophila* B752 strain were also separated on 2D SDS-PAGE (Fig 3-2). The analysis revealed a similar expression pattern to that observed on the 1D SDS-PAGE with most of the highly expressed proteins found in the range of 40-150 kDa. The gel image shows clear and well-separated protein spots although some horizontal streaking from the basic to the acidic end at the top of the gel is present (Fig 3-2).

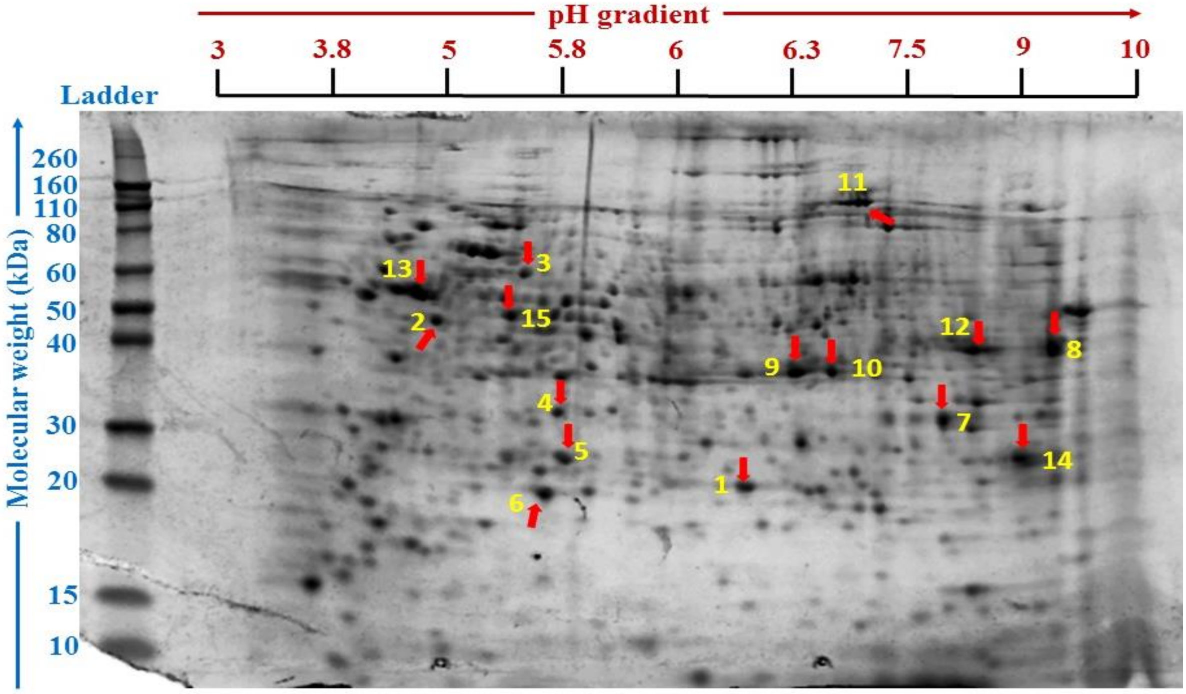


Fig (3-2): 2D SDS-PAGE analysis of intracellular proteins extracted from *E. gracilis* var *saccharophila* (B752). The first dimension separation of proteins was carried out by iso-electric focusing (IEF) of proteins on and IPG strip (pH gradient 3-10). The second dimension separation was detected on SDS-PAGE by the molecular weight of the proteins. Spots chosen for the mass spectrometric analysis are numbered and boxed.

From the gel containing more than 100 distinct spots, 15 intense, well-separated spots were manually excised and used for mass spectrometry analysis to obtain protein identification. This information will then be used for the identification of highly expressed gene of *E. gracilis*.

3.2 Identification of highly expressed proteins using mass spectrometry

Three samples from the 1D gel (Fig 3-1) and all 15 spots from the 2D analysis (Fig 3-2) were analysed by nanoLC ESI MS/MS. Spectral data obtained from the mass spectrometer according to individual mass to charge ratio *m/z* and relative abundance were processed through the Global Proteome Machine (GPM) software (http://www.gpm.org) to predict the identity of the protein based on *m/z*. The proteins identified from various bands from 1D-SDS-PAGE are tabulated below (Table 3-1).

Table 3-1: Proteins identified from 1D-SDS-PAGE using nanoLC ESI MS/MS. The total number of samples analysed was three (Fig 3-1). In all A, B and C sample tubes, their respective three bands were pooled together and analysed (Fig 3-1). Sample A revealed the identity of only one highly similar *E. gracilis* protein while sample B and C revealed the identity of three different proteins.

Sample	UniProt ID	Proteins identified	E-value	pI	Mw (kDa)	Sequence coverage
A	Q8LPA6	Malate synthase-isocitrate lyase	0	6.75	129.75	100%
B	Q5EU90	Trans-2-enoyl-CoA reductase	0	6.08	57.3	93%
	Q39727	Heat shock protein 60	4. E-26	5.6	59.78	100%
	D7PN08	Fatty acyl-coenzyme A reductase	0	8.88	56.36	97%
C	Q43311	Glyceraldehyde-3-phosphate dehydrogenase	6 E-27	7.12	38.17	100%
	B8QU18	NAD-dependent alcohol dehydrogenase	1 E-119	7.98	38.31	100%
	P14963	Elongation factor 1 alpha	0	9	48.6	100%

Table 3-2: Proteins identified from the 2D-SDS-PAGE using nano nanoLC ESI MS/MS. The total number of samples analysed from the 2D-gel was 15 (Fig 3-2).

Spot no	Uniref	Proteins identified	E-value	pI	Mw (kDa)	Sequence coverage
1	Q9XN18	Putative NADH dehydrogenase	3.7E-114	6.25	19.41	100%
2	Q9AQX2	Tubulin alpla-1 chain	0	4.94	49.83	100%
2	O65204	Actin	1E-27	5.1	43.32	96%
3	Q39727	Heat shock protein 60	4E-26	5.6	59.78	100%
3	B2NIV9	6-phosphogluconate dehydrogenase	3E-27	5.69	52.66	100%
4	Q7XYK2	Fructose-1,6 bisphosphatase	1.3E-179	5.43	34.5	96%
4	O24561	Light harvesting chlorophyll a/b binding protein of PSII	1.5E-113	5.33	33.38	98%
5	Q6VEG6	Cytosolic triosephosphate isomerase	9E-78	6.93	28.73	100%
5	P30391	ATP synthase subunit α	2E-66	5.21	29.29	93%
6	W8VZ53	Peroxiredoxin	1E-140	5.56	21.29	100%
7	Q84T13	L-3-hydroxyacyl-CoA dehydrogenase	1E-116	8.67	33.78	92%
7	Q9FPM7	Porin-like protein	1E-178	8.52	30.71	96%
8	P14963	Elongation factor 1-alpha	5E-27	9	48.6	100%
9	B8QU18	NAD-dependent alcohol dehydrogenase	1E-119	7.98	38.31	100%
9	Q43311	Glyceraldehyde-3-phosphate dehydrogenase	6E-27	7.12	38.17	100%
10	B8QU18	NAD-dependent alcohol dehydrogenase	1E-119	6.07	36.73	100%
10	Q43311	Glyceraldehyde-3-phosphate dehydrogenase	6E-27	7.12	38.17	100%
11	Q8LPA6	Malate synthase-isocitrate lyase	0	6.75	129.75	100%
12	P14963	Elongation factor 1 alpha	0	9	48.6	100%
12	Q9SXP9	cAMP-Dependent protein kinase	2E-49	8.11	50.08	58%
13	Q9AQV5	Beta-tubulin	0	4.75	50	100%
14	P48337	DNA-directed RNA polymerase subunit alpha	8E-82	9.77	25.36	96%
15	Q9LEK7	Enolase	0	5.8	46.51	100%

3.3 Identification of highly expressed genes of *E. gracilis* using the expressed sequence tags (ESTs) database

ESTs indicate a transcribed portion of the genome. The information can be used to identify the most highly expressed genes in an organism. After obtaining the mass spectrometric data for highly expressed proteins, the results were compared with the transcriptomic data *i.e.* the expressed sequence tag (EST) database. The Taxonomically Broad EST Database (TBestDB) contains various *E. gracilis*

ESTs. The deposited ESTs were compared on the basis of the number of ESTs to estimate highly expressed proteins, and the result revealed that the Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) had the highest number of ESTs, *i.e.*, 171 (Table 3-3). This result supports the previous result from mass spectrometry (Tables 3-1 and 3-2) as GAPDH was found to be highly expressed in both analyses.

Table 3-3: The top ten highly expressed proteins of *E. gracilis* tabulated by the number of deposited ESTs (http://tbestdb.bcm.umontreal.ca). The descending order of the number of ESTs shows that the GAPDH is the most highly expressed protein.

Uniref	Description	E-Value	Sequence coverage	pI	Mw	No of ESTs
Q43311	Glyceraldehyde-3-phosphate dehydrogenase related cluster	0	100%	7.12	38.17	171
P07436	Tubulin beta-1 chain related cluster	0	91%	4.91	52.13	78
P14963	Elongation factor 1-alpha related cluster	0	97%	9	48.61	68
Q6H798	Putative acetyl-CoA synthetase related cluster	0	87%	5.69	77.59	59
Q25563	Tubulin alpha-13 chain related cluster	0	90%	4.96	49.81	57
Q8LPA6	Malate synthase-isocitrate lyase related cluster	0	99%	6.75	129.75	51
Q875H8	Malic enzyme related cluster	1E-139	95%	6.2	68.94	46
Q42729	Fructose-1,6-bisphosphate aldolase related cluster	1E-180	88%	5.75	39.99	45
Q26937	Heat shock protein 70 related cluster	0	84%	5.64	74.28	43
Q4QJF1	ATP synthase alpha chain related cluster	0	98%	9.73	62.54	31

Based on the results obtained from 2D SDS-PAGE and mass spectrometry, it is now evident that the *gapC* gene encoding GAPDH protein could be considered as a good candidate to be a highly expressed gene. Relying on the fact that the highly expressed gene should possess promoter elements upstream of the translation initiation site, it was decided to isolate the 5'UTR promoter sequence of the *gapC* gene and test the functionality in *E. coli* and *S. cerevisiae*.

3.4 Extraction genomic DNA from *E. gracilis*

The gene responsible for GAPDH protein is the *gapC* gene. Targeting the 5' and 3' DNA of the gene required isolation of high-quality genomic DNA (gDNA) from which the targeted DNA sequences can be amplified.

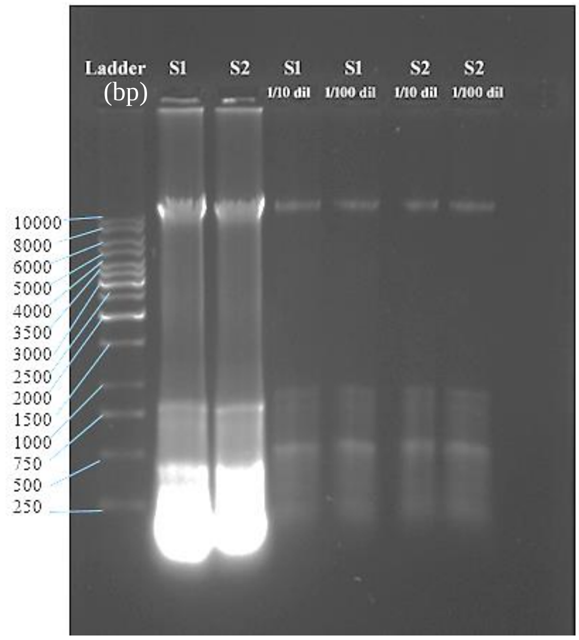


Fig 3-3: The extracted gDNA was visualised on a 0.9% (w/v) agarose gel run at 100V for 60 min. The two samples of gDNA (S1 and S2) and their dilutions (1/10 and 1/100) were loaded into the gel. Each lane was loaded with a volume of 5 µl of the sample. The expected size of genomic DNA is above 10 kb. DNA Ladder (Thermo Scientific GeneRuler 1kb) was loaded into the first well.

Extraction of high molecular weight DNA from *E. gracilis* was achieved successfully. The isolated DNA was intact with no sign of degradation although the samples contained some RNA contamination (Fig 3-3). The extracted DNA was purified further, but the effort was not very successful as more than 50% DNA was lost during purification. The concentration and purity of the extracted gDNA was analysed on NanoDrop 2000. The gDNA concentration was measured as 68.9 ng/µl. The assessment of the purity was on the basis of absorbance ratios at 260/280 nm and 260/230 nm. The absorbance ratio 260/280 nm of S1 was 1.85 and 1.87 with S2, which is within the expected range for pure DNA without any protein or phenol contamination. The ratio 260/230 nm was measured as secondary criteria for DNA purity. For S1 it was found to be 2.46 and 2.31 for S2 which indicates that the extracted DNA is free from organic solvent and protein contamination.

3.5 Isolation the promoter region from the *gapC* gene of *E. gracilis* by PCR

PCR was used to amplify the 5'UTR (promoter) region of *gapC* gene previously deposited by (Henze *et al*., 1995) (GI: 870799). The length of 5'UTR promoter region was of 1016 bp above the translation start site. The length of the target DNA (5'UTR) to be amplified was calculated as 1042 bp including the restriction sites and added nucleotides. The primers used to amplify the 5'UTR were designed by incorporating restriction sites which enable cloning of the 5'UTR region (insert) into the promoterless *E. coli* plasmid OG372 and the *S. cerevisiae* plasmid pRS426 from which the promoter region had been removed (Section 2.8).

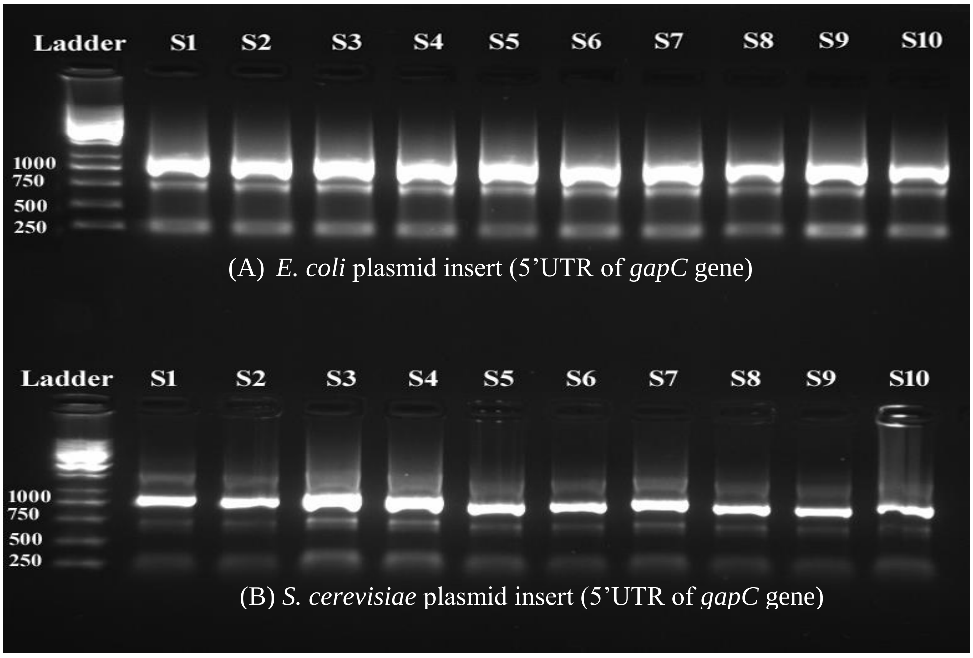

Ladder
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
1000
750
500
250
(A) E. coli plasmid insert (5'UTR of gapC gene)
Ladder
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
1000
750
500
250
(B) S. cerevisiae plasmid insert (5'UTR of gapC gene)

Fig 3-4: PCR amplification of the 5'UTR promoter region of *gapC* gene from the *E. gracilis* genome. The first well of both rows is loaded with a DNA Ladder (Thermo Scientific GeneRuler). The first row in 'A' shows the promoter region amplified including restriction sites compatible with the *E. coli* plasmid (OG372). The second row 'B' depicts the promoter region amplified including restriction sites for the cloning into the plasmid pRS426 of *S. cerevisiae*. Both amplicons are of the size of around 1050 bp, and the expected size of the amplicon is 1042 bp, which is the size of original 5'UTR of *gapC* gene (1015 bp) plus the added restriction sites (27 bp).

3.6 Transformation of *E. coli* and selection of the recombinant colonies

The *E. coli* DH5α cells were transformed with the originally promoter-less OG372 plasmid vector now containing the 5'UTR region of the *E. gracilis gapC* gene (Fig 2-1), amplified above. The purpose of this transformation was to test potential promoter activity of the *E. gracilis* DNA in *E. coli*. As the 5'UTR region was cloned in front of the *β-gal* gene, the emergence of blue transformant colonies on X-gal containing plates would indicate promoter activity.

The result indicated successful transformation, the calculated transformation frequency based on the antibiotic resistant colonies was 1.58×10^5 per µg of DNA. Blue coloured colonies were observed not only on the transformation plates but also colonies on the plates containing the negative control, *i.e.* cells transformed with an intact promoterless plasmid (Appendix 1). The results suggest that the plasmid OG372 though was considered promoterless, it could be inferred that it should have some promoter elements upstream of reporter gene *β-gal* which assist in expressing it (refer to section 4.5 for more details). As there was no clear evidence whether the *gapC* promoter region was functional in *E. coli*, further work was carried out in *S. cerevisiae*.

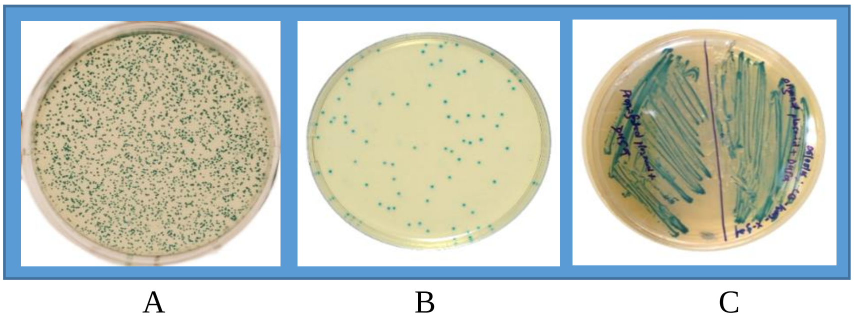
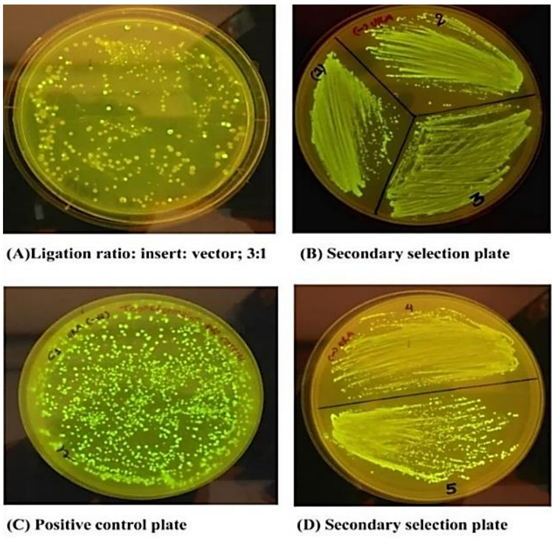


Fig 3-5: *E. coli* transformed with the plasmid containing the 5'UTR promoter region of the *E. gracilis gapC* gene in the X-gal plates containing the antibiotic kanamycin. A: 50 µl of undiluted transformants; B: 50 µl of 1/100 dilution of transformants. C: negative control streaked with transformants containing the original intact plasmid.

3.7 Transformation of *S. cerevisiae* and selection of the recombinants

The plasmid pRS426 (Fig 2-2) of which the original *PGK1* promoter in front of the *eGFP* reporter gene was replaced with the putative *gapC* gene promoter region was transformed to *S. cerevisiae* (BY4742). Transformants were selected on the basis of complementation of the *ura* deficiency.

The transformation was successful as colonies were emerging on the yeast drop-out agar medium (-uracil) plates. The transformation was carried out in different ligation ratios of the insert and the vector (3:1, 1:3 and 1:1). Observation under blue light revealed fluorescent colonies apparently expressing the *eGFP* gene now under the *E. gracilis* promoter region. On the plates with the 3:1 (insert: vector) ligation ratio, thirty-five colonies were fluorescent (Fig 3-6). Four colonies expressing eGFP were observed with 1:3 ligation ratio and six colonies with 1:1 ligation ratio (insert: vector).


(A)Ligation ratio: insert: vector; 3:1

(B) Secondary selection plate

(C) Positive control plate

(D) Secondary selection plate

Fig 3-6: *S. cerevisiae* transformant colonies expressing the eGFP protein. Plate A contains transformants with the vector containing the promoter region of the *gapC* gene of *E. gracilis*. Plates B and D are the colonies taken from plate A for secondary selection. Plate C is a positive control that has the original plasmid containing the yeast *PGK1* promoter.

The eGFP expressing colonies on the plates with the 3:1 insert:vector ligation ratio were marked (A), and bright, distinct and large colonies were streaked onto fresh yeast drop-out agar medium (-uracil) plates for secondary selection (B, D). The secondary selection plates were then observed for fluorescence and all colonies on the plates were expressing eGFP. The results indicated that the newly isolated 5’ UTR region of the *gapC* gene of *E. gracilis* indeed had functional promoter activity in *S. cerevisiae*.

3.8 Confirmation of the insertion and orientation of the 5’UTR region of the *gapC* gene in the *S. cerevisiae* plasmid

The transformant colonies were further confirmed for the presence of the 5’UTR region of the *E. gracilis gapC* gene in the plasmid used for yeast transformation. This was first attempted by colony PCR which proved unsuccessful. This could have been because the plasmid was episomal (non-integrating), and the fact that performing colony PCR with *S. cerevisiae* is not as straightforward as with bacterial colonies because of the yeast cell wall architecture. The failure may also have been due technical and handling issues as PCR is not always successful due to the hindrance of Taq polymerase activity by cell debris and residual culture medium (Amberg *et al*., 2005). Therefore, it was decided to extract the plasmid first and then perform the PCR to verify the presence of the 5’UTR region of the *E. gracilis gapC* gene in the plasmid DNA. Hence plasmid DNA was extracted from the transformant colonies expressing eGFP and as well colonies without eGFP expression. Likewise, the plasmid DNA was extracted from the positive control with a plasmid containing the original *PGK1* promoter. The extracted plasmid DNA from different colonies was used as a template for PCR.

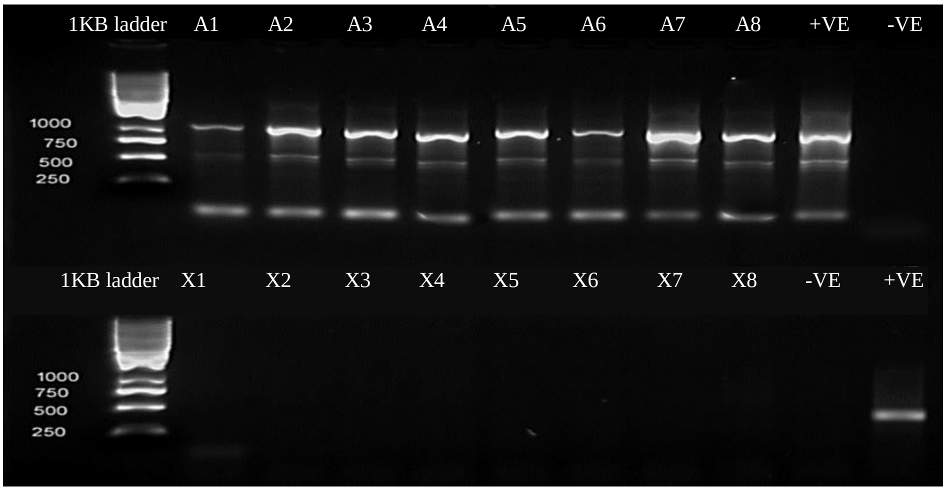


Fig 3-7: Confirmation of the presence of the 5’UTR region of the *gapC* gene in the plasmid used for the transformation of yeast, resulting in fluorescent colonies. The first lanes of the gel were loaded with DNA ladder (GeneRuler 1 kb Plus, Thermo Fisher Scientific, Australia). Lanes A1- A8 represent the PCR product amplified using the plasmid DNA as a template extracted from the *S. cerevisiae*

transformants expressing eGFP. Lanes X1 to X8 are the PCR product amplified using the plasmid DNA as a template extracted from the non-eGFP expressing *S. cerevisiae* colonies.

The gel image (Fig 3-7 upper lane) shows the correct amplifications at around 1050 bp which corresponds to the size of the 5'UTR region of the *gapC* gene plus the *eGFP* gene region in the plasmid which confirms the insertion and orientation of the 5'UTR region of the *gapC* gene. The PCR products were amplified using DNA isolated from the fluorescent colonies as a template and primers (Table 3-4; fwdgapC and reveGFP) targeting the nucleotide position 2457 to 3503 of the plasmid. In contrast, there were no amplification products from the negative control using the same primers. The band seen in the 'positive control' is the 5'UTR region of the *gapC* gene amplified from the *E. gracilis* genome initially (section 3.4).

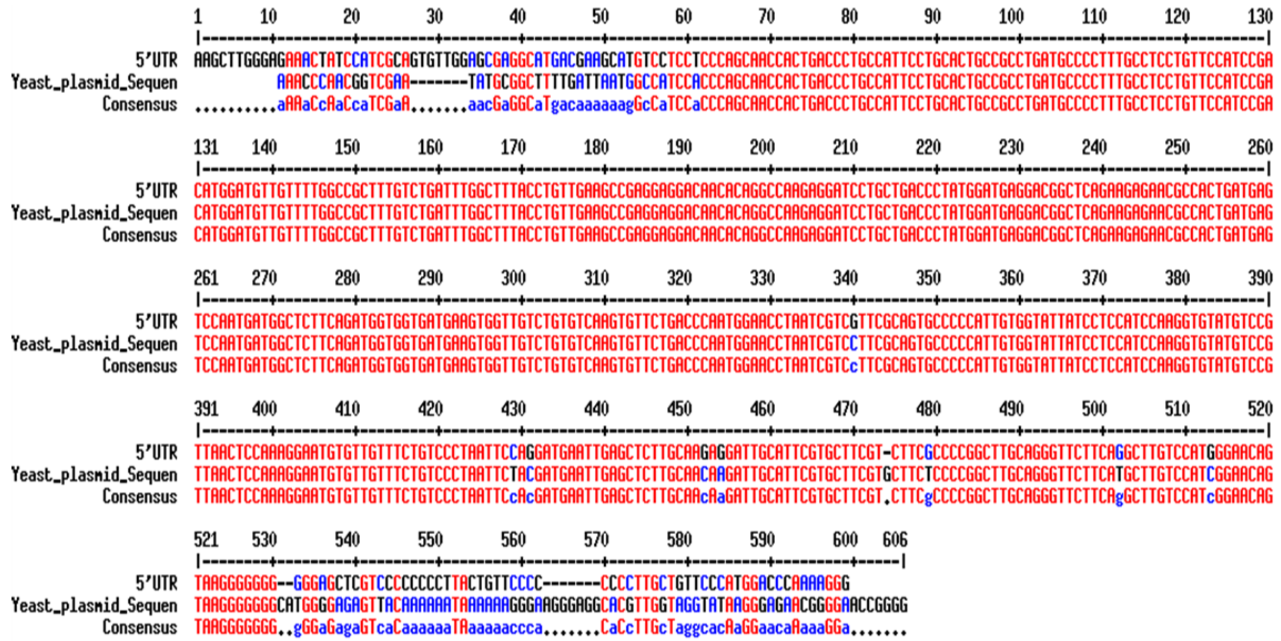
Amplification from the template DNA extracted from non-GPF expressing colonies with primers amplifying the yeast *PGK1* promoter (Table 3-4; fwdPGK1 and revPGK1) did not produce any amplicons (Fig 3-7 lower lane). In the positive control, DNA was used as the template was from the plasmid DNA extracted from the *S. cerevisiae* cells expressing eGFP containing the original plasmid with the *PGK1* promoter. The positive control lane shows correct amplification of a DNA fragment of about 500 bp in size (expected size 507 bp). There were no bands seen in the negative control lane with plasmid DNA extracted from non-eGFP expressing cells transformed with the 5'UTR promoter region of the *gapC* gene, amplified with *PGK1* associated primers (Table 3-4). This result further verified that the 5'UTR region of the *gapC* gene was contained in the pRS426-PGAP-eGFP plasmid (Fig 2-2) and was able to translate the *eGFP* gene to a fluorescing protein.

Table 3-4: Primers used in the above study to confirm the insertion and orientation of the 5'UTR region of the *E. gracilis gapC* gene in the *S. cerevisiae* plasmid.

Primer	Sequence 5′to 3′	Expected PCR product size (bp)
fwdgapC	TGAATTGAGCTCTTGCAAGA	1046
reveGFP	GTTTGAAGGTGATACCTTAG	
fwdPGK1	TTATCGAGAAAGAAATTACCGT	507
revPGK1	ATAATTACTTCCTTGATGATCTG	

3.9 Sequencing the plasmid DNA of the recombinant *S. cerevisiae*

The plasmid DNA extracted from the fluorescing recombinant *S. cerevisiae* cells was sent for sequencing to confirm that the plasmid contained the promoter region of the *gapC* gene isolated from *E. gracilis*. Single primer extension sequencing of the plasmid DNA gave a 589 bp product revealing that the plasmid indeed had the *gapC* promoter (refer to Appendix 7). The obtained plasmid DNA sequence was compared with the original DNA sequence of the 5'UTR region of *gapC* gene. The sequence alignment shows 96% sequence homology with the original 5’UTR DNA sequence of the *gapC* gene of *E. gracilis* (Fig 3-8).



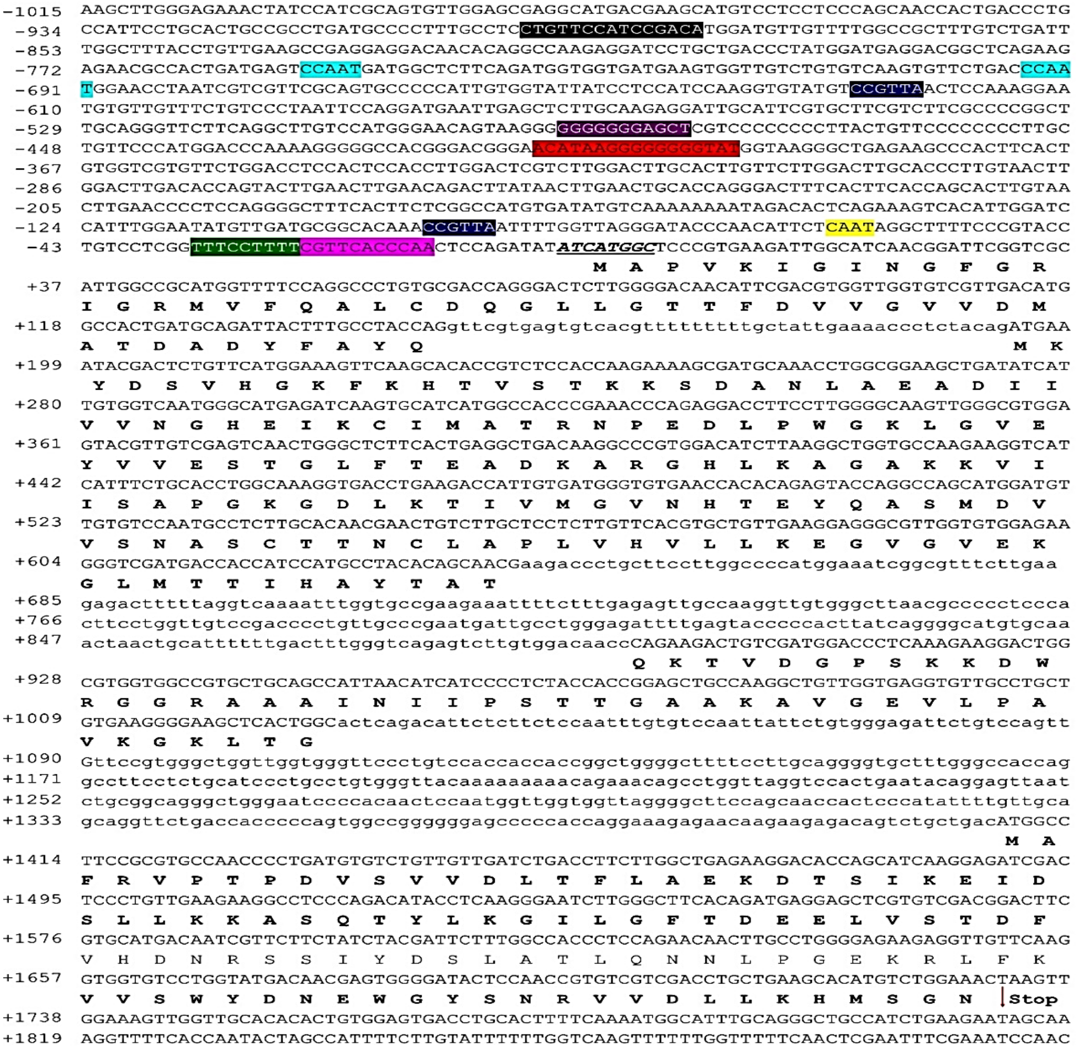
```
                     1        10        20        30        40        50        60        70        80        90       100       110       120       130
                     |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
               5’UTR AAGCTTGGGAGAAACTATCCATCGCAGTGTTGGAGCGAGGCATGACGAAGCATGTCCTCCTCCCAGCAACCACTGACCCTGCCATTCCTGCACTGCCGCCTGATGCCCCTTTGCCTCCTGTTCCATCCGA
Yeast_plasmid_Sequen           AAACCCAACGGTCGAA-------TATGCGGCTTTTGATTAATGGCCATCCACCCAGCAACCACTGACCCTGCCATTCCTGCACTGCCGCCTGATGCCCCTTTGCCTCCTGTTCCATCCGA
           Consensus ..........aAAaCcAaCcaTCGaA.......aacGaGGCaTgacaaaaaagGcCaTCCaCCCAGCAACCACTGACCCTGCCATTCCTGCACTGCCGCCTGATGCCCCTTTGCCTCCTGTTCCATCCGA

                     131      140       150       160       170       180       190       200       210       220       230       240       250       260
                     |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
               5’UTR CATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTGTTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACGGCTCAGAAGAGAACGCCACTGATGAG
Yeast_plasmid_Sequen CATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTGTTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACGGCTCAGAAGAGAACGCCACTGATGAG
           Consensus CATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTGTTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACGGCTCAGAAGAGAACGCCACTGATGAG

                     261      270       280       290       300       310       320       330       340       350       360       370       380       390
                     |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
               5’UTR TCCAATGATGGCTCTTCAGATGGTGGTGATGAAGTGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCGTTCGCAGTGCCCCCATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCG
Yeast_plasmid_Sequen TCCAATGATGGCTCTTCAGATGGTGGTGATGAAGTGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCCTTCGCAGTGCCCCCATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCG
           Consensus TCCAATGATGGCTCTTCAGATGGTGGTGATGAAGTGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCcTTCGCAGTGCCCCCATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCG

                     391      400       410       420       430       440       450       460       470       480       490       500       510       520
                     |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
               5’UTR TTAACTCCAAAGGAATGTGTTGTTTCTGTCCCTAATTCCAGGATGAATTGAGCTCTTGCAAGAGGATTGCATTCGTGCTTCGT-CTTCGCCCCGGCTTGCAGGGTTCTTCAGGCTTGTCCATGGGAACAG
Yeast_plasmid_Sequen TTAACTCCAAAGGAATGTGTTGTTTCTGTCCCTAATTCTACGATGAATTGAGCTCTTGCAACAAGATTGCATTCGTGCTTCGTGCTTCTCCCCGGCTTGCAGGGTTCTTCATGCTTGTCCATCGGAACAG
           Consensus TTAACTCCAAAGGAATGTGTTGTTTCTGTCCCTAATTCcAcGATGAATTGAGCTCTTGCAAcAaGATTGCATTCGTGCTTCGT.CTTCgCCCCGGCTTGCAGGGTTCTTCAgGCTTGTCCATcGGAACAG

                     521      530       540       550       560       570       580       590       600   606
                     |--------+---------+---------+---------+---------+---------+---------+---------+-----|
               5’UTR TAAGGGGGGG--GGGAGCTCGTCCCCCCCCTTACTGTTCCCC-------CCCCTTGCTGTTCCCATGGACCCAAAAGGG
Yeast_plasmid_Sequen TAAGGGGGGGCATGGGGAGAGTTACAAAAAATAAAAAAGGGAAGGGAGGCACGTTGGTAGGTATAAGGGAGAACGGGGAACCGGGG
           Consensus TAAGGGGGGG..gGGaGagaGTcaCaaaaaaTAaaaaaccca.......CaCcTTGcTaggcacAaGGaacaAaaaGGa.......
```

Fig 3-8: Sequence of the plasmid DNA obtained from the recombinant *S. cerevisiae* cells expressing eGFP protein. The DNA sequence was aligned with the original 5’UTR promoter region sequence of the *gapC* gene. The alignment was performed using Multalin version 5.4.1 (Corpet, 1988). The consensus sequence is the calculated order of most frequent residues. Red denotes high consensus; blue denotes low consensus and black denotes neutral consensus.

3.10 Identification of putative transcription factor binding sites (TFBS) in the *gapC* gene promoter

The 5’UTR of the *gapC* gene from *E. gracilis* was able to drive expression of the *eGFP* reporter gene in *S. cerevisiae*. To further study and characterise the promoter, the entire 5’UTR sequence was analysed using MatInspector; Genomatix Software Suite. The result obtained from the MatInspector discovered the transcription factor binding sites (TFBS) similar to plants, fungi and yeast (Appendix 10; Fig 3-9). The result revealed the presence of conserved TFBS such as CAAT box and GC box

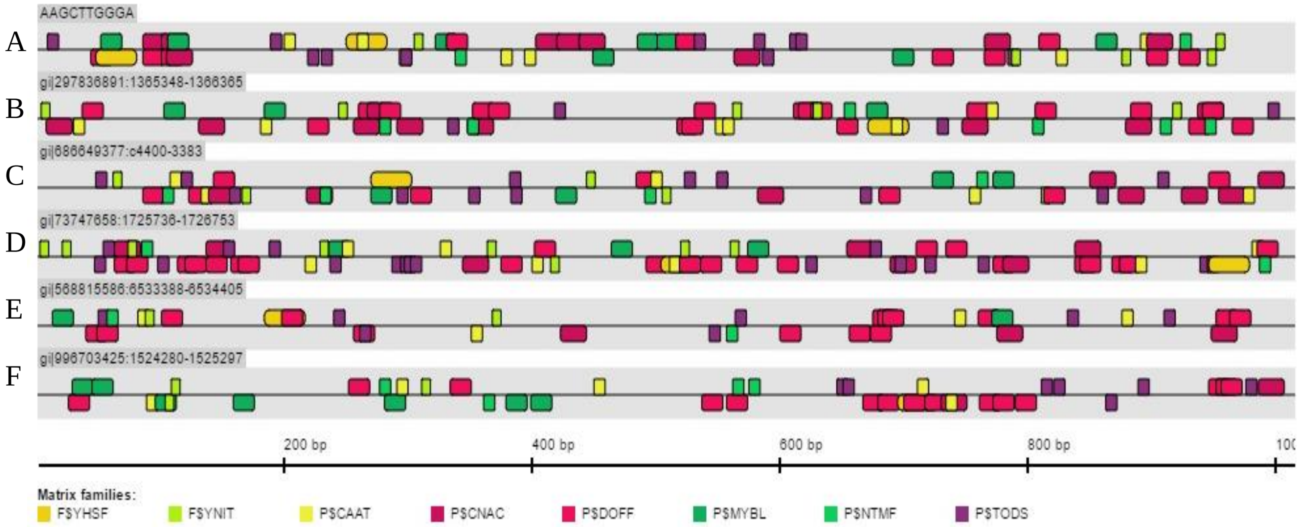
(Fig 3-9) in the 5’UTR sequence of *gapC* gene, however, it lacks the core promoter sequence TATA box. All TFBS obtained from MatInspector are highlighted in the figure below.



```
 -1015  AAGCTTGGGAGAAACTATCCATCGCAGTGTTGGAGCGAGGCATGACGAAGCATGTCCTCCTCCCAGCAACCACTGACCCTG
  -934  CCATTCCTGCACTGCCGCCTGATGCCCCTTTGCCTCCTGTTCCATCCGACATGGATGTTGTTTTGGCCGCTTTGTCTGATT
  -853  TGGCTTTACCTGTTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACGGCTCAGAAG
  -772  AGAACGCCACTGATGAGTCCAATGATGGCTCTTCAGATGGTGGTGATGAAGTGGTTGTCTGTGTCAAGTGTTCTGACCCAA
  -691  TGGAACCTAATCGTCGTTCGCAGTGCCCCCATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCGTTAACTCCAAAGGAA
  -610  TGTGTTGTTTCTGTCCCTAATTCCAGGATGAATTGAGCTCTTGCAAGAGGATTGCATTCGTGCTTCGTCTTCGCCCCGGCT
  -529  TGCAGGGTTCTTCAGGCTTGTCCATGGGAACAGTAAGGGGGGGGGGAGCTCGTCCCCCCCCTTACTGTTCCCCCCCCTTGC
  -448  TGTTCCCATGGACCCAAAAGGGGGCCACGGGACGGGAACATAAGGGGGGGTATGGTAAGGGCTGAGAAGCCCACTTCACT
  -367  GTGGTCGTGTTCTGGACCTCCACTCCACCTTGGACTCGTCTTGGACTTGCACTTGTTCTTGGACTTGCACCCTTGTAACTT
  -286  GGACTTGACACCAGTACTTGAACTTGAACAGACTTATAACTTGAACTGCACCAGGGACTTTCACTTCACCAGCACTTGTAA
  -205  CTTGAACCCCTCCAGGGGCTTTCACTTCTCGGCCATGTGATATGTCAAAAAAAATAGACACTCAGAAAGTCACATTGGATC
  -124  CATTTGGAATATGTTGATGCGGCACAAACCGTTAATTTTGGTTAGGGATACCCAACATTCTCAATAGGCTTTTCCCGTACC
   -43  TGTCCTCGGTTTCCTTTTCGTTCACCCAACTCCAGATATATCATGGCTCCCGTGAAGATTGGCATCAACGGATTCGGTCGC
                                                  M  A  P  V  K  I  G  I  N  G  F  G  R
   +37  ATTGGCCGCATGGTTTTCCAGGCCCTGTGCGACCAGGGACTCTTGGGGACAACATTCGACGTGGTTGGTGTCGTTGACATG
        I  G  R  M  V  F  Q  A  L  C  D  Q  G  L  L  G  T  T  F  D  V  V  G  V  V  D  M
  +118  GCCACTGATGCAGATTACTTTGCCTACCAGgttcgtgagtgtcacgtttttttttgctattgaaaaccctctacagATGAA
        A  T  D  A  D  Y  F  A  Y  Q                                                 M  K
  +199  ATACGACTCTGTTCATGGAAAGTTCAAGCACACCGTCTCCACCAAGAAAAGCGATGCAAACCTGGCGGAAGCTGATATCAT
         Y  D  S  V  H  G  K  F  K  H  T  V  S  T  K  K  S  D  A  N  L  A  E  A  D  I  I
  +280  TGTGGTCAATGGGCATGAGATCAAGTGCATCATGGCCACCCGAAACCCAGAGGACCTTCCTTGGGGCAAGTTGGGCGTGGA
        V  V  N  G  H  E  I  K  C  I  M  A  T  R  N  P  E  D  L  P  W  G  K  L  G  V  E
  +361  GTACGTTGTCGAGTCAACTGGGCTCTTCACTGAGGCTGACAAGGCCCGTGGACATCTTAAGGCTGGTGCCAAGAAGGTCAT
        Y  V  V  E  S  T  G  L  F  T  E  A  D  K  A  R  G  H  L  K  A  G  A  K  K  V  I
  +442  CATTTCTGCACCTGGCAAAGGTGACCTGAAGACCATTGTGATGGGTGTGAACCACACAGAGTACCAGGCCAGCATGGATGT
        I  S  A  P  G  K  G  D  L  K  T  I  V  M  G  V  N  H  T  E  Y  Q  A  S  M  D  V
  +523  TGTGTCCAATGCCTCTTGCACAACGAACTGTCTTGCTCCTCTTGTTCACGTGCTGTTGAAGGAGGGCGTTGGTGTGGAGAA
        V  S  N  A  S  C  T  T  N  C  L  A  P  L  V  H  V  L  L  K  E  G  V  G  V  E  K
  +604  GGGTCGATGACCACCATCCATGCCTACACAGCAACGaagaccctgcttccttggccccatggaaatcggcgtttcttgaa
        G  L  M  T  T  I  H  A  Y  T  A  T
  +685  gagacttttaggtcaaaatttggtgccgaagaaattttctttgagagttgccaaggttgtgggcttaacgccccctccca
  +766  cttcctggttgtccgacccctgttgcccgaatgattgcctgggagattttgagtacccccacttatcaggggcatgtgcaa
  +847  actaactgcatttttttgactttgggtcagagtcttgtggacaaccCAGAAGACTGTCGATGGACCCTCAAAGAAGGACTGG
                                                     Q  K  T  V  D  G  P  S  K  K  D  W
  +928  CGTGGTGGCCGTGCTGCAGCCATTAACATCATCCCCTCTACCACCGGAGCTGCCAAGGCTGTTGGTGAGGTGTTGCCTGCT
        R  G  G  R  A  A  A  I  N  I  I  P  S  T  T  G  A  A  K  A  V  G  E  V  L  P  A
 +1009  GTGAAGGGGAAGCTCACTGGCactcagacattctcttctccaatttgtgtccaattattctgtgggagattctgtccagtt
        V  K  G  K  L  T  G
 +1090  Gttccgtgggctggttggtgggttccctgtccaccaccaccggctggggcttttccttgcaggggtgctttgggccaccag
 +1171  gccttcctctgcatccctgcctgtgggttacaaaaaaaaacagaaacagcctggttaggtccactgaatacaggagttaat
 +1252  ctgcggcagggctgggaatccccacaactccaatggttggtggttaggggcttccagcaaccactcccatatttttgttgca
 +1333  gcaggttctgaccacccccagtggccgggggagcccccaccaggaaagagaacaagaagagacagtctgctgacATGGCC
                                                                                    M  A
 +1414  TTCCGCGTGCCAACCCCTGATGTGTCTGTTGTTGATCTGACCTTCTTGGCTGAGAAGGACACCAGCATCAAGGAGATCGAC
        F  R  V  P  T  P  D  V  S  V  V  D  L  T  F  L  A  E  K  D  T  S  I  K  E  I  D
 +1495  TCCCTGTTGAAGAAGGCCTCCCAGACATACCTCAAGGGAATCTTGGGCTTCACAGATGAGGAGCTCGTGTCGACGGACTTC
        S  L  L  K  K  A  S  Q  T  Y  L  K  G  I  L  G  F  T  D  E  E  L  V  S  T  D  F
 +1576  GTGCATGACAATCGTTCTTCTATCTACGATTCTTTGGCCACCCTCCAGAACAACTTGCCTGGGGAGAAGAGGTTGTTCAAG
        V  H  D  N  R  S  S  I  Y  D  S  L  A  T  L  Q  N  N  L  P  G  E  K  R  L  F  K
 +1657  GTGGTGTCCTGGTATGACAACGAGTGGGGATACTCCAACCGTGTCGTCGACCTGCTGAAGCACATGTCTGGAAACTAAGTT
        V  V  S  W  Y  D  N  E  W  G  Y  S  N  R  V  V  D  L  L  K  H  M  S  G  N  ↓Stop
 +1738  GGAAAGTTGGTTGCACACACTGTGGAGTGACCTGCACTTTTCAAAATGGCATTTGCAGGGCTGCCATCTGAAGAATAGCAA
 +1819  AGGTTTTCACCAATACTAGCCATTTTCTTGTATTTTTTGGTCAAGTTTTTTGGTTTTTCAACTCGAATTTCGAAATCCAAC
```

Fig 3-9: *E. gracilis gapC* gene sequence with the 5’ and 3’ flanking regions and deduced amino acid sequence. Lower case letters indicate an intron sequence. The italic underlined letters denote a possible Kozak sequence including the start codon (Kuo *et al.*, 2013, Yamauchi, 1991). Putative CAAT box is highlighted with yellow colour, and putative CCAAT boxes are highlighted with light blue colour. Putative MYB-related transcription factors are highlighted with dark blue colour. The black highlighted letters denote carbon source responsive elements (CSREs), the red highlighted letters denote a GC box, and the pink highlighted letters denote activator of glycolysis genes. The dark green highlighted letters denote an NAC factor, and purple letters denoted X Core Promoter Element 1 (XCPE1). Further sequence upstream and downstream can be accessed at Genbank Accession No. L39772.

The common transcription factors from binding sites from six different species were analysed using genomatix software suite, Version 3.7. The *E. gracilis* 5’UTR region of *gapC* gene was compared with the 5’UTR region of other five different species. The analysis was carried out to reveal the most highly similar matrix family. The matrix similarity below 0.9 was discarded. The outcome shows that the most of the transcription factors are plant specific (Table 3-5).



Matrix families:

🟧 F$YHSF 🟩 F$YNIT 🟨 P$CAAT 🟥 P$CNAC 🟥 P$DOFF 🟩 P$MYBL 🟩 P$NTMF 🟪 P$TODS

(A) *E. gracilis*; gi|870799; (B) *Arabidopsis lyrata*; gi|297836891; (C) *Trypanosoma grayi*; gi|686649377; (D) *Trypanosoma brucei*; gi|73747658; (E) *Homo sapiens*; gi|568815586; (F) *Oryza sativa*; gi|996703425

Fig 3-10: The common transcription factors binding sites from six different species analysed using genomatix software suite, Version 3.7.

The consensus sequences of the highly similar matrices were compared with the other species which revealed that the most of the consensus sequences were similar to the plant-specific transcription elements. Refer to Appendix 10 for the highly similar transcription factors (matrix similarity >85%) obtained when compared against the transcription factor from plants, human and trypanosomes using MatInspector; Genomatix Software Suite (Cartharius *et al.*, 2005).

Table 3-5: The top most highly similar matrices (100% matrix sim) between the six species are shown in Fig 3-10 and are listed in the following table with the description of each matrix.

Matrix family	Description	Promoter Matches
F$YHSF	Heat shock factors	80.9 % (fungal promoters)
F$YNIT	Activator of nitrogen-regulated genes	76 % (fungal promoters)
P$CAAT	CCAAT binding factors	44.9 % (plant promoters)
P$CNAC	NAC-factors	38.2 % (plant promoters)
P$DOFF	DNA binding with one finger (DOF)	84 % (plant promoters)
P$MYBL	MYB-like transcription factors	95.8 % (plant promoters)
P$NTMF	NAC factors with transmembrane motif	60 % (plant promoters)
P$TODS	Cis regulatory elements	45.2 % (plant promoters)

Chapter 4: Discussion

4.1 *E. gracilis* strain and cultivation

In this study, a variant *E. gracilis* var *saccharophila* was used for the identification of highly expressed proteins produced during cultivation on dark condition supplying suitable growth medium (Section 2.1.1) with a view of isolating a strong nuclear gene promoter with minimal interference from chloroplast proteins. As *E. gracilis* var. *saccharophila* develops chloroplasts and chloroplast-related proteins at a slower rate when cultivated in the dark, it was an optimal strain for isolating a nuclear promoter.

Analysis of the extracted proteins revealed that most of the highly expressed proteins were indeed non-chloroplast proteins. Some of the proteins identified were glyceraldehyde-3-phosphate dehydrogenase, malate synthase-isocitrate lyase, fructose-1,6 bisphosphatase and Elongation factor1.

4.2 Separation of proteins using 2D

In this study, both 1D and 2D SDS-PAGE were used to fractionate proteins extracted from *E. gracilis*. When comparing the two methods, there are several advantages in separating proteins using the 2D approach instead of 1D because of its high resolving power and ability to detect hundreds of proteins on a single gel for their later use them for mass spectrometric analysis.

Mass spectrometer analysis revealed the identity of various proteins. From the fifteen spots selected from the 2D gel (Fig 3-2), some apparently contained more than one protein as several identities were assigned which may be due to similar pI and molecular weight. However, in the main, the identified proteins matched well with their location on the 2D gel regarding molecular weight and pI (Table 4-1). The theoretical molecular weight and pI of each protein identified using mass spectrometry were calculated using the ExPASy (Wilkins *et al.*, 1999) and the results were compared with the location of a corresponding spot on the gel (Table 4-1). There was an excellent match which highlighted the ability of the 2D to fractionate proteins.

Table 4-1: Proteins separated on the various region of 2D gel were compared with the protein identified by mass spectrometry. The range of molecular weights (MW) and pI as indicated by the 2D gel (on the left) and their theoretical MW and pI as predicted by mass spectrometry (on the right) are shown.

Spot	MW (kDa) range	pI range	Proteins identified	Protein MW	Protein pI
1	18-20	6-6.5.0	NADH dehydrogenase	19.41	6.25
2	48-50	4.5-5.0	Tubulin alpla-1 chain	49.8	4.94
3	60-62	5.5-5.8	Heat shock protein 60	59.72	5.60
4	32-35	5.5-5.8	Fructose-1,6 bisphosphatase	34.53	5.43
5	28-30	5-5.5.0	ATP synthase subunit α	29.29	5.21
6	20-22	5-5.5.0	Peroxiredoxin	21.29	5.56
7	32-34	8-8.5.0	L-3-hydroxyacyl-CoA dehydrogenase	33.78	8.67
8	46-48	8.5-9.0	Elongation factor 1-alpha	48.60	9.00
9	38-40	6.5-7.0	Glyceraldehyde-3-phosphate dehydrogenase	38.17	7.12
10	38-40	6.5-7.0	NAD-dependent alcohol dehydrogenase	36.73	6.47
11	125-135	7.0-7.5	Malate synthase-isocitrate lyase	129.75	6.75
12	48-50	7.5-8.0	cAMP-Dependent protein kinase	50.08	8.11
13	50-52	4.5-5.0	Beta-tubulin	50	4.75
14	22-25	9.0-9.5	DNA-directed RNA polymerase subunit alpha	25.36	9.77
15	45-50	5.0-5.8	Enolase	46.51	5.80

4.3 Correlation of highly expressed proteins identified by mass spectrometry with the transcriptomic data obtained from the Expressed sequence tag (EST) database

The ESTs of *E. gracilis* obtained from the TBestDB database were compared with the highly expressed proteins separated by the 2D-SDS-PAGE and subsequently identified using mass spectrometry. The deposited ESTs of *E. gracilis* were arranged in the order from the highest to the lowest, and the top 10 ESTs hits were selected and compared to proteins identified from the 2D gel and mass spectrometric analysis. All proteins with the highest ESTs were identified in the different spots on the 2D gel except one, i.e., malic enzyme (Table 4-2).

Table 4-2: Comparison of proteins with the highest ESTs with proteins identified from the 2D.

Uniref	Description	No of ESTs	Spot in 2D-SDS-PAGE (Fig 3-2)
Q43311	Glyceraldehyde-3-phosphate dehydrogenase	171	Spot 9
P07436	Tubulin beta-1 chain	78	Spot 2
P14963	Elongation factor 1-alpha	68	Spot 8
Q6H798	Acetyl-CoA synthetase	59	Spot 3
Q25563	Tubulin alpha-13 chain	57	Spot 13
Q8LPA6	Malate synthase-isocitrate lyase	51	Spot 11
Q875H8	Malic enzyme	46	-
Q42729	Fructose-1,6-bisphosphate aldolase	45	Spot 4
Q26937	Heat shock protein 70	43	Spot 3
Q4QJF1	ATP synthase alpha chain	31	Spot 5

4.4 PCR amplification of 5'UTR promoter region of *gapC* gene

The complexity of genomic DNA of *E. gracilis* has been noted out previously by O'Neill *et al.* (2015). The complex nature of the *E. gracilis* genomic DNA has probably risen from multiple endosymbiotic events during evolution which may have resulted in the content of more than 80% of repetitive sequences and a large genome of approximately 2 Gbp. The large and complex nature of the *E. gracilis* genomic DNA was also discussed by Montandon and Stutz (1990) who described 40-50 huge chromosomes and highly repetitive sequences of *Euglena*. In addition to that *E. gracilis, the* genome contains the modified nucleotide Base J (β-D-glucosyl-hydroxymethyluracil) which was originally discovered in kinetoplastids (Dooijes *et al.*, 2000). The modified base J hinders DNA polymerase activity which will eventually hamper the PCR reaction making it difficult to optimise the reaction conditions.

The *E. gracilis* genomic DNA also has a high GC content (Yoshida *et al.*, 2016, Ishikawa *et al.*, 2010). While the standard PCR reaction setup works for most DNAs, problems may arise when working with GC-rich genomic DNA. As the pairing between cytosine and guanine creates inter- and intrastrand folds resulting in hairpins and loops (Musso *et al.*, 2006), the template may need high heat to melt, and secondary structures may cause DNA polymerase to stall; also primer annealing may be hindered which (Hube *et al.*, 2005). The addition of organic substances and changing the thermal cycling program may assist to restore correct amplification. The most common additive used for amplification of GC-rich DNA are dimethyl sulfoxide (DMSO); (Pomp and Medrano, 1991), betaine and reducing reagents like DTT (Zhang *et al.*, 2009). Additives like glycerol, formamide and dimethyl sulfoxide (DMSO) help efficient denaturation of GC-rich DNAs and betaine improves the stability of separated strands (Jensen *et al.*, 2010). High temperature may turn out to be useful for separation of strands but it could hamper the polymerase activity and may cause depurination of DNA (Lindahl and Nyberg, 1972). In this study, various PCR optimisations were carried out to obtain correct amplification.

Table 4-3: PCR reaction setup

Component	50 µl reaction	Final Concentration
10X PCR Gold buffer	5 µl	1X
100 mM dNTPs	0.5 µl	1 mM
10 µM Forward Primer	1 µl	0.2 µM
10 µM Reverse Primer	1 µl	0.2 µM
Template DNA	1 µl	100 ng
AmpliTaq Gold® Polymerase 5 U/µl	0.25 µl	1.25 units/50 µl
25mM $MgCl_2$	2 µl	1mM
5X GC Enhancer	2.5 µl	1X
Nuclease-free water	36.75 µl	

Table 4-4: PCR thermocycling parameters

Step	Temp	Time
Initial denaturation	96°C	5 mins
35 cycles	96°C 60°C 72°C	1 min 30 sec 2 min
Final extension	42°C	5 min
Hold	4°C	

For the PCR to be successful, NEB GC enhancer (OneTaq High GC Enhancer: 10 mM Tris-HCI, 25% DMSO 25% Glycerol) was used. In addition to that the initial denaturation was increased up to 96°C for five minutes which is higher temperature compared to normal reaction set up to ensure complete separation of the strands (Table 4-3 and Table 4-4)

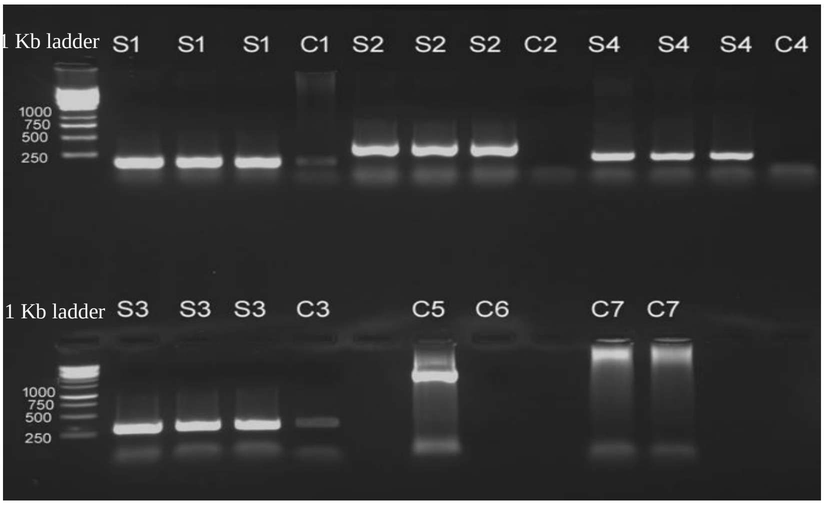


Figure 4-1: Amplification of the various regions of the *gapC* gene of *E. gracilis* to optimise the PCR set up. In total four primers sets (Appendix 8) were used to amplify four regions of the *gapC* gene. S1 denotes three replicates of sample 1 amplified using the primer set 1; S2 denotes three replicates of sample 2 amplified using the primer set 2; S3 denotes three replicates of sample 3 amplified using the primer set 3, and S4 denotes three replicates of sample 4 amplified using the primer set 4. All amplification reactions showed the correct amplification of *E. gracilis* genomic DNA with the expected product size (Appendix 8). C1-C4 are negative controls (without templates), C5 is positive control and C6 is negative control (without primers and template) and C7 is template DNA only. The 1Kb ladder was used as a marker (Thermo Scientific GeneRuler, Australia)

4.5 Highly expressed genes in *E. gracilis*

The result from proteomics analysis of *E. gracilis* discovered many highly expressed proteins. The proteomics data set obtained (Table 3-2) shows that genes encoding e.g. proteins like glyceraldehyde 3-phosphate dehydrogenase (GAPDH), tubulins, elongation factors, and the enzymes malate synthase-isocitrate lyase, enolase are highly expressed. Most of the highly expressed proteins are related to the metabolic cycle of the organism. Comparison of the data obtained here with the ESTs database revealed that the *gapC* gene was the most desirable candidate since the ESTs show GAPDH is the highest hit of deposited sequence of the organism (Table 3-3). The other highly expressed proteins could also be used in future to obtain the promoter sequence if the GAPDH promoter sequence is not functional in *E. gracilis*.

4.5.1 Glyceraldehyde 3-phosphate dehydrogenase (*gapC*), a highly expressed gene in *E. gracilis*

Tracing back from the protein to a gene transcript revealed several different highly expressed genes of *E. gracilis* of which the most highly expressed gene was found to be *gapC* encoding the glyceraldehyde-3-phosphate dehydrogenase protein (GAPDH). *gapC* is a housekeeping gene which has a basic role in cellular metabolic processes. The molecular weight of the GAPDH protein is 37 kDa, and it exists as a tetramer (Tristan *et al.*, 2011). GAPDH has a role in glycolysis as it catalyses the conversion of glyceraldehyde-3-phosphate to 1,3-bisphosphoglycerate. Both cytosolic and plastidial GAPDHs of *E. gracilis* have been purified and characterised (Henze *et al.*, 1995, Leegood *et al.*, 2006). *E. gracilis* plastidial GAPDH utilises an NADP$^+$ cofactor while the cytosolic GAPDH utilises NAD$^+$ cofactor (Grissom and Kahn, 1975, Henze *et al.*, 1995). The highly expressed GAPDH (Uniref: Q43311) identified in this study was the cytosolic NAD$^+$ cofactor-dependent enzyme encoded by the nuclear *gapC* gene. Our finding is also supported by an earlier study conducted by Grissom and Kahn (1975) who discovered that only the dark grown *E. gracilis* cells possessed NAD$^+$; in the presence of light, the culture produced NADP$^+$. Their study also highlighted the fact that the NAD$^+$ dependent GAPDH was related to cell division and appeared to be constitutive. In the current study, the *E. gracilis* cells were grown in the dark which results in the production of NAD$^+$ dependent GAPDH, which is constitutively expressed from the *gapC* gene.

It has been argued that GAPDH in eukaryotes is encoded by an ancient gene copied from the eubacterial genome, transferred to the nuclear genome by several endosymbiotic events (Petersen *et al.*, 2003, Figge *et al.*, 1999, Henze *et al.*, 1995). A study conducted by Henze *et al.* (1995) endosymbiosis during protist evolution was based on the role of GAPDH in both glycolysis and the Calvin cycle of *E. gracilis*. To investigate the evolution, Henze *et al.* (1995) cloned cDNAs corresponding both types of GAPDH from *E. gracilis*. As a result, the molecular structure of the nuclear gene coding for the cytosolic GAPDH (nuclear *gapC* gene; GI: 870799) was reported. The *gapC* gene encoding the cytosolic NAD+ dependent enzyme has four introns, a very unusual secondary structure that does not follow the GT-AG rule, and is flanked by two to three base pairs of direct repeats.

The proteomics data revealed a highly expressed GAPDH protein. Hence it was decided to isolate the 5'UTR promoter region of the *gapC* gene encoding GAPDH protein. However, the highly abundant protein does not necessarily correlate with the promoter activity. However, in our case, the highly expressed protein GAPDH was associated with the strong promoter, as the 5'UTR promoter region of *gapC* gene was able to drive the expression of reporter gene *eGFP*. It should be noted that there are other factors as well which will also determine the promoter activity and that could be protein translation event, protein stability, time to express the protein relative to other promoters and posttranscriptional regulation.

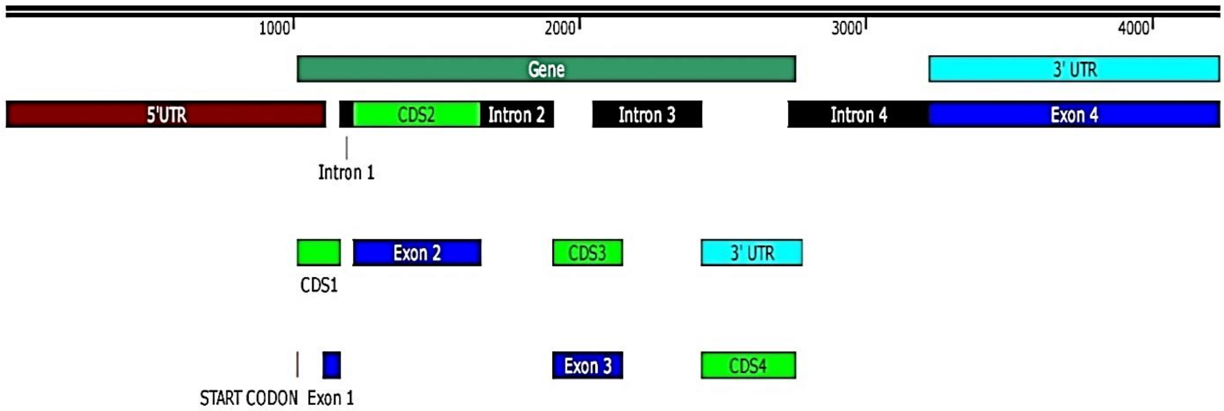


Figure 4-2: Schematic representation of the structure of the *E. gracilis* nuclear *gapC* gene on the sequenced 4.2-kb HindIII fragment (Henze *et al.*, 1995). Sequence analysis showed the presence of three introns in the CDS and one in 3'UTR region. Refer to Appendix 9 for complete CDS and its translation. (Schematic drawing: SnapGene Version 3.2.1).

4.5.1 The yeast *GAP* gene promoter

Several studies of the *GAP* gene promoter have been previously reported in different organisms such as the yeasts *Pichia pastoris* (Olędzka *et al.*, 2003) and *S. cerevisiae* (Rosenberg and Tekamp-Olson, 1992), and the green alga *Dunaliella salina* (Jia *et al.*, 2012). Till date, there are no reports describing isolation and characterisation of an *E. gracilis gap* gene promoter.

The well-studied yeast *GAP* promoter is a constitutive promoter conferring high-level gene expression (Waterham *et al.*, 1997). Their study revealed that the *GAP* promoter was constitutively expressing *β-lactamase* reporter gene however the expression of the reporter gene was slightly variable on the carbon source. Yeast is a suitable host to produce heterologous proteins and biologically significant compounds. Edens *et al.* (1984) reported the use of *GAP* promoter in yeast to produce plant protein thaumatin. The promoter has also been used to express biologically active human proteins like cytochrome P450 (Wu *et al.*, 1991), Golgi UDP-galactose transporters (Sun-Wada *et al.*, 1998), Cu/Zn superoxide dismutase (SOD1) (Yoo *et al.*, 1999) and α_1 -antitrypsin (Brake *et al.*, 1988). Industrial levels of human epidermal growth factor (urogastrone) have been made in *Pichia pastoris* (Urdea *et al.*, 1983). The above examples highlight the usefulness of the *GAP* promoter in recombinant protein expression.

The newly isolated *E. gracilis gapC* gene promoter was shown to be functional in *S. cerevisiae* with successful expression of the eGFP protein although the promoter function has to be further verified in *Euglena*. Drawing an analogy from the yeast work, the *E. gracilis gapC* promoter could be used to express valuable genes in both yeast and *E. gracilis*.

4.6 *E.coli* transformants with leaky expression of the *β-gal* gene on X-gal plates

In this study, the *β-gal* gene in a promoter-less *E. coli* plasmid OG372 (Appendix 1) was used as a reporter for promoter activity. The 5'UTR DNA of the *E. gracilis gapC* gene was cloned upstream of the *β-gal* gene. Therefore, a functional promoter will drive expression of the *β-gal* gene resulting in the formation of blue colonies. This approach did not work as expected as all colonies on X-gal plates were blue along colonies on the control plate, transformed with a plasmid without a promoter. Only white colonies should appear on the control plate containing an empty plasmid since the *β-gal* gene should not have been expressed without a promoter.

The reason for false positives may have been due to non-specific transcriptional elements located upstream of the *β-gal* gene in the plasmid, possibly causing a read-through of not terminated transcripts originating from the kanamycin resistance gene (Appendix 1). Even low levels of expression of the *β-gal* gene result in blue colonies (Sherwood, 2003).

4.7 Testing promoter activity in yeast

The functionality of the isolated promoter region of *gapC* gene from *E. gracilis* was verified in *S. cerevisiae*. The plasmid pRS426 (Appendix 2) which initially possessed the yeast *PGK1* promoter was replaced with the newly isolated *E. gracilis gapC* gene promoter (Fig 2-2).

The yeast plasmid pRS426 containing the *gapC* promoter upstream of the reporter *eGFP* gene and the *URA3* selection marker encoding orotidine-5′-phosphate decarboxylase, a crucial enzyme in pyrimidine biosynthesis in *S. cerevisiae* (Boeke *et al.*, 1984) was transformed into an auxotrophic mutant strain BY4742 (MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0). Transformant selection was carried out by plating the cells onto medium lacking uracil on which only cells which were able to take up the plasmid could grow. Well growing colonies were further tested for fluorescence resulting from successful expression of the *eGFP* reporter gene driven by the *E. gracilis gapC* promoter.

The positive control plates A and B containing transformants carrying the original intact plasmid gave numerous colonies in the yeast drop-out agar medium (-uracil) plates suggesting that the transformation had worked (Fig 4-3). Control plates E and F were plated with cells transformed ligated vector without a promoter. There were five colonies growing on the plate which confirmed successful ligation and transformation (Fig 4-3). Plates C and D transformed with the plasmid containing *gapC* promoter gave 460 colonies (Fig 4-3).

The functionality of the *gapC* promoter was validated upon successful expression of the *eGFP* gene by the *S. cerevisiae* cells. The eGFP protein has the property of fluorescing when excited by UV or short-wave blue light (Chalfie *et al.*, 1994). It has an excitation maximum at 488 nm and minimum at 509 nm (Gomes *et al.*, 2002). Colonies on the positive control plates A and B transformed with the intact plasmid containing the yeast *PGK1* promoter were all fluorescent which means transformation was successful, and that *PGK1* promoter drove the expression of the *eGFP* gene (plate B; Fig 4-3). The plates transformed with the ligated plasmid with the newly isolated *gapC* gene promoter from *E. gracilis* were also fluorescent in the blue light (plate D; Fig 4-3). This result suggests that the *gapC* promoter was functional in yeast driving expression of the *eGFP* gene. The negative control plates with transformants containing the ligated plasmid without any promoter sequence were not fluorescent (plate F; Fig 4-3). The result illustrates that only a functional promoter could drive the expression of the reporter gene.

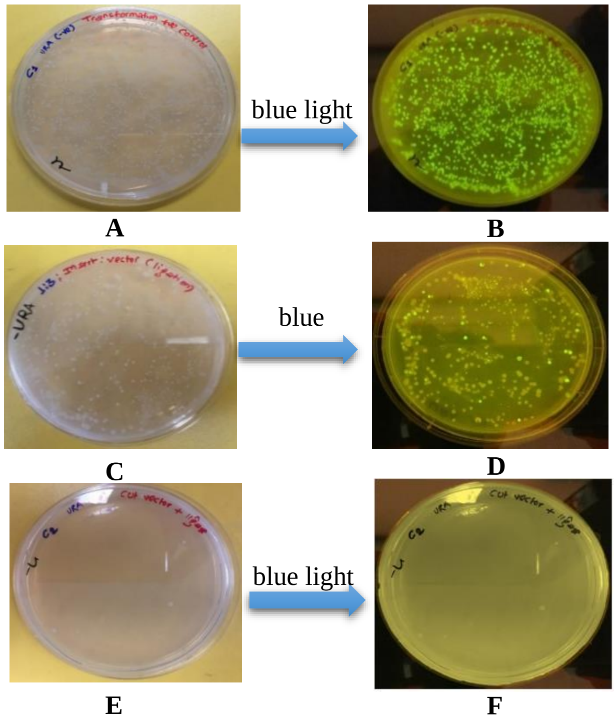


Fig 4-3: The *S. cerevisiae* colonies transformed with the plasmid pRS426 (plates C and D) containing the reporter *eGFP* gene under promoter *gapC* were fluorescent when exposed to blue light. From the approximately 460 colonies on the plates, 25 colonies were fluorescent. This result confirms successful transformation and functionality of promoter *gapC*. Plates A and B are the positive controls transformed with the original plasmid pRS426 with *PGK1* promoter. All colonies were fluorescent when exposed to the blue light. The plates E and F are the negative controls of cells transformed with the plasmid without any promoter sequence. Only five colonies appeared on the plates, and none of them were fluorescent as they lacked a promoter. This result demonstrates successful ligation and transformation. the eGFP expression in *S. cerevisiae* transformed with the original plasmid containing

the yeast *PGK1* promoter and transformants containing the plasmid harboring the newly isolated *E. gracilis gapC* promoter will be performed to further explore the strength of the *gapC* promoter using fluorescent microscopy (Wei and Dai, 2014), flow cytometry (Ducrest *et al.*, 2002) and microplate assays (Collins *et al.*, 1998).

4.8 Heterologous promoters in *S. cerevisiae*

In this study, *S. cerevisiae* was transformed with a yeast shuttle vector containing a non-native *E. gracilis gapC* promoter. Successful transformation was achieved and the transformed cells were expressing *eGF*P gene under the foreign promoter. While native yeast promoters like *PGK1*, *ADH1*, *PYK1*, *TEF1* and *ACT1* are frequently used to express heterologous genes in yeast (Young *et al.*, 2016), the use of a non-native promoter in yeast is very rare.

There is a limited number of reports on functional heterologous promoters in *S. cerevisiae*. For example, a non-native Cauliflower Mosaic Virus (CaMV) 35S promoter was used to express bacterial beta-glucuronidase (*GUS*) reporter gene in *S. cerevisiae*. The study revealed the accurate translation of GUS-mRNA into GUS protein although the expression level was moderate in comparison to the native *ADH1* promoter (Hirt *et al.*, 1990). The study illustrated the fact that the heterologous promoter worked in *S. cerevisiae*. Expression of the *eGFP* reporter gene under the non-native *gapC* promoter in *S. cerevisiae* suggests that *gapC* promoter could be further tested to drive the expression of the valuable gene in yeast.

The fact that the heterologous *E. gracilis gapC* gene promoter is functional in yeast suggests that the transcriptional machinery of *S. cerevisiae* and *E. gracilis* share some similarities. The promoter characterisation further revealed that yeast transcriptional factor binding sites such as the yeast GC-Box, yeast activator of glycolysis genes binding sites and yeast stress response elements binding site (Appendix 10) were available in the *gapC* promoter. Functionality testing of the *gapC* promoter in *E. gracilis* will follow after establishing a transformation platform for the organism for the final confirmation of promoter activity in its native environment. The choice of yeast to test the heterologous promoter *gapC* of *E. gracilis*, was because of lack of transformation platform for *Euglena*. Furthermore, ease of culture, highly characterised genome, established transformation platform and ease of selection of tranformants made us to choose the a yeast to carry out functionality test of *gapC* promoter in it.

4.9 Characterisation of the *E. gracilis gapC* gene promoter (*gapC*)

To study and characterise the promoter, the entire 5’UTR sequence was analysed using MatInspector; Genomatix Software Suite (Cartharius *et al.*, 2005). This 5’UTR (1015 bp) region of the *gapC* gene contains two CCAAT boxes, one CAAT box and one GC box, implying its function as a promoter. A Kozak translation initiation sequence is also present (in italic and underlined, Fig 3-9). Two MYB like transcription factor binding sites are present (Fig 3-9; blue highlighted) with 100% matrix similarity. The *myb* gene family encodes a nuclear protein which has a role as transcriptional transactivators. They also function as regulators of plant stress response (Ambawat *et al.*, 2013). MYB factors family of proteins comprise the conserved MYB DNA-binding domain found in eukaryotes. In plants, the R2R3-type domain represents the MYB-protein subfamily, and R2R3-type MYB genes control the formation of secondary metabolites. Cytosolic GAPDH is involved in glycolysis which may be regulated by these MYB like transcription factors (Stracke *et al.*, 2001). Carbon source responsive elements (CSREs) were found in *gapC* promoter (Fig 4-5; black highlighted) which initiate the transcription of many genes involved in gluconeogenesis (Bitter *et al.*, 1991). CSREs are found to be active when glucose concentration is low, so it can be inferred that the functionality of CSRE transcriptional elements in the *gapC* promoter is consistent with participation of the *gapC* in gluconeogenesis. GAPDH has a significant role in both glycolysis and gluconeogenesis by reversibly catalysing the oxidation and phosphorylation of intermediate glyceraldehyde 3-phosphate dehydrogenase to the energy-rich intermediate 1,3-bisphosphoglycerate (Guan and Xiong, 2011).

The identified *gapC* promoter sequence lacks a TATA box which is one of the conserved promoter element in eukaryotes. TATA box is typically located 25-35 bp upstream of the transcription start site. The known TATA-less promoters have a common character of having several GC boxes that bind to the transcriptional activator (Pugh and Tjian, 1991). The identified *gapC* promoter sequence possess a GC box (Fig 3-9; red highlighted) which may have a role in initiating the transcription of the *gapC* gene. In previous studies, the GC elements were found to possess related features as enhancers and bind to transcription factors (Klug and Cummings, 2009). An activator of glycolytic genes was found to be located 13 bp upstream of the translation start site in the *gapC* promoter (Fig 3-9; pink highlighted). The activator is known as RAP1 transcriptional regulator. It has an ability to undertake various cellular functions like transcription regulation (activation/repression) and telomere growth (Shore, 1994).

The NAC transcription factor binding site was found to located 24 bp upstream of the translation start site of the *gapC* promoter (Fig 3-9; dark green highlighted). These plant-specific transcriptional factors are known to regulate various plant specific development programs and are one of the key transcription factor families in plants. They also have a role in DNA binding and dimerisation, activation and repression of transcription; they are also found to be a prominent factor in stress signalling pathways. Since *E. gracilis* possess some features characteristic of plants, NAC may have a role in activation of the *gapC* gene. X Core Promoter Element 1 (XCPE1) also exist in the *gapC* promoter (Fig 3-9; purple highlighted) which has a role in driving RNA polymerase II transcription from a TATA-less promoter (Tokusumi *et al.*, 2007). Though the identified *gapC* promoter is a TATA-less promoter, it has other transcriptional factors like XCPE1 and GC box which are found to regulate transcription from a TATA-less promoter.

4.9.1 TATA-less promoters

Transcription initiation in eukaryotes are not only dependent on the TATA box. In their research, (Pugh and Tjian, 1991) disclosed the fact that although a promoter lacks a TATA box, the TFIID complex that possesses a TATA box binding protein can conduct transcription initiation forming a multisubunit complex comprising of TBP and multiple TBP-associated factors (TAFs). Their findings also suggested that the specificity protein 1 (Spl) and CCAAT box transcription factor (CTF) regulate transcription of a promoter lacks TATA box.

Promoters that lack the TATA element have been found to have a downstream promoter element (DPE) which functions cooperatively with the initiator (INR). DPE-dependent promoters were found to be used in the absence of TATA Box in *Drosophila* (Kutach and Kadonaga, 2000). Characterisation of *S. cerevisiae* promoters by Seizl *et al.* (2011) showed that only 15% of yeast promoters contained a TATA box, while more than one-third of the yeast promoters possess GA element (GAE), which regulates transcription without a TATA box. Initiator elements (INRs) are often found at the transcription start site and are sufficient to direct the RNA Pol II transcription without a TATA box (Martinez *et al.*, 1994). In the study conducted by (Yang *et al.*, 2007), they indicated that about 76% of human core promoters lack TATA-like elements and possess Sp1 binding sites. Moreover, they discovered that about 30% of human core promoters containing the consensus INR are TATA-less. They also suggest that "housekeeping" genes generally are TATA-less, which is the case of the *gapC* gene of *E. gracilis*. In conclusion, although the *gapC* promoter lacks a TATA box, it has other transcriptional elements like GC box, XCPE1, CAAT box, NAC and MYB transcription factor binding sites which was recognised by *S. cerevisiae* and hence, the *gapC* drove the expression of the *eGFP* gene.

It now evident that the identified *gapC* promoter from *E. gracilis* is functional in *S. cerevisiae.* The identified transcription factor binding sites (TFBS) of *gapC* was recognised by the yeast transcription machinery and helps to express the *eGFP* gene. As discussed earlier, the identified TFBS are highly similar to the plant specific promoter element and which may be resultant of *Euglena* possessing the characteristics of plants as well. Hence it can be inferred from the overall results that the *gapC* being native to *E. gracilis,* it should be functional in *Euglena* as well.

Chapter 5: Conclusion and future directions

E. gracilis is a promising host to produce nutraceuticals, cosmetics, vitamins, minerals, amino acids, paramylon, wax esters and more. Considering the potential, very little is currently known about the molecular biology of the organism. Although *E. gracilis* transcriptomes are now available, complete genome sequence data are still missing. Also, tools and methods for genetic transformation of *Euglena* are sparse. Lack of an efficient nuclear transformation method may be the reason creating an obstacle for genetic engineering of the organism. Moreover, lack of information of strong promoters and regulatory elements are restricting progress towards making *E. gracilis* an industrially viable organism.

Proteomics approach to identify the highly expressed gene of *E. gracilis* has led to the outcome and revealed that the *gapC* gene encoding GAPDH protein is the most highly expressed gene. The potential promoter region 5'UTR of the *gapC* gene was indeed a functional promoter as it drives the expression of *eGFP* gene in *S. cerevisiae*. The identified *E. gracilis gapC* gene promoter (*gapC*) was further characterised by the presence of the core promoter elements and cis-acting elements. Though the *gapC* lacked core promoter element TATA box, it has other binding sites for significant transcriptional factors like XCPE1 and GC box which are found to regulate transcription in a TATA-less promoter. The most of the identified transcription factor in *E. gracilis* matched to the plant-specific transcription elements which may be because of possessing characteristics of plants as well.

Further work needs to be carried out to prove that the identified promoter *gapC* is functional in *E. gracilis*. A transcription analysis in yeast would also provide an idea of the strength of the promoter compared to e.g. the well-characterised yeast *GAP* promoter. A future goal stemming from this research would be to use the identified promoter for assembling it into a *Euglena* expression vector along with other components such as the transcription terminator (potentially the 3'UTR of the *gapC* gene), multiple cloning sites and a transformation selection marker. After validation of the promoter in *E. gracilis* the next step would be to express valuable genes in *Euglena*, for instance with a view to increase production of paramylon, vitamins, amino acids and biomass. Translation of the research into industrial applications will be the ultimate outcome of the entire project.

References

AEBERSOLD, R. & MANN, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207.

AMBAWAT, S., SHARMA, P., YADAV, N. R. & YADAV, R. C. 2013. MYB transcription factor genes as regulators for plant responses: an overview. *Physiology and Molecular Biology of Plants,* 19**,** 307-321.

AMBERG, D. C., BURKE, D. J. & STRATHERN, J. N. 2005. Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual, 2005 Edition (Cold Spring).

ARMALEO, D., YE, G.-N., KLEIN, T. M., SHARK, K. B., SANFORD, J. C. & JOHNSTON, S. A. 1990. Biolistic nuclear transformation of *Saccharomyces cerevisiae* and other fungi. *Current Genetics,* 17**,** 97-103.

BARSANTI, L., PASSARELLI, V., EVANGELISTA, V., FRASSANITO, A. M. & GUALTIERI, P. 2011. Chemistry, physico-chemistry and applications linked to biological activities of β-glucans. *Natural Product Reports,* 28**,** 457-466.

BARSANTI, L., VISMARA, R., PASSARELLI, V. & GUALTIERI, P. 2001. Paramylon (β-1, 3-glucan) content in wild type and WZSL mutant of *Euglena gracilis*. Effects of growth conditions. *Journal of Applied Phycology,* 13**,** 59-65.

BITTER, G. A., CHANG, K. K. & EGAN, K. M. 1991. A multi-component upstream activation sequence of the *Saccharomyces cerevisiae* glyceraldehyde-3-phosphate dehydrogenase gene promoter. *Molecular and General Genetics MGG,* 231**,** 22-32.

BOEKE, J. D., LA CROUTE, F. & FINK, G. R. 1984. A positive selection for mutants lacking orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Molecular and General Genetics MGG,* 197**,** 345-346.

BORST, P. & SABATINI, R. 2008. Base J: discovery, biosynthesis, and possible functions. *Annual Review Microbiology,* 62**,** 235-251.

BOYNTON, J. E., GILLHAM, N. W., HARRIS, E. H., HOSLER, J. P., JOHNSON, A. M., JONES, A. R., RANDOLPH-ANDERSON, B. L., ROBERTSON, D., KLEIN, T. M. & SHARK, K. B. 1988. Chloroplast transformation in *Chlamydomonas* with high velocity microprojectiles. *Science,* 240**,** 1534-1538.

BRAKE, A. J., HALLEWELL, R. A. & ROSENBERG, S. 1988. Expression of α-1 antitrypsin in yeast Google Patents [http://www.google.ch/patents/US4752576].

BUCHER, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology,* 212**,** 563-578.

CANDIANO, G., BRUSCHI, M., MUSANTE, L., SANTUCCI, L., GHIGGERI, G. M., CARNEMOLLA, B., ORECCHIA, P., ZARDI, L. & RIGHETTI, P. G. 2004. Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. *Electrophoresis,* 25**,** 1327-1333.

CARTHARIUS, K., FRECH, K., GROTE, K., KLOCKE, B., HALTMEIER, M., KLINGENHOFF, A., FRISCH, M., BAYERLEIN, M. & WERNER, T. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics,* 21**,** 2933-2942.

CHALFIE, M., TU, Y., EUSKIRCHEN, G., WARD, W. W. & PRASHER, D. C. 1994. Green fluorescent protein as a marker for gene expression. *Science,* 263**,** 802-5.

CHOW, K.-C. & TUNG, W. 1999. Electrotransformation of *Chlorella vulgaris*. *Plant Cell Reports,* 18**,** 778-780.

COLLINS, L. A., TORRERO, M. N. & FRANZBLAU, S. G. 1998. Green Fluorescent Protein Reporter Microplate Assay for High-Throughput Screening of Compounds against *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy,* 42**,** 344-347.

CORPET, F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research,* 16**,** 10881-10890.

CURACH, N. C., TE'O, V. S., GIBBS, M. D., BERGQUIST, P. L. & NEVALAINEN, K. H. 2004. Isolation, characterization and expression of the hex1 gene from *Trichoderma reesei*. *Gene,* 331**,** 133-140.

DANIELL, H. 1997. Transformation and foreign gene expression in plants mediated by microprojectile bombardment. *Recombinant Gene Expression Protocols***,** 463-489.

DOETSCH, N. A., FAVREAU, M. R., KUSCUOGLU, N., THOMPSON, M. D. & HALLICK, R. B. 2001. Chloroplast transformation in *Euglena gracilis*: splicing of a group III twintron transcribed from a transgenic psbK operon. *Current Genetics,* 39**,** 49-60.

DOOIJES, D., CHAVES, I., KIEFT, R., DIRKS-MULDER, A., MARTIN, W. & BORST, P. 2000. Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. *Nucleic acids research,* 28**,** 3017-3021.

DUCREST, A. L., AMACKER, M., LINGNER, J. & NABHOLZ, M. 2002. Detection of promoter activity by flow cytometric analysis of GFP reporter expression. *Nucleic Acids Research,* 30**,** e65.

DUNN, M., RAMIREZ-TRUJILLO, J. & HERNANDEZ-LUCAS, I. 2009. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. *Microbiology,* 155**,** 3166-3175.

ECONOMOU, C., WANNATHONG, T., SZAUB, J. & PURTON, S. 2014. A simple, low-cost method for chloroplast transformation of the green alga *Chlamydomonas reinhardtii*. *Chloroplast Biotechnology: Methods and Protocols,* 1132 401-411.

EDENS, L., BOM, I., LEDEBOER, A. M., MAAT, J., TOONEN, M. Y., VISSER, C. & VERRIPS, C. T. 1984. Synthesis and processing of the plant protein thaumatin in yeast. *Cell,* 37**,** 629-633.

FIGGE, R. M., SCHUBERT, M., BRINKMANN, H. & CERFF, R. 1999. Glyceraldehyde-3-phosphate dehydrogenase gene diversity in eubacteria and eukaryotes: evidence for intra-and inter-kingdom gene transfer. *Molecular biology and evolution,* 16**,** 429-440.

GIETZ, R. D. & SCHIESTL, R. H. 2007. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature protocols,* 2**,** 31-34.

GOMES, I., FILIPOVSKA, J., JORDAN, B. A. & DEVI, L. A. 2002. Oligomerization of opioid receptors. *Methods,* 27**,** 358-365.

GONG, Y., HU, H., GAO, Y., XU, X. & GAO, H. 2011. Microalgae as platforms for production of recombinant proteins and valuable compounds: progress and prospects. *Journal of Industrial Microbiology and Biotechnology,* 38**,** 1879-1890.

GRECO, M., SÁEZ, C. A., BROWN, M. T. & BITONTI, M. B. 2014. A Simple and Effective Method for High Quality Co-Extraction of Genomic DNA and Total RNA from Low Biomass *Ectocarpus siliculosus*, the Model Brown Alga. *Public Library of Science,* 9**,** e96470.

GRISSOM, F. E. & KAHN, J. S. 1975. Glyceraldehyde-3-phosphate dehydrogenases from *Euglena gracilis*: Purification and physical and chemical characterization. *Archives of biochemistry and biophysics,* 171**,** 444-458.

GUAN, K.-L. & XIONG, Y. 2011. Regulation of intermediary metabolism by protein acetylation. *Trends in biochemical sciences,* 36**,** 108-116.

GUO, H. & XIONG, J. 2006. A specific and versatile genome walking technique. *Gene,* 381**,** 18-23.

HALLICK, R. B., HONG, L., DRAGER, R. G., FAVREAU, M. R., MONFORT, A., ORSAT, B., SPIELMANN, A. & STUTZ, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Research,* 21**,** 3537-3544.

HALLMANN, A. 2007. Algal transgenics and biotechnology. *Transgenic Plant Journal,* 1**,** 81-98.

HANAHAN, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *Journal of molecular biology,* 166**,** 557-580.

HANAHAN, D. 1985. Techniques for transformation of E. coli. *DNA cloning,* 1**,** 109-135.

HENZE, K., BADR, A., WETTERN, M., CERFF, R. & MARTIN, W. 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proceedings of the National Academy of Sciences,* 92**,** 9122-9126.

HERNANDEZ-GARCIA, C. M. & FINER, J. J. 2014. Identification and validation of promoters and cis-acting regulatory elements. *Plant Science,* 217**,** 109-119.

HIRT, H., KOGL, M., MURBACHER, T. & HEBERLE-BORS, E. 1990. Evolutionary conservation of transcriptional machinery between yeast and plants as shown by the efficient expression from the CaMV 35S promoter and 35S terminator. *Current Genetics,* 17**,** 473-9.

HUBE, F., REVERDIAU, P., IOCHMANN, S. & GRUEL, Y. 2005. Improved PCR method for amplification of GC-rich DNA sequences. *Molecular biotechnology,* 31**,** 81-84.

HUTNER, S., BACH, M. K. & ROSS, G. 1956. A Sugar-Containing Basal Medium for Vitamin B12-Assay with *Euglena;* Application to Body Fluids. *The Journal of Protozoology,* 3**,** 101-112.

INOUE, H., NOJIMA, H. & OKAYAMA, H. 1990. High efficiency transformation of *Escherichia coli* with plasmids. *Gene,* 96**,** 23-28.

ISHIKAWA, T., TAJIMA, N., NISHIKAWA, H., GAO, Y., RAPOLU, M., SHIBATA, H., SAWA, Y. & SHIGEOKA, S. 2010. *Euglena gracilis* ascorbate peroxidase forms an intramolecular dimeric structure: its unique molecular characterization. *Biochemical Journal,* 426**,** 125-134.

JENSEN, M. A., FUKUSHIMA, M. & DAVIS, R. W. 2010. DMSO and betaine greatly improve amplification of GC-rich constructs in de novo synthesis. *Public Library of Science,* 5**,** e11024.

JIA, Y., LI, S., ALLEN, G., FENG, S. & XUE, L. 2012. A novel glyceraldehyde-3-phosphate dehydrogenase (GAPDH) promoter for expressing transgenes in the halotolerant alga *Dunaliella salina*. *Current microbiology,* 64**,** 506-513.

KEMPNER, E. & MILLER, J. 1968. The molecular biology of *Euglena gracilis*: V. Enzyme localization. *Experimental cell research,* 51**,** 150-156.

KINDLE, K. L., RICHARDS, K. L. & STERN, D. B. 1991. Engineering the chloroplast genome: techniques and capabilities for chloroplast transformation in *Chlamydomonas reinhardtii*. *Proceedings of the National Academy of Sciences,* 88**,** 1721-1725.

KLEBS, G. 1893. *Flagellatenstudien*, Akadem. Verlag-Ges.

KLUG, W. S. & CUMMINGS, M. R. 2009. *Concepts of Genetics: Ninth Edition,* San Francisco, Pearson Benjamin Cummings, 463–464.

KOIZUMI, N., SAKAGAMI, H., UTSUMI, A., FUJINAGA, S., TAKEDA, M., ASANO, K., SUGAWARA, I., ICHIKAWA, S., KONDO, H. & MORI, S. 1993. Anti-HIV (human immunodeficiency virus) activity of sulfated paramylon. *Antiviral Research,* 21**,** 1-14.

KONDO, Y., KATO, A., HOJO, H., NOZOE, S., TAKEUCHI, M. & OCHI, K. 1992. Cytokine-Related Immunopotentiating Activities of Paramylon, a. BETA.-(1. RAR. 3)-D-Glucan from *Euglena gracilis*. *Journal of Pharmacobio-Dynamics,* 15**,** 617-621.

KONDRASHOV, F. A., KOONIN, E. V., MORGUNOV, I. G., FINOGENOVA, T. V. & KONDRASHOVA, M. N. 2006. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biology direct,* 1**,** 1.

KRAJČOVIČ, J., VESTEG, M. & SCHWARTZBACH, S. D. 2015. Euglenoid flagellates: A multifaceted biotechnology platform. *Journal of Biotechnology,* 202**,** 135-145.

KUMAR, S. V., MISQUITTA, R. W., REDDY, V. S., RAO, B. J. & RAJAM, M. V. 2004. Genetic transformation of the green alga—*Chlamydomonas reinhardtii* by *Agrobacterium tumefaciens*. *Plant Science,* 166**,** 731-738.

KUO, R. C., ZHANG, H., ZHUANG, Y., HANNICK, L. & LIN, S. 2013. Transcriptomic Study Reveals Widespread Spliced Leader Trans-Splicing, Short 5′-UTRs and Potential Complex Carbon Fixation Mechanisms in the Euglenoid Alga *Eutreptiella* sp. *Public Library of Science,* 8**,** e60826.

KUTACH, A. K. & KADONAGA, J. T. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Molecular and cellular biology,* 20**,** 4754-4764.

LACHNEY, C. L. & LONERGAN, T. A. 1985. Regulation of cell shape in *Euglena gracilis*. III. Involvement of stable microtubules. *Journal of cell science,* 74**,** 219-237.

LAGRANGE, T., KAPANIDIS, A. N., TANG, H., REINBERG, D. & EBRIGHT, R. H. 1998. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes and Development,* 12**,** 34-44.

LEEGOOD, R. C., SHARKEY, T. D. & VON CAEMMERER, S. 2006. *Photosynthesis: physiology and metabolism,* Volume 9. Springer Science & Business Media. Chapter 2 , 27-28.

LINDAHL, T. & NYBERG, B. 1972. Rate of depurination of native deoxyribonucleic acid. *Biochemistry,* 11**,** 3610-3618.

LURIA, S. & BURROUS, J. W. 1957. Hybridization between *Escherichia coli* and *Shigella*. *Journal of bacteriology,* 74**,** 461.

MANTOVANI, R. 1998. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Research,* 26**,** 1135-1143.

MARTIN, W., SOMERVILLE, C. & LOISEAUX-DE GOER, S. 1992. Molecular phylogenies of plastid origins and algal evolution. *Journal of molecular evolution,* 35**,** 385-404.

MARTINEZ, E., CHIANG, C.-M., GE, H. & ROEDER, R. 1994. TATA-binding protein-associated factor (s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. *European Molecular Biology Organization,* 13**,** 3115.

MATSUDA, F., HAYASHI, M. & KONDO, A. 2011. Comparative profiling analysis of central metabolites in *Euglena gracilis* under various cultivation conditions. *Bioscience, Biotechnology, and Biochemistry,* 75**,** 2253-2256.

MEDERIC, C., BERTAUX, O., ROUZEAU, J. D. & VALENCIA, R. 1987. Isolation of high molecular weight DNA from whole *Euglena* cells. *Protoplasma,* 141**,** 139-148.

MONTANDON, P.-E. & STUTZ, E. 1990. Structure and expression of the *Euglena gracilis* nuclear gene coding for the translation elongation factor EF-1 α. *Nucleic acids research,* 18**,** 75-82.

MUELLER, P. R. & WOLD, B. 1989. In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. *Science,* 246**,** 780-786.

MUSSO, M., BOCCIARDI, R., PARODI, S., RAVAZZOLO, R. & CECCHERINI, I. 2006. Betaine, dimethyl sulfoxide, and 7-deaza-dGTP, a powerful mixture for amplification of GC-rich DNA sequences. *The Journal of Molecular Diagnostics,* 8**,** 544-550.

O'BRIEN, E. A., KOSKI, L. B., ZHANG, Y., YANG, L., WANG, E., GRAY, M. W., BURGER, G. & LANG, B. F. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Research,* 35**,** 445-451.

O'NEILL, E. C., TRICK, M., HILL, L., REJZEK, M., DUSI, R. G., HAMILTON, C. J., ZIMBA, P. V., HENRISSAT, B. & FIELD, R. A. 2015. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Molecular BioSystems,* 11**,** 2808-2820.

O'NEILL, E. C., TRICK, M., HENRISSAT, B. & FIELD, R. A. 2015. *Euglena* in time: Evolution, control of central metabolic processes and multi-domain proteins in carbohydrate and natural product biochemistry. *Perspectives in Science,* 6**,** 84-93.

OCHMAN, H., GERBER, A. S. & HARTL, D. L. 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics,* 120**,** 621-623.

OGAWA, T., TAMOI, M., KIMURA, A., MINE, A., SAKUYAMA, H., YOSHIDA, E., MARUTA, T., SUZUKI, K., ISHIKAWA, T. & SHIGEOKA, S. 2015. Enhancement of photosynthetic capacity in *Euglena gracilis* by expression of cyanobacterial fructose-1, 6-/sedoheptulose-1, 7-bisphosphatase leads to increases in biomass and wax ester production. *Biotechnology for Biofuels,* 8**,** 1-11.

OGBONNA, J., ICHIGE, E. & TANAKA, H. 2002. Interactions between photoautotrophic and heterotrophic metabolism in photoheterotrophic cultures of *Euglena gracilis*. *Applied Microbiology and Biotechnology,* 58**,** 532-538.

OLĘDZKA, G., DĄBROWSKI, S. & KUR, J. 2003. High-level expression, secretion, and purification of the thermostable aqualysin I from Thermus aquaticus YT-1 in *Pichia pastoris*. *Protein expression and purification,* 29**,** 223-229.

PANCHOLI, V. 2001. Multifunctional α-enolase: its role in diseases. *Cellular and Molecular Life Sciences CMLS,* 58**,** 902-920.

PARKINSON, J. & BLAXTER, M. 2009. Expressed Sequence Tags: An Overview. *In:* PARKINSON, J. (ed.) *Expressed Sequence Tags (ESTs): Generation and Analysis.* Totowa, NJ: Chapter 1. Human Press, 1-12.

PETERSEN, J., BRINKMANN, H. & CERFF, R. 2003. Origin, evolution, and metabolic role of a novel glycolytic GAPDH enzyme recruited by land plant plastids. *Journal of Molecular Evolution,* 57**,** 16-26.

POMP, D. & MEDRANO, J. F. 1991. Organic solvents as facilitators of polymerase chain reaction. *Biotechniques,* 10**,** 58-9.

PRIBNOW, D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences,* 72**,** 784-788.

PRINGSHEIM, E. 1955. Kleine Mitteilungen über Flagellaten und Algen. *Archiv für Mikrobiologie,* 21**,** 401-410.

PUGH, B. F. & TJIAN, R. 1991. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes & Development,* 5**,** 1935-1945.

RISHI, A., NELSON, N. D. & GOYAL, A. 2004. Genome walking of large fragments: an improved method. *Journal of Biotechnology,* 111**,** 9-15.

RODRÍGUEZ-ZAVALA, J., ORTIZ-CRUZ, M., MENDOZA-HERNÁNDEZ, G. & MORENO-SÁNCHEZ, R. 2010. Increased synthesis of α-tocopherol, paramylon and tyrosine by *Euglena gracilis* under conditions of high biomass production. *Journal of Applied Microbiology,* 109**,** 2160-2172.

ROSENBERG, S. & TEKAMP-OLSON, P. 1992. *Enhanced yeast transcription employing hybrid GAPDH promoter region constructs.,* 5089398**,** United States Patent. [https://www.google.ch/patents/US5089398].

ROSENFELD, J., CAPDEVIELLE, J., GUILLEMOT, J. C. & FERRARA, P. 1992. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Analytical biochemistry,* 203**,** 173-9.

ROSENTHAL, A. & JONES, D. 1990. Genomic walking and sequencing by oligo-cassette mediated polymerase chain reaction. *Nucleic Acids Research,* 18**,** 3095-3096.

SANFORD, J. C. 1990. Biolistic plant transformation. *Physiologia Plantarum,* 79**,** 206-209.

ŠANTEK, B. I., FELSKI, M., FRIEHS, K., LOTZ, M. & FLASCHEL, E. 2009. Production of paramylon, a β-1, 3-glucan, by heterotrophic cultivation of *Euglena gracilis* on a synthetic medium. *Engineering in Life Sciences,* 9**,** 23-28.

SEIZL, M., HARTMANN, H., HOEG, F., KURTH, F., MARTIN, D. E., SÖDING, J. & CRAMER, P. 2011. A conserved GA element in TATA-less RNA polymerase II promoters. *Public Library of Science,* 6**,** e27595.

SEXTON, T. B., CHRISTOPHER, D. A. & MULLET, J. E. 1990. Light-induced switch in barley psbD-psbC promoter utilization: a novel mechanism regulating chloroplast gene expression. *The European Molecular Biology Organization Journal,* 9**,** 4485-4494.

SHERWOOD, A. L. 2003. Virtual elimination of false positives in blue-white colony screening. *BioTechniques,* 34**,** 644-647.

SHORE, D. 1994. RAP1: a protean regulator in yeast. *Trends in Genetics,* 10**,** 408-412.

SIKORSKI, R. S. & HIETER, P. 1989. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics,* 122**,** 19-27.

SILVER, J. & KEERIKATTE, V. 1989. Novel use of polymerase chain reaction to amplify cellular DNA adjacent to an integrated provirus. *Journal of Virology,* 63**,** 1924-1928.

SINGH, M. V. & WEIL, P. A. 2002. A method for plasmid purification directly from yeast. *Analytical biochemistry,* 307**,** 13-17.

SMALE, S. T. & BALTIMORE, D. 1989. The "initiator" as a transcription control element. *Cell,* 57**,** 103-113.

SMALE, S. T. & KADONAGA, J. T. 2003. The RNA polymerase II core promoter. *Annual Review of Biochemistry,* 72**,** 449-479.

SMITH, P., KROHN, R., HERMANSON, G., MALLIA, A., GARTNER, F. & PROVENZANO, M. 1985. Measurement of protein using bicinchoninic acid *Analytical Biochemistry,* 150**,** 76-85.

STRACKE, R., WERBER, M. & WEISSHAAR, B. 2001. The R2R3-MYB gene family in Arabidopsis thaliana. *Current Opinion Plant Biology,* 4**,** 447-56.

SUN-WADA, G.-H., YOSHIOKA, S., ISHIDA, N. & KAWAKITA, M. 1998. Functional expression of the human UDP-galactose transporters in the yeast *Saccharomyces cerevisiae*. *Journal of biochemistry,* 123**,** 912-917.

TAHEDL, H. & HÄDER, D.-P. 2001. Automated biomonitoring using real time movement analysis of *Euglena gracilis*. *Ecotoxicology and Environmental Safety,* 48**,** 161-169.

TANNREUTHER, G. W. 1923. Nutrition and reproduction in *Euglena*. *Archiv für Entwicklungsmechanik der Organismen,* 52**,** 367-383.

THOMPSON, E. L., O'CONNOR, W., PARKER, L., ROSS, P. & RAFTOS, D. A. 2015. Differential proteomic responses of selectively bred and wild-type Sydney rock oyster populations exposed to elevated CO2. *Molecular Ecology,* 24**,** 1248-62.

TOKUSUMI, Y., MA, Y., SONG, X., JACOBSON, R. H. & TAKADA, S. 2007. The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Molecular and cellular biology,* 27**,** 1844-1858.

TONOOKA, Y. & FUJISHIMA, M. 2009. Comparison and critical evaluation of PCR-mediated methods to walk along the sequence of genomic DNA. *Applied microbiology and biotechnology,* 85**,** 37-43.

TRIGLIA, T., PETERSON, M. G. & KEMP, D. J. 1988. A procedure for in vitro amplification of DNA segments that lie outside the boundaries of known sequences. *Nucleic Acids Research,* 16**,** 8186-8196.

TRISTAN, C., SHAHANI, N., SEDLAK, T. W. & SAWA, A. 2011. The diverse functions of GAPDH: views from different subcellular compartments. *Cellular signalling,* 23**,** 317-323.

URDEA, M. S., MERRYWEATHER, J. P., MULLENBACH, G. T., COIT, D., HEBERLEIN, U., VALENZUELA, P. & BARR, P. J. 1983. Chemical synthesis of a gene for human epidermal growth factor urogastrone and its expression in yeast. *Proceedings of the National Academy of Sciences,* 80**,** 7461-7465.

WANG, W., SUN, J., NIMTZ, M., DECKWER, W.-D. & ZENG, A.-P. 2003. Protein identification from two-dimensional gel electrophoresis analysis of *Klebsiella pneumoniae* by combined use of mass spectrometry data and raw genome sequences. *Proteome Science,* 1**,** 6-15.

WATANABE, T., SHIMADA, R., MATSUYAMA, A., YUASA, M., SAWAMURA, H., YOSHIDA, E. & SUZUKI, K. 2013. Antitumor activity of the β-glucan paramylon from *Euglena* against preneoplastic colonic aberrant crypt foci in mice. *Food and function,* 4**,** 1685-1690.

WATERHAM, H. R., DIGAN, M. E., KOUTZ, P. J., LAIR, S. V. & CREGG, J. M. 1997. Isolation of the *Pichia pastoris* glyceraldehyde-3-phosphate dehydrogenase gene and regulation and use of its promoter. *Gene,* 186**,** 37-44.

WEBER, K. & OSBORN, M. 1969. The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis. *Journal of Biological Chemistry,* 244**,** 4406-4412.

WEI, T. & DAI, H. 2014. Quantification of GFP signals by fluorescent microscopy and flow cytometry. *Methods in Molecular Biology,* 1163**,** 23-31.

WILKINS, M. R., GASTEIGER, E., BAIROCH, A., SANCHEZ, J. C., WILLIAMS, K. L., APPEL, R. D. & HOCHSTRASSER, D. F. 1999. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology,* 112**,** 531-552.

WU, D.-A., HU, M.-C. & CHUNG, B.-C. 1991. Expression and functional study of wild-type and mutant human cytochrome P450c21 in *Saccharomyces cerevisiae*. *DNA and cell biology,* 10**,** 201-209.

XIAO, W. 2006. *Yeast Protocols,* Humana Press, Totowa, NJ.

YAMAUCHI, K. 1991. The sequence flanking translational initiation site in protozoa. *Nucleic acids research,* 19**,** 2715-2720.

YANG, C., BOLOTIN, E., JIANG, T., SLADEK, F. M. & MARTINEZ, E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene,* 389**,** 52-65.

YASUHIRA, S. & SIMPSON, L. 1997. Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *Journal of molecular evolution,* 44**,** 341-347.

YOO, H. Y., KIM, S. S. & RHO, H. M. 1999. Overexpression and simple purification of human superoxide dismutase (SOD1) in yeast and its resistance to oxidative stress. *Journal of biotechnology,* 68**,** 29-35.

YOSHIDA, Y., TOMIYAMA, T., MARUTA, T., TOMITA, M., ISHIKAWA, T. & ARAKAWA, K. 2016. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BioMed Central Genomics,* 17**,** 1-5.

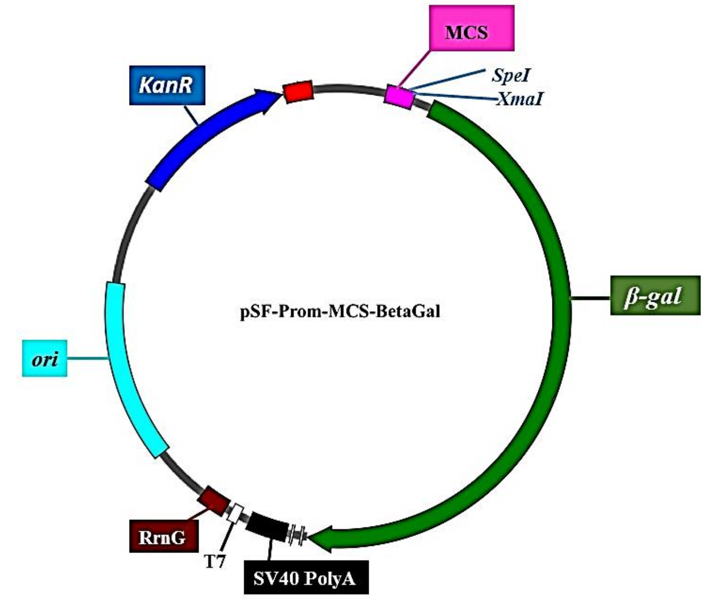
YOUNG, E. M., GORDON, D. B. & VOIGT, C. 2016. Composability and design of parts for large-scale pathway engineering in yeast. US Patent 20,160,083,722.

ZHANG, Z. & GURR, S. J. 2000. Walking into the unknown: a 'step down'PCR-based technique leading to the direct sequence analysis of flanking genomic DNA. *Gene,* 253**,** 145-150.

ZHANG, Z., YANG, X., MENG, L., LIU, F., SHEN, C. & YANG, W. 2009. Enhanced amplification of GC-rich DNA with two organic reagents. *Biotechniques,* 47**,** 775-779.

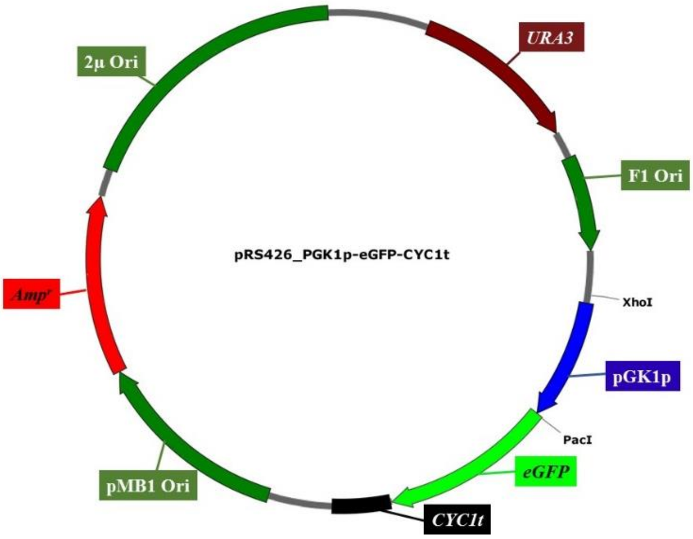
Appendices

Appendix 1



The structure of the pSF-PromMCS-BetaGal (OG372) plasmid without a promoter. The plasmid is of 6918 bp in size and contains a kanamycin resistance gene for transformant selection, a pUC origin of replication and a T7 terminator. (Schematic generated using SnapGene®).

Appendix 2



Structure of the pRS426-PGK1p-eGFP-CYC1t plasmid schematic generated using SnapGene®). The plasmid contains the *PGK1* promoter downstream of the F1 origin of replication. The *eGPF* reporter gene is downstream of the pGK1 promoter. The plasmid contains an ampicillin resistance gene (*Amp*^{r}) for selection of *E. coli* transformants during plasmid propagation and *URA* gene for selection of transformants.

Appendix 3

SpeI

```
1       ACTAGTAAGCTTGGGAGAAACTATCCATCGCAGTGTTGGAGCGAGGCATGACGAAGCATG      60
61      TCCTCCTCCCAGCAACCACTGACCCTGCCATTCCTGCACTGCCGCCTGATGCCCCTTTGC 120
121     CTCCTGTTCCATCCGACATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTG   180
181     TTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACG      240
241     GCTCAGAAGAGAACGCCACTGATGAGTCCAATGATGGCTCTTCAGATGGTGGTGATGAAG      300
301     TGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCGTTCGCAGTGCCCCC360
361     ATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCGTTAACTCCAAAGGAATGTGTTGTT 420
421     TCTGTCCCTAATTCCAGGATGAATTGAGCTCTTGCAAGAGGATTGCATTCGTGCTTCGTC 480
481     TTCGCCCCGGCTTGCAGGGTTCTTCAGGCTTGTCCATGGGAACAGTAAGGGGGGGGGGAG      540
541     CTCGTCCCCCCCCTTACTGTTCCCCCCCCTTGCTGTTCCCATGGACCCAAAAGGGGGCCA 600
601     CGGGACGGGAACACTAAGGGGGGGGTATGGTAAGGGCTGAGAAGCCCACTTCACTGTGGT      660
661     CGTGTTCTGGACCTCCACTCCACCTTGGACTCGTCTTGGACTTGCACTTGTTCTTGGACT  720
721     TGCACCCTTGTAACTTGGACTTGACACCAGTACTTGAACTTGAACAGACTTATAACTTGA      780
781     ACTGCACCAGGGACTTTCACTTCACCAGCACTTGTAACTTGAACCCCTCCAGGGGCTTTC      840
841     ACTTCTCGGCCATGTGATATGTCAAAAAAAATAGACACTCAGAAAGTCACATTGGATCCA      900
901     TTTGGAATATGTTGATGCGGCACAAACCGTTAATTTTGGTTAGGGATACCCAACATTCTC960
961     AATAGGCTTTTCCCGTACCTGTCCTCGGTTTCCTTTTCGTTCACCCAACTCCAGATATAT   1020
1021    CCCCGGG
```

XmaI

The 5’UTR region of *gapC* gene insert for the *E. coli* plasmid

Appendix 4

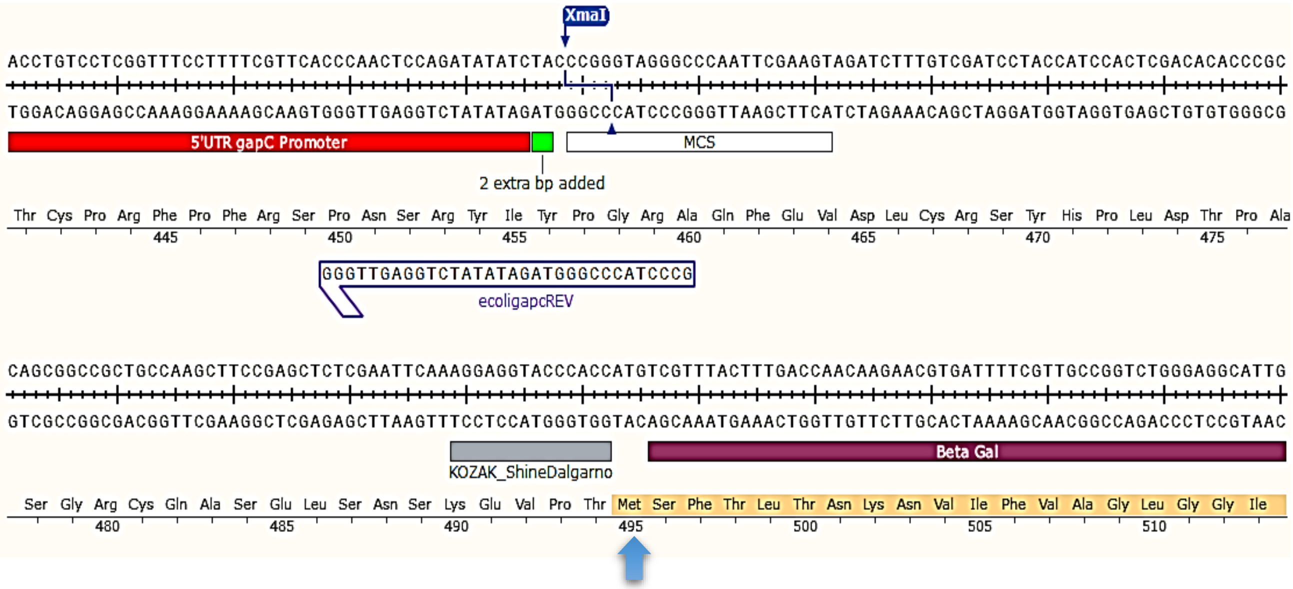
XhoI

```
1       CTCGAGAAGCTTGGGAGAAACTATCCATCGCAGTGTTGGAGCGAGGCATGACGAAGCATG      60
61      TCCTCCTCCCAGCAACCACTGACCCTGCCATTCCTGCACTGCCGCCTGATGCCCCTTTGC 120
121     CTCCTGTTCCATCCGACATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTG   180
181     TTGAAGCCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACG      240
241     GCTCAGAAGAGAACGCCACTGATGAGTCCAATGATGGCTCTTCAGATGGTGGTGATGAAG      300
301     TGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCGTTCGCAGTGCCCCC360
361     ATTGTGGTATTATCCTCCATCCAAGGTGTATGTCCGTTAACTCCAAAGGAATGTGTTGTT 420
421     TCTGTCCCTAATTCCAGGATGAATTGAGCTCTTGCAAGAGGATTGCATTCGTGCTTCGTC 480
481     TTCGCCCCGGCTTGCAGGGTTCTTCAGGCTTGTCCATGGGAACAGTAAGGGGGGGGGGAG      540
541     CTCGTCCCCCCCCTTACTGTTCCCCCCCCTTGCTGTTCCCATGGACCCAAAAGGGGGCCA 600
601     CGGGACGGGAACACTAAGGGGGGGGTATGGTAAGGGCTGAGAAGCCCACTTCACTGTGGT      660
661     CGTGTTCTGGACCTCCACTCCACCTTGGACTCGTCTTGGACTTGCACTTGTTCTTGGACT  720
721     TGCACCCTTGTAACTTGGACTTGACACCAGTACTTGAACTTGAACAGACTTATAACTTGA      780
781     ACTGCACCAGGGACTTTCACTTCACCAGCACTTGTAACTTGAACCCCTCCAGGGGCTTTC      840
841     ACTTCTCGGCCATGTGATATGTCAAAAAAAATAGACACTCAGAAAGTCACATTGGATCCA      900
901     TTTGGAATATGTTGATGCGGCACAAACCGTTAATTTTGGTTAGGGATACCCAACATTCTC960
961     AATAGGCTTTTCCCGTACCTGTCCTCGGTTTCCTTTTCGTTCACCCAACTCCAGATATAT   1020
1021    CTTAATTAA
```

PacI

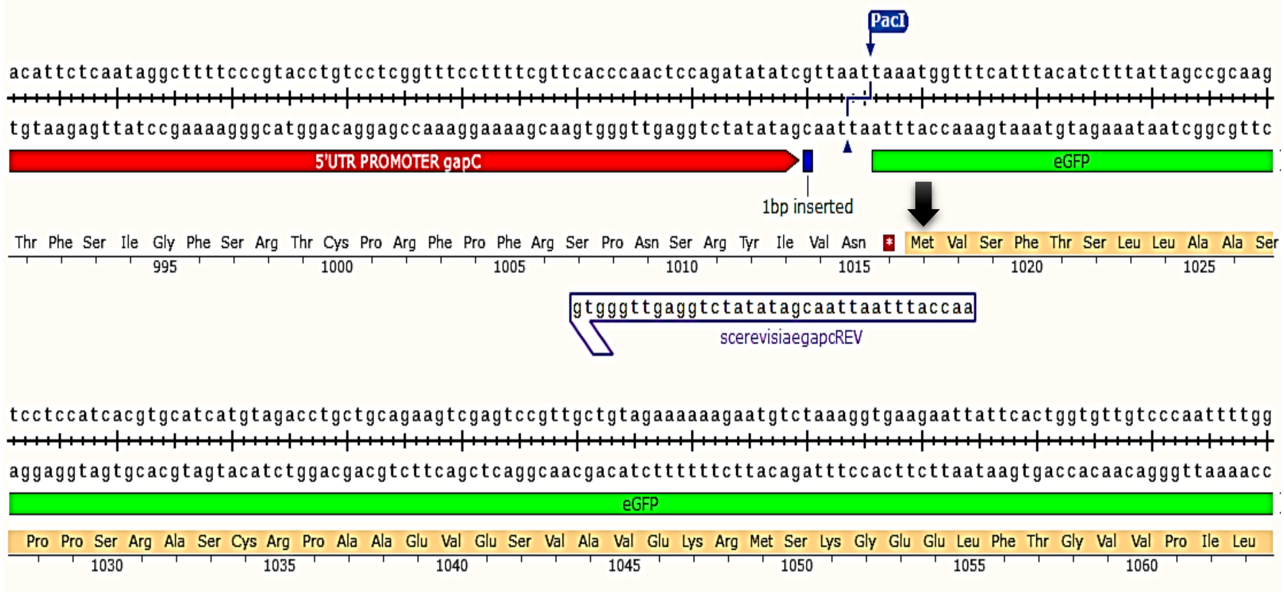
The 5’UTR region of *gapC* gene insert for the *S. cerevisiae* plasmid

Appendix 5


XmaI
ACCTGTCCTCGGTTTCCTTTTCGTTCACCCAACTCCAGATATATCTACCCGGGTAGGGCCCAATTCGAAGTAGATCTTTGTCGATCCTACCATCCACTCGACACACCCGC
TGGACAGGAGCCAAAGGAAAAGCAAGTGGGTTGAGGTCTATATAGATGGGCCCATCCCGGGTTAAGCTTCATCTAGAAACAGCTAGGATGGTAGGTGAGCTGTGTGGGCG
5'UTR gapC Promoter
MCS
2 extra bp added
Thr Cys Pro Arg Phe Pro Phe Arg Ser Pro Asn Ser Arg Tyr Ile Tyr Pro Gly Arg Ala Gln Phe Glu Val Asp Leu Cys Arg Ser Tyr His Pro Leu Asp Thr Pro Ala
445
450
455
460
465
470
475
GGGTTGAGGTCTATATAGATGGGCCCATCCCG
ecoligapcREV
CAGCGGCCGCTGCCAAGCTTCCGAGCTCTCGAATTCAAAGGAGGTACCCACCATGTCGTTTACTTTGACCAACAAGAACGTGATTTTCGTTGCCGGTCTGGGAGGCATTG
GTCGCCGGCGACGGTTCGAAGGCTCGAGAGCTTAAGTTTCCTCCATGGGTGGTACAGCAAATGAAACTGGTTGTTCTTGCACTAAAAGCAACGGCCAGACCCTCCGTAAC
KOZAK_ShineDalgarno
Beta Gal
Ser Gly Arg Cys Gln Ala Ser Glu Leu Ser Asn Ser Lys Glu Val Pro Thr Met Ser Phe Thr Leu Thr Asn Lys Asn Val Ile Phe Val Ala Gly Leu Gly Gly Ile
480
485
490
495
500
505
510

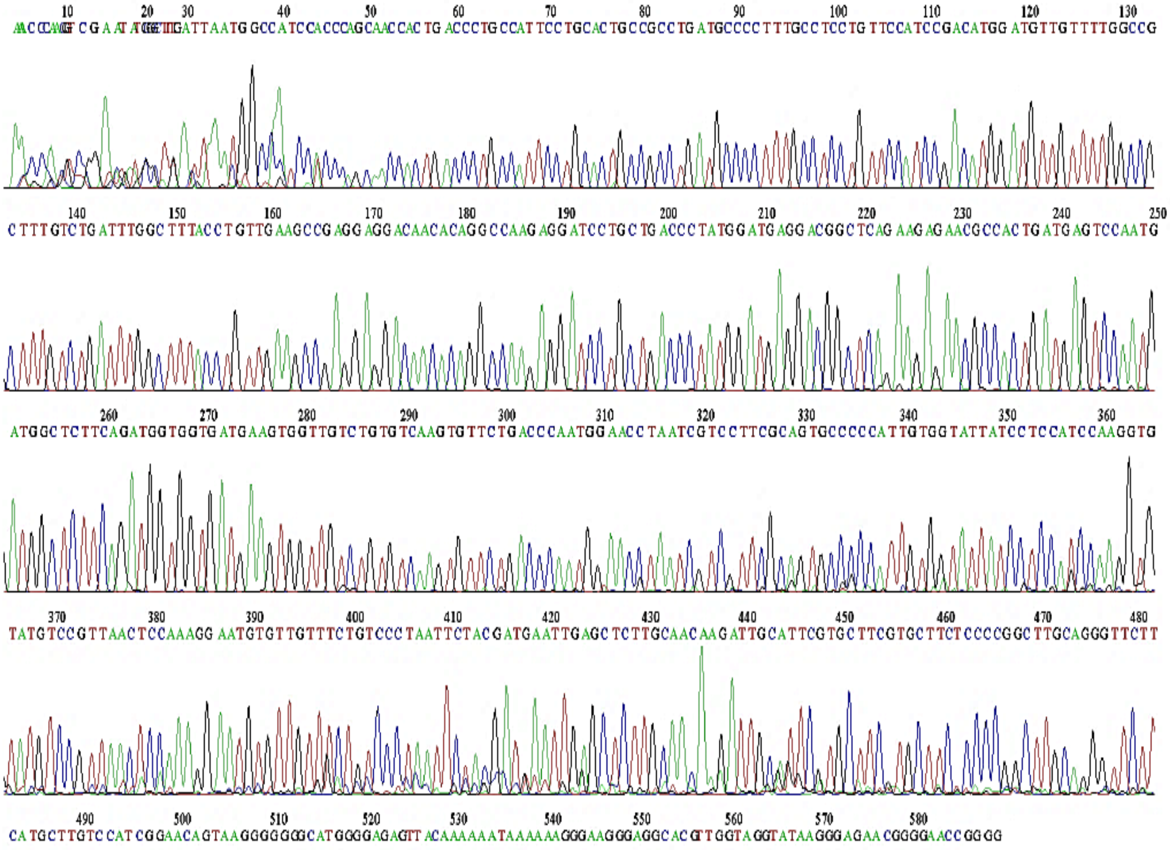
Construction of the expression vector (pRS426) with the insert 5'UTR promoter region of the *gapC* gene of *E. gracilis*. To ensure that the insert is in frame with the reporter *beta-gal* gene two extra bp were added while designing reverse primers. After adding extra two bp, the codon for the *beta-gal* gene is in frame (represented by blue arrow) which means that the insert is in frame.

Appendix 6


PacI
acattctcaataggcttttcccgtacctgtcctcggtttccttttcgttcacccaactccagatatatcgttaattaaatggtttcatttacatctttattagccgcaag
tgtaagagttatccgaaaagggcatggacaggagccaaaggaaaagcaagtgggttgaggtctatatagcaattaatttaccaaagtaaatgtagaaataatcggcgttc
5'UTR PROMOTER gapC
eGFP
1bp inserted
Thr Phe Ser Ile Gly Phe Ser Arg Thr Cys Pro Arg Phe Pro Phe Arg Ser Pro Asn Ser Arg Tyr Ile Val Asn * Met Val Ser Phe Thr Ser Leu Leu Ala Ala Ser
995
1000
1005
1010
1015
1020
1025
gtgggttgaggtctatatagcaattaatttaccaa
scerevisiaegapcREV
tcctccatcacgtgcatcatgtagacctgctgcagaagtcgagtccgttgctgtagaaaaagaatgtctaaaggtgaagaattattcactggtgttgtcccaattttgg
aggaggtagtgcacgtagtacatctggacgacgtcttcagctcaggcaacgacatctttttcttacagatttccacttcttaataagtgaccacaacagggttaaaacc
eGFP
Pro Pro Ser Arg Ala Ser Cys Arg Pro Ala Ala Glu Val Glu Ser Val Ala Val Glu Lys Arg Met Ser Lys Gly Glu Glu Leu Phe Thr Gly Val Val Pro Ile Leu
1030
1035
1040
1045
1050
1055
1060

Construction of the expression vector (OG372) with the insert 5'UTR promoter region of the *gapC* gene of *E. gracilis*. To ensure that the insert is in frame with the reporter *eGFP* gene one extra bp was added while designing reverse primers. After adding extra one bp, the codon for the *eGFP* gene is in frame (represented by black arrow) which means that the insert is in frame.

Appendix 7



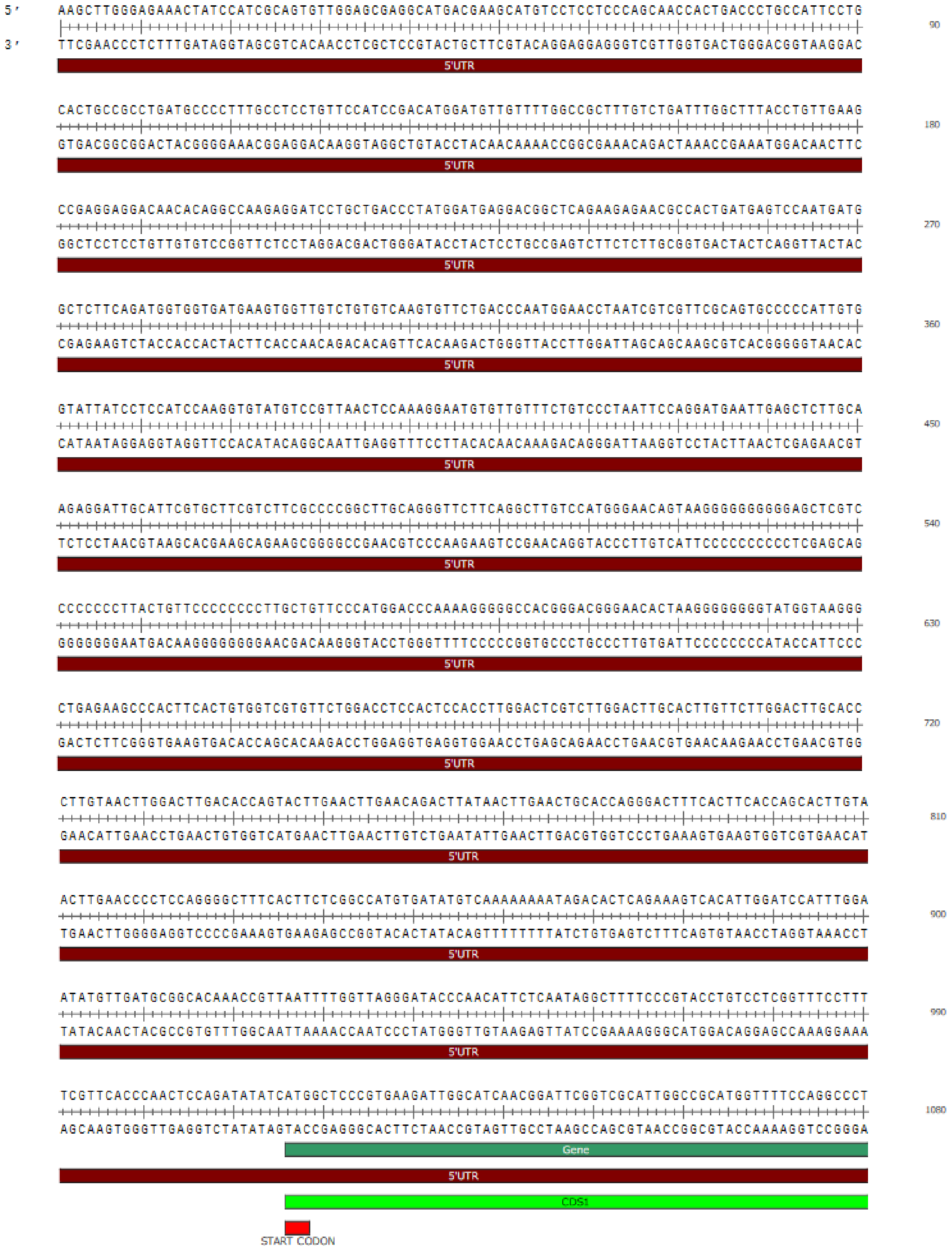
Conirmation of insertion of the 5’UTR promoter region of *gapC* gene in the plasmid pRS426.

Appendix 8

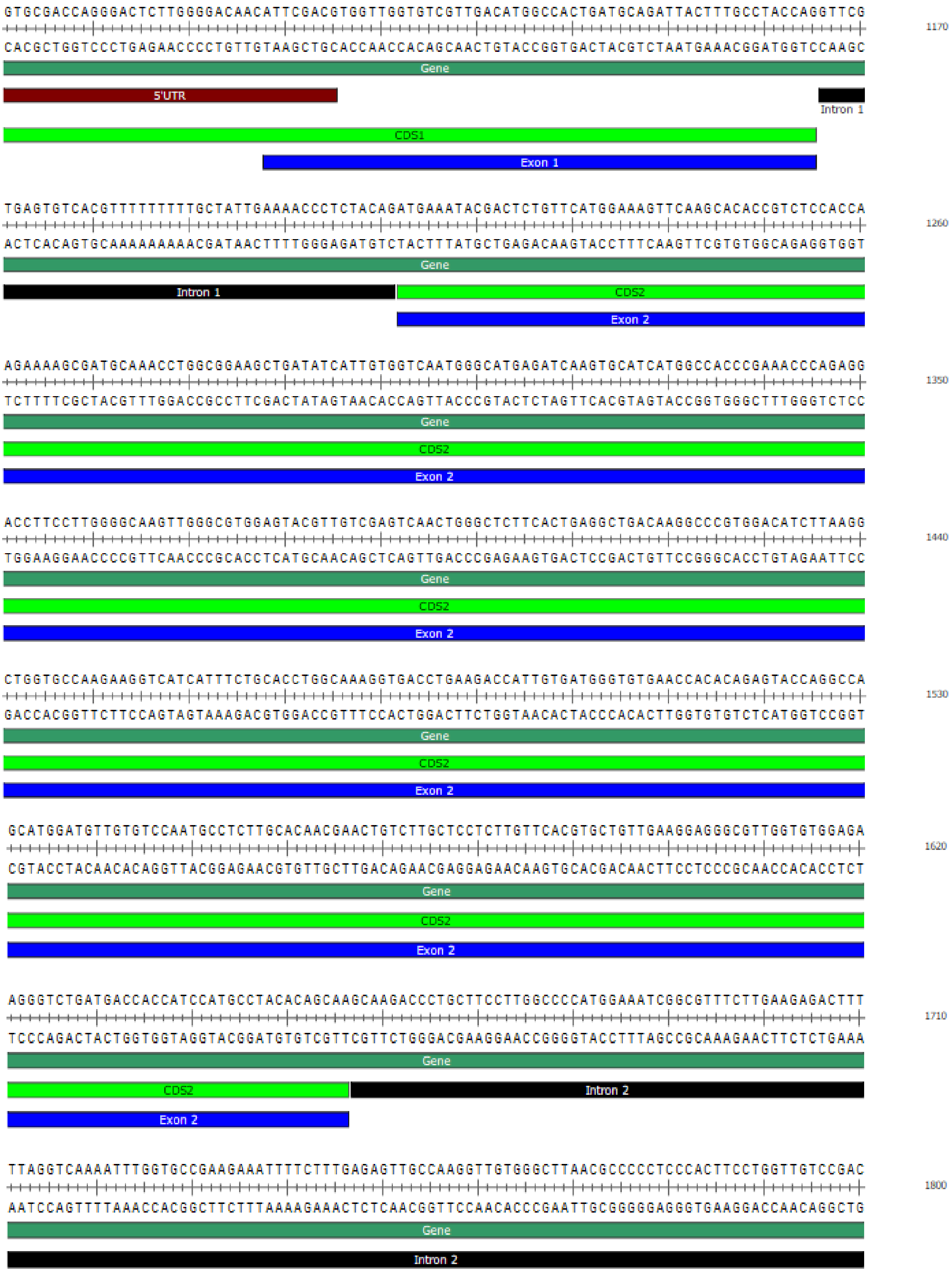
Primers		**Amplicon size**
Primer set 1: for amplification of sample 1		
Fwd_5UTR_primer_1	5’CTTGGGAGAAACTATCCATCGCAG 3’	155 bp
Rev_sequencing_primer_1	5’ AGACAAAGCGGCCAAAACAACATC 3’	
Primer set 2: for amplification of sample 2		
Fwd_5UTR_test_primer_2	5’ TCACTGTGGTCGTGTTCTGGACCT 3’	336 bp
Rev_5UTR_test_primer_2	5’ GAGGACAGGTACGGGAAAAGCCTA 3’	
Primer set 3: for amplification of sample 3		
Fwd_gapC_*E_gracilis*_3	5’ AAGTGCATCATGGCCACCCGAAAC 3’	323 bp
Rev_gapC_*E_gracilis*_3	5’ TGGATGGTGGTCATCAGACCCTTC 3’	
Primer set: for amplification of sample 4		
Fwd_Primer_test_4	5’ GGAGAAGGGTCTGATGACCA 3’	226 bp
Rev_Primer_test_4	5’ TCAAAATCTCCCAGGCAATC 3’	

List of primers used for amplifying the various regions of the *gapC* gene of *E. gracilis.*

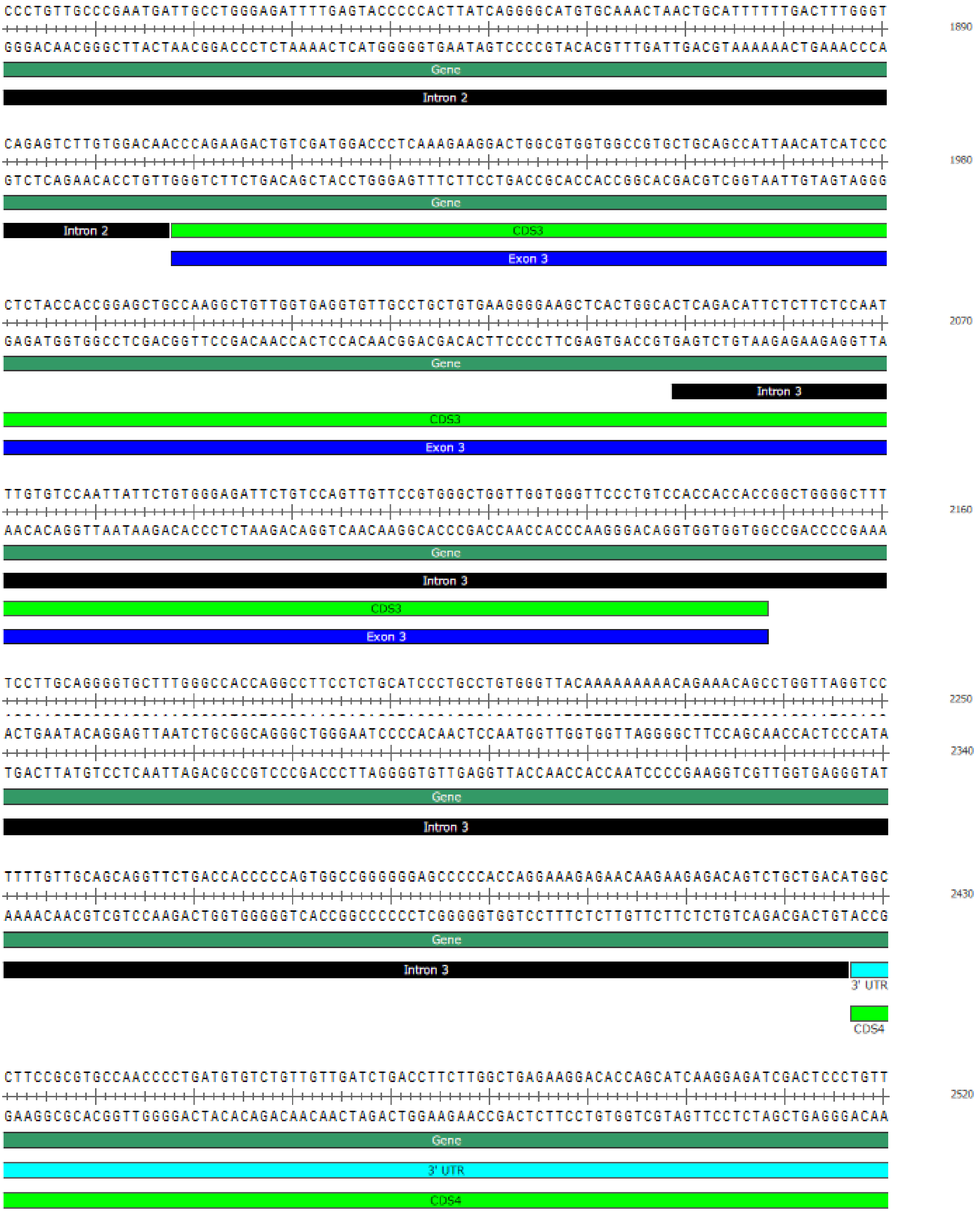
Appendix 9



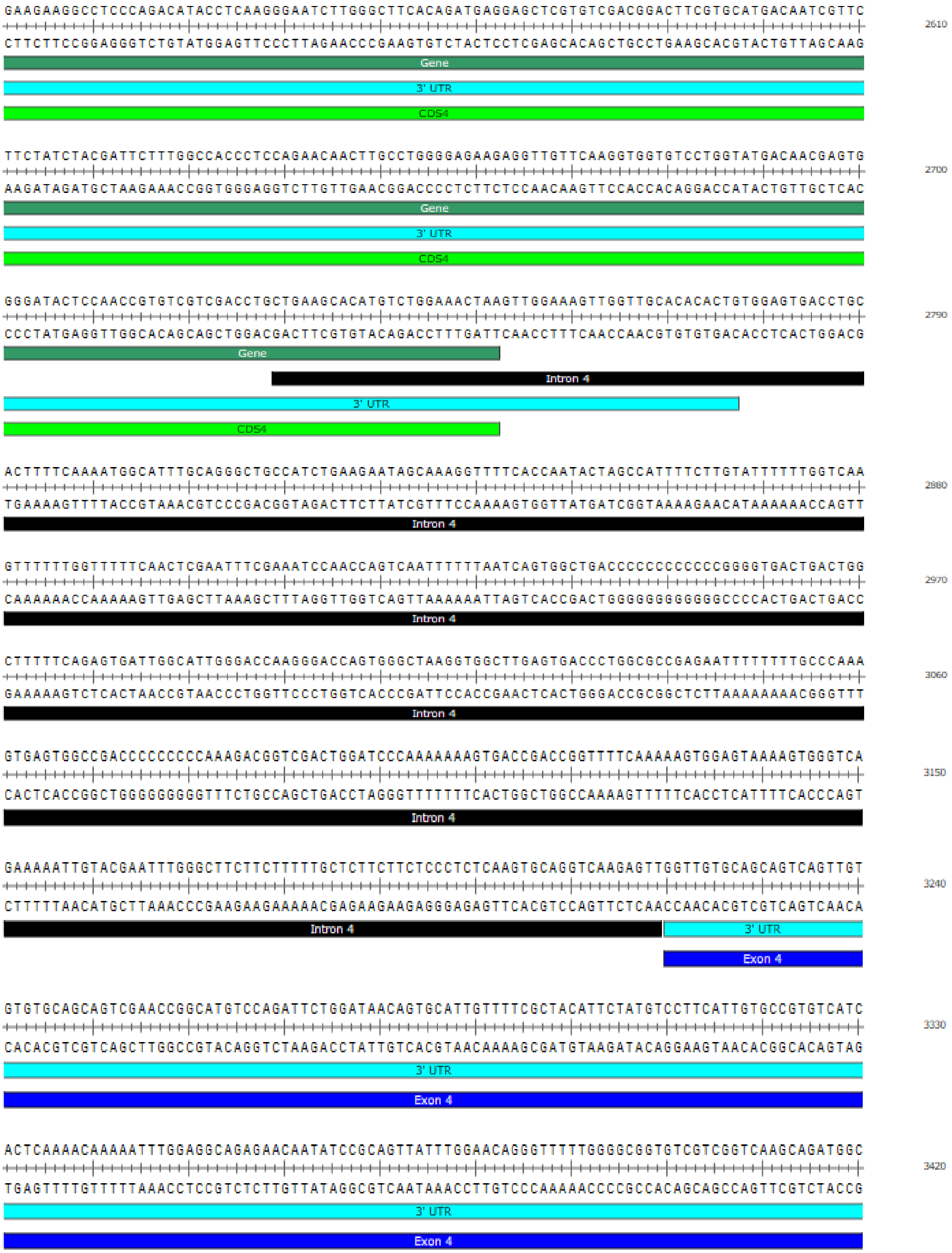
```
5'  AAGCTTGGGAGAAACTATCCATCGCAGTGTTGGAGCGAGGCATGACGAAGCATGTCCTCCCAGCAACCACTGACCCTGCCATTCCTG   90
3'  TTCGAACCCTCTTTGATAGGTAGCGTCACAACCTCGCTCCGTACTGCTTCGTACAGGAGGAGGGTCGTTGGTGACTGGGACGGTAAGGAC
    5'UTR

    CACTGCCGCCTGATGCCCCTTTGCCTCCTGTTCCATCCGACATGGATGTTGTTTTGGCCGCTTTGTCTGATTTGGCTTTACCTGTTGAAG   180
    GTGACGGCGGACTACGGGGAAACGGAGGACAAGGTAGGCTGTACCTACAACAAAACCGGCGAAACAGACTAAACCGAAATGGACAACTTC
    5'UTR

    CCGAGGAGGACAACACAGGCCAAGAGGATCCTGCTGACCCTATGGATGAGGACGGCTCAGAAGAGAACGCCACTGATGAGTCCAATGATG   270
    GGCTCCTCCTGTTGTGTCCGGTTCTCCTAGGACGACTGGGATACCTACTCCTGCCGAGTCTTCTCTTGCGGTGACTACTCAGGTTACTAC
    5'UTR

    GCTCTTCAGATGGTGGTGATGAAGTGGTTGTCTGTGTCAAGTGTTCTGACCCAATGGAACCTAATCGTCGTTCGCAGTGCCCCCATTGTG   360
    CGAGAAGTCTACCACCACTACTTCACCAACAGACACAGTTCACAAGACTGGGTTACCTTGGATTAGCAGCAAGCGTCACGGGGGTAACAC
    5'UTR

    GTATTATCCTCCATCCAAGGTGTATGTCCGTTAACTCCAAAGGAATGTGTTGTTTCTGTCCCTAATTCCAGGATGAATTGAGCTCTTGCA   450
    CATAATAGGAGGTAGGTTCCACATACAGGCAATTGAGGTTTCCTTACACAACAAAGACAGGGATTAAGGTCCTACTTAACTCGAGAACGT
    5'UTR

    AGAGGATTGCATTCGTGCTTCGTCTTCGCCCCGGCTTGCAGGGTTCTTCAGGCTTGTCCATGGGAACAGTAAGGGGGGGGGGAGCTCGTC   540
    TCTCCTAACGTAAGCACGAAGCAGAAGCGGGGCCGAACGTCCCAAGAAGTCCGAACAGGTACCCTTGTCATTCCCCCCCCCCTCGAGCAG
    5'UTR

    CCCCCCCTTACTGTTCCCCCCCCTTGCTGTTCCCATGGACCCAAAAGGGGGCCACGGGACGGGAACACTAAGGGGGGGGTATGGTAAGGG   630
    GGGGGGGAATGACAAGGGGGGGGAACGACAAGGGTACCTGGGTTTTCCCCCGGTGCCCTGCCCTTGTGATTCCCCCCCCATACCATTCCC
    5'UTR

    CTGAGAAGCCCACTTCACTGTGGTCGTGTTCTGGACCTCCACTCCACCTTGGACTCGTCTTGGACTTGCACTTGTTCTTGGACTTGCACC   720
    GACTCTTCGGGTGAAGTGACACCAGCACAAGACCTGGAGGTGAGGTGGAACCTGAGCAGAACCTGAACGTGAACAAGAACCTGAACGTGG
    5'UTR

    CTTGTAACTTGGACTTGACACCAGTACTTGAACTTGAACAGACTTATAACTTGAACTGCACCAGGGACTTTCACTTCACCAGCACTTGTA   810
    GAACATTGAACCTGAACTGTGGTCATGAACTTGAACTTGTCTGAATATTGAACTTGACGTGGTCCCTGAAAGTGAAGTGGTCGTGAACAT
    5'UTR

    ACTTGAACCCCTCCAGGGGCTTTCACTTCTCGGCCATGTGATATGTCAAAAAAAATAGACACTCAGAAAGTCACATTGGATCCATTTGGA   900
    TGAACTTGGGGAGGTCCCCGAAAGTGAAGAGCCGGTACACTATACAGTTTTTTTTATCTGTGAGTCTTTCAGTGTAACCTAGGTAAACCT
    5'UTR

    ATATGTTGATGCGGCACAAACCGTTAATTTTGGTTAGGGATACCCAACATTCTCAATAGGCTTTTCCCGTACCTGTCCTCGGTTTCCTTT   990
    TATACAACTACGCCGTGTTTGGCAATTAAAACCAATCCCTATGGGTTGTAAGAGTTATCCGAAAAGGGCATGGACAGGAGCCAAAGGAAA
    5'UTR

    TCGTTCACCCAACTCCAGATATATCATGGCTCCCGTGAAGATTGGCATCAACGGATTCGGTCGCATTGGCCGCATGGTTTTCCAGGCCCT   1080
    AGCAAGTGGGTTGAGGTCTATATAGTACCGAGGGCACTTCTAACCGTAGTTGCCTAAGCCAGCGTAACCGGCGTACCAAAAGGTCCGGGA
                             Gene
    5'UTR
                             CDS1
                          START CODON
```



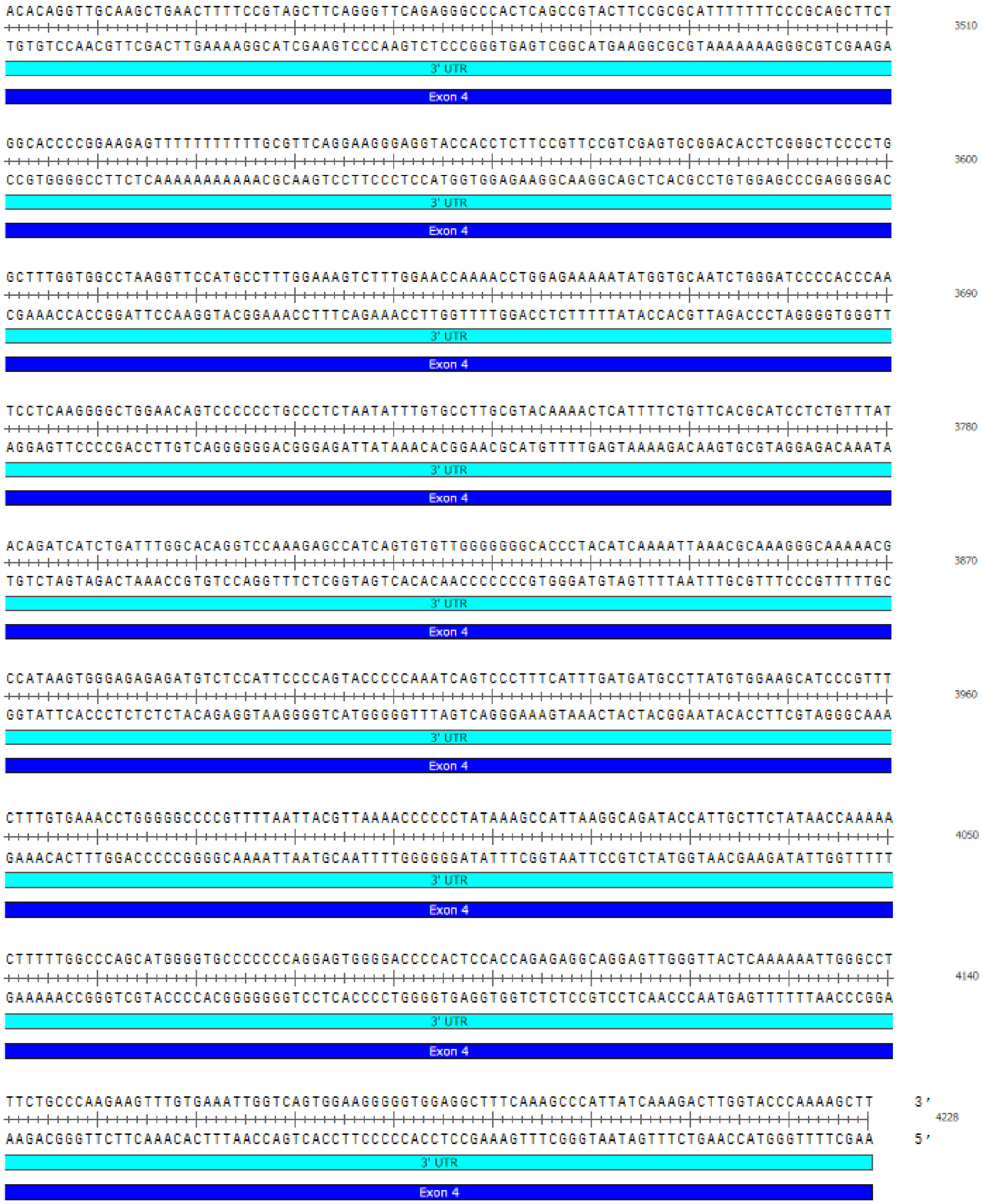
```
GTGCGACCAGGGACTCTTGGGGACAACATTCGACGTGGTTGGTGTCGTTGACATGGCCACTGATGCAGATTACTTTGCCTACCAGGTTCG
CACGCTGGTCCCTGAGAACCCCTGTTGTAAGCTGCACCAACCACAGCAACTGTACCGGTGACTACGTCTAATGAAACGGATGGTCCAAGC   1170

TGAGTGTCACGTTTTTTTTTGCTATTGAAAACCCTCTACAGATGAAATACGACTCTGTTCATGGAAAGTTCAAGCACACCGTCTCCACCA
ACTCACAGTGCAAAAAAAAACGATAACTTTTGGGAGATGTCTACTTTATGCTGAGACAAGTACCTTTCAAGTTCGTGTGGCAGAGGTGGT   1260

AGAAAAGCGATGCAAACCTGGCGGAAGCTGATATCATTGTGGTCAATGGGCATGAGATCAAGTGCATCATGGCCACCCGAAACCCAGAGG
TCTTTTCGCTACGTTTGGACCGCCTTCGACTATAGTAACACCAGTTACCCGTACTCTAGTTCACGTAGTACCGGTGGGCTTTGGGTCTCC   1350

ACCTTCCTTGGGGCAAGTTGGGCGTGGAGTACGTTGTCGAGTCAACTGGGCTCTTCACTGAGGCTGACAAGGCCCGTGGACATCTTAAGG
TGGAAGGAACCCCGTTCAACCCGCACCTCATGCAACAGCTCAGTTGACCCGAGAAGTGACTCCGACTGTTCCGGGCACCTGTAGAATTCC   1440

CTGGTGCCAAGAAGGTCATCATTTCTGCACCTGGCAAAGGTGACCTGAAGACCATTGTGATGGGTGTGAACCACACAGAGTACCAGGCCA
GACCACGGTTCTTCCAGTAGTAAAGACGTGGACCGTTTCCACTGGACTTCTGGTAACACTACCCACACTTGGTGTGTCTCATGGTCCGGT   1530

GCATGGATGTTGTGTCCAATGCCTCTTGCACAACGAACTGTCTTGCTCCTCTTGTTCACGTGCTGTTGAAGGAGGGCGTTGGTGTGGAGA
CGTACCTACAACACAGGTTACGGAGAACGTGTTGCTTGACAGAACGAGGAGAACAAGTGCACGACAACTTCCTCCCGCAACCACACCTCT   1620

AGGGTCTGATGACCACCATCCATGCCTACACAGCAAGCAAGACCCTGCTTCCTTGGCCCCATGGAAATCGGCGTTTCTTGAAGAGACTTT
TCCCAGACTACTGGTGGTAGGTACGGATGTGTCGTTCGTTCTGGGACGAAGGAACCGGGGTACCTTTAGCCGCAAAGAACTTCTCTGAAA   1710

TTAGGTCAAAATTTGGTGCCGAAGAAATTTTCTTTGAGAGTTGCCAAGGTTGTGGGCTTAACGCCCCCTCCCACTTCCTGGTTGTCCGAC
AATCCAGTTTTAAACCACGGCTTCTTTAAAAGAAACTCTCAACGGTTCCAACACCCGAATTGCGGGGGAGGGTGAAGGACCAACAGGCTG   1800
```



```
CCCTGTTGCCCGAATGATTGCCTGGGAGATTTTGAGTACCCCCACTTATCAGGGGCATGTGCAAACTAACTGCATTTTTTGACTTTGGGT   1890
GGGACAACGGGCTTACTAACGGACCCTCTAAAACTCATGGGGGTGAATAGTCCCCGTACACGTTTGATTGACGTAAAAAACTGAAACCCA
Gene
Intron 2

CAGAGTCTTGTGGACAACCCAGAAGACTGTCGATGGACCCTCAAAGAAGGACTGGCGTGGTGGCCGTGCTGCAGCCATTAACATCATCCC   1980
GTCTCAGAACACCTGTTGGGTCTTCTGACAGCTACCTGGGAGTTTCTTCCTGACCGCACCACCGGCACGACGTCGGTAATTGTAGTAGGG
Gene
Intron 2          CDS3
                  Exon 3

CTCTACCACCGGAGCTGCCAAGGCTGTTGGTGAGGTGTTGCCTGCTGTGAAGGGGAAGCTCACTGGCACTCAGACATTCTCTTCTCCAAT   2070
GAGATGGTGGCCTCGACGGTTCCGACAACCACTCCACAACGGACGACACTTCCCCTTCGAGTGACCGTGAGTCTGTAAGAGAAGAGGTTA
Gene
                                                                     Intron 3
CDS3
Exon 3

TTGTGTCCAATTATTCTGTGGGAGATTCTGTCCAGTTGTTCCGTGGGCTGGTTGGTGGGTTCCCTGTCCACCACCACCGGCTGGGGCTTT   2160
AACACAGGTTAATAAGACACCCTCTAAGACAGGTCAACAAGGCACCCGACCAACCACCCAAGGGACAGGTGGTGGTGGCCGACCCCGAAA
Gene
Intron 3
CDS3
Exon 3

TCCTTGCAGGGGTGCTTTGGGCCACCAGGCCTTCCTCTGCATCCCTGCCTGTGGGTTACAAAAAAAAACAGAAACAGCCTGGTTAGGTCC   2250

ACTGAATACAGGAGTTAATCTGCGGCAGGGCTGGGAATCCCCACAACTCCAATGGTTGGTGGTTAGGGGCTTCCAGCAACCACTCCCATA   2340
TGACTTATGTCCTCAATTAGACGCCGTCCCGACCCTTAGGGGTGTTGAGGTTACCAACCACCAATCCCCGAAGGTCGTTGGTGAGGGTAT
Gene
Intron 3

TTTTGTTGCAGCAGGTTCTGACCACCCCCAGTGGCCGGGGGGAGCCCCCACCAGGAAAGAGAACAAGAAGAGACAGTCTGCTGACATGGC   2430
AAAACAACGTCGTCCAAGACTGGTGGGGGTCACCGGCCCCCCTCGGGGGTGGTCCTTTCTCTTGTTCTTCTCTGTCAGACGACTGTACCG
Gene
Intron 3                                                                                3' UTR
                                                                                        CDS4

CTTCCGCGTGCCAACCCCTGATGTGTCTGTTGTTGATCTGACCTTCTTGGCTGAGAAGGACACCAGCATCAAGGAGATCGACTCCCTGTT   2520
GAAGGCGCACGGTTGGGGACTACACAGACAACAACTAGACTGGAAGAACCGACTCTTCCTGTGGTCGTAGTTCCTCTAGCTGAGGGACAA
Gene
3' UTR
CDS4
```



```
GAAGAAGGCCTCCCAGACATACCTCAAGGGAATCTTGGGCTTCACAGATGAGGAGCTCGTGTCGACGGACTTCGTGCATGACAATCGTTC   2610
CTTCTTCCGGAGGGTCTGTATGGAGTTCCCTTAGAACCCGAAGTGTCTACTCCTCGAGCACAGCTGCCTGAAGCACGTACTGTTAGCAAG
Gene
3' UTR
CDS4

TTCTATCTACGATTCTTTGGCCACCCTCCAGAACAACTTGCCTGGGGAGAAGAGGTTGTTCAAGGTGGTGTCCTGGTATGACAACGAGTG   2700
AAGATAGATGCTAAGAAACCGGTGGGAGGTCTTGTTGAACGGACCCCTCTTCTCCAACAAGTTCCACCACAGGACCATACTGTTGCTCAC
Gene
3' UTR
CDS4

GGGATACTCCAACCGTGTCGTCGACCTGCTGAAGCACATGTCTGGAAACTAAGTTGGAAAGTTGGTTGCACACACTGTGGAGTGACCTGC   2790
CCCTATGAGGTTGGCACAGCAGCTGGACGACTTCGTGTACAGACCTTTGATTCAACCTTTCAACCAACGTGTGTGACACCTCACTGGACG
Gene
Intron 4
3' UTR
CDS4

ACTTTTCAAAATGGCATTTGCAGGGCTGCCATCTGAAGAATAGCAAAGGTTTTCACCAATACTAGCCATTTTCTTGTATTTTTTGGTCAA   2880
TGAAAAGTTTTACCGTAAACGTCCCGACGGTAGACTTCTTATCGTTTCCAAAAGTGGTTATGATCGGTAAAAGAACATAAAAAACCAGTT
Intron 4

GTTTTTTGGTTTTTCAACTCGAATTTCGAAATCCAACCAGTCAATTTTTTAATCAGTGGCTGACCCCCCCCCCCCGGGGTGACTGACTGG   2970
CAAAAAACCAAAAAGTTGAGCTTAAAGCTTTAGGTTGGTCAGTTAAAAAATTAGTCACCGACTGGGGGGGGGGGGCCCCACTGACTGACC
Intron 4

CTTTTTCAGAGTGATTGGCATTGGGACCAAGGGACCAGTGGGCTAAGGTGGCTTGAGTGACCCTGGCGCCGAGAATTTTTTTTGCCCAAA   3060
GAAAAAGTCTCACTAACCGTAACCCTGGTTCCCTGGTCACCCGATTCCACCGAACTCACTGGGACCGCGGCTCTTAAAAAAAACGGGTTT
Intron 4

GTGAGTGGCCGACCCCCCCCCAAAGACGGTCGACTGGATCCCAAAAAAAGTGACCGACCGGTTTTCAAAAAGTGGAGTAAAAGTGGGTCA   3150
CACTCACCGGCTGGGGGGGGGTTTCTGCCAGCTGACCTAGGGTTTTTTTCACTGGCTGGCCAAAAGTTTTTCACCTCATTTTCACCCAGT
Intron 4

GAAAAATTGTACGAATTTGGGCTTCTTCTTTTTGCTCTTCTTCTCCCTCTCAAGTGCAGGTCAAGAGTTGGTTGTGCAGCAGTCAGTTGT   3240
CTTTTTAACATGCTTAAACCCGAAGAAGAAAAACGAGAAGAAGAGGGAGAGTTCACGTCCAGTTCTCAACCAACACGTCGTCAGTCAACA
Intron 4                                                              3' UTR
                                                                      Exon 4

GTGTGCAGCAGTCGAACCGGCATGTCCAGATTCTGGATAACAGTGCATTGTTTTCGCTACATTCTATGTCCTTCATTGTGCCGTGTCATC   3330
CACACGTCGTCAGCTTGGCCGTACAGGTCTAAGACCTATTGTCACGTAACAAAAGCGATGTAAGATACAGGAAGTAACACGGCACAGTAG
3' UTR
Exon 4

ACTCAAAACAAAAATTTGGAGGCAGAGAACAATATCCGCAGTTATTTGGAACAGGGTTTTTGGGGCGGTGTCGTCGGTCAAGCAGATGGC   3420
TGAGTTTTGTTTTTAAACCTCCGTCTCTTGTTATAGGCGTCAATAAACCTTGTCCCAAAAACCCCGCCACAGCAGCCAGTTCGTCTACCG
3' UTR
Exon 4
```



```
ACACAGGTTGCAAGCTGAACTTTTCCGTAGCTTCAGGGTTCAGAGGGCCCACTCAGCCGTACTTCCGCGCATTTTTTTCCCGCAGCTTCT
                                                                                          3510
TGTGTCCAACGTTCGACTTGAAAAGGCATCGAAGTCCCAAGTCTCCCGGGTGAGTCGGCATGAAGGCGCGTAAAAAAAGGGCGTCGAAGA

GGCACCCCGGAAGAGTTTTTTTTTTTGCGTTCAGGAAGGGAGGTACCACCTCTTCCGTTCCGTCGAGTGCGGACACCTCGGGCTCCCCTG
                                                                                          3600
CCGTGGGGCCTTCTCAAAAAAAAAAACGCAAGTCCTTCCCTCCATGGTGGAGAAGGCAAGGCAGCTCACGCCTGTGGAGCCCGAGGGGAC

GCTTTGGTGGCCTAAGGTTCCATGCCTTTGGAAAGTCTTTGGAACCAAAACCTGGAGAAAAATATGGTGCAATCTGGGATCCCCACCCAA
                                                                                          3690
CGAAACCACCGGATTCCAAGGTACGGAAACCTTTCAGAAACCTTGGTTTTGGACCTCTTTTTATACCACGTTAGACCCTAGGGGTGGGTT

TCCTCAAGGGGCTGGAACAGTCCCCCCTGCCCTCTAATATTTGTGCCTTGCGTACAAAACTCATTTTCTGTTCACGCATCCTCTGTTTAT
                                                                                          3780
AGGAGTTCCCCGACCTTGTCAGGGGGGACGGGAGATTATAAACACGGAACGCATGTTTTGAGTAAAAGACAAGTGCGTAGGAGACAAATA

ACAGATCATCTGATTTGGCACAGGTCCAAAGAGCCATCAGTGTGTTGGGGGGGCACCCTACATCAAAATTAAACGCAAAGGGCAAAAACG
                                                                                          3870
TGTCTAGTAGACTAAACCGTGTCCAGGTTTCTCGGTAGTCACACAACCCCCCCGTGGGATGTAGTTTTAATTTGCGTTTCCCGTTTTTGC

CCATAAGTGGGAGAGAGATGTCTCCATTCCCCAGTACCCCCAAATCAGTCCCTTTCATTTGATGATGCCTTATGTGGAAGCATCCCGTTT
                                                                                          3960
GGTATTCACCCTCTCTCTACAGAGGTAAGGGGTCATGGGGGTTTAGTCAGGGAAAGTAAACTACTACGGAATACACCTTCGTAGGGCAAA

CTTTGTGAAACCTGGGGGCCCCGTTTTAATTACGTTAAAACCCCCCTATAAAGCCATTAAGGCAGATACCATTGCTTCTATAACCAAAAA
                                                                                          4050
GAAACACTTTGGACCCCCGGGGCAAAATTAATGCAATTTTGGGGGGATATTTCGGTAATTCCGTCTATGGTAACGAAGATATTGGTTTTT

CTTTTTGGCCCAGCATGGGGTGCCCCCCCAGGAGTGGGGACCCCACTCCACCAGAGAGGCAGGAGTTGGGTTACTCAAAAAATTGGGCCT
                                                                                          4140
GAAAAACCGGGTCGTACCCCACGGGGGGGTCCTCACCCCTGGGGTGAGGTGGTCTCTCCGTCCTCAACCCAATGAGTTTTTTAACCCGGA

TTCTGCCCAAGAAGTTTGTGAAATTGGTCAGTGGAAGGGGGTGGAGGCTTTCAAAGCCCATTATCAAAGACTTGGTACCCAAAAGCTT    3'
                                                                                           4228
AAGACGGGTTCTTCAAACACTTTAACCAGTCACCTTCCCCCACCTCCGAAAGTTTCGGGTAATAGTTTCTGAACCATGGGTTTTCGAA    5'
```

Euglena gracilis glyceraldehyde-3-phosphate dehydrogenase (*gapC*) gene, complete cds with defined 5'UTR, 3'UTR, exons and introns.

Appendix 10

Matrix Family	Detailed Family Information	Start position	End position	Strand	Core sim.	Matrix sim.	Sequence
P$MYBS	MYB proteins with single DNA binding repeat	11	27	+	1	0.91	gaaactATCCatcgcag
V$CARE	Calcium-response elements	32	42	+	1	0.951	ggagcGAGGca
V$LTSM	Localised tandem sequence motif	114	128	+	1	0.891	cctcctgttcCATCc
V$EVI1	EVI1-myleoid transforming protein	248	264	+	1	1	cgccactGATGagtcca
F$GATA	Fungal GATA binding factors	251	265	+	0.853	0.86	cactGATGagtccaa
P$OPAQ	Opaque-2 like transcriptional activators	252	268	+	1	0.89	actgaTGAGtccaatga
P$NCS2	Nodulin consensus sequence 2	266	280	+	1	0.871	tgatggCTCTtcaga
V$EBOX	E-box binding factors	303	319	+	0.951	0.831	tgtgTCAAgtgttctga
V$NKXH	NKX homeodomain factors	303	321	+	1	1	tgtgtcAAGTgttctgacc
P$CAAT	CCAAT binding factors	319	327	+	1	0.981	acCCAAtgg
F$MMAT	M-box interacting with Mat1-Mc	352	362	+	1	0.916	cccATTGtggt
V$LTSM	Localised tandem sequence motif	362	376	+	1	0.88	tattatcctcCATCc
P$MYBL	MYB-like proteins	382	398	+	1	1	gtatgtccGTTAactcc
V$HNF1	Hepatic Nuclear Factor 1	382	398	+	1	0.91	gtatgtccGTTAactcc
P$TERE	Tracheary-element-regulating cis-elements, conferring TE-specific expression	395	405	+	1	0.87	ctccAAAGgaa
P$PCDR	Factors involved in programmed cell death response	406	412	+	1	1	TGTGttg
V$DLXF	Distal-less homeodomain transcription factors	414	432	+	1	0.916	ttctgtcccTAATtccagg
V$NKXH	NKX homeodomain factors	418	436	+	1	0.91	gtcccTAATtccaggatga
V$STAT	Signal transducer and activator of transcription	422	440	+	1	0.961	ctaaTTCCaggatgaattg
V$P53F	p53 tumor suppressor	488	512	+	0.885	0.914	gcagggttcttcaggCTTGtccatg
F$YSTR	Yeast stress response elements	516	530	+	1	1	acagtaaGGGGgggg
V$NDPK	Nucleoside diphosphate kinase	520	536	+	0.949	0.911	taAGGGgggggggagct
V$KLFS	Krueppel like transcription factors	520	538	+	0.857	0.951	taaggGGGGggggagctcg
V$PLAG	Pleomorphic adenoma gene	521	543	+	1	1	aaGGGGgggggagctcgtcccc
V$XBBF	X-box binding factors	567	585	+	0.75	0.919	ctgttcccaTGGAcccaaa
F$YSTR	Yeast stress response elements	580	594	+	1	1	cccaaaaGGGGgcca
V$PLAG	Pleomorphic adenoma gene	585	607	+	1	1	aaGGGGgccacgggacgggaaca
F$YSTR	Yeast stress response elements	605	619	+	1	1	acactaaGGGGgggg
F$YMIG	Yeast GC-Box Proteins	604	622	+	1	0.891	aacactaagggGGGGgtat

Matrix Family	Detailed Family Information	Start position	End position	Strand	Core sim.	Matrix sim.	Sequence
V$GREF	Glucocorticoid responsive and related elements	911	929	+	0.89	0.861	gcgGCACaaaccgttaatt
V$ZF12	C2H2 zinc finger transcription factors 12	915	929	+	1	0.818	cacaaaccgTTAAtt
P$MYBL	MYB-like proteins	915	931	+	1	1	cacaaaccGTTAatttt
V$OVOL	OVO homolog-like transcription factors	917	931	+	1	0.961	caaaccGTTAatttt
V$HOMF	Homeodomain transcription factors	918	936	+	1	0.951	aaaccgtTAATtttggtta
V$NKX6	NK6 homeobox transcription factors	920	934	+	1	0.917	accgTTAAttttggt
P$MYBL	MYB-like proteins	925	941	+	1	0.912	taattttgGTTAgggat
V$ZF10	C2H2 zinc finger transcription factors 10	931	945	+	1	0.861	tggTTAGggataccc
P$NTMF	NAC factors with transmembrane motif	983	991	+	1	1	tTTCCtttt
F$YRAP	Yeast activator of glycolyse genes / repressor of mating type l	980	1002	+	1	0.851	cggtttccttttcgttCACCcaa
V$SREB	Sterol regulatory element binding proteins	992	1006	+	1	0.921	cgtTCACccaactcc
F$G1TF	G1 specifc transcription factors	797	805	+	1	1	cacCAGCac
F$YSTR	Yeast stress response elements	819	833	+	1	1	ccctccaGGGGcttt
V$NEUR	NeuroD, Beta2, HLH domain	841	855	+	1	0.921	cggCCATgtgatatg
V$TALE	TALE homeodomain class recognizing TG motifs	846	862	+	1	1	atgtgatatGTCAaaaa

The highly similar transcription factors compared against the plant, algal and fungal promoter elements using MatInspector; Genomatix Software Suite (Cartharius *et al.*, 2005). The result obtained shows only those matrices above 85% similarity.